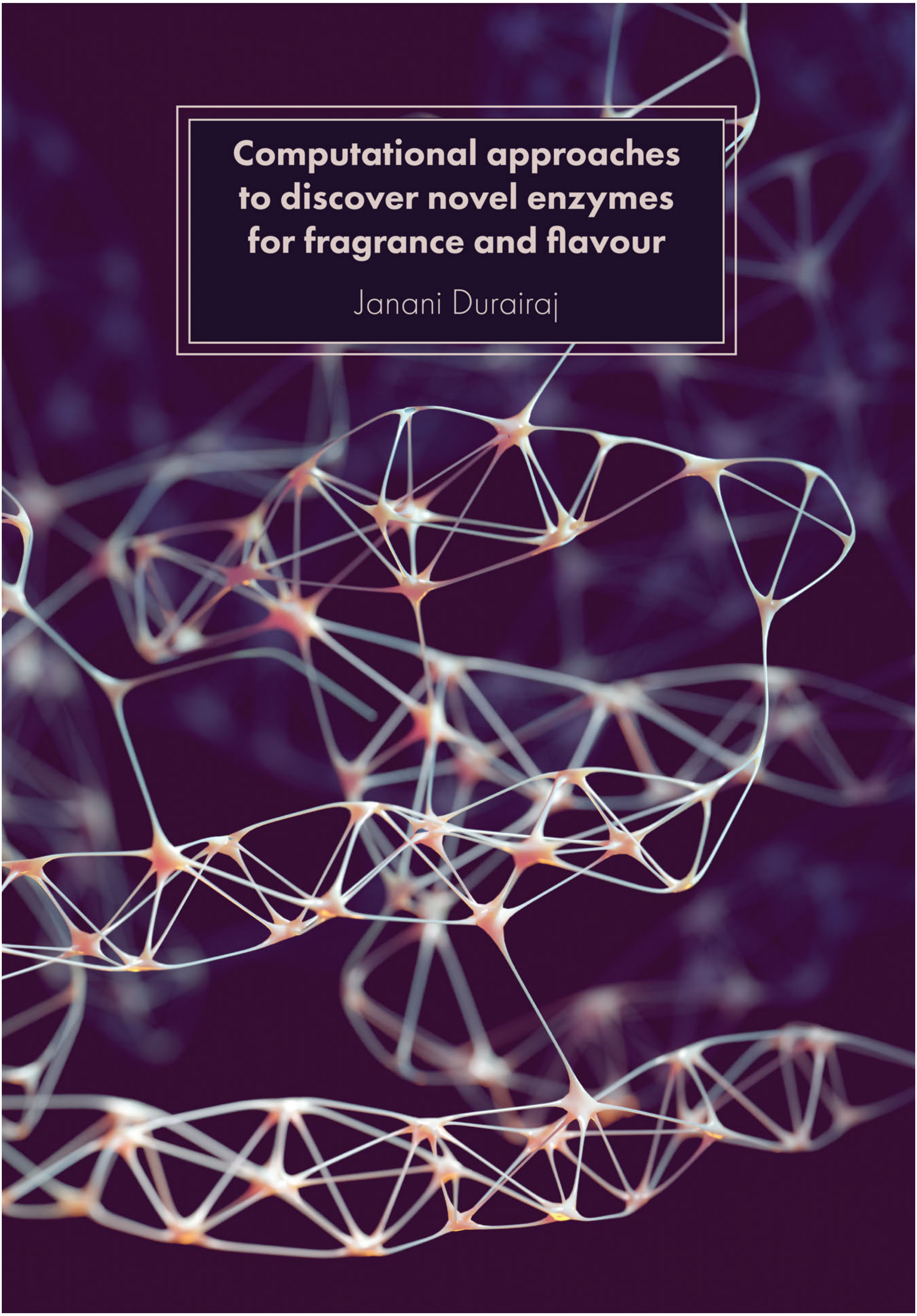
# Computational approaches to discover novel enzymes for fragrance and flavour

Janani Durairaj



**Propositions**

1. Contemporary computational techniques are not sufficiently advanced to predict plant specialized metabolic reactions.
   (this thesis)

2. Protein structure-based approaches will become as widespread as sequence-based in bioinformatics.
   (this thesis)

3. Scientists are responsible for stopping the spread of pseudoscience.

4. The days of the science polymath have passed.

5. Technology is held to a far higher standard than humans, be it self-driving cars or vaccines.

6. In an increasingly border-less world, polyculturalism is a precondition to reducing social inequality.

Propositions belonging to the thesis, entitled

Computational approaches to discover novel enzymes for fragrance and flavour

Janani Durairaj
Wageningen, 15 September
2021

# Computational approaches to discover novel enzymes for fragrance and flavour

Janani Durairaj

**Thesis committee**

**Promotor**
Prof. Dr D. de Ridder
Professor of Bioinformatics
Wageningen University & Research

**Co-promotor**
Dr A.D.J van Dijk
Assistant professor, Bioinformatics Group
Wageningen University & Research

**Other members**
Prof. M.E. Schranz, Wageningen University & Research
Prof. Dr A.E. Osbourn, John Innes Centre, UK
Prof. Dr M.J.T. Reinders, Delft University of Technology
Prof. Dame J. Thornton, European Bioinformatics Institute, UK

This research was conducted under the auspices of the Graduate School Experimental Plant Sciences

# Computational approaches
# to discover novel enzymes
# for fragrance and flavour

Janani Durairaj

**Thesis**
submitted in fulfilment of the requirements for the degree of doctor
at Wageningen University
by the authority of the Rector Magnificus,
Prof. Dr A.P.J. Mol,
in the presence of the
Thesis Committee appointed by the Academic Board
to be defended in public
on Wednesday 15 September 2021
at 4 p.m. in the Aula.

Janani Durairaj
Computational approaches to discover novel enzymes for fragrance and flavour

PhD thesis, Wageningen University, Wageningen, The Netherlands (2021)
With references, with summary in English

ISBN: 978-94-6395-923-0
DOI: https://doi.org/10.18174/551031

# CONTENTS

# CHAPTER 1

# General introduction

# 1.1 The aromatic world of plants

## 1.1.1 Plant specialized metabolites

Plants have existed on this earth for over 400 million years and along the way have diversified into the incredibly vast kingdom we observe today. Needless to say, humans rely on plants for nutrition, their carbon dioxide storage potential, and as a source of raw material for numerous products. In addition, with over 400,000 vascular plant species[1] collectively producing between 200,000 and 1 million chemical compounds[2], the plant kingdom's dazzling chemical diversity is also of great value and potential to humankind. For millennia, humans have been finding new uses for plants and plant compounds far beyond their value as nourishment - as pharmaceutical agents, preservatives[3], flavouring compounds[4] and dietary supplements[5], pesticides[6] and insecticides[7], cosmetics and perfumes[8], and even as a way to alleviate the ever-growing and unsustainable demand for plastics[9] (see Osbourn & Lanzotti[10] for a more comprehensive review of these applications).

Medicinal plant usage is thought to date as far back as 60,000 BC[11], with the oldest written evidence of drug preparation from plants, a Sumerian clay slab, believed to be around 5000 years old[12]. With the isolation of morphine from opium poppy in 1817 and quinine to treat malaria from the bark of the cinchona tree in 1820, the field of natural product chemistry bloomed and prospered. Now, 25%-50% of marketed drugs are of natural origin[13,14], derived from over 10,000 plant species[15]. Pyrethrins derived from *Chrysanthemum cinerariifolium*[16], azadirachtin from neem plants[17] and, prior to its numerous negative side effects being discovered, nicotine from nightshades[18] have been used to varying degrees as sources of insecticides, with the current demand worldwide for the pyrethrum flowers in excess of 25,000 tons annually, satisfied by the estimated 150 million flowers still hand-harvested daily in Kenya, Tanzania, and Ecuador[19]. Vanillin and caffeine are two popular examples of plant-derived molecules used in food additives and flavouring agents. Various molecules associated with floral scents have found their way in to the growing fragrance and aromatherapy industries, either individually or as part of volatile (essential) oils. Given that a vast majority of plant species have never even been described, much less surveyed for their chemical constituents, it is plausible and likely that many new sources of valuable plant-derived materials remain to be discovered.

A large percentage of this chemical diversity in plants consists of specialized metabolites (SMs, previously secondary metabolites), compounds not required for the primary biochemical pathways underlying cell growth and reproduction. Plants have been in an intense and ongoing evolutionary battle with their environment, and thus have evolved to produce a vast number of SMs specifically involved in environmental adaptation, such as phenolics, terpenes, alkaloids etc. To compensate for the relatively stationary nature of plants, these compounds often make up plant odour and colour, allowing them to be sensed from far and wide. SMs typically make up less than 1% of the total carbon in a plant species[13] and arise from a limited number of simple chemical scaffolds which are then chemically modified by an array of diverse enzyme families to produce the vast and intricate language of natural products on display.

SMs help defend against invading herbivores and pathogens, either directly[20] or by recruiting their natural enemies as allies[21]; control seed germination in unfavourable conditions[22]; regulate symbiosis[23]; impede competing plant species[24]; attract pollinators to ensure successful propagation[25]; and even act as UV absorbing compounds to prevent leaf damage by light[26]. From a human perspective, SMs have found uses in antibacterial, antiviral, and anti-fungal drugs, chemotherapeutic agents, agents against inflammation, diabetes and heart diseases, in crop protection and in consumer fragrances, to name just a few.

Despite the fact that we have evidence of the function of quite a few SMs as discussed above, such evidence has not yet been found - or perhaps does not exist - for a majority of these compounds. One explanation of the staggering SM diversity comes from the Screening Hypothesis[27], which recognizes that potent biological activity is a rare property for any molecule, and hypothesizes that organisms are compelled to generate as much chemical variability as possible to increase the probability of finding a molecule with a certain function. Thus, the appearance of specialized metabolism may have originated from the interplay of creating large numbers of chemical structures and screening these compounds in their environment for new useful functions. Other explanations consider synergistic effects between compounds[28], complex genetic correlations between defence traits[29], multi-functional roles for metabolites[30], diversity in the numbers and types of enemies, and plant-herbivore co-evolution[31] as possible contributors to the immense assortment of plant SMs.

### 1.1.2 Terpenes and terpenoids

Terpenes and their modified derivatives, terpenoids, form one of the largest families of SMs, with many roles in plants as toxins, attractants, and signalling agents[32], all derived from the same five-carbon isoprenoid units coupled together into 10-carbon (derived from the geranyl diphosphate substrate), 15-carbon (from the farnesyl diphosphate FPP substrate), and even 2000-500,000-carbon chains, as is the case for rubber. These chains are produced and modified by different SM enzyme families, the most prominent of which are the terpene synthases (TPSs). Many terpene volatiles are direct products of TPSs, but others are formed through transformation of the initial products by oxidation, dehydrogenation, acylation, and other reaction types, each of which is catalysed by a group (or several groups) of related SM enzymes families.

The terpenoid-related SM producing enzyme families that we know of are quite distinct from each other in terms of sequence, structure, general organization, and distribution across plant species[33]. However, they all seem to serve the same aim, namely to provide the potential to produce many different chemical structures by means of limited genetic resources. This compels these families to share certain properties that set them apart from enzymes involved in primary metabolism. This includes their proclivity to act on multiple substrates, catalyse multiple reactions, or produce multiple products[34]. In addition, these enzymes tend to have little correlation between their levels of sequence similarity and the chemical similarity of their respective substrates, intermediates and products[33]. The high levels of sequence diversity in SM enzymes is intrinsically linked with the inbuilt flexibility with regard to the chemical scaffolds they can modify or produce[35]. Orthologous genes may encode enzymes for different SMs[36] and repeated independent evolution of SM enzymes has been observed from homologous genes which are not necessarily orthologous[35]. Many of these enzyme families work in an assembly line fashion, to produce terpenes and further modify these into more complex terpenoids with altered properties[37].

The name terpene originates from turpentine, derived from terpenoid resins found in coniferous trees[38]. Since then humans have found numerous industrial uses for this metabolite class, ranging from the manufacture of biopolymers and inks[39], flavours and fragrances[40,41], pharmaceuticals and cosmetic products[42–44], biofuels[45], and natural rubber[46]. Traditionally, terpenoid compounds are derived from harvested plant parts via steam distillation, solvent extraction, and cold pressing. However, the amount of terpenes produced by a given plant species is typically minute, requiring large-scale unsustainable plant harvesting to extract industrial levels of product. This is compounded by an ever-increasing demand for these compounds, and drastically shrinking natural resources due to over-exploitation[47], modern agricultural practices[48], climate change and natural disasters[49]. On the other hand, the complexity of these molecules makes chemical synthesis inherently difficult and expensive, even disregarding the environmentally unfriendly production processes involved. To counter these issues, bioproduction of terpenoids via microbial fermentation or via biocatalysis of isolated enzymes and natural precursors provides an excellent alternative.

Development of such bioproduction routes involves the isolation and characterization of plant TPSs and other enzymes involved in the biosynthetic pathway to identify those producing the desired terpene products, followed by optimization of these enzymes to allow them to perform favourably in their new environment. In this thesis I mainly focus on the plant sesquiterpene synthase (STS) enzyme family, consisting of TPSs which utilize the C15 FPP as substrate to collectively produce over 300 known monocyclic, bicyclic, and tricyclic sesquiterpenes[50].

### 1.1.3 A history of sequence-structure-function relationships in sesquiterpene synthases

A variety of isoprenoids and terpenoids were successfully isolated from crude plant extracts many decades ago with DerMarderosian et al.[51] providing a comprehensive overview. Research over many years has also established common carbocationic reaction mechanisms for plant STSs[50], describing the formation of seven parent cations that determine the overall structure of the final sesquiterpenes. However, despite our detailed knowledge of these mechanisms, relatively little is known about their structural basis - i.e. how do various residues and structural features of these diverse proteins mediate substrate binding and the numerous cyclization, rearrangement, and modification reactions that follow?

The amino acid sequences of these enzymes provide some answers to this question. Plant STSs are generally 550-580 amino acids long and lack the characteristic N-

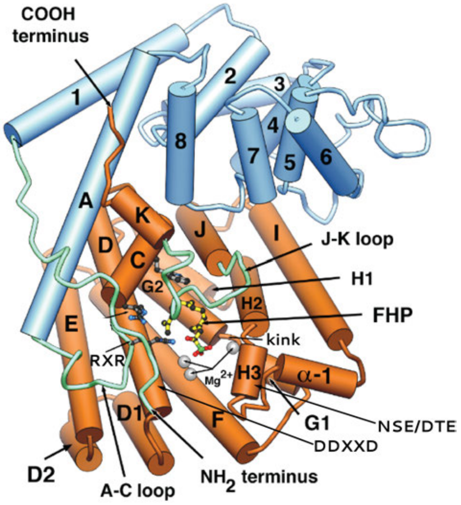


Figure 1.1: Schematic view of the TEAS-farnesyl hydroxy-phosphonate (FHP) complex. Blue rods represent $\alpha$-helices in the N-terminal domain; orange rods represent $\alpha$-helices in the C-terminal domain. Loop regions shown in green are disordered in the native TEAS structure. Motifs and the helix kink discussed in the text are labelled.
Image adapted from Starks et al.[52]

terminal transit peptide found in mono-TPSs. This causes the translation products of STS genes to localize within the cytosol where they come into contact with varying concentrations of their preferred substrate, FPP. The diphosphate moiety of this substrate is captured by a conserved RXR motif and divalent metal ions like $Mg^{2+}$ or $Mn^{2+}$ which are themselves bound by motifs DDXXD and NSE/DTE, at the entrance of the active site[52] (Figure 1.1). This allows the hydrophobic moiety to enter the active site cavity and undergo cyclizations and rearrangements to produce sesquiterpene products. Plant STSs share between 30-99% sequence identity but much of this corresponds to phylogenetic or taxonomic relationships between plant species rather than the types of sesquiterpenes produced[50]. Examples exist on both ends of the spectrum, with nearly identical enzymes producing chemically dissimilar products and enzymes sharing less than 30% sequence identity producing the exact same products. Thus, protein structural information was needed to understand the mechanisms behind these enzymes. The first experimentally solved plant terpene synthase structure was 5-*epi*-aristolochene synthase from *Nicotiana benthamiana* (TEAS), released in 1997[52]. This was eventually followed by numerous plant STS structures, plant mono-TPS structures (which sometimes share very high degrees of sequence identity and structural similarity with plant STSs), and bacterial and fungal STS structures which share little sequence identity but have the same structural fold and are capable of producing a number of the same sesquiterpenes. Mutational studies swapping residues or stretches of sequence between two closely related STSs also provided valuable knowledge on areas in the protein connected with catalytic activity. In this thesis we focus on plant STSs, which have a tertiary structure as shown in Figure 1.1, consisting entirely of $\alpha$-helices connected by short loops and turns, organized into two structural domains. The active site is a hydrophobic pocket formed by six $\alpha$-helices with two loops on the surface closing it off.

The determination of the TEAS structure both alone and in complex with two FPP substrate analogs helped elucidate the function of the highly conserved DDXXD motif in coordinating two of the $Mg^{2+}$ ions required for diphosphate binding[52]. The third $Mg^{2+}$ ion is ligated by residues in the more variable NSE/DTE motif. Inspection of changes in folding upon binding the substrate analog indicated that the C-terminal J-K loop becomes ordered on substrate binding to seal the active-site pocket from surrounding solvent, positioning the residues in another conserved STS motif, RXR, in close proximity to the C1 hydroxyl group of the substrate. The combined positive charges of the $Mg^{2+}$ ions and the two Arginines from the motif direct the diphosphate away from the hydrophobic pocket. Similarly, two residues in the so-called "kink" of helix G direct the cationic end of the farnesyl chain into the hydrophobic active site, poised for modification and further carbocationic rearrangements[52]. Modelling a related vetispiradiene synthase from *Hyoscyamus muticus* (HVS) with TEAS as template indicated highly similar active site residue positioning, with chimeric constructs of TEAS and HVS producing a mixture of 5-*epi*-artistolochene and vetispiradiene[53]. From this it was hypothesized that specificity depends on the determination of the active site conformation by the surrounding layers of residues. This was followed up with the identification of nine residues responsible for the conversion of specificity from TEAS to HVS, via mutational analysis[53,54]. Over the years, the structures of multiple TEAS mutants with different substrate analogs have been solved, each leading to novel insights into the underlying mechanisms involved. The Y520F mutant demonstrated the presence of a neutral intermediate, germacrene A, which is re-protonated and modified again to produce the final product[55]. Solving TEAS with *cis* and *trans* substrate analogs shed light on the similarities in catalytic mechanisms for producing cisoid and transoid sesquiterpenes[56]. Another seminal study[57] solved multiple crystal structures of TEAS and a four-residue mutant in complex with two substrate analogs and three sesquiterpene molecules, substantiating the role of the J-K loop in shielding the active site to allow for multiple reactions to occur before final product release. Product profiles of these proteins were compared across a variety of pH and temperature conditions, demonstrating that minor product concentrations can depend on extrinsic environmental conditions.

The structure of $\delta$-cadinene synthase (DCS) from *Gossypium arboreum* (tree cotton) was solved in 2009, containing a second DDXXD motif instead of the more common NSE/DTE motif found in 90% of characterized plant STSs[58]. A number of studies on cadinene synthases from different species offer evidence for different reaction schemes[59–61], demonstrating that some sesquiterpenes, such as those derived from the cadalane skeleton, can arise from multiple different reaction paths.

2011 saw the release of the structure of a three-domain bisabolene synthase from a coniferous species, *Abies grandis*[62], establishing the evolutionary link between di- and sesqui-TPSs. This species contains the famed 52 product $\gamma$-humulene synthase and 34 product $\delta$-selinene synthase[63], both of which were also the target of extensive mutational and directed evolution studies to determine residues connected to STS promiscuity[64,65].

In 2013, the crystal structures of an *Artemisia annua* $\alpha$-bisabolol synthase and a mutant producing $\gamma$-humulene were solved[66], pinpointing five residues involved in this product shift that spans across quite distinct parent carbocations. Mutational studies across other synthases from *A. annua*[67,68] helped discover more residues and their involvement in modifying products.

Over the years, more plant STS structures have been solved[69,70] and more mutant studies have been performed[71,72]. Similar timelines of sequence-structure-function relationships can be identified for other SM enzyme families involved in terpenoid production, such as the upstream isoprenoid synthases and prenyltransferases that produce TPS substrates, and the downstream families, such as the cytochrome P450s, dehydrogenases etc. involved in post-modification of TPS-produced terpenes into more complex terpenoids[37]. However, while such approaches have been successful in finding residues influencing specificity for a single enzyme, the small scale of studies in the light of the large diversity of SM enzymes makes it impossible to say if these studies find aspects shared across all enzymes in these families. In order to pinpoint residues and patterns important to overall SM specificity, larger scale techniques are needed. Therefore, researchers have increasingly been turning to computational methods to take SM enzymes across species and specificities into account.

### 1.1.4 The need for bioinformatics

While the floral aroma profiles of various plant species have been generated with relative ease, the identification and characterization of the enzymes responsible for these emissions is a more difficult task. Functional characterization of the product profiles of TPS enzymes typically consists of sequence similarity-based identification and cloning of putative TPS genes from next-generation sequencing results, followed by heterologous expression of recombinant genes in expression systems such as *Escherichia coli* with TPS substrates introduced, and finally detection and identification of mass spectra from a GC-MS assay. The product peaks can be reliably identified by comparison to internal standards. However, often such standards are not available and identification is done by comparison to spectra and retention times from literature, which is not as accurate. Further verification and determination of chirality can be done using the more expensive NMR technique, but this is rarely performed. In addition, SM enzymes are widely influenced by various environmental factors[73]. Unfortunately, this variability is also observed in the lab - experimental conditions can play a major role in the outcome of characterization studies[57]. For example, issues with cloning or unfavourable experimental conditions could result in nonfunctional enzymes or enzymes producing undetectable level of products. Rearrangements of enzyme products due to pH, temperature, or other compounds present in the experimental setting is also possible and can lead to misidentification of the actual products[57]. Large-scale screening experiments depend on the design of high-throughput assays which often are not available or, if they are, cannot sufficiently differentiate between the different terpenes. Hence, functional characterization of SM enzymes is a time-consuming, laborious process containing an unavoidable level of noise. In addition, many enzymes of interest produce very low amounts of terpenes, in their native plant species but especially so in microbial fermentation settings. Here they are incorporated into organisms for which they are not optimized with respect to codon distributions, specificity and efficiency of the pathways involved in generating the required substrates and cofactors, pH and temperature conditions, and more[74]. While improved screening systems may help in this regard to select more active enzymes, this depends on the availability of such enzymes and the transfer of their favourable properties to different production strains and systems. This illustrates the need for actually understanding the residues and regions involved in the various biosynthetic reactions producing terpenoids, such that each of the players can be engineered to produce desired levels of desired products.

The rapid growth of next-generation sequencing and initiatives such as the 1000 plants transcriptome project[75] have greatly increased the data available on plant protein sequences. Many thousands of these sequences are uncharacterised putative TPSs. Functional characterization of all these proteins would be impossible in a reasonable time frame, especially as these numbers will continue growing as more plants are sequenced. Thus, narrowing down interesting and relevant enzymes using computational techniques for product specificity prediction is necessary. These techniques will have to take into account the above-mentioned aspects of sparse and noisy characterized data. They will also have to pinpoint motifs and residues responsible for biosynthetic pathway determination - insights which can help design mutants and engineer enzymes with desired properties such as higher activity, specificity, thermostability and more. Due to the high sequence diversity shown by STSs, this thesis posits that protein structures and structural bioinformatics techniques are necessary to obtain higher performance and deeper biological insights.

## 1.2 Computational approaches to explore proteins

### 1.2.1 Representation

Protein bioinformatics is a fast-growing and thriving field dealing with algorithms and data structures to explore, compare and contrast (groups of) proteins. The first step in this exploration typically consists of choosing a format to represent proteins that can be understood by computers, sometimes referred to as an *embedding*. This is followed by the usage of algorithms that take the embedding as input and return various results and insights for user interpretation.

Proteins consist of multiple levels of information, stored in the primary, secondary, and tertiary structures leading to multiple different representation techniques. The most common of these is the primary structure, i.e. the one-dimensional amino acid sequence. This by itself forms the input for various algorithms based on $k$-mer counting, hidden Markov models[76] or multiple sequence alignment[77] to transform the sequence into an embedding. These embeddings can be further used to find remote homologs[78], inspect conserved and variable residue positions linked to biological mechanisms[79], generate phylogenetic trees describing evolutionary relationships[80], extract motifs specific to certain groups and subgroups useful for delineating catalytic sites[81], and categorize novel protein sequences found while mining genomes and transcriptomes[82]. Multiple sequence alignments can also be used to perform correlated mutation analysis[83], based on the theory of residue co-evolution which postulates that a mutation in one residue involved in an interaction leads to preferential evolutionary selection of interaction partners with complementary mutations maintaining the interaction. This technique can be used to predict within-protein interacting residues in families with scant structural information[84], interactions across protein-protein complexes[85], and catalytic interactions such as between residues contacting a ligand[86]. Many of these embeddings and algorithms are explored in this thesis for the plant STS enzyme family.

Numerous protein families have divergent protein sequences and yet share highly similar structures, topologies, and folds, since structure tends to evolve slower than sequence[87]. Furthermore, protein tertiary structure typically leads to a wealth of information not found in sequence - three-dimensional atom coordinates, spatial interactions, solvent accessibility, residue dynamics and electrostatics, and more. However, protein structures are often studied on an individual basis or at most across a couple of highly similar proteins. This is due to a combination of factors - hitherto relative scarcity of structural data compared to sequence, the multitude of interconnected high-dimensional information that is challenging to embed, explore and interpret across multiple proteins, and the lack of availability of accessible and fast computational tools to ease this embedding and exploration process. The scarcity aspect is being alleviated with an accelerating increase in the number of experimentally determined structures[88] and rapid progress in computational structure prediction techniques[89]. This further drives the demand for better tools and algorithms to explore and utilize this rich data source to extract mechanistic insights and predictions. For example, algorithms to embed structures for similarity search across a database of protein structures are required to be very fast, as researchers typically expect instant results on search[90–92]. Typical disadvantages of these techniques are their lack of adaptation to proteins from within the same family, as they are usually designed to distinguish between diverse proteins, and their lack of interpretability in terms of which regions of structure are comparable between two proteins. For proteins from the same family or sharing a high degree of structural similarity, multiple structure alignment provides a more accurate means of comparison and allows for the use of techniques analogous to those using sequence representations, such as structural motif detection, identification of conserved and variable residues etc. These alignments can then be used to represent various structural features such as residue dynamics, electrostatics, depths, and accessibility, along with sequence-derived physicochemical properties. For tasks involving protein interaction or interfaces, another common technique is to use descriptors of the molecular surface[93–95], and even joint representations of binding partners, be they other proteins[96], small molecule ligands[97], or nucleic acids[98].

Until recently, the main *in silico* approach to determine molecular function of proteins, for instance as described by the Gene Ontology (GO) classification scheme[99], was through homology-based functional annotation transfer - i.e. for a new query protein, a search is made for similar sequences or structures (using the embeddings described above) to find candidates likely to share a common evolutionary origin, and functional annotations of these candidates are transferred to the query protein. This approach is often hindered by the lack of homologs in existing public databases for a variety of query proteins, and, more critically, the fact that high conservation does not always equal function conservation[100], and that proteins with low similarity may still share the same function[101]. This becomes particularly relevant as the level of detail of the function being annotated increases, i.e. as we move down the hierarchy from general molecular function to the specific catalytic reactions and specificities involved. Hence, it is not straightforward to transfer function globally without taking into account information specific to residues and regions of the protein relevant to the function being considered. These challenges make homology-based annotation transfer problematic[102], especially as annotation errors can propagate and be amplified as more and more proteins are annotated by transfer[103]. Furthermore, global similarity to a small set of similar proteins does not further understanding of the inner workings of proteins sharing a function, an aspect that becomes increasingly important in fields such as drug discovery and biotechnology where such understanding opens doors for specialized protein engineering.

Thus, to further link proteins and their embeddings with specific functional characteristics of the proteins under consideration, and to find more intricate patterns within distinct and relevant residues, it has become increasingly common to reach for the set of algorithms and techniques collectively referred to as machine learning.

### 1.2.2 Machine learning

Machine learning (ML) is defined as "the study of computer algorithms that improve automatically through experience and by the use of data" [104]. Typically, these algorithms make use of statistics to find patterns in datasets and are often used to link these patterns to specific outcomes or groupings[105]. In the context of proteins, ML approaches can broadly be divided into *protein family based* and *protein universe based* techniques. These two categories differ in the kinds of prediction problems they are applied to, the kinds of algorithms used, and the kinds of representations and embedding used as input.

Protein family based ML is used to predict properties of the members of individual protein families consisting of hundreds to thousands of experimentally characterized proteins, such as each of the above-mentioned SM enzyme families. There are wide range of algorithms at our disposal for these tasks, including but not limited to k-nearest neighbours algorithms (k-NNs)[106], support vector machines (SVMs)[107], Gaussian processes[108], and ensemble methods such as Random Forests[109] and gradient boosting trees[110]. These have been successfully applied to a variety of questions ranging from predicting catalytic activity[111], the effect of mutations[112] and variants[113], interactions with other proteins[114], nucleic acids[115,116], and peptides[117], thermostability[118], drug-target binding affinity[119], and ligand specificity[120]. Since the proteins under consideration are close from an evolutionary perspective, multiple protein alignment is commonly used as a starting point to generate the input embeddings for these tasks. While sequence alignment has generally been much more popular than structure alignment, the existence of SM enzyme-like families which share the same structural fold despite having little primary sequence similarity necessitates the use of structure-based alignment methods. Protein family ML often has to deal with sparse datasets and rely on algorithms which can handle a large number of features measured across a small number of data points. In addition, many approaches in this field aim to interpret prediction results to derive insights about underlying mechanisms and residues which may be important for function. Such predictions and insights further drive experimental research to explore novel and relevant protein family space.

The larger-scale protein universe based ML typically uses tens of thousands of proteins from diverse superfamilies to learn global properties of proteins, such as secondary and tertiary structure and folding, interactions, broad function classes etc. Deep learning (DL) is a common choice for such problems, as it is known to drastically outperform other techniques in the presence of large amounts of data. Much work in this area has been done using protein sequences as input since the growth of protein structure and modelling data is relatively recent. Unlike protein family ML, alignment is generally not an option in such techniques since most proteins in the dataset are evolutionarily remote, thus most described embedding techniques for large-scale ML depend on learning alignment-free patterns across diverse protein sequences or on generating on-the-fly alignments of sub-groups of data during the learning process. Recent examples of global sequence embeddings have been shown to capture amino acid characteristics and other physiological properties of proteins as a whole[121–124]. Structure-guided sequence embeddings have also started to appear[125], providing a compromise between scarce structure data and abundant sequence data. Global unsupervised embeddings can also be adapted and applied to protein family ML successfully but currently suffer from a lack of interpretability.

### 1.2.3 Successes of machine learning on proteins

Over the past decade, protein ML has moved far beyond theoretical studies into numerous real-world applications, some of which are described in this section with special emphasis on approaches related to protein structures. In protein structure prediction, be it secondary structure, backbone angles, contacts, folds, or full-atom structure, ML has become indispensable and forms the backbone of a number of popular tools and algorithms. A majority of the more recent predictors in this field use deep learning, as is common in such protein universe problems.

Secondary structure prediction has come a long way since its start in 1951[126], with recent methods[127,128] achieving prediction accuracies over 80%, a steady increase from the 70% accuracy reported in 1993[129]. The driving force behind the increase in prediction performance in many cases is attributed to better features in the representations used[126] - while early methods used amino acid features derived from single residues[130,131], gains were seen by incorporating sequence profiles[132] that implicitly include conserved structural information across homologous sequences[129,133,134], known secondary structure information from homologous proteins[135–138], and predicted solvent accessibility and backbone torsion angles[127,139,140] (also areas where ML and DL have become the norm[141–143]). Residue-residue contact prediction, based on the underlying biological theory of co-evolution[144], also saw a shift from more traditional statistical approaches[83,145] to more accurate deep learning based techniques[146–150], which even go as far as distance matrix prediction[147,151].

All-atom structure prediction is typically divided into template-based and template-free approaches. Template-based modelling or homology modelling uses previously determined structures of related proteins as the reference upon which to model the target. This includes methods for the 1) detection of, and 2) alignment to, a related protein of known structure, followed by 3) modelling of the backbone, loops and side chains and 4) subsequent evaluation of these models to return the most reasonable ones[152]. Though still relatively uncommon, ML has been used in each of these steps[153–156]. Template-free modelling, on the other hand, which does not rely on global similarity to a known structure and hence can be applied to proteins with novel folds, now makes heavy use of ML and DL, with the recent release of AlphaFold2 making headlines for its breakthrough results in the Critical Assessment of Structure Prediction (CASP14) competition. In fact, all the top-performing CASP13[157] structure prediction methods rely on deep convolutional neural networks for predicting residue contacts or distances, predicting backbone torsion angles and/or ranking the final models; for a recent review on the underlying techniques used, see Kuhlman & Bradley[89].

Some significant applications of protein universe ML are in the field of drug discovery, where such techniques have become integral[158]. Their contributions start from the computational modelling of putative receptor targets, which often involves secondary structure prediction, solvent accessibility prediction, and/or residue contact map prediction, as discussed above. Subsequently, binding sites in the target structure and putative drug candidates are identified using cavity/pocket prediction techniques, prediction of "druggable" regions, and protein-ligand binding site[120] prediction. This is typically followed by molecular docking to evaluate protein-ligand interaction and affinity between the target and a variety of drug candidates. In the case of unknown target proteins or to identify off-target binding candidates, reverse/inverse docking[159,160] is used to create embeddings of drugs and search across protein structure databases for good docking solutions. In these contexts, ML approaches are used to improve scoring functions of binding affinity and plausible docking poses[161–164]. Unsupervised clustering techniques are used to organize and prioritize large databases of receptors.

Other areas in which structure-based protein universe ML has taken over include prediction of general function[165], protein-protein[166] and protein-ligand interactions[167], interfaces[168] and hot spots[169], stability changes in protein mutants[112,170], catalytic turnover rates[111], post-translational modifications, protein dynamics[171], and amino acid sequences for *de novo* protein design[89].

Protein family ML has mostly been sequence-based so far, both due to a lack of solved crystal structures and high degree of diversity in sequence-based approaches. In drug discovery settings, the superfamily of G-protein coupled receptors (GPCRs) has been a very important target due to the role these proteins play in physiological processes covering vision, olfaction, neuronal signal transmission, cell differentiation, pain, muscle contraction, and hormone secretion, to name a few. Sequence-based ML models have been designed for predicting interactions with ligands[172] and drugs[173], predicting N-linked glycosylation sites[174], and distinguishing GPCRs from non-GPCRs[175]. Since GPCRs are membrane proteins and typical protein universe techniques for structure and interface prediction are usually trained on soluble proteins[176], more specialized approaches have been developed for predicting GPCR structure[177] and oligomerization[178]. Arguably, the second most important drug targets after GPCRs are the kinases[179]. With over 7,000 structures solved covering 308 kinases across 8 groups and complexed with over 3000 unique ligands and inhibitors, structure-based ML approaches are more prevalent for addressing challenges within this all-important superfamily. These include methods to predict inhibition[180], binding affinity[181] in specific kinase families, and conformational change between the so-called active and inactive conformations[182,183].

In the field of natural products and specialized metabolism in plants, bacteria, and fungi, ML has slowly been gaining popularity over more traditional approaches involving similarity search or analysis of a few, closely related proteins. ML has been used for successful identification of SM genes across a plant genome, while also identifying genomic features relevant to individual SM enzyme families[184]. Unsupervised ML can help identify clusters of co-expressed SM genes[185]. ML is also being incorporated to understand and engineer specific SM enzymes, with an early example where principal component analysis (PCA) was performed on enzyme mutants to identify quantitative structure-function relationships[186]. Codexis' ProSAR[187] attempts to determine the contribution of each residue to activity based on a training library of mutants, and iteratively designs the next library step for directed evolution using the influential positions found. In 2013, a Gaussian process model to predict thermostability was used to engineer highly thermostable cytochrome p450s[188]. Companies make use of ML for prioritizing strain candidates that perform well in their specific fermentation setups[189], and researchers have described efforts using ML for pathway optimization[190].

Predicting product specificity in SM enzyme families such as the STSs presents a number of challenges due to factors discussed throughout the previous sections. These include the complex nature of the relation between sequence similarity and functional similarity, their individual promiscuity and collective capability to produce hundreds of different molecules, and the sparsity of available experimental characterization data combined with the unavoidable noise in this data. Thus, we explore the design of a number of structural bioinformatics and ML techniques in this thesis and use these on STSs in an attempt to reveal novel insights in these elusive enzymes and to select new STSs for use in fragrance and flavour applications.

## 1.3 Thesis overview

This thesis brings protein structure bioinformatics and machine learning to enzymes involved in plant specialized metabolism, with a focus on sesquiterpene synthases (STSs).

In **Chapter 2** we collect and review experimentally characterized STSs from literature and analyse them from a sequence perspective. We conclude that phylogeny plays a larger role in sequence similarity than product specificity, necessitating the use of structural information to go further with predicting function. This is explored in **Chapter 3**, where we combine structural modelling, sequence co-evolution, and machine learning to predict the first step in STS product formation and pinpoint various structural regions involved in determining this step. To make better use of structure-derived features in a machine learning context, we create a novel multiple structure alignment algorithm, Caretta, in **Chapter 4** which is aimed at protein families with diverse sequences sharing a structural fold. Caretta combines alignment with automatic structure feature extraction in a visual and interactive tool that enables easy exploration of protein structures from different perspectives. In **Chapter 5** we delve into topological differences between proteins with Geometricus, a *k*-mer counting alternative for structures that uses the concept of rotation-invariant moments to define "shape-mers". **Chapter 6** combines the Geometricus algorithm with Caretta to greatly speed up multiple structure alignment, allowing for analyses involving many thousands of proteins. We demonstrate the use of both algorithms across different levels of protein hierarchy, and for pinpointing relevant residues and structural regions. All of these advances culminate in **Chapter 7**, where we develop a novel framework combining aligned structural features from proteins with chemical compound descriptors, to predict product specificity in STSs, and **Chapter 8**, where we develop an interactive data visualization portal allowing protein biologists to explore the interconnected properties of their protein family of interest from both a sequence and structure perspective.

I conclude this thesis in **Chapter 9**, with a discussion of the challenges and obstacles in understanding natural product enzymes and how a feedback loop between computational approaches and experimental design can help solve some of these challenges. I welcome the new era of large structure-rich datasets brought about by recent advances in structural bioinformatics, and discuss the wide range of opportunities this opens up in the field.

# References

[1] Willis, K. J. et al. (2017). *State of the World's Plants Report-2017*. Royal Botanic Gardens.

[2] Fang, C., Fernie, A. R., & Luo, J. (2019). Exploring the diversity of plant metabolism. *Trends in Plant Science*, *24*, 83–98.

[3] Meyer, A., Suhr, K., Nielsen, P., Holm, F. et al. (2002). Natural food preservatives. *Minimal Processing Technologies in the Food Industry*, (pp. 124–174).

[4] Caputi, L., & Aprea, E. (2011). Use of terpenoids as natural flavouring compounds in food industry. *Recent Patents on Food, Nutrition & Agriculture*, *3*, 9–16.

[5] Prasain, J., & Barnes, S. (2009). Recent advances in traditional medicines and dietary supplements. *Plant-derived Natural Products*, (pp. 533–546).

[6] S Mann, R., & E Kaufman, P. (2012). Natural product pesticides: Their development, delivery and use against insect vectors. *Mini-reviews in Organic Chemistry*, *9*, 185–202.

[7] Isman, M. B., & Akhtar, Y. (2007). Plant natural products as a source for developing environmentally acceptable insecticides. In *Insecticides Design Using Advanced Technologies* (pp. 235–248). Springer.

[8] Frey, C. (2005). *Natural Flavors and Fragrances: Chemistry, Analysis, and Production*. ACS Publications.

[9] Mooney, B. P. (2009). The second green revolution? Production of plant-based biodegradable plastics. *Biochemical Journal*, *418*, 219–232.

[10] Osbourn, A. E., & Lanzotti, V. (2009). *Plant-Derived Natural Products*. Springer.

[11] Lietava, J. (1992). Medicinal plants in a Middle Paleolithic grave Shanidar IV? *Journal of Ethnopharmacology*, *35*, 263–266.

[12] Petrovska, B. B. (2012). Historical review of medicinal plants' usage. *Pharmacognosy Reviews*, *6*, 1.

[13] Bourgaud, F., Gravot, A., Milesi, S., & Gontier, E. (2001). Production of plant secondary metabolites: A historical perspective. *Plant Science*, *161*, 839–851.

[14] Newman, D. J., & Cragg, G. M. (2007). Natural products as sources of new drugs over the last 25 years. *Journal of Natural Products*, *70*, 461–477.

[15] McChesney, J. D., Venkataraman, S. K., & Henri, J. T. (2007). Plant natural products: Back to the future or into extinction? *Phytochemistry*, *68*, 2015–2022.

[16] Casida, J. E. (1980). Pyrethrum flowers and pyrethroid insecticides. *Environmental Health Perspectives*, *34*, 189–202.

[17] Isman, M. B., Koul, O., Luczynski, A., & Kaminski, J. (1990). Insecticidal and antifeedant bioactivities of neem oils and their relationship to azadirachtin content. *Journal of Agricultural and Food Chemistry*, *38*, 1406–1411.

[18] Soloway, S. (1976). Naturally occurring insecticides. *Environmental Health Perspectives*, *14*, 109–117.

[19] Levy, L. W. (1981). A large-scale application of tissue culture: The mass propagation of pyrethrum clones in Ecuador. *Environmental and Experimental Botany*, *21*, 389–395.

[20] Chen, M.-S. (2008). Inducible direct plant defense against insect herbivores: A review. *Insect Science*, *15*, 101–114.

[21] Rosenthal, G. A., & Berenbaum, M. R. (2012). *Herbivores: Their Interactions with Secondary Plant Metabolites: Ecological and Evolutionary Processes* volume 2. Academic Press.

[22] Bouwmeester, H. J., Matusova, R., Zhongkui, S., & Beale, M. H. (2003). Secondary metabolite signalling in host–parasitic plant interactions. *Current Opinion in Plant Biology*, *6*, 358–364.

[23] Wani, Z. A., Ashraf, N., Mohiuddin, T., & Riyaz-Ul-Hassan, S. (2015). Plant-endophyte symbiosis, an ecological perspective. *Applied Microbiology and Biotechnology*, *99*, 2955–2965.

[24] Rizvi, S., Haque, H., Singh, V., & Rizvi, V. (1992). A discipline called allelopathy. In *Allelopathy* (pp. 1–10). Springer.

[25] Kessler, D., & Baldwin, I. T. (2007). Making sense of nectar scents: The effects of nectar secondary metabolites on floral visitors of *Nicotiana attenuata*. *The Plant Journal*, *49*, 840–854.

[26] Li, J., Ou-Lee, T.-M., Raba, R., Amundson, R. G., & Last, R. L. (1993). Arabidopsis flavonoid mutants are hypersensitive to UV-B irradiation. *The Plant Cell*, *5*, 171–179.

[27] Firn, R. D., & Jones, C. G. (2003). Natural products–a simple model to explain chemical diversity. *Natural Product Reports*, *20*, 382–391.

[28] Rasmann, S., & Agrawal, A. A. (2009). Plant defense against herbivory: Progress in identifying synergism, redundancy, and antagonism between resistance traits. *Current Opinion in Plant Biology*, *12*, 473–478.

[29] Carmona, D., Lajeunesse, M. J., & Johnson, M. T. (2011). Plant traits that predict resistance to herbivores. *Functional Ecology*, *25*, 358–367.

[30] Neilson, E. H., Goodger, J. Q., Woodrow, I. E., & Møller, B. L. (2013). Plant chemical defense: At what cost? *Trends in Plant Science*, *18*, 250–258.

[31] Speed, M. P., Fenton, A., Jones, M. G., Ruxton, G. D., & Brockhurst, M. A. (2015). Coevolution can explain defensive secondary metabolite diversity in plants. *New Phytologist*, *208*, 1251–1263.

[32] Gershenzon, J., & Dudareva, N. (2007). The function of terpene natural products in the natural world. *Nature Chemical Biology*, *3*, 408–414.

[33] Pichersky, E., Noel, J. P., & Dudareva, N. (2006). Biosynthesis of plant volatiles: Nature's diversity and ingenuity. *Science*, *311*, 808–811.

[34] Kreis, W., & Munkert, J. (2019). Exploiting enzyme promiscuity to shape plant specialized metabolism. *Journal of Experimental Botany*, *70*, 1435–1445.

[35] Pichersky, E., & Gang, D. R. (2000). Genetics and biochemistry of secondary metabolites in plants: An evolutionary perspective. *Trends in Plant Science*, *5*, 439–445.

[36] van der Hoeven, R. S., Monforte, A. J., Breeden, D., Tanksley, S. D., & Steffens, J. C. (2000). Genetic Control and Evolution of Sesquiterpene Biosynthesis in Lycopersicon esculentum and L. hirsutum. *The Plant Cell*, *12*, 2283–2294.

[37] Dudareva, N., Pichersky, E., & Gershenzon, J. (2004). Biochemistry of Plant Volatiles. *Plant Physiology*, *135*, 1893–1902.

[38] Kekulé, A. (1866). *Lehrbuch der organischen Chemie oder der Chemie der Kohlenstoffverbindungen*. F. Enke.

[39] Bohlmann, J., & Keeling, C. I. (2008). Terpenoid biomaterials. *The Plant Journal*, *54*, 656–669.

[40] Philippe, R. N., De Mey, M., Anderson, J., & Ajikumar, P. K. (2014). Biotechnological production of natural zero-calorie sweeteners. *Current Opinion in Biotechnology*, *26*, 155–161.

[41] Celedon, J. M., & Bohlmann, J. (2016). Chapter Three - Genomics-Based Discovery of Plant Genes for Synthetic Biology of Terpenoid Fragrances: A Case Study in Sandalwood oil Biosynthesis. In S. E. O'Connor (Ed.), *Methods in Enzymology* (pp. 47–67). Academic Press volume 576.

[42] Pateraki, I., Andersen-Ranberg, J., Jensen, N. B., Wubshet, S. G., Heskes, A. M., Forman, V., Hallström, B., Hamberger, B., Motawia, M. S., Olsen, C. E. et al. (2017). Total biosynthesis of the cyclic AMP booster forskolin from Coleus forskohlii. *eLife*, *6*, e23001.

[43] Zager, J. J., Lange, I., Srividya, N., Smith, A., & Lange, B. M. (2019). Gene networks underlying cannabinoid and terpenoid accumulation in cannabis. *Plant Physiology*, *180*, 1877–1897.

[44] Paddon, C. J. et al. (2013). High-level semi-synthetic production of the potent antimalarial artemisinin. *Nature*, *496*, 528–532.

[45] Peralta-Yahya, P. P., Zhang, F., del Cardayre, S. B., & Keasling, J. D. (2012). Microbial engineering for the production of advanced biofuels. *Nature*, *488*, 320–328.

[46] Van Beilen, J. B., & Poirier, Y. (2008). Production of renewable polymers from crop plants. *The Plant Journal*, *54*, 684–701.

[47] Chen, S.-L., Yu, H., Luo, H.-M., Wu, Q., Li, C.-F., & Steinmetz, A. (2016). Conservation and sustainable use of medicinal plants: Problems, progress, and prospects. *Chinese Medicine*, *11*, 37.

[48] Dagulo, L., Danyluk, M. D., Spann, T. M., Valim, M. F., Goodrich-Schneider, R., Sims, C., & Rouseff, R. (2010). Chemical characterization of orange juice from trees infected with citrus greening (Huanglongbing). *Journal of Food Science*, *75*, C199–C207.

[49] Bomgardner, M. M. (2012). The sweet smell of microbes. *Chemical & Engineering News*, *90*, 25–29.

[50] Degenhardt, J., Köllner, T. G., & Gershenzon, J. (2009). Monoterpene and sesquiterpene synthases and the origin of terpene skeletal diversity in plants. *Phytochemistry*, *70*, 1621–1637.

[51] DerMarderosian, A., Beutler, J. A. et al. (2002). *The review of natural products: the most complete source of natural product information*. Ed. 3. Facts and Comparisons.

[52] Starks, C. M., Back, K., Chappell, J., & Noel, J. P. (1997). Structural basis for cyclic terpene biosynthesis by tobacco 5-*epi*-aristolochene synthase. *Science*, *277*, 1815–1820.

[53] Greenhagen, B. T., O'Maille, P. E., Noel, J. P., & Chappell, J. (2006). Identifying and manipulating structural determinates linking catalytic specificities in terpene synthases. *Proceedings of the National Academy of Sciences*, *103*, 9826–9831.

[54] O'Maille, P. E., Malone, A., Dellas, N., Andes Hess, B., Smentek, L., Sheehan, I., Greenhagen, B. T., Chappell, J., Manning, G., & Noel, J. P. (2008). Quantitative exploration of the catalytic landscape separating divergent plant sesquiterpene synthases. *Nature Chemical Biology*, *4*, 617–623.

[55] Rising, K. A., Starks, C. M., Noel, J. P., & Chappell, J. (2000). Demonstration of germacrene A as an intermediate in 5-*epi*-aristolochene synthase catalysis. *Journal of the American Chemical Society*, *122*, 1861–1866.

[56] Noel, J. P., Dellas, N., Faraldos, J. A., Zhao, M., Hess, B. A., Smentek, L., Coates, R. M., & O'Maille, P. E. (2010). Structural elucidation of cisoid and transoid cyclization pathways of a sesquiterpene synthase using 2-fluorofarnesyl diphosphates. *ACS Chemical Biology*, *5*, 377–392.

[57] Koo, H. J., Vickery, C. R., Xu, Y., Louie, G. V., O'Maille, P. E., Bowman, M., Nartey, C. M., Burkart, M. D., & Noel, J. P. (2016). Biosynthetic potential of sesquiterpene synthases: Product profiles of Egyptian Henbane premnaspirodiene synthase and related mutants. *The Journal of Antibiotics*, *69*, 524–533.

[58] Gennadios, H. A., Gonzalez, V., Di Costanzo, L., Li, A., Yu, F., Miller, D. J., Allemann, R. K., & Christianson, D. W. (2009). Crystal structure of $(+)$-$\delta$-cadinene synthase from Gossypium arboreum and evolutionary divergence of metal binding motifs for catalysis. *Biochemistry*, *48*, 6175–6183.

[59] Chen, X.-Y., Chen, Y., Heinstein, P., & Davisson, V. J. (1995). Cloning, expression, and characterization of $(+)$-$\delta$-cadinene synthase: A catalyst for cotton phytoalexin biosynthesis. *Archives of Biochemistry and Biophysics*, *324*, 255–266.

[60] Bülow, N., & König, W. A. (2000). The role of germacrene D as a precursor in sesquiterpene biosynthesis: Investigations of acid catalyzed, photochemically and thermally induced rearrangements. *Phytochemistry*, *55*, 141–168.

[61] Faraldos, J. A., Miller, D. J., González, V., Yoosuf-Aly, Z., Cascón, O., Li, A., & Allemann, R. K. (2012). A 1,6-ring closure mechanism for $(+)$-$\delta$-cadinene synthase? *Journal of the American Chemical Society*, *134*, 5900–5908.

[62] McAndrew, R. P., Peralta-Yahya, P. P., DeGiovanni, A., Pereira, J. H., Hadi, M. Z., Keasling, J. D., & Adams, P. D. (2011). Structure of a three-domain sesquiterpene synthase: a prospective target for advanced biofuels production. *Structure*, *19*, 1876–1884.

[63] Steele, C. L., Crock, J., Bohlmann, J., & Croteau, R. (1998). Sesquiterpene synthases from grand fir (*Abies grandis*): Comparison of constitutive and wound-induced activities, and cDNA isolation, characterization, and bacterial expression of $\delta$-selinene synthase and $\gamma$-humulene synthase. *Journal of Biological Chemistry*, *273*, 2078–2089.

[64] Little, D. B., & Croteau, R. B. (2002). Alteration of product formation by directed mutagenesis and truncation of the multiple-product sesquiterpene synthases $\delta$-selinene synthase and $\gamma$-humulene synthase. *Archives of Biochemistry and Biophysics*, *402*, 120–135.

[65] Yoshikuni, Y., Ferrin, T. E., & Keasling, J. D. (2006). Designed divergent evolution of enzyme function. *Nature*, *440*, 1078–1082.

[66] Li, J.-X., Fang, X., Zhao, Q., Ruan, J.-X., Yang, C.-Q., Wang, L.-J., Miller, D. J., Faraldos, J. A., Allemann, R. K., Chen, X.-Y., & Zhang, P. (2013). Rational engineering of plasticity residues of sesquiterpene synthases from *Artemisia annua*: Product specificity and catalytic efficiency. *The Biochemical Journal*, *451*, 417–426.

[67] Salmon, M., Laurendon, C., Vardakou, M., Cheema, J., Defernez, M., Green, S., Faraldos, J. A., & O'Maille, P. E. (2015). Emergence of terpene cyclization in *Artemisia annua*. *Nature Communications*, *6*, 6143.

[68] Li, Z., Gao, R., Hao, Q., Zhao, H., Cheng, L., He, F., Liu, L., Liu, X., Chou, W. K. W., Zhu, H., & Cane, D. E. (2016). The T296V mutant of amorpha-4,11-diene synthase is defective in allylic diphosphate isomerization but retains the ability to cyclize the intermediate (3R)-nerolidyl diphosphate to amorpha-4,11-diene. *Biochemistry*, *55*, 6599–6604.

[69] Blank, P. N., Shinsky, S. A., & Christianson, D. W. (2019). Structure of sesquisabinene synthase 1, a terpenoid cyclase that generates a strained [3.1.0] bridged-bicyclic product. *ACS Chemical Biology*, *14*, 1011–1019.

[70] Bank, R. P. D. RCSB PDB - 5JO7: Henbane premnaspirodiene synthase (HPS), also known as Henbane vetispiradiene synthase (HVS) from *Hyoscyamus muticus*. https://www.rcsb.org/structure/5JO7.

[71] Singh, S., Thulasiram, H. V., Sengupta, D., & Kulkarni, K. (2021). Dynamic coupling analysis on plant sesquiterpene synthases provides leads for the identification of product specificity determinants. *Biochemical and Biophysical Research Communications*, *536*, 107–114.

[72] Di Girolamo, A., Durairaj, J., van Houwelingen, A., Verstappen, F., Bosch, D., Cankar, K., Bouwmeester, H., de Ridder, D., van Dijk, A. D. J., & Beekwilder, J. (2020). The santalene synthase from *Cinnamomum camphora*: Reconstruction of a sesquiterpene synthase from a monoterpene synthase. *Archives of Biochemistry and Biophysics*, *695*, 108647.

[73] Verma, N., & Shukla, S. (2015). Impact of various factors responsible for fluctuation in plant secondary metabolites. *Journal of Applied Research on Medicinal and Aromatic Plants*, *2*, 105–113.

[74] Frister, T., Hartwig, S., Alemdar, S., Schnatz, K., Thöns, L., Scheper, T., & Beutel, S. (2015). Characterisation of a recombinant patchoulol synthase variant for biocatalytic production of terpenes. *Applied Biochemistry and Biotechnology*, *176*, 2185–2201.

[75] Matasci, N. et al. (2014). Data access for the 1,000 Plants (1KP) project. *GigaScience*, *3*.

[76] Koski, T. (2001). *Hidden Markov models for bioinformatics* volume 2. Springer.

[77] Edgar, R. C., & Batzoglou, S. (2006). Multiple sequence alignment. *Current Opinion in Structural Biology*, *16*, 368–373.

[78] Karplus, K., Barrett, C., & Hughey, R. (1998). Hidden Markov models for detecting remote protein homologies. *Bioinformatics*, *14*, 846–856.

[79] Brandt, B. W., Feenstra, K. A., & Heringa, J. (2010). Multi-Harmony: Detecting functional specificity from sequence alignment. *Nucleic Acids Research*, *38*, W35–W40.

[80] Pal, S. K., Bandyopadhyay, S., & Ray, S. S. (2006). Evolutionary computation in bioinformatics: A review. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, *36*, 601–615.

[81] Jones, P. et al. (2014). InterProScan 5: Genome-scale protein function classification. *Bioinformatics*, *30*, 1236–1240.

[82] Bateman, A. et al. (2004). The Pfam protein families database. *Nucleic Acids Research*, *32*, D138–D141.

[83] Morcos, F., Pagnani, A., Lunt, B., Bertolino, A., Marks, D. S., Sander, C., Zecchina, R., Onuchic, J. N., Hwa, T., & Weigt, M. (2011). Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proceedings of the National Academy of Sciences*, *108*, E1293–E1301.

[84] Kamisetty, H., Ovchinnikov, S., & Baker, D. (2013). Assessing the utility of coevolution-based residue–residue contact predictions in a sequence- and structure-rich era. *Proceedings of the National Academy of Sciences*, *110*, 15674–15679.

[85] Lovell, S. C., & Robertson, D. L. (2010). An Integrated View of Molecular Coevolution in Protein–Protein Interactions. *Molecular Biology and Evolution*, *27*, 2567–2575.

[86] Little, D. Y., & Chen, L. (2009). Identification of coevolving residues and coevolution potentials emphasizing structure, bond formation and catalytic coordination in protein evolution. *PLoS One*, *4*, e4762.

[87] Illergård, K., Ardell, D. H., & Elofsson, A. (2009). Structure is three to ten times more conserved than sequence—A study of structural response in protein cores. *Proteins: Structure, Function, and Bioinformatics*, *77*, 499–508.

[88] Burley, S. K., Berman, H. M., Kleywegt, G. J., Markley, J. L., Nakamura, H., & Velankar, S. (2017). Protein Data Bank (PDB): the single global macromolecular structure archive. *Protein Crystallography*, (pp. 627–641).

[89] Kuhlman, B., & Bradley, P. (2019). Advances in protein structure prediction and design. *Nature Reviews Molecular Cell Biology*, *20*, 681–697.

[90] Budowski-Tal, I., Nov, Y., & Kolodny, R. (2010). FragBag, an accurate representation of protein structure, retrieves structural neighbors from the entire PDB quickly and accurately. *Proceedings of the National Academy of Sciences*, *107*, 3481–3486.

[91] Liu, Y., Ye, Q., Wang, L., & Peng, J. (2018). Learning structural motif representations for efficient protein structure search. *Bioinformatics*, *34*, i773–i780.

[92] Guzenko, D., Burley, S. K., & Duarte, J. M. (2020). Real time structural search of the protein data bank. *PLoS Computational Biology*, *16*, e1007970.

[93] Kihara, D., Sael, L., Chikhi, R., & Esquivel-Rodriguez, J. (2011). Molecular surface representation using 3D Zernike descriptors for protein shape comparison and docking. *Current Protein & Peptide Science*, *12*, 520–530.

[94] Yin, S., Proctor, E. A., Lugovskoy, A. A., & Dokholyan, N. V. (2009). Fast screening of protein surfaces using geometric invariant fingerprints. *Proceedings of the National Academy of Sciences of the United States of America*, *106*, 16622–16626.

[95] Zhu, X., Xiong, Y., & Kihara, D. (2015). Large-scale binding ligand prediction by improved patch-based method Patch-Surfer2.0. *Bioinformatics*, *31*, 707–713.

[96] Gainza, P., Sverrisson, F., Monti, F., Rodolà, E., Boscaini, D., Bronstein, M. M., & Correia, B. E. (2020). Deciphering interaction fingerprints from protein molecular surfaces using geometric deep learning. *Nature Methods*, *17*, 184–192.

[97] Wójcikowski, M., Kukiełka, M., Stepniewska-Dziubinska, M. M., & Siedlecki, P. (2019). Development of a protein–ligand extended connectivity (PLEC) fingerprint and its application for binding affinity predictions. *Bioinformatics*, *35*, 1334–1341.

[98] Liu, Z.-P., Wu, L.-Y., Wang, Y., Zhang, X.-S., & Chen, L. (2010). Prediction of protein–RNA binding sites by a random forest method with combined features. *Bioinformatics*, *26*, 1616–1622.

[99] Ashburner, M. et al. (2000). Gene Ontology: Tool for the unification of biology. *Nature Genetics*, *25*, 25–29.

[100] Rost, B. (2002). Enzyme function less conserved than anticipated. *Journal of Molecular Biology*, *318*, 595–608.

[101] Joshi, T., & Xu, D. (2007). Quantitative assessment of relationship between sequence similarity and function similarity. *BMC Genomics*, *8*, 222.

[102] Schnoes, A. M., Brown, S. D., Dodevski, I., & Babbitt, P. C. (2009). Annotation error in public databases: Misannotation of molecular function in enzyme superfamilies. *PLoS Computational Biology*, *5*, e1000605.

[103] Gilks, W. R., Audit, B., de Angelis, D., Tsoka, S., & Ouzounis, C. A. (2005). Percolation of annotation errors through hierarchically structured protein sequence databases. *Mathematical Biosciences*, *193*, 223–234.

[104] Mitchell, T. M. et al. (1997). *Machine learning* volume 45. McGraw Hill.

[105] Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. *Science*, *349*, 255–260.

[106] Peterson, L. E. (2009). K-nearest neighbor. *Scholarpedia*, *4*, 1883.

[107] Noble, W. S. (2006). What is a support vector machine? *Nature Biotechnology*, *24*, 1565–1567.

[108] Rasmussen, C. E. (2004). Gaussian Processes in Machine Learning. In O. Bousquet, U. von Luxburg, & G. Rätsch (Eds.), *Advanced Lectures on Machine Learning: ML Summer Schools 2003, Canberra, Australia, February 2 - 14, 2003, Tübingen, Germany, August 4 - 16, 2003, Revised Lectures* Lecture Notes in Computer Science (pp. 63–71). Berlin, Heidelberg: Springer.

[109] Breiman, L. (2001). Random Forests. *Machine Learning*, *45*, 5–32.

[110] Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, *29*, 1189–1232.

[111] Heckmann, D., Lloyd, C. J., Mih, N., Ha, Y., Zielinski, D. C., Haiman, Z. B., Desouki, A. A., Lercher, M. J., & Palsson, B. O. (2018). Machine learning applied to enzyme turnover numbers reveals protein structural correlates and improves metabolic models. *Nature Communications*, *9*, 5252.

[112] Fang, J. (2020). A critical review of five machine learning-based algorithms for predicting protein stability changes upon mutation. *Briefings in Bioinformatics*, *21*, 1285–1292.

[113] Yang, K. K., Wu, Z., & Arnold, F. H. (2019). Machine-learning-guided directed evolution for protein engineering. *Nature Methods*, *16*, 687–694.

[114] Keskin, O., Tuncbag, N., & Gursoy, A. (2016). Predicting protein-protein interactions from the molecular to the proteome level. *Chemical Reviews*, *116*, 4884–4909.

[115] Puton, T., Kozlowski, L., Tuszynska, I., Rother, K., & Bujnicki, J. M. (2012). Computational methods for prediction of protein–RNA interactions. *Journal of Structural Biology*, *179*, 261–268.

[116] Kauffman, C., & Karypis, G. (2012). Computational tools for protein–DNA interactions. *WIREs Data Mining and Knowledge Discovery*, *2*, 14–28.

[117] Audie, J., & Swanson, J. (2013). Advances in the prediction of protein-peptide binding affinities: Implications for peptide-based drug discovery. *Chemical Biology & Drug Design*, *81*, 50–60.

[118] Modarres, H. P., Mofrad, M. R., & Sanati-Nezhad, A. (2016). Protein thermostability engineering. *RSC Advances*, *6*, 115252–115270.

[119] Ain, Q. U., Aleksandrova, A., Roessler, F. D., & Ballester, P. J. (2015). Machine-learning scoring functions to improve structure-based binding affinity prediction and virtual screening. *WIREs Computational Molecular Science*, *5*, 405–424.

[120] Zhao, J., Cao, Y., & Zhang, L. (2020). Exploring the computational methods for protein-ligand binding site prediction. *Computational and Structural Biotechnology Journal*, *18*, 417–426.

[121] Alley, E. C., Khimulya, G., Biswas, S., AlQuraishi, M., & Church, G. M. (2019). Unified rational protein engineering with sequence-based deep representation learning. *Nature Methods*, *16*, 1315–1322.

[122] Bepler, T., & Berger, B. (2019). Learning protein sequence embeddings using information from structure. *ArXiv e-prints*.

[123] Heinzinger, M., Elnaggar, A., Wang, Y., Dallago, C., Nechaev, D., Matthes, F., & Rost, B. (2019). Modeling aspects of the language of life through transfer-learning protein sequences. *BMC Bioinformatics*, *20*, 723.

[124] Rives, A., Meier, J., Sercu, T., Goyal, S., Lin, Z., Liu, J., Guo, D., Ott, M., Zitnick, C. L., Ma, J., & Fergus, R. (2021). Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences*, *118*.

[125] Sledzieski, S., Singh, R., Cowen, L., & Berger, B. (2021). Sequence-based prediction of protein-protein interactions: a structure-aware interpretable deep learning model. *bioRxiv*.

[126] Yang, Y., Gao, J., Wang, J., Heffernan, R., Hanson, J., Paliwal, K., & Zhou, Y. (2018). Sixty-five years of the long march in protein secondary structure prediction: The final stretch? *Briefings in Bioinformatics*, *19*, 482–494.

[127] Heffernan, R., Paliwal, K., Lyons, J., Dehzangi, A., Sharma, A., Wang, J., Sattar, A., Yang, Y., & Zhou, Y. (2015). Improving prediction of secondary structure, local backbone angles, and solvent accessible surface area of proteins by iterative deep learning. *Scientific Reports*, *5*, 11476.

[128] Wang, S., Peng, J., Ma, J., & Xu, J. (2016). Protein secondary structure prediction using deep convolutional neural fields. *Scientific Reports*, *6*, 18962.

[129] Rost, B., & Sander, C. (1993). Improved prediction of protein secondary structure by use of sequence profiles and neural networks. *Proceedings of the National Academy of Sciences of the United States of America*, *90*, 7558–7562.

[130] Finkelstein, A., & Ptitsyn, O. (1971). Statistical analysis of the correlation among amino acid residues in helical, $\beta$-stractural and non-regular regions of globular proteins. *Journal of Molecular Biology*, *62*, 613–624.

[131] Chou, P. Y., & Fasman, G. D. (1974). Prediction of protein conformation. *Biochemistry*, *13*, 222–245.

[132] Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Research*, *25*, 3389–3402.

[133] Cuff, J. A., & Barton, G. J. (2000). Application of multiple sequence alignment profiles to improve protein secondary structure prediction. *Proteins: Structure, Function, and Bioinformatics*, *40*, 502–511.

[134] Dor, O., & Zhou, Y. (2007). Achieving 80% ten-fold cross-validated accuracy for secondary structure prediction by large-scale training. *Proteins*, *66*, 838–845.

[135] Cheng, H., Sen, T. Z., Jernigan, R. L., & Kloczkowski, A. (2007). Consensus Data Mining (CDM) protein secondary structure prediction server: Combining GOR V and Fragment Database Mining (FDM). *Bioinformatics*, *23*, 2628–2630.

[136] Li, D., Li, T., Cong, P., Xiong, W., & Sun, J. (2012). A novel structural position-specific scoring matrix for the prediction of protein secondary structures. *Bioinformatics*, *28*, 32–39.

[137] Saraswathi, S., Fernández-Martínez, J. L., Kolinski, A., Jernigan, R. L., & Kloczkowski, A. (2012). Fast learning optimized prediction methodology (FLOPRED) for protein secondary structure prediction. *Journal of Molecular Modeling*, *18*, 4275–4289.

[138] Magnan, C. N., & Baldi, P. (2014). SSpro/ACCpro 5: Almost perfect prediction of protein secondary structure and relative solvent accessibility using profiles, machine learning and structural similarity. *Bioinformatics*, *30*, 2592–2597.

[139] Faraggi, E., Zhang, T., Yang, Y., Kurgan, L., & Zhou, Y. (2012). SPINE X: Improving protein secondary structure prediction by multistep learning coupled with prediction of solvent accessible surface area and backbone torsion angles. *Journal of Computational Chemistry*, *33*, 259–267.

[140] Yaseen, A., & Li, Y. (2014). Context-based features enhance protein secondary structure prediction accuracy. *Journal of Chemical Information and Modeling*, *54*, 992–1002.

[141] Cheng, J., Randall, A. Z., Sweredoski, M. J., & Baldi, P. (2005). SCRATCH: A protein structure and structural feature prediction server. *Nucleic Acids Research*, *33*, W72–76.

[142] Shen, Y., & Bax, A. (2013). Protein backbone and sidechain torsion angles predicted from NMR chemical shifts using artificial neural networks. *Journal of Biomolecular NMR*, *56*, 227–241.

[143] Mataeimoghadam, F., Newton, M. A. H., Dehzangi, A., Karim, A., Jayaram, B., Ranganathan, S., & Sattar, A. (2020). Enhancing protein backbone angle prediction by using simpler models of deep neural networks. *Scientific Reports*, *10*, 19430.

[144] Göbel, U., Sander, C., Schneider, R., & Valencia, A. (1994). Correlated mutations and residue contacts in proteins. *Proteins: Structure, Function, and Bioinformatics*, *18*, 309–317.

[145] Marks, D. S., Colwell, L. J., Sheridan, R., Hopf, T. A., Pagnani, A., Zecchina, R., & Sander, C. (2011). Protein 3d structure computed from evolutionary sequence variation. *PLoS One*, *6*, e28766.

[146] Ma, J., Wang, S., Wang, Z., & Xu, J. (2015). Protein contact prediction by integrating joint evolutionary coupling analysis and supervised learning. *Bioinformatics*, *31*, 3506–3513.

[147] Xu, J. (2019). Distance-based protein folding powered by deep learning. *Proceedings of the National Academy of Sciences*, *116*, 16856–16865.

[148] Jones, D. T., & Kandathil, S. M. (2018). High precision in protein contact prediction using fully convolutional neural networks and minimal sequence features. *Bioinformatics*, *34*, 3308–3315.

[149] Wang, S., Sun, S., Li, Z., Zhang, R., & Xu, J. (2017). Accurate de novo prediction of protein contact map by ultra-deep learning model. *PLoS Computational Biology*, *13*, e1005324.

[150] Liu, Y., Palmedo, P., Ye, Q., Berger, B., & Peng, J. (2018). Enhancing evolutionary couplings with deep convolutional neural networks. *Cell Systems*, *6*, 65–74.e3.

[151] Ovchinnikov, S., Park, H., Varghese, N., Huang, P.-S., Pavlopoulos, G. A., Kim, D. E., Kamisetty, H., Kyrpides, N. C., & Baker, D. (2017). Protein structure determination using metagenome sequence data. *Science*, *355*, 294–298.

[152] Fiser, A. (2010). Template-based protein structure modeling. *Computational Biology*, (pp. 73–94).

[153] Cheng, J., & Baldi, P. (2006). A machine learning information retrieval approach to protein fold recognition. *Bioinformatics*, *22*, 1456–1463.

[154] Makigaki, S., & Ishida, T. (2020). Sequence alignment using machine learning for accurate template-based protein structure prediction. *Bioinformatics*, *36*, 104–111.

[155] Nguyen, S. P., Li, Z., Xu, D., & Shang, Y. (2017). New deep learning methods for protein loop modeling. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, *16*, 596–606.

[156] Wang, Z., Tegge, A. N., & Cheng, J. (2009). Evaluating the absolute quality of a single protein model using structural features and support vector machines. *Proteins: Structure, Function, and Bioinformatics*, *75*, 638–647.

[157] Kryshtafovych, A., Schwede, T., Topf, M., Fidelis, K., & Moult, J. (2019). Critical assessment of methods of protein structure prediction (CASP)—Round XIII. *Proteins: Structure, Function, and Bioinformatics*, *87*, 1011–1020.

[158] Lima, A. N., Philot, E. A., Trossini, G. H. G., Scott, L. P. B., Maltarollo, V. G., & Honorio, K. M. (2016). Use of machine learning approaches for novel drug discovery. *Expert Opinion on Drug Discovery*, *11*, 225–239.

[159] Lee, M., & Kim, D. (2012). Large-scale reverse docking profiles and their applications. *BMC Bioinformatics*, *13*, S6.

[160] Grinter, S. Z., Liang, Y., Huang, S.-Y., Hyder, S. M., & Zou, X. (2011). An inverse docking approach for identifying new potential anti-cancer targets. *Journal of Molecular Graphics & Modelling*, *29*, 795–799.

[161] Li, H., Leung, K.-S., Wong, M.-H., & Ballester, P. J. (2015). Improving AutoDock Vina using Random Forest: The growing accuracy of binding affinity prediction by the effective exploitation of larger data sets. *Molecular Informatics*, *34*, 115–126.

[162] Jiménez, J., Škalič, M., Martínez-Rosell, G., & De Fabritiis, G. (2018). KDEEP: Protein–Ligand Absolute Binding Affinity Prediction via 3D-Convolutional Neural Networks. *Journal of Chemical Information and Modeling*, *58*, 287–296.

[163] Ballester, P. J., & Mitchell, J. B. O. (2010). A machine learning approach to predicting protein–ligand binding affinity with applications to molecular docking. *Bioinformatics*, *26*, 1169–1175.

[164] Shen, C., Ding, J., Wang, Z., Cao, D., Ding, X., & Hou, T. (2020). From machine learning to deep learning: Advances in scoring functions for protein–ligand docking. *WIREs Computational Molecular Science*, *10*, e1429.

[165] Zhou, N., Jiang, Y., Bergquist, T. R., Lee, A. J., Kacsoh, B. Z., Crocker, A. W., Lewis, K. A., Georghiou, G., Nguyen, H. N., Hamid, M. N. et al. (2019). The CAFA challenge reports improved protein function prediction and new functional annotations for hundreds of genes through experimental screens. *Genome Biology*, *20*, 1–23.

[166] Bertoni, M., Kiefer, F., Biasini, M., Bordoli, L., & Schwede, T. (2017). Modeling protein quaternary structure of homo- and hetero-oligomers beyond binary interactions by homology. *Scientific Reports*, *7*, 10480.

[167] Liu, T., & Altman, R. B. (2009). Prediction of calcium-binding sites by combining loop-modeling with machine learning. *BMC Structural Biology*, *9*, 72.

[168] Maheshwari, S., & Brylinski, M. (2015). Prediction of protein–protein interaction sites from weakly homologous template structures using meta-threading and machine learning. *Journal of Molecular Recognition*, *28*, 35–48.

[169] Lise, S., Archambeau, C., Pontil, M., & Jones, D. T. (2009). Prediction of hot spot residues at protein-protein interfaces by combining machine learning and energy-based methods. *BMC Bioinformatics*, *10*, 365.

[170] Masso, M., & Vaisman, I. I. (2008). Accurate prediction of stability changes in protein mutants by combining machine learning with structure based computational mutagenesis. *Bioinformatics*, *24*, 2002–2009.

[171] Noé, F., De Fabritiis, G., & Clementi, C. (2020). Machine learning for protein folding and dynamics. *Current Opinion in Structural Biology*, *60*, 77–84.

[172] Seo, S., Choi, J., Ahn, S. K., Kim, K. W., Kim, J., Choi, J., Kim, J., & Ahn, J. (2018). Prediction of GPCR-ligand binding using machine learning algorithms. *Computational and Mathematical Methods in Medicine*, *2018*, e6565241.

[173] Hu, J., Li, Y., Yang, J.-Y., Shen, H.-B., & Yu, D.-J. (2016). GPCR–drug interactions prediction using random forest with drug-association-matrix-based post-processing procedure. *Computational Biology and Chemistry*, *60*, 59–71.

[174] Xie, H.-L., Fu, L., & Nie, X.-D. (2013). Using ensemble SVM to identify human GPCRs N-linked glycosylation sites based on the general form of Chou's PseAAC. *Protein Engineering, Design and Selection*, *26*, 735–742.

[175] Liao, Z., Ju, Y., & Zou, Q. (2016). Prediction of G protein-coupled receptors with SVM-Prot features and Random Forest. *Scientifica*, *2016*, e8309253.

[176] Barreto, C. A. V., Baptista, S. J., Preto, A. J., Matos-Filipe, P., Mourão, J., Melo, R., & Moreira, I. (2020). Chapter Four - Prediction and targeting of GPCR oligomer interfaces. In J. Giraldo, & F. Ciruela (Eds.), *Progress in Molecular Biology and Translational Science* (pp. 105–149). Academic Press volume 169.

[177] Wu, H., Wang, K., Lu, L., Xue, Y., Lyu, Q., & Jiang, M. (2017). Deep conditional random field approach to transmembrane topology prediction and application to GPCR three-dimensional structure modeling. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, *14*, 1106–1114.

[178] Townsend-Nicholson, A., Altwaijry, N., Potterton, A., Morao, I., & Heifetz, A. (2019). Computational prediction of GPCR oligomerization. *Current Opinion in Structural Biology*, *55*, 178–184.

[179] Cohen, P. (2002). Protein kinases — the major drug targets of the twenty-first century? *Nature Reviews Drug Discovery*, *1*, 309–315.

[180] Miljković, F., Rodríguez-Pérez, R., & Bajorath, J. (2019). Machine learning models for accurate prediction of kinase inhibitors with different binding modes. *Journal of Medicinal Chemistry*, *63*, 8738–8748.

[181] de Ávila, M. B., Xavier, M. M., Pintro, V. O., & de Azevedo, W. F. (2017). Supervised machine learning techniques to predict binding affinity. A study for cyclin-dependent kinase 2. *Biochemical and Biophysical Research Communications*, *494*, 305–310.

[182] McSkimming, D. I., Rasheed, K., & Kannan, N. (2017). Classifying kinase conformations using a machine learning approach. *BMC Bioinformatics*, *18*, 86.

[183] Ung, P. M.-U., Rahman, R., & Schlessinger, A. (2018). Redefining the protein kinase conformational space with machine learning. *Cell Chemical Biology*, *25*, 916–924.e2.

[184] Moore, B. M., Wang, P., Fan, P., Leong, B., Schenck, C. A., Lloyd, J. P., Lehti-Shiu, M. D., Last, R. L., Pichersky, E., & Shiu, S.-H. (2019). Robust predictions of specialized metabolism genes through machine learning. *Proceedings of the National Academy of Sciences*, *116*, 2344–2353.

[185] Dang, T.-T. T., Franke, J., Carqueijeiro, I. S. T., Langley, C., Courdavault, V., & O'Connor, S. E. (2018). Sarpagan bridge enzyme has substrate-controlled cyclization and aromatization modes. *Nature Chemical Biology*, *14*, 760–763.

[186] Damborský, J. (1998). Quantitative structure-function and structure-stability relationships of purposely modified proteins. *Protein Engineering, Design and Selection*, *11*, 21–30.

[187] Fox, R. J. et al. (2007). Improving catalytic function by ProSAR-driven enzyme evolution. *Nature Biotechnology*, *25*, 338–344.

[188] Romero, P. A., Krause, A., & Arnold, F. H. (2013). Navigating the protein fitness landscape with Gaussian processes, . *110*, E193–E201.

[189] Ramzi, A. B., Baharum, S. N., Bunawan, H., & Scrutton, N. S. (2020). Streamlining natural products biomanufacturing with omics and machine learning driven microbial engineering. *Frontiers in Bioengineering and Biotechnology*, *8*.

[190] Jervis, A. J. et al. (2019). Machine learning of designed translational control allows predictive pathway optimization in *Escherichia coli*. *ACS Synthetic Biology*, *8*, 127–136.

CHAPTER 2

# An analysis of characterized plant sesquiterpene synthases

Janani Durairaj*, Alice Di Girolamo*, Harro J. Bouwmeester, Dick de Ridder, Jules Beekwilder, and Aalt D.J. van Dijk

* authors contributed equally

This chapter has been published as

Durairaj, Janani, et al. “An analysis of characterized plant sesquiterpene synthases.” *Phytochemistry* 158 (2019): 157-165.

## Abstract

Plants exhibit a vast array of sesquiterpenes, C15 hydrocarbons which often function as herbivore-repellents or pollinator-attractants. These in turn are produced by a diverse range of sesquiterpene synthases. A comprehensive analysis of these enzymes in terms of product specificity has been hampered by the lack of a centralized resource of sufficient functionally annotated sequence data. To address this, we have gathered 262 plant sesquiterpene synthase sequences with experimentally characterized products. The annotated enzyme sequences allowed for an analysis of terpene synthase motifs, leading to the extension of one motif and recognition of a variant of another. In addition, putative terpene synthase sequences were obtained from various resources and compared with the annotated sesquiterpene synthases. This analysis indicated regions of terpene synthase sequence space which so far are unexplored experimentally. Finally, we present a case describing mutational studies on residues altering product specificity, for which we analysed conservation in our database. This demonstrates an application of our database in choosing likely-functional residues for mutagenesis studies aimed at understanding or changing sesquiterpene synthase product specificity.

## 2.1 Introduction

The terpenome represents a huge, ancient and diverse family of natural products. In addition to terpenes, it also encompasses steroids and carotenoids, comprising more than 60,000 members[1]. These compounds all derive from the same 5-carbon precursor units, coupled together linearly and then cyclized, rearranged, and modified in various ways. Terpenes serve many roles in plants, for example as toxins against herbivores or pathogens, or as attractants for pollinators[2]. In turn, terpenes extracted from plants are used by mankind for a range of applications - as pharmaceutical agents, insecticides, preservatives, fragrances, and flavours[3].

Terpenes are built from 5-carbon isoprenoid units, and they mainly exist as monoterpenes (C10), sesquiterpenes (C15) or diterpenes (C20), based on the number of such units used. In each case, a linear substrate loses a diphosphate group, usually cyclizes and then undergoes a variety of carbocation rearrangements. Though the exact number of sesquiterpenes found in nature is hard to determine, Tian et al.[4] estimated computationally that the number of sesquiterpene intermediates far outnumber those of monoterpenes, due to the increase in chain length.

Interestingly, sesquiterpenes found in nature can be divided into seven groups based on their parent cation and the first cyclization step in their formation[5]. Hence the extreme diversity of chemical compounds with desirable fragrances or medicinal properties is based on just seven initial carbocations. This makes the enzymes catalysing their formation both interesting and difficult to characterize functionally.

Each plant species is capable of synthesizing a number of sesquiterpenes using a specialized class of enzymes called sesquiterpene synthases (STSs). First, a farnesyl diphosphate synthase, produces the C15 substrate for STSs, farnesyl diphosphate

(FPP), from the C5-unit isopentenyl diphosphate (IPP) and its isomer dimethylallyl diphosphate (DMAPP)[6]. STSs then create the myriad of sesquiterpenes found in nature by catalysing carbocation formation from the linear FPP followed by a series of cyclizations and rearrangements (Figure 2.1). Products are formed from intermediate carbocations after deprotonation, phosphorylation, or hydration[4].

The STSs themselves represent a very diverse set of enzymes with a wide range of sequence similarities, despite having a common structural fold shared by plant, animal, fungal, and bacterial terpene synthases (TPSs)[7]. Hence, prediction of enzyme function from sequence is highly challenging in the case of STSs. Moreover, sequence diversity in STSs is not dependent on the products formed. This problem has been addressed so far by inspection of TPS structures[7] and by mutational analyses that attempt to change the product of a synthase with the smallest number of residue changes[8]. The former, though an attractive approach, is limited especially in plants due to the sparsity of experimentally determined structures, while the latter often leads to unnatural enzymes with lower catalytic activity than their wild-type parents. Characterization of multiple TPSs from the same species by the same study has allowed for some small-scale sequence comparison of those synthases[9,10]. However, no previous attempts have been made to compare all experimentally characterized plant STS sequences according to the products that they form. We have collated a curated database of plant STSs with characterized products from literature. This database can be accessed at `www.bioinformatics.nl/sesquiterpene/synthasedb`.

With this database and aforementioned product grouping scheme, the active domain sequences of 262 plant STSs were analysed in terms of the precursor carbocations of their products. These were also compared with the many yet-uncharacterized putative TPS enzymes. Residues from previous product-changing mutational studies were mapped on our database of enzymes, indicating conservation of the corresponding positions across groups of sequences forming different product cations. This demonstrates the usefulness of our database in finding residues involved in STS product specificity.

## 2.2 Results and Discussion

### 2.2.1 Database of characterized STSs

To obtain a comprehensive set of annotated STSs, our starting point was the SwissProt database, a subset of UniProt[11] in which a curated and annotated set of proteins is available. This provided a set of 104 STSs. In addition, we manually reviewed literature linked to enzymes with the characteristic TPS domain in TremBl, the uncurated subset of UniProt. In this way, the number of curated plant STS sequences with experimentally characterized product data in the database was more than doubled.

We present a database of 262 manually curated characterized plant STSs, shown in Table 2.1. The enzymes originate from a hundred different plant species and collectively account for the production of 117 different sesquiterpenes. Such a large number of possible products makes it difficult to find enough enzymes with the same product for a meaningful analysis of product specificity. To solve this, the sequences were divided into seven groups, making use of the sesquiterpene precursor carbocation scheme as specified by Degenhardt et al.[5], described in Figure 2.1. The reaction cascade of an STS is initiated by metal-mediated removal of the diphosphate anion in the FPP substrate, leading to the formation of a transoid (2*E*,6*E*)-farnesyl cation (farnesyl cation) which can undergo cyclization either via *10-exo-trig* or *11-endo-trig* cyclizations on the C10-C11 double bond to the resulting cations 1 or 2 respectively However, the farnesyl cation can also isomerize to form a cisoid (2*Z*,6*E*)-farnesyl cation (nerolidyl cation). The nerolidyl cation, in addition to a C1-attack (either via *10-exo-trig* or *11-endo-trig*) on the C10-C11 double bond to form cations 3 or 4, can also undergo cyclization at its C6-C7 double bond either via *6-exo-trig* or *7-endo-trig*, forming cations 5 or 6. These carbocations undergo multiple further skeletal rearrangements, cyclizations, hydride or methyl shifts, and other modifications to form the end products of the enzyme[5]. Along with this myriad of cyclic products, acyclic sesquiterpenes can also be formed from either the farnesyl or the nerolidyl cation through proton loss or addition of water[5,12,13]. This schematic of carbocations derived from FPP can be used to divide sesquiterpenes produced by plants into seven groups - both based on their parent cation (farnesyl or nerolidyl) and the first cyclization that occurs (by attack of the carbocation on the 10,1-; 11,1-; 6,1-; or 7,1-double bond; or acyclic). For an STS enzyme, the carbocation of its major product is then used to determine its group in Table 2.1.

This division of STSs is in general straightforward even when multiple products are formed by one enzyme. Specifically, of the 98 sequences which also have minor products (Supp. Table 2.1), only 17 have minor products whose precursor carbocation differs from the major product's. Nine of these produce acyclic products in addition to their major product. This could be the result of incomplete cyclization caused by premature termination of intermediates[14]. Eight enzymes in the database either produce (-)-germacrene D or they produce germacrene D and the chirality was not determined during the enzyme's characterization. (-)-germacrene D can be formed via a 10,1- or a 11,1- cyclization of the farnesyl cation (cation 1 or 2). Though each enzyme is likely to only follow one cyclization route to form its product, this route has so far not been determined, so these sequences are shown separately in Table 2.1 and in the remainder of the text. The existence of other sesquiterpenes which can be formed via different cyclization routes cannot be ruled out, however in our analysis we stick to the cyclization routes provided by IUBMB's *Enzyme Nomenclature* Supplement 24 (2018)[15] in order to determine the precursor carbocation for a given sesquiterpene.

The database contains 233 angiosperm STSs, 16 gymnosperm enzymes from coniferous species and 13 enzymes from nonseed plants such as mosses and ferns. As described by Jia et al.[16], the latter species have TPSs which are more related to microbial TPSs than those from spermatophytes. Information on each of the 262 enzymes, including the sequence, species, UniProt ID, products (major and minor), product type, and PubMed ID of the paper detailing its experimental characterization, is available as a web service at `www.bioinformatics.nl/sesquiterpene/synthasedb`. The service supports searching, sorting and downloading of all or subsets of the data.

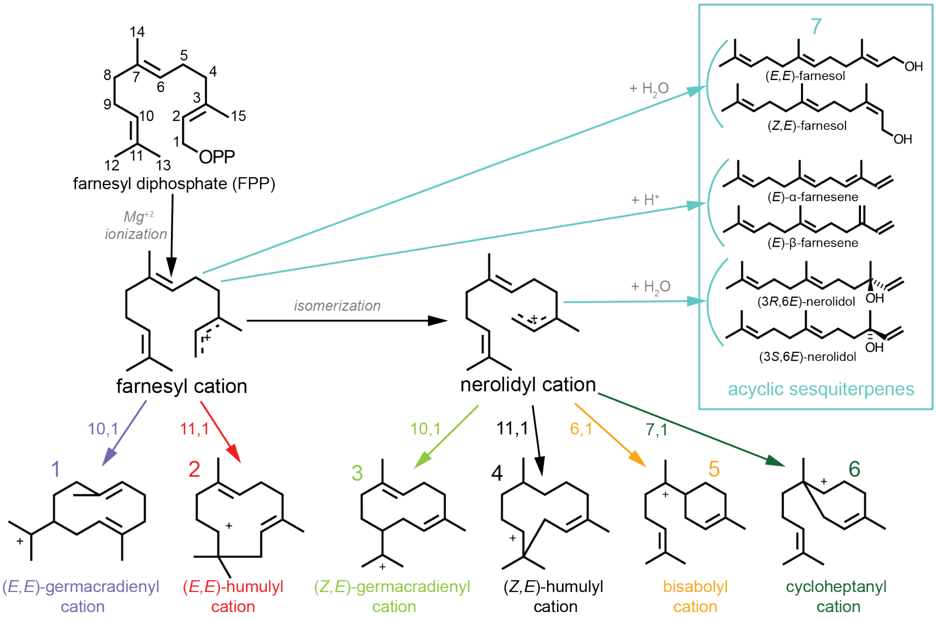


Figure 2.1: The reaction mechanism of sesquiterpene production starts with farnesyl diphosphate (FPP). Loss of the diphosphate moiety (OPP) leads to farnesyl cation formation. The farnesyl cation can subsequently be converted to the nerolidyl cation. Possible cyclizations for both cations are indicated in the figure. The subsequently formed cyclic cations undergo further modifications and rearrangements to form sesquiterpenes. An alternative route is to form acyclic sesquiterpenes from either the farnesyl or the nerolidyl cation as indicated in the box. These different product-precursors are used to classify the different sesquiterpenes and their synthases.

| Major Product Group | Cation/Cyclization | No. of sequences | | | | No. of species | | | | No. of products |
|---|---|---|---|---|---|---|---|---|---|---|
| | | A | G | N | Total | A | G | N | Total | |
| **1** | 10,1 / farnesyl | 77 | 1 | 3 | 81 | 44 | 1 | 3 | 48 | 43 |
| **2** | 11,1 / farnesyl | 42 | 3 | 3 | 48 | 32 | 3 | 3 | 38 | 11 |
| **3** | 10,1 / nerolidyl | 19 | 1 | 1 | 21 | 16 | 1 | 1 | 18 | 20 |
| **4** | 11,1 / nerolidyl | 0 | 4 | 0 | 4 | 0 | 4 | 0 | 4 | 3 |
| **5** | 6,1 / nerolidyl | 44 | 3 | 2 | 49 | 23 | 3 | 2 | 28 | 32 |
| **6** | 7,1 / nerolidyl | 0 | 0 | 1 | 1 | 0 | 0 | 1 | 1 | 1 |
| **7** | acyclic | 43 | 4 | 3 | 50 | 23 | 4 | 3 | 30 | 6 |
| - | (-)-germacrene D | 8 | 0 | 0 | 8 | 6 | 0 | 0 | 6 | 1 |
| **Total** | | 233 | 16 | 13 | 262 | 84 | 8 | 9 | 101 | 117 |

Table 2.1: Number of characterized plant STS sequences, species, and products covered in each product group. (-)-germacrene D synthases are shown separately as discussed in the text. A=Angiosperms, G=Gymnosperms, N=Nonseed

On average, the enzymes comprise $553 \pm 56$ residues. The tertiary structure of STS enzymes usually consists of two alpha-helical domains[17]. The N-terminal domain is considered relictual in plant STSs and is not present at all in nonseed plant STSs[16], while the C-terminal domain, consisting of an $\alpha$-helical bundle, is catalytically active[7,18]. The hydrophobic active site pocket in this domain is formed by six $\alpha$-helices, closed by two loops. Supp. Table 2.2 gives a list of plant STS structures from the Protein Data Bank (PDB)[19]. The C-terminal sub-sequences containing the active site are obtained from each enzyme in the database using information from Pfam[20], and consist of $266 \pm 7$ residues. N-terminal sub-sequences were extracted only from the spermatophyte enzymes in the database, again using information from Pfam, and consist of $173 \pm 12$ residues. In spermatophyte STSs, residues distal to the active site have been shown to contribute to product specificity potentially by influencing active site geometry[21]. These residues may reside in the extremities of the C-terminal domain, or in the N-terminal domain.

Supp. Figure 2.1 shows the pairwise sequence identity scores for each pair of C-terminal domain sub-sequences for the enzymes in the database, hierarchically clustered and coloured by product cation type. It can be seen that many pairs of sequences have less than 40% sequence identity. Similarly, Supp. Figure 2.2 shows the hierarchical clustering of concatenated N-terminal and C-terminal sub-sequences for spermatophyte enzymes. Both clusterings appear very comparable.

The phylogenetic tree of C-terminal sub-sequences of all 262 enzymes (Figure 2.2) shows some grouping of spermatophyte enzymes based on their product precursor. In general, the neighbour of an enzyme is from the same or related species, and if there are enough examples from the same species then some product-based grouping is seen. For example, the clades containing mostly enzymes from *Zea mays* on the right are separated based on the product carbocation of the enzyme even while being grouped by the species. However, this is not a consistent trend - enzymes from *Vitis* and *Santalum* at the top of the tree group mainly by species and not by product type. In fact, the three *Santalum* synthase sequences marked in Figure

2.2, making products derived from three different cyclic carbocations, have more than 90% in common. In any case, the product group of an enzyme from a species not present in the tree is nearly impossible to predict, while enzymes from species which are less represented in the tree can also be difficult to classify. In addition, clades forming predominantly one product carbocation are seen in many different parts of the tree, showing that strongly varying sequences can catalyse the same cyclization reaction and even produce the same product, such as the two marked $\beta$-caryophyllene synthases from *Arabidposis lyrata* and *Zea perennis* which have a sequence identity less than 30%. Hence, phylogenetic analysis is biased and cannot be an accurate predictor of TPS product specificity. Supp. Figure 2.3, shows a similar tree considering both N-terminal and C-terminal sub-sequences concatenated together, for spermatophyte STS sequences only. N-terminal domain information again does not seem to affect the structure of the tree. Even though this does not rule out the possibility that residues in the N-terminal domain influence product specificity, it indicates that including the N-terminal domain in the large scale sequence analysis that we perform does not add information compared to using only the C-terminal domain. Since product and intermediate formation occur in the active site pocket, it may be easier to find sequence-function determinants in the C-terminal domain. Hence, from this point on we concentrate on the C-terminal sub-sequences of TPSs.

The clade containing all the nonseed plant STSs in Figure 2.2 is clearly separate from the spermatophyte sequences. The enzyme from *Anthoceros punctatus*, a bryophyte, is the only sequence in the database producing a 7,1/nerolidyl-derived product ($\beta$-acoradiene) and is hence an out-group both in terms of species as well as product carbocation. Comparing nonseed plant sequences to the more typical plant TPS sequences would be futile, both due to their homology with microbial enzymes and their low numbers in the database, hence they are excluded from the remainder of the analysis.

### 2.2.2 Chemical similarities between sesquiterpenes

Each of the seven possible sesquiterpene precursors (Figure 2.1) usually undergoes a wide range of further rearrangements, cyclizations, and modifications, catalysed by the STS enzyme, to finally result in a sesquiterpene product. To start exploring the enzyme grouping scheme, we initially investigated whether similarities between the final sesquiterpene chemical structures would reflect the parent carbocations involved in their production. To this end, chemical similarities between sesquiterpenes with the same parent cation were compared to similarities between those without. Chemical similarities were measured using Dice similarity[22] between extended connectivity fingerprints, as described by Rogers & Hahn[23]. Similarities between 165 sesquiterpenes are plotted using multi-dimensional scaling (MDS), in Figure 2.3A, with the colour representative of the precursor cation. These 165 compounds collectively represent every enantiomer of the 117 sesquiterpenes produced by the enzymes in our database, since many of the experimental characterization studies used to build the database did not resolve the chirality of the STS's product. MDS is a technique used to visualize the level of similarity of individual objects in a dataset using a distance matrix, such that the between-object distances are preserved as well as possible. Therefore,

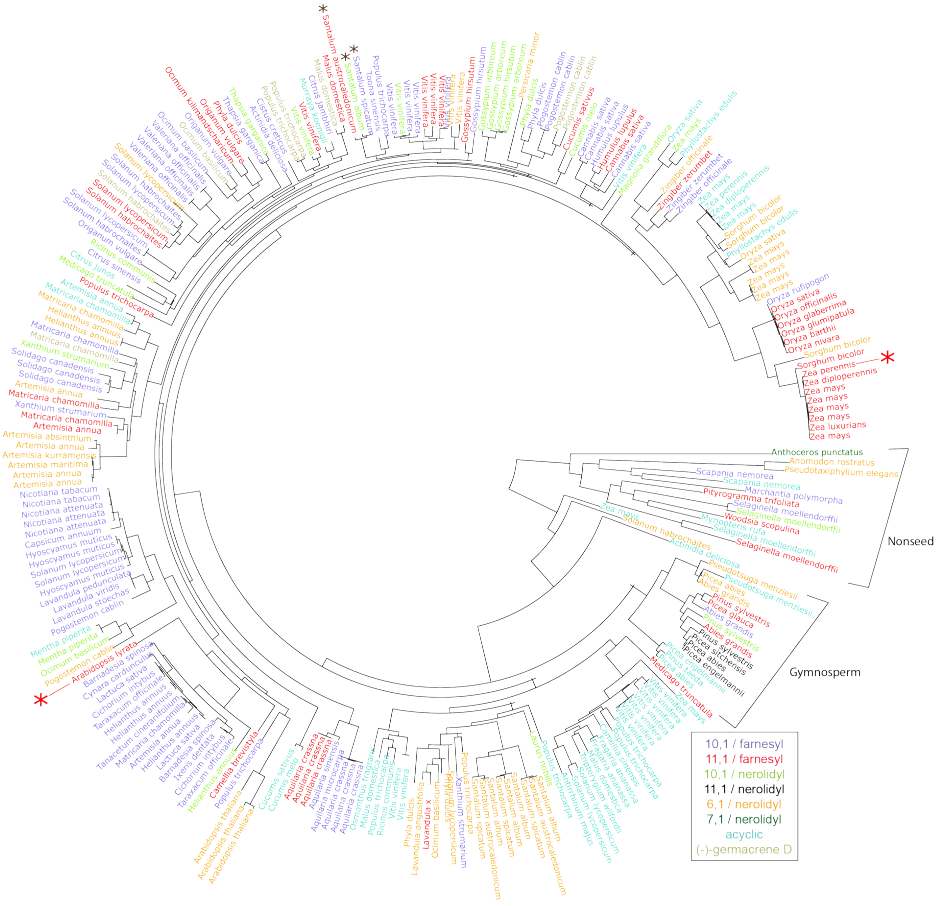


Figure 2.2: Phylogenetic tree of C-terminal sub-sequences for characterized plant STSs, coloured according to the major product's initial carbocation (see Figure 2.1). Nonseed and gymnosperm clades are indicated separately. Red and brown asterisks mark cases discussed in the text: red - two $\beta$-caryophyllene synthases from *Arabidopsis lyrata* and *Zea perennis* which have less than 30% pairwise sequence identity; brown - three synthases from *Santalum* with higher than 90% sequence identity.

two objects appearing close to each other in the MDS plot represent sesquiterpenes which likely have a high chemical similarity, while those further away have lower similarity. Acyclic sesquiterpenes are clearly distinguishable in the plot, as they are linear in nature. Interestingly, many products derived from the 6,1-cyclized cation (cation 5) are also distinct from those derived from 10,1- or 11,1-cyclized cations despite further cyclizations and rearrangements after this first step. They cluster midway between the acyclic and other cyclic products, which makes sense given the presence of an acyclic tail portion in cation 5. The sesquiterpenes formed from the other cyclic cations seem less distinguishable.

### 2.2.3 Characterized sequence space

Though a manual literature search gave us access to more functionally characterized TPS sequences, there is a large and steadily growing number of protein sequences present in various databases which have not been characterized at all. Many of these proteins are potential TPSs which contain the characteristic, catalytic site containing, C-terminal domain. Comparing uncharacterized and characterized enzymes may give indications of the nature of an uncharacterized enzyme, in particular about the cyclization route it is likely to take, thereby assisting in the setup of experiments for functional characterization.

To explore this, an MDS plot was made of C-terminal sub-sequences of the 249 spermatophyte enzymes in our database with those of 6278 other spermatophyte TPS-like sequences, obtained from sequenced genomes and transcriptomes. These 6278 sequences are, to the best of our knowledge, uncharacterized. Figure 2.3B shows this plot where the colours represent the product precursor carbocation of characterized STSs and the uncharacterized sequences are shown in grey. Similar sequences are depicted closer together in the plot.

Figure 2.3B has a few commonalities with the MDS plot of chemical similarities between sesquiterpenes, Figure 2.3A. Many sequences catalysing acyclic products as well as those derived from cation 5 cluster separately from the others. In fact, the enzymes making nerolidol, an acyclic sesquiterpene, cluster separately at the bottom right of the plot (light blue), leading us to hypothesize that perhaps many of the other uncharacterized STSs in this area also catalyse the formation of nerolidol. A second similarity is that enzymes forming products derived from 10,1- and 11,1-cyclized cations are difficult to distinguish. This again confirms, as was seen in the phylogenetic tree (Figure 2.2), that overall sequence similarity by itself cannot be an accurate guide to product specificity.

The uncharacterized sequences depicted in Figure 2.3B could be mono-, di-, or sesquiterpene synthases. Supp. Figure 2.4 shows 57 monoterpene synthases and 20 diterpene synthases from SwissProt, along with the 249 STSs in our database. Despite the skewed numbers, a separation between mono- and sesquiterpene synthases can be seen, indicating areas of the sequence space where more STSs are likely to be found.

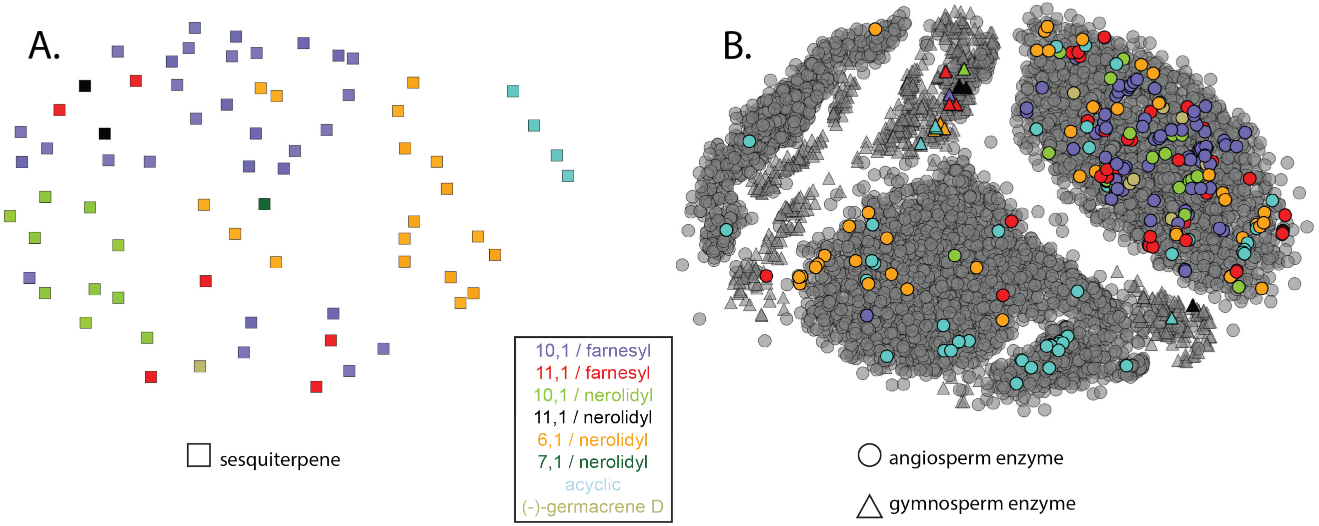


Figure 2.3: A. MDS plot of 165 sesquiterpenes found in nature, based on chemical fingerprint similarities. Each square represents a sesquiterpene and the more chemically similar two sesquiterpenes are, the closer they are placed in the plot. Colours are based on the sesquiterpene's precursor carbocation. B. MDS plot of TPS C-terminal domain sub-sequences with colouring based on STS major product carbocation. Unknown proteins which are likely to be TPSs are shown in grey. The more similar two sequences are, the closer they are in the plot.

Product specificity is even harder to identify in the case of gymnosperm synthases, as insufficient data is available to separate enzymes with different product cations. It has been noted before that gymnosperm TPSs resemble each other more than they do their angiosperm counterparts, regardless of catalytic activity[24,25]. The enzymes from these species may be more informative if analysed separately, but this would require more gymnosperm sequences to be functionally annotated.

### 2.2.4 Comparing known TPS motifs across sequences

A database such as ours allows for a comparison of residues in previously studied structural elements across many STS sequences. A thorough study of TPS structures has led to the identification of several motifs important for catalytic activity[7]. In the case of STSs, the hydrophobic moiety of the STS substrate, FPP, is directed into the active site cavity, to undergo the cyclizations and rearrangements described in Figure 2.1. Research on STS structures has proposed that the diphosphate moiety is captured by the motif RXR and divalent metal ions like $Mg^{2+}$ or $Mn^{2+}$, which are themselves bound by motifs DDXXD and NSE/DTE, at the entrance of the active site[26]. Here, we compare these three motifs across the sequences in our database. Figure 2.4A shows the motifs discussed below on a tobacco aristolochene synthase structure[26]. Figure 2.4B shows each motif on a schematic representation of the alignment of all C-terminal sub-sequences in the database.

#### Aspartate-rich DDXXD motif conserved in plant STSs

The most conserved motif of TPSs is the metal binding aspartate-rich motif found both in plant and microbial TPSs as well as in isoprenyl diphosphate synthases. Numerous studies performed on this motif, both site-directed mutagenesis and X-ray crystallography analysis, show that it is involved in binding the divalent metal ions in the active site entrance[27]. The canonical form of the motif, **D**DXX**(D,E)**, where bold-faced residues indicate those proposed to bind $Mg^{2+}$ or $Mn^{2+}$, is found in 247 of the 249 spermatophyte enzymes. Of the remaining two, one is a (+)-germacrene-D synthase from *Solidago canadensis* with an Asn replacing the first Asp[28]. The other is a bicyclogermacrene synthase from *Matricaria chamomilla* with an Asn replacing the second Asp[29]. These examples indicate that either one of the first two Aspartates may be sufficient for maintaining catalytic activity.

#### Expanded NSE/DTE motif found in most sequences

The opposite site of the active site entry is also involved in metal-binding, due to the presence of a second, less-defined motif, termed the NSE/DTE motif[30]. An early form of this motif, as detailed by Christianson [30] had a consensus of (L,V)(V,L,A)**(N,D)**D(L ,I,V)X**(S,T)**XXX**E**, where the residues in bold coordinate $Mg^{2+}$ ions. However, searching for a motif with this consensus only captured 38 of the 249 spermatophyte sequences in our database, indicating that it may be too restrictive given the current knowledge of sequences. When only the metal-binding portion of the motif is considered, the consensus sequence **(N,D)**DXX**(S,T,G)**XXX**E** covers 219 spermatophyte sequences in the database. The possibility of Gly in the second metal-binding

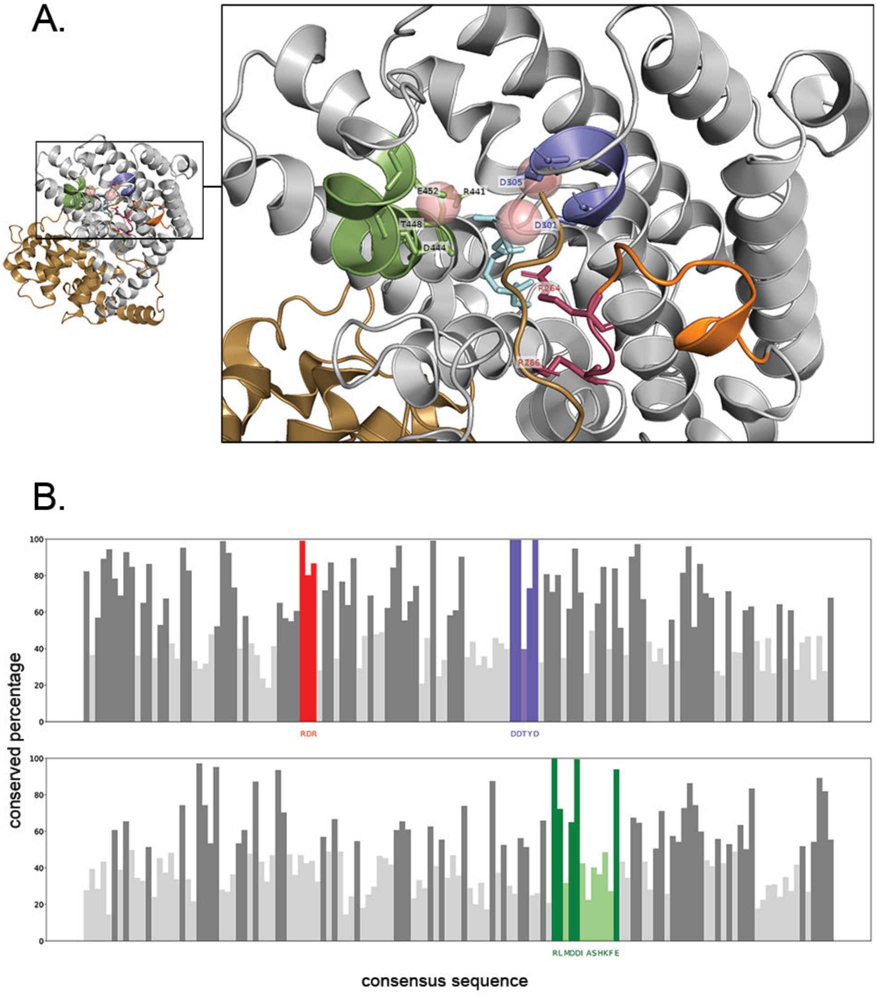


Figure 2.4: A. Known TPS motifs - RXR (red), DDXXD (purple) and NSE/DTE (green) shown on the structure of tobacco 5-epi-aristolochene synthase (PDB ID: 5EAT). The C-terminal domain is in grey while the N-terminal domain is in brown. Pink spheres represent $Mg^{2+}$ ions. A substrate analog, farnesyl hydroxyphosphonate (FHP) is in blue. The A-C loop is coloured in orange. The two conserved Arginines in the RXR motif, the metal-binding residues in the DDXXD and NSE/DTE motifs, and the Arginine in the expanded NSE/DTE motif discussed in the text are shown in stick representation. B. The same motifs shown on a schematic of the alignment of all spermatophyte C-terminal sub-sequences from the database. Each bar represents the percentage conservation of the consensus amino acid in the corresponding position of the alignment. Lighter coloured bars represent positions where the consensus amino acid is <50% conserved.

| Motif | No. of Sequences |
|---|---:|
| **D**DXX**T**XXX**E** | 57 |
| **D**DXX**S**XXX**E** | 55 |
| **N**DXX**S**XXX**E** | 44 |
| **D**DXX**G**XXX**E** | 25 |
| **N**DXX**T**XXX**E** | 22 |
| **N**DXX**G**XXX**E** | 16 |
| **D**DXX(**D**, **E**) | 20 |
| Other | 11 |

Table 2.2: Division of the different versions of the second metal-binding motif among characterized spermatophyte STS sequences. Sequences with motifs not covered by either motif consensus sequence **(N,D)**DXX**(S,T,G)**XXX**E** or **D**DXX**(D,E)** are classified as "Other".

position is justified by Zhou & Peters[31], with the proposal that Gly may allow a water molecule to substitute for the hydroxyl group of Ser/Thr. Some TPSs however, are known to have a second, catalytically active, aspartate rich motif instead of the NSE/DTE motif[32–34] with the same consensus as the first, **D**DXX**(D,E)**. This occurs in 20 sequences. Table 2.2 shows the distribution of the sequences over the different versions of the second motif.

A highly conserved Arg is found 3 residues upstream of all versions of the NSE/DTE motif or second aspartate-rich motif, in all of the spermatophyte sequences in the database. All 6278 uncharacterized spermatophyte TPS sequences also have an arginine in this position. Hence, an extended form of the motif may be more relevant for spermatophyte STSs, with the consensus RXX**(N,D)**DXX**(S,T,G)**XXX**E** or RXX**D**DXX**(D,E)**.

### RXR motif not conserved in nerolidol synthases

The RXR motif is found about 35 amino acids upstream of the DDXXD motif, located on a flexible loop in the structure, termed the A-C loop. This loop has been shown to become ordered upon FPP binding[26]. The two Arg residues in the motif were proposed to be involved in the complexation of diphosphate after ionization of the substrate, thereby preventing nucleophilic attack on any of the carbocationic intermediates[26]. 215 of the 249 spermatophyte plant sequences have the canonical RXR motif, while 18 of the remaining have an altered RXQ motif in the same region. Interestingly, these 18 enzymes all catalyse the formation of nerolidol, an acyclic sesquiterpene. This indicates that RXQ may be unable to capture diphosphate to the same extent as RXR, causing a premature quenching of an intermediate carbocation by water before cyclization has occurred[5].

### 2.2.5 Comparing residues involved in product specificity across sequences

Many studies have addressed the importance of specific residues located in the active site of TPSs via mutational analyses. Some of the best characterized TPSs derive from *Artemisia annua*, which is the source of many medicinal terpenes. Some of the STSs from *A. annua* have served as examples to identify residues involved in critical steps in the cyclization cascade. In this section three examples of *A. annua* STSs are described, for which residues involved in product specificity were experimentally investigated. We use these as a case-study to illustrate how the large set of characterized STSs that we make available can potentially be used to guide such experimental investigations. These examples are:

1. Salmon et al.[35] tested a wide library of mutants for the ($E$)-$\beta$-farnesene synthase (UniProt: Q9FXY7) from *A. annua*, an STS catalysing the formation of an acyclic product. They discovered that a single substitution, Tyr402Leu, confers to the synthase a cyclase activity, resulting in zingiberene and $\beta$-bisabolene as the most abundant products. Both these sesquiterpenes derive from cation 5.

   In sequences catalysing the formation of 10,1 and 11,1 cyclized products (cations 1, 2, 3 and 4), this position is highly conserved (88-100%) in the database as a Tyr, and Leu does not occur. However, STSs producing cation 5 and those producing acyclic products have relatively lower conservation in this position (70% Tyr and 53% Phe respectively) and Leu is found 14% of the time in cation 5. Thus, conservation patterns in this position are indicative of the corresponding residue's contribution to product specificity.

2. In another study, Li et al.[36] studied the effect of mutations on the cyclization reaction of the bisabolol synthase from *A. annua* (UniProt: M4HZ33). A possible reaction mechanism involves formation of a nerolidyl cation, followed by the formation of cation 5 by a 1,6 ring closure, and deprotonation to produce the final product bisabolol[37]. The authors identified a mutation that interfered with this 1,6 ring closure and showed that the substitution Leu399Thr changed the product specificity, to $\gamma$-humulene, derived from cation 2, a 11,1 cyclization of the farnesyl cation[36].

   Interestingly, a Leu at this position is quite rare; it is present in only four sequences in the database, all four of which belong to the group of sequences producing cation 5. Instead, this position is highly conserved (>95%) as either a Ser or a Thr in the database.

3. Amorpha-4,11-diene is a bicyclic sesquiterpene produced from the 6,1-cyclized bisabolyl cation, cation 5 in Table 2.1. Li et al.[38] did a mutational analysis of the amorpha-4,11-diene synthase from *A. annua* (UniProt: Q9AR04), and showed that the residue Thr296 can cause a loss of cyclization activity when mutated.

   This residue is 82% conserved as either a Ser or a Thr in cyclic STSs. Importantly, in acyclic STSs the most common amino acid is a Tyr, with a conservation of 38%. Acyclic STSs even have amino acids such as Gln, Gly and Ile in this position, never seen in the cyclic STSs in the database. The variability and low conservation score indicates that changing this position in cyclic STSs away from a Ser or Thr could result in the formation of acyclic products, as shown by Li et al.[38].

In summary, analysis of these *A. annua* examples of residues involved in the first cyclization step in STSs indicates that conservation patterns across all the annotated enzymes are consistent with the functional roles of these residues. This suggests it would be possible to obtain residues potentially involved in product specificity from this database. Such a data-driven approach is in contrast to how these mutational studies have traditionally been guided, i.e. by comparison of two or three sequences from the same or related species. Therefore, a potential application of our database is in guiding site-directed mutagenesis studies in a way which avoids species bias and hence may reveal additional residues involved in product specificity. One such residue position obtained by studying conservation patterns has been discussed above in Section 2.2.4, namely the second arginine in the RXR motif. This position was found to be glutamine in most nerolidol synthases, something not seen in any of the cyclic synthases. Mutating this residue in cyclic synthases and monitoring for acyclic products, and vice versa, could confirm the residue's role in the cyclization of sesquiterpene products.

## 2.3 Conclusion

We compiled a manually curated set of experimentally characterized plant STSs along with their major products. This database is the largest centralized resource of annotated plant STSs to date and allows for thorough sequence-based analysis of these diverse enzymes. The enzymes in the database are grouped according to the carbocationic origin and cyclization of their major product. Such a division alleviates the task of functional analysis and comparison between the enzymes. Using the database we were able to extend and find variants of existing STS motifs. In addition, residues from previous mutational studies, when mapped onto the enzymes in the database, were found to have detectable conservation patterns that differed from group to group. Such properties of residues can be extrapolated and used to guide further mutational studies. The database as a whole helps to understand the current state of STS sequence space characterization, and provides a starting point for future efforts to predict product specificity.

## 2.4 Methods

### 2.4.1 Literature search for characterized STSs

To find potentially characterized STSs, an HMM search was performed using hmmer (version 3.1b2)[39] on the UniProt database[40] using the HMM of the C-terminal domain of TPSs from Pfam[20] (Pfam ID: PF03936). Protein sequences with a hit having an E-value $< 10^{-10}$ and a total protein length between 350 and 650 residues were selected. The UniProt IDs of these sequences were then linked to PubMed IDs, either directly through programmatic access of UniProt if the PubMed ID was present, or through a programmatic text search of the title and authors given in UniProt, using the PubMed API[41]. The PubMed articles thus obtained were searched manually for evidence of experimental characterization of sesquiterpenes through in-vivo or in-vitro GC-MS studies, and the corresponding UniProt IDs were collected.

For each UniProt ID found, the major product described in the corresponding paper was stored. Minor products with GC-MS peaks at least quarter the height of the major product peak were stored as well.

### 2.4.2 Measuring chemical similarities

The diagram of the sesquiterpene grouping scheme was made using ChemDoodle (version 9)[42]. The InChI strings for 165 sesquiterpenes were obtained from PubChem[43] using the Python wrapper for the PubChem REST API[44], PubChemPy (version 1.0.4). To measure the similarity between different sesquiterpenes, rdkit (Release 2017.09.3) was used[45]. A circular chemical fingerprint, called the Morgan fingerprint, with a radius of 2 angstroms, as explained by Rogers & Hahn[23], was obtained for each sesquiterpene. The similarity between every pair of fingerprints was then calculated using Dice similarity[22]. The distance was given as $1 - similarity$. The distance matrix of all sesquiterpenes was then used to create a multi-dimensional scaling (MDS) plot using the Python scikit-learn library (version 0.19.1)[46], and then plotted using matplotlib (version 2.1.2)[47].

### 2.4.3 Aligning sequences

For characterized spermatophyte plant STS sequences, the C-terminal catalytically active portion and the N-terminal portion of the enzyme were found with hmmer HMM searches (version 3.1b2)[39] using the TPS C-terminal Pfam domain (Pfam ID: PF03936) and the TPS N-terminal Pfam domain (Pfam ID: PF01397) respectively. These were then separately aligned using Clustal Omega (version 1.2.4)[48], with all heuristic features off and the respective Pfam domains as a guide for alignment. From these separate alignments, a concatenated N+C alignment was formed, covering both domains.

For some of the nonseed plant STS sequences however, a C-terminal Pfam domain search returned <200 residues instead of the usual 250-270. Aligning the full nonseed sequences using the spermatophyte C-terminal sub-sequence alignment as a profile showed the position of the C-terminal portion for these sequences, so this was used to extract the required C-terminal sub-sequences for nonseed plants. An alignment consisting of both seed and nonseed characterized C-terminal sub-sequences was constructed using Clustal Omega with the same parameters as above.

### 2.4.4 Phylogenetic tree construction

A phylogenetic tree was built and visualized for the characterized spermatophyte and nonseed plant enzymes in the database using the ETE toolkit (version 3.1.1)[49]. The previously explained alignment of all C-terminal sub-sequences was used, with columns having >50% gaps removed using trimAL[50]. The best protein model from

JTT, WAG, VT, LG and MtREV was chosen using ProtTest[51], and finally a RaxML maximum likelihood tree was built[52]. Similarly, a phylogenetic tree for the spermatophyte sequences was built with the same approach using the concatenated N+C alignment.

### 2.4.5 Finding mono-, di-, and uncharacterized TPSs

Characterized plant mono- and diterpene synthases were obtained from SwissProt[11] using a C-terminal TPS Pfam domain hmmer (verson 3.1b2)[39] HMM search followed by collecting the sequences from plant species for which the catalytic activity was mentioned. These were not manually checked.

Uncharacterized TPS C-terminal sub-sequences were then obtained from plant species in TremBl[11], Ensembl Plants (release 38)[53], and the 1000 Plants Transcriptome Project[54] again using a Pfam domain search. Only those sequences where the search returned a sub-sequence having both DDXX(D,E) and (N,D)DXX(S,T,G)XXXE or two DDXX(D,E) motifs within it, and whose sub-sequence length was within two standard deviations of the mean C-terminal sub-sequence length of characterized STS enzymes were retained. In both sets, sequences from nonseed plant species were discarded.

### 2.4.6 Measuring sequence similarities

A distance matrix of all spermatophyte TPS C-terminal sub-sequences: characterized mono-, di- and sesquiterpene synthases as well as uncharacterized enzymes, was constructed using the pairwise sequence k-tuple measure described by Wilbur & Lipman[55], implemented in Clustal Omega (version 1.2.4)[48]. This distance matrix was then used to construct an MDS plot using scikit-learn (version 0.19.1)[46] and plotted using matplotlib (version 2.1.2)[47]. A cluster-map of sequence identities between characterized STS enzymes was made using the distance matrix of just these enzymes and complete hierarchical clustering using scipy (version 1.0.0)[56] and seaborn (version 0.8.1)[57].

### 2.4.7 Visualizing an STS structure

The tobacco 5-epi-aristolochene synthase structure from the Protein Data Bank (PDB)[58] with PDB ID 5EAT was used to visualize known TPS motifs, along with $Mg^{2+}$ ions and farnesyl hydroxyphosphonate (FHP) substrate analog. Visualization was done using PyMOL 2.1[59].

## Acknowledgements

This work is part of the research programme Novel Enzymes for Flavour and Fragrance with project number TTW 15043 which is financed by the Netherlands Organisation for Scientific Research (NWO).

## Supplementary information

All supplementary sections, figures, and tables are available online at `https://doi.org/10.1016/j.phytochem.2018.10.020`

## References

[1] Buckingham, J. (1997). *Dictionary of Natural Products, Supplement 4* volume 11. CRC Press.

[2] Gershenzon, J., & Dudareva, N. (2007). The function of terpene natural products in the natural world. *Nature Chemical Biology*, *3*, 408–414.

[3] Schempp, F. M., Drummond, L., Buchhaupt, M., & Schrader, J. (2017). Microbial cell factories for the production of terpenoid flavor and fragrance compounds. *Journal of Agricultural and Food chemistry*, *66*, 2247–2258.

[4] Tian, B., Poulter, C. D., & Jacobson, M. P. (2016). Defining the product chemical space of monoterpenoid synthases. *PLoS Computational Biology*, *12*, e1005053.

[5] Degenhardt, J., Köllner, T. G., & Gershenzon, J. (2009). Monoterpene and sesquiterpene synthases and the origin of terpene skeletal diversity in plants. *Phytochemistry*, *70*, 1621–1637.

[6] Ogura, K., Koyama, T., & Sagami, H. (1997). Polyprenyl diphosphate synthases. In *Cholesterol* (pp. 57–87). Springer.

[7] Gao, Y., Honzatko, R. B., & Peters, R. J. (2012). Terpenoid synthase structures: A so far incomplete view of complex catalysis. *Natural Product Reports*, *29*, 1153–1175.

[8] Segura, M. J., Jackson, B. E., & Matsuda, S. P. (2003). Mutagenesis approaches to deduce structure–function relationships in terpene synthases. *Natural Product Reports*, *20*, 304–317.

[9] Martin, D. M., Aubourg, S., Schouwey, M. B., Daviet, L., Schalk, M., Toub, O., Lund, S. T., & Bohlmann, J. (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*, 226.

[10] Martin, D. M., Fäldt, J., & Bohlmann, J. (2004). Functional characterization of nine Norway spruce TPS genes and evolution of gymnosperm terpene synthases of the TPS-d subfamily. *Plant Physiology*, *135*, 1908–1927.

[11] Boeckmann, B., Bairoch, A., Apweiler, R., Blatter, M.-C., Estreicher, A., Gasteiger, E., Martin, M. J., Michoud, K., O'Donovan, C., Phan, I., Pilbout, S., & Schneider, M. (2003). The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Research*, *31*, 365–370.

[12] Bairoch, A. (2000). The ENZYME database in 2000. *Nucleic Acids Research*, *28*, 304–305.

[13] Christianson, D. W. (2017). Structural and chemical biology of terpenoid cyclases. *Chemical Reviews*, *117*, 11570–11648.

[14] Köllner, T. G., Gershenzon, J., & Degenhardt, J. (2009). Molecular and biochemical evolution of maize terpene synthase 10, an enzyme of indirect defense. *Phytochemistry*, *70*, 1139–1145.

[15] Webb, E. C. et al. (1992). *Enzyme Nomenclature 1992. Recommendations of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology on the Nomenclature and Classification of Enzymes*. Ed. 6. Academic Press.

[16] Jia, Q., Köllner, T. G., Gershenzon, J., & Chen, F. (2018). MTPSLs: New terpene synthases in nonseed plants. *Trends in Plant Science*, *23*, 121–128.

[17] Cao, R., Zhang, Y., Mann, F. M., Huang, C., Mukkamala, D., Hudock, M. P., Mead, M. E., Prisic, S., Wang, K., Lin, F.-Y. et al. (2010). Diterpene cyclases and the nature of the isoprene fold. *Proteins: Structure, Function, and Bioinformatics*, *78*, 2417–2432.

[18] Joly, A., & Edwards, P. A. (1993). Effect of site-directed mutagenesis of conserved aspartate and arginine residues upon farnesyl diphosphate synthase activity. *Journal of Biological Chemistry*, *268*, 26983–26989.

[19] Burley, S. K., Berman, H. M., Kleywegt, G. J., Markley, J. L., Nakamura, H., & Velankar, S. (2017). Protein Data Bank (PDB): the single global macromolecular structure archive. *Protein Crystallography*, (pp. 627–641).

[20] Bateman, A. et al. (2004). The Pfam protein families database. *Nucleic Acids Research*, *32*, D138–D141.

[21] Greenhagen, B. T., O'Maille, P. E., Noel, J. P., & Chappell, J. (2006). Identifying and manipulating structural determinates linking catalytic specificities in terpene synthases. *Proceedings of the National Academy of Sciences*, *103*, 9826–9831.

[22] Willett, P., Barnard, J. M., & Downs, G. M. (1998). Chemical similarity searching. *Journal of chemical information and computer sciences*, *38*, 983–996.

[23] Rogers, D., & Hahn, M. (2010). Extended-connectivity fingerprints. *Journal of Chemical Information and Modeling*, *50*, 742–754.

[24] Chen, F., Tholl, D., Bohlmann, J., & Pichersky, E. (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*, *66*, 212–229.

[25] Trapp, S. C., & Croteau, R. B. (2001). Genomic organization of plant terpene synthases and molecular evolutionary implications. *Genetics*, *158*, 811–832.

[26] Starks, C. M., Back, K., Chappell, J., & Noel, J. P. (1997). Structural basis for cyclic terpene biosynthesis by tobacco 5-*epi*-aristolochene synthase. *Science*, *277*, 1815–1820.

[27] Aaron, J. A., & Christianson, D. W. (2010). Trinuclear metal clusters in catalysis by terpenoid synthases. *Pure and Applied Chemistry*, *82*, 1585–1597.

[28] Prosser, I., Altug, I. G., Phillips, A. L., König, W. A., Bouwmeester, H. J., & Beale, M. H. (2004). Enantiospecific (+)-and (-)-germacrene D synthases, cloned from goldenrod, reveal a functionally active variant of the universal isoprenoid-biosynthesis aspartate-rich motif. *Archives of Biochemistry and Biophysics*, *432*, 136–144.

[29] Son, Y.-J., Kwon, M., Ro, D.-K., & Kim, S.-U. (2014). Enantioselective microbial synthesis of the indigenous natural product (-)-$\alpha$-bisabolol by a sesquiterpene synthase from chamomile (*Matricaria recutita*). *Biochemical Journal*, *463*, 239–248.

[30] Christianson, D. W. (2006). Structural biology and chemistry of the terpenoid cyclases. *Chemical Reviews*, *106*, 3412–3442.

[31] Zhou, K., & Peters, R. J. (2009). Investigating the conservation pattern of a putative second terpene synthase divalent metal binding motif in plants. *Phytochemistry*, *70*, 366–369.

[32] Gennadios, H. A., Gonzalez, V., Di Costanzo, L., Li, A., Yu, F., Miller, D. J., Allemann, R. K., & Christianson, D. W. (2009). Crystal structure of (+)-$\delta$-cadinene synthase from Gossypium arboreum and evolutionary divergence of metal binding motifs for catalysis. *Biochemistry*, *48*, 6175–6183.

[33] Steele, C. L., Crock, J., Bohlmann, J., & Croteau, R. (1998). Sesquiterpene synthases from grand fir (*Abies grandis*): Comparison of constitutive and wound-induced activities, and cDNA isolation, characterization, and bacterial expression of $\delta$-selinene synthase and $\gamma$-humulene synthase. *Journal of Biological Chemistry*, *273*, 2078–2089.

[34] Little, D. B., & Croteau, R. B. (2002). Alteration of product formation by directed mutagenesis and truncation of the multiple-product sesquiterpene synthases $\delta$-selinene synthase and $\gamma$-humulene synthase. *Archives of Biochemistry and Biophysics*, *402*, 120–135.

[35] Salmon, M., Laurendon, C., Vardakou, M., Cheema, J., Defernez, M., Green, S., Faraldos, J. A., & O'Maille, P. E. (2015). Emergence of terpene cyclization in *Artemisia annua*. *Nature Communications*, *6*, 6143.

[36] Li, J.-X., Fang, X., Zhao, Q., Ruan, J.-X., Yang, C.-Q., Wang, L.-J., Miller, D. J., Faraldos, J. A., Allemann, R. K., Chen, X.-Y., & Zhang, P. (2013). Rational engineering of plasticity residues of sesquiterpene synthases from *Artemisia annua*: Product specificity and catalytic efficiency. *The Biochemical Journal*, *451*, 417–426.

[37] Benedict, C. R., Lu, J.-L., Pettigrew, D. W., Liu, J., Stipanovic, R. D., & Williams, H. J. (2001). The cyclization of farnesyl diphosphate and nerolidyl diphosphate by a purified recombinant $\delta$-cadinene synthase. *Plant Physiology*, *125*, 1754–1765.

[38] Li, Z., Gao, R., Hao, Q., Zhao, H., Cheng, L., He, F., Liu, L., Liu, X., Chou, W. K., Zhu, H. et al. (2016). The T296V mutant of amorpha-4, 11-diene synthase is defective in allylic diphosphate isomerization but retains the ability to cyclize the intermediate (3*R*)-nerolidyl diphosphate to amorpha-4, 11-diene. *Biochemistry*, *55*, 6599–6604.

[39] Eddy, S. R. (1998). Profile hidden Markov models. *Bioinformatics*, *14*, 755–763.

[40] Consortium, U. (2016). UniProt: The universal protein knowledgebase. *Nucleic Acids Research*, *45*, D158–D169.

[41] Wheeler, D. L., Barrett, T., Benson, D. A., Bryant, S. H., Canese, K., Chetvernin, V., Church, D. M., DiCuccio, M., Edgar, R., Federhen, S. et al. (2006). Database resources of the national center for biotechnology information. *Nucleic Acids Research*, *35*, D5–D12.

[42] Todsen, W. L. (2014). ChemDoodle 6.0. *Journal of Chemical Information and Modeling*, *54*, 2391–2393.

[43] Bolton, E. E., Wang, Y., Thiessen, P. A., & Bryant, S. H. (2008). PubChem: Integrated platform of small molecules and biological activities. In *Annual Reports in Computational Chemistry* (pp. 217–241). Elsevier volume 4.

[44] Kim, S., Thiessen, P. A., Bolton, E. E., & Bryant, S. H. (2015). PUG-SOAP and PUG-REST: Web services for programmatic access to chemical information in PubChem. *Nucleic Acids Research*, *43*, W605–W611.

[45] Landrum, G. et al. (2006). RDKit: Open-source cheminformatics, . URL: `http://www.rdkit.org/`.

[46] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V. et al. (2011). Scikit-Learn: Machine learning in Python. *the Journal of Machine Learning Research*, *12*, 2825–2830.

[47] Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. *Computing in Science & Engineering*, *9*, 90–95.

[48] Sievers, F., Wilm, A., Dineen, D., Gibson, T. J., Karplus, K., Li, W., Lopez, R., McWilliam, H., Remmert, M., Söding, J., Thompson, J. D., & Higgins, D. G. (2011). Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Molecular Systems Biology*, *7*, 539.

[49] Huerta-Cepas, J., Serra, F., & Bork, P. (2016). ETE 3: Reconstruction, analysis, and visualization of phylogenomic data. *Molecular Biology and Evolution*, *33*, 1635–1638.

[50] Capella-Gutiérrez, S., Silla-Martínez, J. M., & Gabaldón, T. (2009). trimAl: A tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics*, *25*, 1972–1973.

[51] Abascal, F., Zardoya, R., & Posada, D. (2005). ProtTest: Selection of best-fit models of protein evolution. *Bioinformatics*, *21*, 2104–2105.

[52] Stamatakis, A. (2014). RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, *30*, 1312–1313.

[53] Kersey, P. J., Allen, J. E., Allot, A., Barba, M., Boddu, S., Bolt, B. J., Carvalho-Silva, D., Christensen, M., Davis, P., Grabmueller, C. et al. (2017). Ensembl Genomes 2018: An integrated omics infrastructure for non-vertebrate species. *Nucleic Acids Research*, *46*, D802–D808.

[54] Matasci, N. et al. (2014). Data access for the 1,000 Plants (1KP) project. *GigaScience*, *3*.

[55] Wilbur, W. J., & Lipman, D. J. (1983). Rapid similarity searches of nucleic acid and protein data banks. *Proceedings of the National Academy of Sciences*, *80*, 726–730.

[56] Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J. et al. (2020). SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nature Methods*, *17*, 261–272.

[57] Waskom, M. et al. (2017). Mwaskom/seaborn: V0.8.1 (September 2017), .

[58] Bernstein, F. C., Koetzle, T. F., Williams, G. J., Meyer, E. F., Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T., & Tasumi, M. (1977). The Protein Data Bank: A computer-based archival file for macromolecular structures. *Journal of Molecular Biology*, *112*, 535–542.

[59] DeLano, W. L. et al. (2002). PyMOL: An open-source molecular graphics tool. *CCP4 Newsletter on Protein Crystallography*, *40*, 82–92.

CHAPTER 3

# Integrating structure-based machine learning and co-evolution to investigate specificity in plant sesquiterpene synthases

Janani Durairaj, Elena Melillo, Harro J. Bouwmeester, Dick de Ridder, Jules Beekwilder, and Aalt D.J. van Dijk

This chapter has been published as

Durairaj, Janani, et al. "Integrating structure-based machine learning and co-evolution to investigate specificity in plant sesquiterpene synthases." *PLoS Computational Biology* 17.3 (2021): e1008197.

# Abstract

Sesquiterpene synthases (STSs) catalyse the formation of a large class of plant volatiles called sesquiterpenes. While thousands of putative STS sequences from diverse plant species are available, only a small number of them have been functionally characterized. Sequence identity-based screening for desired enzymes, often used in biotechnological applications, is difficult to apply here as STS sequence similarity is strongly affected by species. This calls for more sophisticated computational methods for functionality prediction. We investigate the specificity of precursor cation formation in these elusive enzymes. By inspecting multi-product STSs, we demonstrate that STSs have a strong selectivity towards one precursor cation. We use a machine learning approach combining sequence and structure information to accurately predict precursor cation specificity for STSs across all plant species. We combine this with a co-evolutionary analysis on the wealth of uncharacterized putative STS sequences, to pinpoint residues and distant functional contacts influencing cation formation and reaction pathway selection. These structural factors can be used to predict and engineer enzymes with specific functions, as we demonstrate by predicting and characterizing two novel STSs from *Citrus bergamia*.

## 3.1 Introduction

One of the largest and most structurally diverse family of plant-derived natural products is the isoprenoid or terpenoid family, with over 60,000 members comprising mono-, sesqui-, di-, tri-, and sesterterpenes, along with steroids and carotenoids[1]. These phytochemicals serve plants in defence against pathogens or herbivores and as attractants of pollinators[2]. They are also of high economic value to humankind due to their widespread use in pharmaceutical agents, insecticides, preservatives, fragrances, and flavoring[3]. The immense diversity of the terpenoid family derives from the polymerization and rearrangement of a varying number of simple 5-carbon isoprenoid units. Monoterpenes are 10-carbon (C10) compounds built up of two such units, sesquiterpenes are composed of three and hence are C15 compounds, diterpenes (C20) are composed of four, and so on. Sesquiterpenes are especially interesting due to their high diversity. Their formation is catalysed from the C15 substrate, farnesyl pyrophosphate (FPP), by sesquiterpene synthases (STSs), a class of enzymes found in plants, fungi and bacteria[4].

Recently, we published a database of over 250 experimentally characterized STSs from over one hundred plant species, collectively responsible for the formation of over a hundred different sesquiterpenes[5]. These compounds all derive from the same substrate, FPP, through a branching tree of reactions such as cyclizations, hydride shifts, methyl shifts, rearrangements, re- and de-protonations to give rise to the immense existing variety in sesquiterpene structures. Apart from the functionally characterized STSs in the database, there are thousands of putative STSs in sequenced plant genomes and transcriptomes whose product specificity is unknown. In addition, many STSs in our database are multi-product enzymes, further complicating the matter of product specificity prediction. As a first contribution, we show that multi-product STSs usually catalyse products specific to a single pathway, indicating selectivity towards one precursor cation. Finding residue positions related to this cation choice across all STSs can reveal important aspects of the underlying mechanisms. However, our previous sequence-based analysis showed that these enzymes are very diverse, and sequence similarity is heavily influenced by phylogeny[5]. While an approach using hidden Markov models derived from sequences is available to predict what kind of terpene synthase (mono-, di-, tri-, sesqui- etc.) a particular enzyme may be[6], this kind of sequence-based grouping was not seen within STSs making products derived from a particular cation or cyclization[5]. As a result, previous studies directed at identifying determinants of catalytic specificity in STSs mainly used mutational approaches between and within a few enzymes from the same or closely related species[7–10]. While such approaches have been successful in finding residues influencing product specificity, their small scale in light of the large diversity of STSs makes it likely that they miss aspects shared across all plant STSs. However, terpene synthases across plants, animals, fungi, and bacteria all share a common structural fold[11]. Protein structures typically evolve at a slower pace than sequences, which means they can contain a wealth of information not easily retrieved from the corresponding sequences.

Here, we combine homology modelling to incorporate STS structural information and machine learning to tease out contributions of different residues to cation specificity. We show that structure-based prediction performs well across all plant species, including on STS enzymes that were published recently and were not used for the construction of the predictor. Such structure- or model-based machine learning has been explored before in other enzyme families and prediction tasks[12–15], and is challenging. One major challenge is the immense number of features produced, as each protein has many hundreds of residues, each of which has its own set of structural features. This poses a problem in cases like the current one, where labelled, experimentally characterized data is sparse. Here we used a novel hierarchical classification approach where many classifiers are first trained on each feature across all residues, after which the most predictive residues are selected. The final classifier is only trained on the feature values of these predictive residues. Thus, we are able to prune noisy and irrelevant features in order to pinpoint residue positions correlating with cation specificity. These selected residues are likely intrinsically linked to the catalytic mechanism of an STS and contribute to the enzymatic formation of the precursor cation. Many of these residues are also not found when relying on sequence-derived features alone, emphasizing the importance of structure in understanding catalytic activity.

In addition, while the current characterized sequence space may be small, there are many thousands of uncharacterized putative terpene synthases whose sequences can provide valuable information about evolution and conservation, especially in regions where reliable structural information is not available. A correlated mutations analysis on all putative terpene synthases indicates co-evolving residue partners for our set of cation-specific residues which are implicated in shared functional activity (such as intermediate binding or coordination), favouring their co-evolution. Examining these residues and pairs in the context of each other and co-crystallized substrate analogs reveals important aspects of the STS reaction mechanism.

Apart from the independent test set of recently characterized enzymes, we also present a use-case of our predictor for STS specificity screening by predicting and characterizing bisabolyl cation synthases from *Citrus bergamia*, which further demonstrated the accuracy of the predictor. As the number of experimentally characterized STSs grows, this accuracy will further increase, potentially allowing for more fine-grained product specificity prediction.

The three-pronged approach presented here combines a modest amount of labelled sequence data, a very small amount of experimental structure data, and large amounts of unlabelled sequence data using homology modelling, interpretable machine learning, and co-evolutionary analysis to predict and investigate the underlying mechanisms of cation specificity in STSs. This approach can also be useful for exploring specificity in other enzyme families with characteristics similar to the STSs.

## 3.2 Results and Discussion

### 3.2.1 Sesquiterpene synthases follow a single branch of the reaction tree

The reaction cascade of an STS can take two directions. As is depicted in Figure 3.1, all reactions are initiated by a metal-mediated removal of the diphosphate anion in the (*E*,*E*)-FPP substrate, leading to the formation of a transoid (2*E*,6*E*)-farnesyl cation (farnesyl cation). The farnesyl cation may then isomerize to form a cisoid (2*Z*,6*E*)-farnesyl cation (nerolidyl cation). These two cations may be quenched by water or undergo a proton loss to form acyclic products (acyclic-F and acyclic-N). However, both farnesyl and nerolidyl cations can undergo cyclization at the C10-C11 bond, while the nerolidyl cation can also cyclize at the C6-C7 bond. The resulting cyclic cations can undergo further hydride shifts, methyl shifts, cyclizations, rearrangements, re- and de-protonations to form the final products of the enzyme[16]. Thus, the farnesyl and nerolidyl cations form the roots of a branching tree of hundreds of diverse intermediates and end products.

Many STSs are multi-product enzymes, with two of the more extreme examples being $\delta$-selinene and $\gamma$-humulene synthases from *Abies grandis*, which produce 52 and 34 sesquiterpenes respectively. In order to determine whether cation specificity is maintained across minor products, we looked at the reaction pathways of the sesquiterpenes produced by the multi-product enzymes in our previously assembled database[5]. In their review, Vattekkatte et al.[17] looked into multi-product mono-, sesqui-, and triterpene synthases with respect to factors affecting their promiscuity, such as substrate isomers, metal cofactors and pH. However, they did not specifically address the similarity of an enzyme's minor products to the major product. The collation of characterized STSs in our database provides us with 96 multi-product STSs across a wide variety of species, to better analyse and address this question.

For each sesquiterpene, the route taken in the reaction tree, up to the depth shown in Figure 3.1 was determined as explained in Materials and Methods. Out of the 96 enzymes with more than one product, 79 (82%) had products from the same

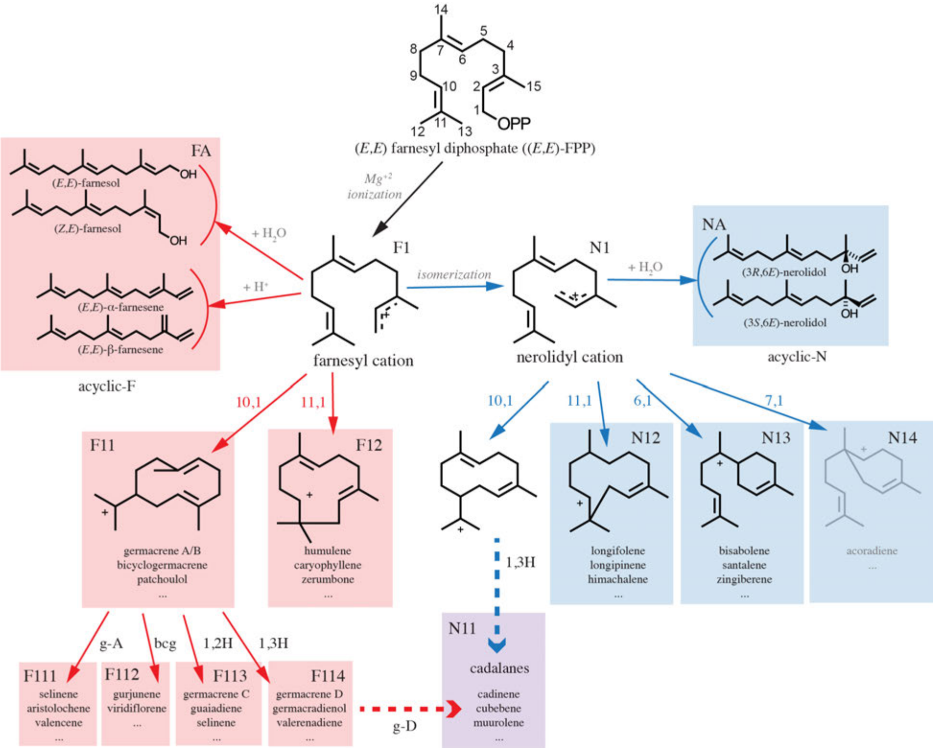


Figure 3.1: The reaction mechanism of sesquiterpene production starts with farnesyl diphosphate ((*E*,*E*)-FPP). Loss of the diphosphate moiety (OPP) leads to farnesyl cation formation. The farnesyl cation can subsequently be converted to the nerolidyl cation. Acyclic sesquiterpenes (acyclic-F and acyclic-N) are formed from these two cations by proton loss or reaction with water molecules. Possible cyclizations for both cations are indicated in the figure. The subsequently formed cyclic cations undergo modifications and rearrangements to form cyclic sesquiterpenes. Some of these sesquiterpenes (g-A and bcg) themselves act as neutral intermediates which can be re-protonated and undergo further reactions to form more products. Products are also formed from specific charged intermediates such as a 1,2- or 1,3-hydride shift of the 10,1-cyclized farnesyl cation (1,2H, 1,3H) and the cadalane skeleton (cadalanes), which can be formed via either of the two precursor cations, or via acid-induced rearrangement of germacrene D. The 7,1-cyclization of the nerolidyl cation, shown in grey, is not found in plant-derived sesquiterpenes. g-A = germacrene A, g-D = germacrene D, bcg = bicyclogermacrene

branch of the tree, three were 10,1-farnesyl synthases with products from different sub-branches, seven had products from the same cation but a different initial cyclization, and twelve synthases had products from different cations, including the aforementioned multi-product *Abies grandis* $\gamma$-humulene synthase. Of these twelve multi-cation STSs, however, eight had an acyclic farnesyl product in addition to nerolidyl-derived compounds. The ease of formation of acyclic farnesyl products (a single step from the farnesyl cation) indicates that they can be formed even by a nerolidyl synthase as the farnesyl cation is the precursor of the nerolidyl cation. Thus, there are only four examples of true multi-cation STSs ($<$5% of the experimentally characterized multi-product enzymes).

This analysis indicates that STSs are, in the vast majority of cases, optimized for the production of sesquiterpenes from a single, well-defined reaction route, by careful control of intermediates right from the commencement of the reaction, at the precursor cation formation step. This insight can be helpful in STS engineering: changing the reaction specificity of an existing STS to products in the same reaction pathway may be easier to accomplish, with fewer mutations, than the introduction of a new reaction pathway. For instance, the 412 active mutants made by O'Maille et al. [18], exploring the mutation space of tobacco 5-*epi*-aristolochene synthase and *Hyoscyamus muticus* vetispiradiene synthase, in many cases resulted in an increased production of germacrene A along with the original product 5-*epi*-aristolochene, which is derived from germacrene A. Given that even multi-product STSs make sesquiterpenes from the same cation, understanding and predicting this cation specificity can greatly narrow down the possible products of a given enzyme.

### 3.2.2 Structure-based cation prediction helps overcomes species bias

STS enzymes all have similar tertiary structures consisting entirely of $\alpha$-helices and short connecting loops and turns. Each structure is typically organized into two domains, with the C-terminal domain containing the active site. The conserved nature of STS enzyme structures across the plant kingdom indicates that applying machine learning on attributes derived from these structures may explain more about cation and product specificity in STSs than sequence-derived attributes, which are more phylogeny-specific. However, due to the lack of available crystal structures for all the characterized enzymes, we turn to homology modelling to make up the deficit. Six crystal structures of STS enzymes were used for multi-template homology modelling of the C-terminal domains of 247 characterized plant STSs. Table 3.1 describes these six structures, three of which are farnesyl synthases, two nerolidyl synthases, and one is a cadalane-type synthase. Supp. Section 3.1 provides more detail on the modelling results, by comparing multi-template models to those created using the single closest template, and by comparing models of the six experimental structures to themselves. Models of the full enzyme sequences were also made but found to be sub-optimal due to the lack of a defined sequence alignment in regions surrounding the C-terminal domain. These results indicate that the final C-terminal domain models are accurate and capture the characteristics of the true structures in this region.

Table 3.1: **The six structures used for multi-template modelling**

| Name | PDB ID | Resolution | Species | Product | Cation |
|---|---|---|---|---|---|
| GACS | 3G4F | 2.65Å | *Gossypium arboreum* | $(+)$-$\delta$-cadinene | cadalane |
| AGBS | 3SDU | 1.89Å | *Abies grandis* | $\alpha$-bisabolene | nerolidyl |
| AABS | 4FJQ | 2.00Å | *Artemisia annua* | $\alpha$-bisabolol | nerolidyl |
| AAHS | 4GAX | 1.99Å | *Artemisia annua* | $\gamma$-humulene | farnesyl |
| HMVS | 5JO7 | 2.15Å | *Hyoscyamus muticus* | vetispiradiene | farnesyl |
| TEAS | 5EAU | 2.15Å | *Nicotiana tabacum* | 5-*epi*-aristolochene | farnesyl |

In order to assess the effect of using features derived from modelled structures compared to purely sequence-based approaches we compared results across three classifiers. One is a simple rule-based classifier, Clf-id, that assigns a test sequence the same class as its closest training sequence based on sequence identity. While this approach is a good baseline and often used in biotechnological applications, machine learning-based models have two advantages over this simple model. Firstly, they are capable of incorporating more complex features, such as the sequence and structure features described in Section 3.4.4, as well as recognizing more complex patterns in these features, allowing for more accurate predictions that generalize across proteins. Secondly, trained machine learning models can be inspected to understand the patterns used for prediction[19]. In this case, this can help gain insight into the contributions of different residues to cation specificity. Therefore, the other two classifiers use the hierarchical machine learning framework described in Materials and Methods with only sequence features (Clf-seq) and with sequence and structure features (Clf-str) respectively. Our classification frameworks make use of gradient boosting trees due to their good out-of-box performance and capability of handling missing feature values caused by deletions in some enzymes.

The dataset consists of 176 farnesyl cation-specific STSs and 72 nerolidyl cation-specific STSs. The remaining 25 STSs are not used for training as they either form products from both cations or only cadalane-type compounds. The cadalane skeleton (Figure 3.1) can be formed by either of the two precursor cations[20] or in acidic conditions of *in vitro* assays from rearrangements of germacrene D[21]. These two alternatives make it difficult to judge whether a cadalane STS goes through the farnesyl or the nerolidyl pathway.

| Scheme | Random Split | | | Genus Split | | | Clade Split | | |
|---|---|---|---|---|---|---|---|---|---|
| Clf- | bAcc | AUC | AUPRC | bAcc | AUC | AUPRC | bAcc | AUC | AUPRC |
| id | 0.88 ± 0.05 | **0.88 ± 0.06** | 0.88 ± 0.05 | 0.72 ± 0.11 | 0.72 ± 0.11 | 0.69 ± 0.16 | 0.51 | 0.51 | 0.46 |
| seq | 0.88 ± 0.04 | 0.83 ± 0.05 | 0.94 ± 0.02 | 0.69 ± 0.07 | 0.88 ± 0.07 | 0.75 ± 0.16 | 0.51 | 0.62 | 0.54 |
| str | **0.90 ± 0.04** | 0.86 ± 0.03 | **0.94 ± 0.02** | **0.73 ± 0.07** | **0.89 ± 0.07** | **0.77 ± 0.13** | **0.64** | **0.75** | **0.59** |

Table 3.2: 1. Clf-id - sequence-identity rule-based classifier, 2. Clf-seq - classification framework using sequence features, 3. Clf-str - classification framework using sequence and structure features. Each column section shows the results of a different validation scheme: randomized 5-fold cross validation (Random Split), genus-based cross validation (Genus Split), and training on 177 dicot STSs and testing on 48 monocot and conifer STSs (Clade Split). For each scheme, balanced accuracy (bAcc), area under the ROC curve (AUC), and area under the precision-recall curve (AUPRC) are presented. The Random Split and Genus Split are repeated 5 and 10 times respectively, leading to the reported standard deviation values.

Table 3.2 shows the performance of these three classifiers using increasingly difficult validation schemes: a random five-fold cross-validation (Random Split), a leave-10-genera-out based scheme (Genus Split), and, finally, training on 177 dicot STSs (124 farnesyl, 53 nerolidyl) with 48 monocot and coniferous STSs (29 farnesyl, 19 nerolidyl) in the test set (Clade Split). Due to the imbalanced nature of the dataset, we use a variety of different metrics to measure performance. These are further described in the Materials and Methods. While Clf-str outperforms the sequence-based approaches by a small margin in the random cross-validation results, the improvement is much more striking in the phylogenetic validation schemes. As STS sequence similarity is biased more towards phylogeny than functional activity, Clf-id and Clf-seq make more errors when testing on species far away from those in the training set. Since Clf-str uses structure-derived information, it is less affected by this bias. This indicates that the structure-based classification framework is more suited to be applied across all plant species, including under-explored species, without losing out on predictive performance. Supp. Figure 3.1 shows the predicted nerolidyl percentages for each enzyme with Clf-str (using the probabilities returned by the genus-based split for each enzyme in the dataset). A clear separation is seen between farnesyl and nerolidyl-cation specific enzymes. However, because of the much lower number of nerolidyl-cation specific enzymes in our dataset, the nerolidyl predicted probabilities for nerolidyl-cation specific enzymes (average 53% ± 30%) are generally lower than the farnesyl predicted probabilities of farnesyl-cation specific enzymes (average 88% ± 19%, calculated as 100 - nerolidyl predicted probability percentage).

As a consequence of its superior performance, the structure-based classifier likely finds features and residues that are important for cation specificity across all plant species - something we can look into to understand generic STS cation determinants.

Thirty cation-specific residues were selected from Clf-str, as described in Materials and Methods. Figure 3.2 visualizes the characterized STS enzymes with respect to the features values of the cation-specific residues, coloured by cation and cyclization specificity. Though imperfect, a separation of farnesyl and nerolidyl cation-specific STSs can be seen. Most cadalane STSs lie on the farnesyl side, with only two being predicted as nerolidyl cation-specific STSs in the Genus Split results. This can indicate that many cadalane synthases in fact make their products through a germacrene D intermediate, or, if the measurements were conducted *in vitro*, then perhaps acidic assay conditions led to spontaneous product rearrangements, thus the interpretation of Figure 3.2 in terms of STSs producing only cadalane products is unclear. While nerolidol synthases (N-acyclic in Figures 3.1 and 3.2) cluster separately from the rest, farnesene and farnesol synthases (F-acyclic in Figures 3.1 and 3.2) are found all across the reduced space. Due to the ease of formation of these acyclic farnesyl products, it is possible that ancestral versions of these enzymes did indeed produce nerolidyl-derived compounds, but this capability was later lost.

A further test of Clf-str was performed on 42 STS enzymes characterized from August 2017-January 2020, not included in the first release of the characterized STS database[5], 31 of which come from species not present in the current set. This new set consists of 24 farnesyl cation-specific STSs, 16 nerolidyl cation-specific STSs,

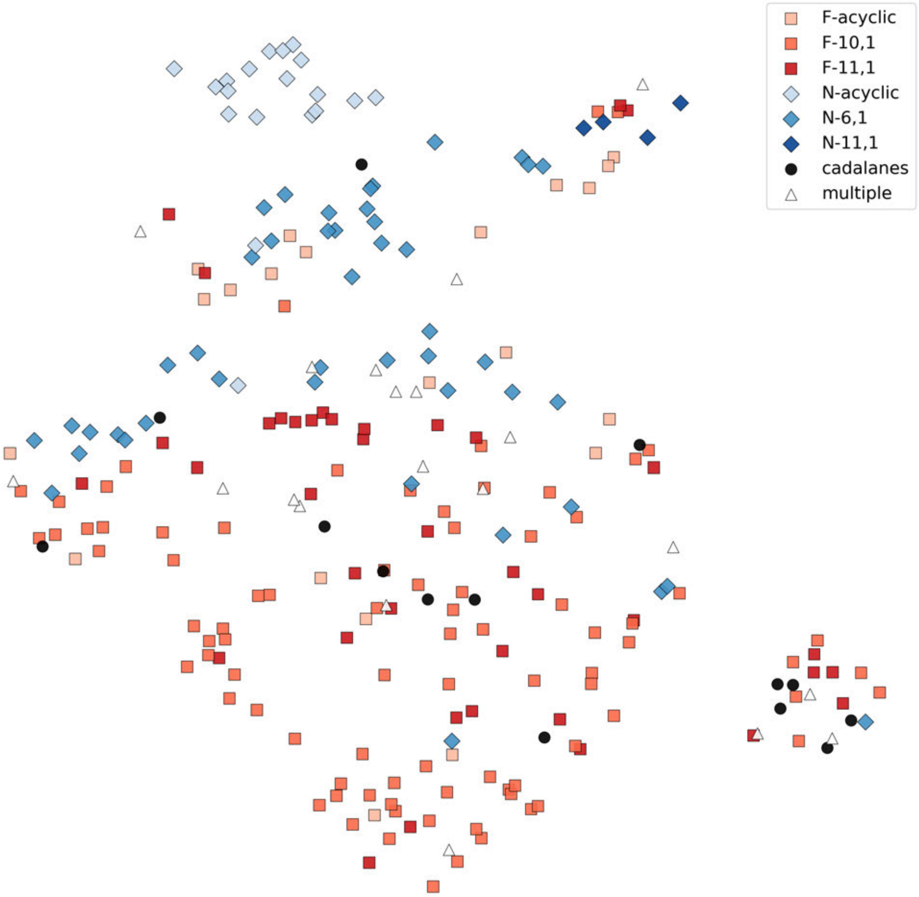


Figure 3.2: Characterized STSs visualized using the feature values of the cation-specific residues followed by dimensionality reduction using UMAP[22], which positions STSs with similar feature values closer to each other. Squares represent farnesyl cation-specific STSs and diamonds represent nerolidyl cation-specific STSs. Each STS is also coloured by its cyclization specificity. Enzymes catalysing products from different precursor cations are marked as triangles.

three STSs producing only cadalane compounds, and one STS which produces both farnesol and nerolidol. Clf-str correctly predicted all the nerolidyl cation-specific STSs and all but two of the farnesyl cation-specific STSs. Both the cadalane and the acyclic STSs were predicted as farnesyl cation-specific STSs. These enzymes are listed in Supp. Table 3.1 and have been added to the second version of the characterized STS database, found at `www.bioinformatics.nl/sesquiterpene/synthasedb`.

### 3.2.3 Residues in five structural regions contribute to cation specificity

The cation-specific residues according to our structure-based predictor are indicated in Figure 3.3A on the tobacco *epi*-aristolochene synthase (TEAS) structure. They are roughly found in five different structural regions, labelled A-E. Also shown are the residues in the three known terpene synthase motifs, namely RXR, DDXXD, and NSE/DTE, as well as the magnesium ions and substrate analog. Figure 3.3B shows the sequence composition of these thirty residues across farnesyl and nerolidyl cation-specific STSs. While the sequence logos (Figure 3.3B) show significant differences in some predictive positions, others have very similar amino acid distributions across the two cations, indicating that their differences lie solely in some combination of their structural features likely due to their structural interaction with neighbouring residues. Thus, these residues would not have been identifiable from sequence-based analysis alone, further demonstrating the power of the integrative approach presented here. Supp. Figure 3.2 shows residue scores across the 10 folds in the genus-based split. The scoring is consistent irrespective of the training set used, indicating that these residues are indeed catalytically important across all plant species.

To obtain more information about these thirty residues, we turned to the wealth of uncharacterized putative terpene synthase enzymes in sequenced plant genomes and transcriptomes. The products of these putative enzymes are unknown, so they cannot be used to train a classifier; however, the sequences themselves still carry valuable information about conservation and divergence. We used co-evolutionary analysis to inspect these sequences in the context of the cation-specific residues. Co-evolutionary analysis is a statistical technique applied on protein sequence alignments based on the underlying biological theory of residue co-evolution[23]. This theory postulates that if there is a mutation in one residue involved in an interaction, then proteins in which its interaction partner is mutated as well, in a way that maintains their interaction, are preferentially selected by evolution. While this technique is most often used to find potentially interacting residues within a protein in protein families with scant structural information, an alternative scenario of co-evolution can play out in the case of functionally related residues[24]. For instance, two residues which contact a substrate or an intermediate, while not interacting directly, may still co-evolve to maintain their shared interactions with the substrate.

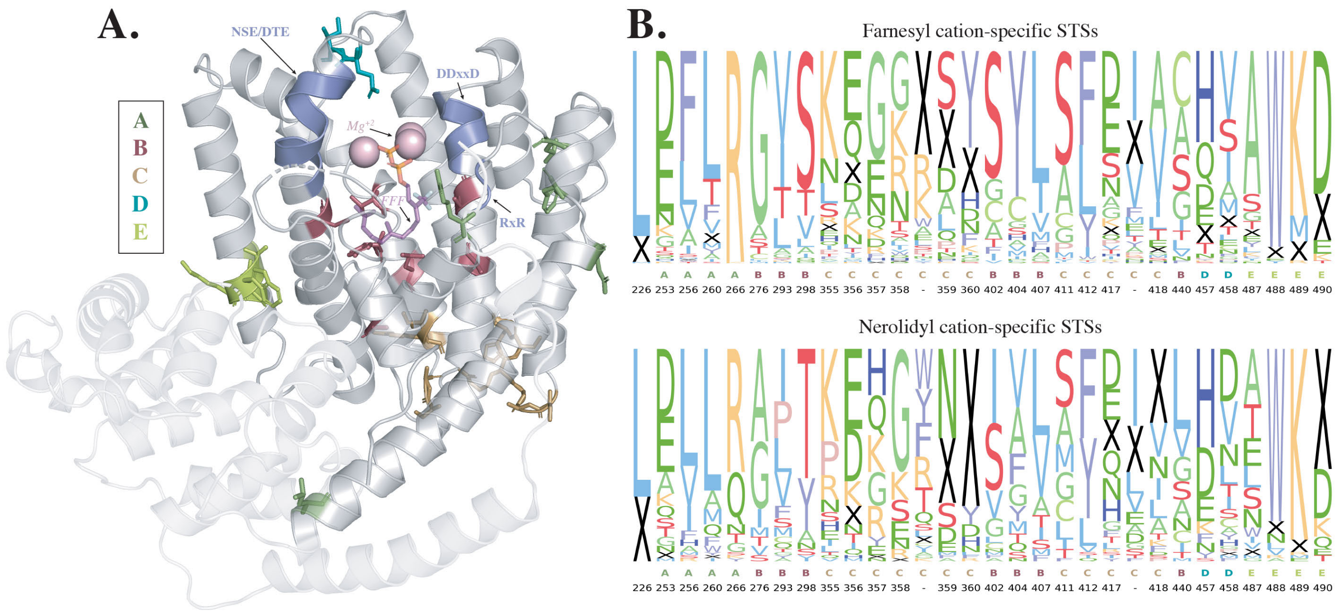


Figure 3.3: **A.** Thirty cation-specific residues found by the structure-based Clf-str predictor on the tobacco *epi*-aristolochene synthase (TEAS) structure, coloured by region. Terpene synthase motifs are labelled in purple, $Mg^{2+}$ ions in pink, and a substrate analog in orange and dark pink. The N-terminal domain is shaded with a lower opacity. **B.** Sequence conservation of Clf-str cation-specific residues across farnesyl and nerolidyl cation-specific STSs, labelled by region and residue position in the TEAS structure. The height of a letter represents its frequency in that position. An insertion/deletion is represented by a black 'X'. Residue positions which are deleted in the TEAS structure are represented by '-'s and correspond to residue 627 and residue 687 respectively in the *Abies grandis* $\alpha$-bisabolene synthase structure (PDB ID: 3SDU). Note that if, for a given position, the residues in both logos are similar, this indicates that in spite of similarity in sequence at this position, the farnesyl and nerolidyl cation-specific enzymes are structurally different.

We used 8344 putative terpene synthase N- and C-terminal domains obtained from sequenced plant genomes and transcriptomes to perform a co-evolutionary analysis as described in Materials and Methods. Supp. Figure 3.3B and Supp. Figure 3.3C show the predicted contact map from this analysis compared to the pairwise minimum $\beta$-carbon Euclidean distance matrix across the six structures in Table 3.1. When looking at the top 1500 predicted contacts (Supp. Figure 3.3A), 328 have residues at least 7 positions apart in the sequence, indicating long range interactions across different structural regions. Only 78 (24%) of these are not capable of physical interaction ($>$11 Å apart) in all of the six STS crystal structures. 10 of these predicted pairs, shown in Figure 3.4, have at least one residue among the thirty cation-specific residues. Below, we discuss specific examples of these residues and pairs in context of the five regions predicted to be involved in cation specificity.

Residues in region A (coloured dark green in Figures 3.3 and 3.4) lie in the A-C loop, close to the conserved RXR motif, with one residue forming the second Arg in the motif itself. This motif has been implicated in the complexation of the diphosphate moiety, preventing nucleophilic attacks on any of the intermediate carbocations[25]. As this is one of the first steps to occur in order for the resulting charged intermediate to undergo cyclization and further reactions, it can play a crucial role in determining how the newly formed cation is positioned, thereby determining whether a farnesyl cation is formed or a nerolidyl cation. In previous work we showed that many nerolidol (N-acyclic) synthases have a mutation in this motif, from RXR to RXQ (as can be seen in the sequence logo; Figure 3.3B, position 266), indicating that changes in and around this motif can indeed affect the products formed.

The six residues in region B (coloured red in Figures 3.3 and 3.4) all lie right in the centre of the active site cavity, in helix D (G276, T293, S298, in TEAS), around the kink region in helix G2 (T402, Y404, L407) and in helix H2 (C440), enveloping the descending substrate from all sides. The residues in this region are very close to both the substrate analog co-crystallized with TEAS as well as the analog co-crystallised with *Abies grandis* $\alpha$-bisabolene synthase, as depicted in Figure 3.4**C**. This proximity has led to a more thorough exploration of these residues in the context of product specificity, than in other regions of the structure. For instance, Yoshikuni et al. [8], 2006 explored plasticity residues in the active site of the promiscuous *Abies grandis* $\gamma$-humulene synthase. Among the many mutants they made, those that converted the major product from the farnesyl-derived $\gamma$-humulene to nerolidyl-derived products such as $\beta$-bisabolene, $\alpha$-longipinene, longifolene, and sibirene, contained mutations in the residues corresponding to T402, Y404 and C440 in TEAS - three cation-specific residues according to our predictor. Two of these residues (Y404 and C440) have also been explored by Salmon et al. [26] when mutating the acyclic $\beta$-farnesene synthase from *Artemisia annua* to a cyclic nerolidyl cation-derived enzyme.

Similarly, Li et al. [27], 2013 demonstrated that a single mutation in the kink in the G2 helix can change the product specificity of an *Artemisia annua* STS from $\alpha$-bisabolol, a nerolidyl-derived sesquiterpene, to the farnesyl-derived $\gamma$-humulene. T402 from this kink has co-evolved with S298 in the parallel helix D. As depicted in Figure 3.4B (column 1), while these two positions are very often both Serine in farnesyl cation-specific

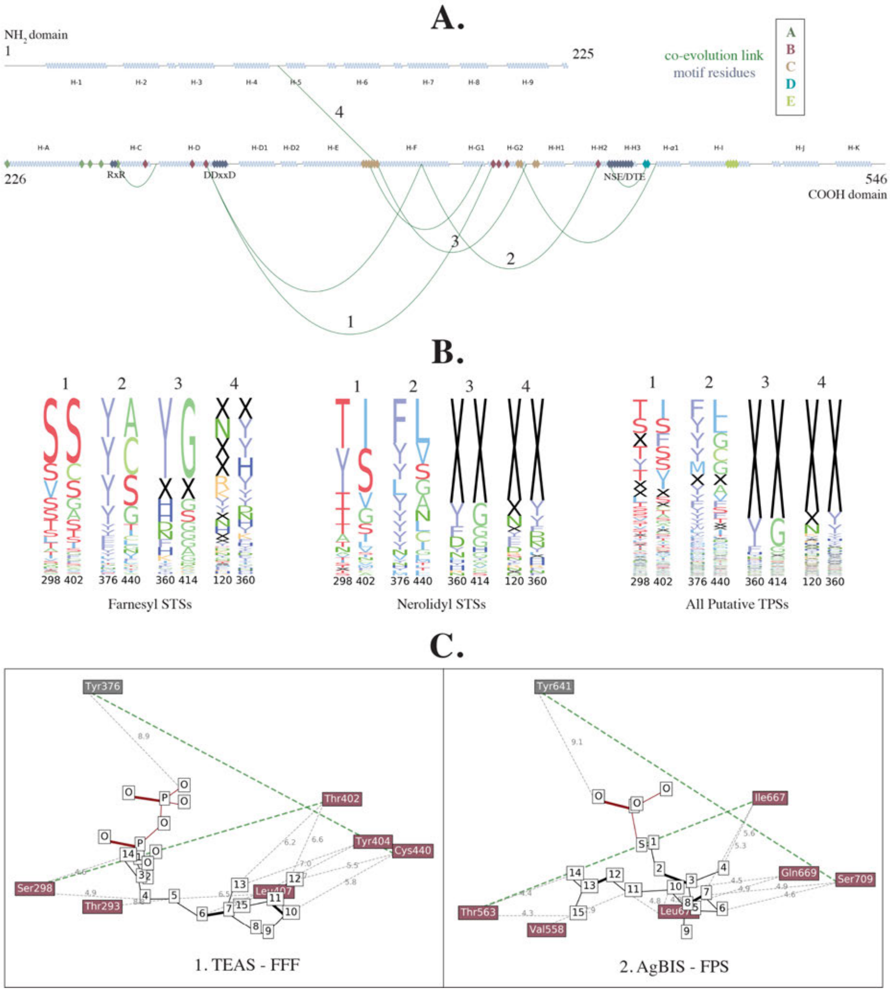


Figure 3.4: **A.** Tobacco *epi*-aristolochene synthase (TEAS) secondary structure with distal cation-specific co-evolutionary contacts (green arcs), motif residues (purple), and cation-specific residues (coloured by region). Helix naming as in Starks et al. [25] **B.** Sequence-pair conservation of four cation-specific contacts discussed in the text, across farnesyl and nerolidyl cation-specific STSs, and all putative terpene synthases. The height of a pair of letters represents the frequency of the pair appearing in those two positions, with 'X' representing gaps. **C**. Diagrams indicating the proximity of residues labelled B in Figure 3.3B, as well as the residues that they co-evolve with, to substrate analogs trifluorofarnesyl diphosphate (FFF) co-crystallized with TEAS (left) and farnesyl thiodiphosphate (FPS) co-crystallized with *Abies grandis* $\alpha$-bisabolene synthase (AgBIS) (right). Carbon atoms are numbered (white boxes) as in the FFF subtrate analog moeity in PDB ID 5EAU. The closest distance (in Å) between each residue's $\beta$-carbon and a substrate atom is labelled in gray. Two co-evolving contacts (labelled 1 and 2 in **A**) are colored in green.

STSs, in nerolidyl cation-specific STSs the commonly occurring pairs are Thr-Ile or Tyr-Ser. The dipole of T402 has been implicated along with T401 in directing the cationic end of the farnesyl chain into the active site, preparing it for a C10 attack[25]. Isoleucine, which is not often found to be a catalytic residue due to its inert nature, cannot perform this task in nerolidyl cation-specific STSs. Another contact is between the cation-specific residue C440 and Y376 (numbered 2 in Figure 3.4B). A mutational analysis on a multi-product maize STS by Köllner et al.[28] demonstrated the importance of Y376 in the formation of bicyclic products such as sesquithujene and bergamotene, derived from the nerolidyl cation. The residue positioned three residues downstream of Y376 was identified by Köllner et al.[29] in 2009 to be involved in controlling the ratio of $\alpha$-bergamotene to the acyclic $\beta$-farnesene in maize STS orthologs. Therefore, the combined effects of position 376 and 440 are likely required for the formation of the nerolidyl cation followed by a second cyclization to bicyclic nerolidyl sesquiterpenes. An alignment of TEAS with the examples discussed here is depicted in Supp. Figure 3.4. These examples demonstrate that residues found important by our structure-based predictor are indeed involved in catalytic and functional activity. They also establish the power of an integrative machine learning approach to pinpoint residue positions important across a variety of species, a combination of what one would find from each of the individual studies referenced above. A Fisher's exact test for the significance of the number of residues found both by our predictor and in literature returned a $p$-value of $9.8e^{-07}$.

The 12 residues in region C (coloured orange in Figures 3.3 and 3.4) encompass the entire E-F loop and parts of the G2-H1 loop at the very bottom of the active site cavity. An interesting residue here is H360, the last residue in the E-F loop. Sequence conservation shows that this position is very often deleted in nerolidyl cation-specific synthases, while farnesyl cation-specific synthases usually have bulky residues such as Tyrosine and Histidine (Figure 3.3B, position 360). Two of its co-evolving partners (numbered 3 and 4 in Figure 3.4B), one from the parallel helix G2 and one from the 4-5 loop in the N-terminal region, are also primarily deleted in nerolidyl cation-specific STSs but present in farnesyl cation-specific STSs, albeit usually as Glycine in helix G2. While the connection with the N-terminal domain is surprising, the parallel residue in the C-terminal domain, when present, may physically interact at some point during the reaction or in other plant STSs, not captured in the six crystal structures currently available[30]. A deletion can break this interaction, which in turn can have an effect on the positioning of helix G2 in the active site and thereby the positioning of the cation-specific residues that lie within it. These subtle alterations in cavity shape may in turn affect which kinds of intermediates fit comfortably inside the cavity.

Two consecutive high scoring residues (region D, coloured blue in Figures 3.3 and 3.4), lie in the H3-$\alpha$1 loop, close to the catalytic NSE/DTE motif. This motif is involved in coordinating $Mg^{2+}$ ions along with the DDXXD motif on the opposite side[31]. This region lies at the entrance of the active site cavity and is in an optimum position to contact the substrate as it enters the cavity. In addition, the inability to crystallize this region in three of the six crystal structures indicates that this loop is very flexible[32].

Residues in region E (coloured light green in Figures 3.3 and 3.4) lie in helix I, near the end of the C-terminal domain and close to helix 7 and helix 8 in the N-terminal domain.

Overall, these results show that cation-specific residues in regions labelled A, B, and D lie within areas known to participate directly in the catalytic reaction. These residues were predicted by our machine learning approach without using any knowledge on their functional properties. Some of these residues have been mutated before and were shown to be important for cation specificity. This indicates that the other residues are also likely to perform similarly crucial roles, perhaps also in STSs that have not been used so far in mutagenesis experiments. Residues labelled C and E lie quite far from the active site and could be involved in subtle alterations of the cavity shape or in stabilising contacts with the N-terminal domain. Though this domain is known to be important for plant STS reactions, its exact function has not been fully explored. However, just as O'Maille et al.[18] showed that residues distant from the active site can still be functionally crucial, these distal residues are likely to have multifaceted and interdependent roles in cation specificity that only such large-scale computational approaches can recognize. Further experiments and mutational studies in these regions are required to confirm and elaborate their involvement in the STS reaction mechanism. Meanwhile, the structure-based predictor, as well as the cation-specific sequence and contact conservation information described can be used to screen through the many thousands of uncharacterized putative STSs with a particular cation specificity in mind as demonstrated in the next section.

### 3.2.4 Bisabolyl cation synthases from *Citrus bergamia* 'Femminello'

One potential application of the cation-specificity predictor presented here is to screen for enzymes with a desired specificity. We demonstrate this application to find STSs catalysing the formation of products derived from the bisabolyl cation from 23 terpene synthase-like sequences extracted from the transcriptome of *Citrus bergamia* 'Femminello' (described in Materials and Methods). Using the hidden Markov model approach detailed by Priya et al.[6], 11 sequences out of these 23 were predicted to be STSs (as opposed to mono- or diterpene synthases). We used the cation specificity predictor on these 11 and sorted by decreasing order of predicted nerolidyl cation specificity, selecting enzymes with predicted probability percentage above 10%, based on the predicted percentages of the characterized database (Supp. Figure 3.1).

Two enzymes clustered close to the nerolidol cluster in Figure 3.2 and were thus excluded, resulting in four enzymes with >10% predicted nerolidyl cation specificity. Three of these could be experimentally characterized, submitted to GenBank with identifiers MT636927, MT636928 and MW384854 respectively. MT636927 and MT636928 produced bisabolyl cation-derived products. MT636927 has 55% predicted nerolidyl specificity and produced trans-$\alpha$-bergamotene, $\beta$ bisabolene, and $\alpha$ bisabolol. MT636928 has 11% predicted nerolidyl specificity, and produced zingiberene. MW384854 has 26% predicted nerolidyl specificity but produced the farnesyl-cation derived caryophyllene. The chromatograms and the fragmentation patterns of the identified peaks and the reference compounds can be found in Supp. Figure 3.5 and Supp. Section 3.2.

Sequence identity based screening, on the other hand, predicts all 11 enzymes as farnesyl cation specific showing that based on only sequence identity, we cannot prioritize candidate genes for production of bisabolyl cation-derived products. Thus, the cation specificity predictor can be used for effective screening of STSs with desired intermediate specificity, saving time, labour and costs required for extensive experimental characterization. Considering that the bisabolyl cation is one of the least represented intermediates in our dataset, expanding the number of experimentally characterized enzymes used for training can further increase the accuracy of our results, and even allow for more fine-grained product specificity prediction.

## 3.3 Conclusion

The availability of growing numbers of characterized and putative sesquiterpene synthases opens doors for the application of computational analyses in order to obtain insights about this large and amazingly diverse family of enzymes. While STSs collectively produce many hundreds of compounds, these are all rearrangements of two precursor carbocations deriving from a single substrate. We show that multiproduct STS enzymes catalyse the formation of products deriving from the same cation, indicating that cation specificity is determined early in the reaction. A combination of structure-based supervised machine learning and unsupervised co-evolution gives us a set of structural regions implicated in cation specificity determination as well as possible functional relationships between residues in these regions and other parts of the STS structure. The predictor itself can be used for cation-specificity screening, while the residues and corresponding linkages discussed here can be used to design mutational studies with a higher likelihood of maintaining catalytic activity while changing cation specificity. Such an integrative approach can also be applied to other diverse enzyme families in order to uncover large-scale interdependent relationships between catalytic residues influencing product specificity. As the number of characterized STSs from across the plant kingdom increases, more specific predictors can be designed, in order to screen STSs at the cyclization or even product level.

## 3.4 Materials and methods

### 3.4.1 Reaction pathway determination

The reaction pathway for each sesquiterpene in the database was determined using the scheme detailed in IUBMB's *Enzyme Nomenclature* Supplement 24 (2018)[33] up to the depth specified in Figure 3.1. For example, the sesquiterpene viridiflorene would be labelled F112 as it derives from bicyclogermacrene which itself is labelled F11. Sesquiterpenes derived from the cadalane skeleton, namely cadinanes, cubebenes, copaenes, amorphenes, sativenes, muurolenes, ylangenes, and their alcoholic variants, are marked as cadalanes as they can form from multiple reaction pathways.

Two sesquiterpenes share a reaction path if the pathway annotation of one is a non-strict prefix of the other's. For example, sesquiterpenes labelled F1, F11, and F113 belong on the same reaction path while those labelled F111, F112, and F12 do not. If multiple cadalane-type compounds are produced by one enzyme, they are assumed to come from the same path. These rules are used to calculate the number of multi-product enzymes with products following the same reaction path.

STSs were labelled as farnesyl or nerolidyl according to the group that their products belong to. STSs making cadalane products along with additional non-cadalane products are labelled with the cation of these other products. Multi-product STSs producing compounds from different cations, as well as cadalane STSs without any non-cadalane product are considered separately and are not used for training.

### 3.4.2 Sequence extraction and alignment

N-terminal and C-terminal domain sequences were extracted from all spermatophyte plant STSs from the database using HMMER[34] and the Pfam[35] domains PF01397 and PF03936 respectively. All N-terminal and C-terminal sequence alignments were made using Clustal Omega[36], using the corresponding Pfam domain HMM to guide the alignment. A combined N- and C-terminal domain HMM was built by aligning each half of the common seed sequences from both respective Pfam domains, stacking the resulting alignments together, and using the hmmbuild tool in HMMER[34]. This HMM is referred to as Terpene_synth_N_C.

### 3.4.3 Homology modelling

For each STS, 500 multi-template homology models were created of the C-terminal domain region using MODELLER[37], with six STS structures from the PDB[38] as templates, as listed in Table 3.1. These were aligned to each sequence using the C-terminal PF03936 Pfam domain[35] as a guide, using Clustal Omega[36]. The top three models were selected based on their N-DOPE score for feature extraction.

For comparison, 500 models were also made using a single template for each enzyme; the template chosen was the one having the maximum sequence identity to the enzyme being modelled. Similarly, models were made for each of the six template structures using the other five structures as templates. Models of full STS sequences (including the N-terminal domain) were also made using a similar multi-template approach with the custom Terpene_synth_N_C HMM to guide the alignment to the templates. Results for these three additional approaches are presented in Supp. Section 3.1.

### 3.4.4 Feature extraction

Sequence and structure features were extracted from each STS as described below and aligned according to the C-terminal domain alignment. Gaps in the alignment were represented as NaNs for continuous features and as a separate category for categorical features.

Sequence features

For each STS sequence, PSIBLAST[39] was run on the non-redundant protein database (nr)[40] and used to calculate a position-specific scoring matrix (PSSM) and a position-specific frequency matrix (PSFM). The information content of each column in the PSSM was also calculated. SCRATCH[41] was used to predict the secondary structure and surface accessibility of each residue. Finally, the raw amino acid sequence was also used as a feature source. Categorical features were one-hot encoded.

Structure features

Structural features were extracted for each of the top three homology models for each STS. All atom-level features were converted into $\alpha$-carbon, $\beta$-carbon, and mean residue features. For Gly, the $\alpha$-carbon was used for the $\beta$-carbon features as well. ProDy[42] was used to calculate the 50-mode Gaussian Network Model (GNM) and Anisotropic Network Model (ANM) atom fluctuations using the `calcGNM`/`calcANM` functions followed by the `calcSqFlucts` function. APBS[43] was used to calculate the Coulomb and Born electrostatics of a modelled structure. PDB2PQR[43] was first used to generate a PQR file from each PDB file, followed by running the `born` command with an `epsilon` (solvent dielectric constant) of 80 and the `coulomb` command with the `-e` option. DSSP features are calculated using ProDy[42] to give hydrogen bond energies, surface accessibility, dihedral angles ($\alpha$), bend angles ($\kappa$), $\phi$, and $\psi$ backbone torsion angles, and tco angles (cosine angle between the C=O of residue $i$ and the C=O of residue $i-1$). Residue depths were extracted using BioPython[44] from the PDB files of the top three models.

### 3.4.5 Classification framework

A classification framework using Gradient boosting trees (as depicted in Supp. Figure 3.6) was built for different sets of features. The framework is trained in three steps:

1. A separate gradient boosting tree is trained for each kind of feature for all residues. XGBoost[45] was used with default parameter settings for these intermediate classifiers (`n_trees = 100`, `learning_rate = 0.1`, `gamma = 0`, `subsample = 1`, `colsample_bytree = 1`, `colsample_bylevel = 1`). These simple settings are sufficient as these classifiers are only used to find predictive residues, as described in the next step.

2. The sum of normalized weights for each residue across all the trained feature models from Step 1 is used as a scoring measure to select the top thirty residues.

3. A final gradient boosting forest with much stricter parameter settings (`n_trees = 2000`, `learning_rate = 0.005`, `gamma = 0.01`, `subsample = 0.7`, `colsample_bytree = 0.1`, `colsample_bylevel = 0.1`) is trained using XGBoost[45] on all the feature values of the top residues picked in Step 2. These parameter settings are chosen to make a more conservative classifier that avoids overfitting in three ways: reduced model complexity by regularization (using the `gamma` parameter), robustness to noise by random selection in each intermediate tree of both data points (the `subsample` parameter) and features (the `colsample`

parameters), and a slow learning rate combined with a large number of trees to increase the power of the ensemble.

For testing, the features of the selected thirty residue positions in the test enzymes are fed into the trained classifier.

Clf-seq and Clf-str are two classifiers built using this framework utilizing only sequence features and both sequence and structure features, respectively. Clf-id is a simple rule-based classifier that does not use this framework and instead returns the class of the closest training set sequence based on sequence identity.

### 3.4.6 Validation and testing

Three validation schemes are used to test a classifier.

1. Random Split: A random five-fold cross-validation with 80%-20% train-test split.
2. Genus Split: A scheme in which cases from 65 genera are used for training and the rest for testing, repeated 10 times with different sets.
3. Clade Split: All dicot STSs are used for training and monocot and conifer STSs for testing.

Three different metrics are used to measure the performance of each classifier, using the definitions of $TP$ and $TN$ as the number of nerolidyl cation-specific synthases and number of farnesyl cation-specific synthases predicted correctly at a certain threshold of predicted probability, and $FP$ and $FN$ as the number of nerolidyl cation-specific synthases and number of farnesyl cation-specific synthases predicted incorrectly at a certain threshold. All metrics are calculated using the scikit-learn Python library[46].

1. Balanced accuracy (bAcc): $\frac{1}{2}\left(\frac{TP}{TP+FN}+\frac{TN}{TN+FP}\right)$ at a threshold of $0.5$.
2. Area Under the Receiver Operating Characteristic Curve (AUC): Calculated as the area under the plot of the fraction of $TP$ out of the total number of nerolidyl cation-specific synthases vs. the fraction of $FP$ out of the total number of farnesyl cation-specific synthases, at various threshold settings.
3. Area Under the Precision-Recall Curve (AUPRC): Calculated as the area under the plot of the precision $(TP/(TP+FP))$ vs. the recall $(TP/(TP+FN)$ at various threshold settings.

42 newly characterized synthases from literature (listed in Supp. Table 3.1) are used as the final independent test set.

### 3.4.7 Selecting cation-specific residues

The normalized weights across all feature classifiers were summed across all the folds of the Genus Split and the resulting thirty highest scoring positions represent the set of cation-specific residues. The sequence and structural features of these residues were used to visualize the set of characterized STSs. This was done by applying UMAP[22] to reduce the dimensionality to 2.

### 3.4.8 Co-evolution analysis on plant terpene synthase-like proteins

An HMM search was performed using HMMER[34] and the custom Terpene_synth_N_C HMM across all plant UniProt proteins[47] and all plant transcriptome sequences from the OneKP transcriptome dataset[48]. Only those with sequence length at least one standard deviation away from the mean sequence length of the characterized STSs from the database[5] were retained. The resulting set of uncharacterized sequences were aligned with Clustal Omega[36] using the same HMM and 10 guide-tree/HMM iterations (`iter = 10`). Alignment positions not present in any of the six structures in Table 3.1 were discarded.

CCMPred[49] was used to perform co-evolution analysis on this alignment. The top 1500 predicted contacts were selected based on their confidence scores (Supp. Figure 3.3A). Contacts containing one residue from the cation-specific positions, at least 11 Å apart in any of the six structures in Table 3.1 and seven residues apart in sequence were retained.

### 3.4.9 Visualization of cation-specific residues and contacts

Cation-specific residues and contacts were visualized in multiple ways.

- **3D Structure** - PyMOL[50] was used to visualize the three-dimensional structure of tobacco 5-epi-aristolochene synthase (TEAS, PDB ID: 5EAU) and label terpene synthase motif residues and cation specific residues.

- **Sequence and Co-evolution Conservation Logos** - The positions of predictive residues in farnesyl and nerolidyl cation-specific STSs were used to generate two sequence conservation logos based on the percentage of appearance of each amino acid at each position. The sequence conservation of four co-evolving residue pairs was also visualized across farnesyl and nerolidyl cation-specific STSs and the set of putative terpene synthases. These figures were made with matplotlib[51].

- **Co-evolutionary Links** - The cation-specific residues and contacts as well as terpene synthase motif residues were visualized on the secondary structure of the N-terminal and C-terminal domain portions of the tobacco aristolochene synthase (TEAS) structure found by the two respective Pfam domains (PF01397 and PF03936), using matplotlib[51]. Helices are labelled as described by Starks et al.[25].

- **Substrate Analog Proximity** - Substrate analogs trifluorofarnesyl diphosphate (FFF) and farnesyl thiodiphosphate (FPS) were extracted from tobacco *epi*-aristolochene synthase PDB ID: 5EAU, and *Abies grandis* $\alpha$-bisabolene synthase PDB ID: 3SAE respectively. Their positions in both structures were obtained by superposing the two structures to each other using the `align` command in PyMOL[50]. Distances between a subset of the cation-specific residues and the atoms of the substrate analogs were visualized using matplotlib[51]. The atoms in both analogs are numbered according to the numbering of FFF.

### 3.4.10 *Citrus bergamia* 'Femminello' STSs

The cation specificity predictor was employed to select four STSs among the putative terpenes synthases from *C. bergamia* with the highest nerolidyl cation specificity. The sequences were codon optimised, synthesised and expressed in *Rhodobacter sphaeroides*, as described earlier in Beekwilder et al.[52]. The analysis of the products coming from the engineered strains was performed on the GC Agilent 7890B coupled to the MS Agilent 5977B. The used column is an HP-5MS 30m x 250um x 0.25um. The resulting chromatograms and the fragmentation patterns of the identified peaks and the reference compounds can be found in Supp. Figure 3.5 and Supp. Section 3.2.

### 3.4.11 Data and code availability

The characterized STS sequence, product, and species data used can be found at `https://www.bioinformatics.nl/sesquiterpene/synthasedb/`

Code used for modelling, feature extraction, and building the various cation prediction classifiers presented here can be found at `https://git.wur.nl/durai001/sts_cation_prediction`, along with data for:

1. homology models of the C-terminal domains of characterized STSs using a single, closest template from the template structures in Table 3.1.
2. multi-template homology models of the C-terminal domain, built using all six templates.
3. multi-template homology models of the full (N- and C-terminal domain) structures using the same six templates.
4. an alignment of 8344 putative terpene synthases
5. the predicted contact matrix returned by CCMPred on the above alignment

Protein visualization code, for displaying sequence logos, co-evolution logos, and secondary structure elements, can be found at `https://git.wur.nl/durai001/clemmys`.

## Acknowledgments

We thank Miguel Correa Marrero for valuable feedback on early drafts and stimulating discussions on the approaches discussed in the text. We also thank Kenneth Paul Rivadeneira Guadamud for collecting 21 of the recently characterized sequences used in the test set.

## Supplementary information

All supplementary sections, figures, and tables are available online at `https://doi.org/10.1371/journal.pcbi.1008197`

# References

[1] Buckingham, J. (1997). *Dictionary of Natural Products, Supplement 4* volume 11. CRC Press.

[2] Gershenzon, J., & Dudareva, N. (2007). The function of terpene natural products in the natural world. *Nature Chemical Biology*, *3*, 408–414.

[3] Schempp, F. M., Drummond, L., Buchhaupt, M., & Schrader, J. (2017). Microbial cell factories for the production of terpenoid flavor and fragrance compounds. *Journal of Agricultural and Food chemistry*, *66*, 2247–2258.

[4] Chen, F., Tholl, D., Bohlmann, J., & Pichersky, E. (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*, *66*, 212–229.

[5] Durairaj, J., Di Girolamo, A., Bouwmeester, H. J., de Ridder, D., Beekwilder, J., & van Dijk, A. D. J. (2019). An analysis of characterized plant sesquiterpene synthases. *Phytochemistry*, *158*, 157–165.

[6] Priya, P., Yadav, A., Chand, J., & Yadav, G. (2018). Terzyme: A tool for identification and analysis of the plant terpenome. *Plant Methods*, *14*, 4.

[7] Greenhagen, B. T., O'Maille, P. E., Noel, J. P., & Chappell, J. (2006). Identifying and manipulating structural determinates linking catalytic specificities in terpene synthases. *Proceedings of the National Academy of Sciences*, *103*, 9826–9831.

[8] Yoshikuni, Y., Ferrin, T. E., & Keasling, J. D. (2006). Designed divergent evolution of enzyme function. *Nature*, *440*, 1078–1082.

[9] Kampranis, S. C., Ioannidis, D., Purvis, A., Mahrez, W., Ninga, E., Katerelos, N. A., Anssour, S., Dunwell, J. M., Degenhardt, J., Makris, A. M. et al. (2007). Rational conversion of substrate and product specificity in a Salvia monoterpene synthase: Structural insights into the evolution of terpene synthase function. *The Plant Cell*, *19*, 1994–2005.

[10] Segura, M. J., Jackson, B. E., & Matsuda, S. P. (2003). Mutagenesis approaches to deduce structure–function relationships in terpene synthases. *Natural Product Reports*, *20*, 304–317.

[11] Gao, Y., Honzatko, R. B., & Peters, R. J. (2012). Terpenoid synthase structures: A so far incomplete view of complex catalysis. *Natural Product Reports*, *29*, 1153–1175.

[12] Berliner, N., Teyra, J., Çolak, R., Lopez, S. G., & Kim, P. M. (2014). Combining structural modeling with ensemble machine learning to accurately predict protein fold stability and binding affinity effects upon mutation. *PLoS One*, *9*, e107353.

[13] Ferraro, E., Via, A., Ausiello, G., & Helmer-Citterich, M. (2006). A novel structure-based encoding for machine-learning applied to the inference of SH3 domain specificity. *Bioinformatics*, *22*, 2333–2339.

[14] Cang, Z., & Wei, G.-W. (2018). Integration of element specific persistent homology and machine learning for protein-ligand binding affinity prediction. *International Journal for Numerical Methods in Biomedical Engineering*, *34*, e2914.

[15] Romero, P. A., Krause, A., & Arnold, F. H. (2013). Navigating the protein fitness landscape with Gaussian processes, . *110*, E193–E201.

[16] Degenhardt, J., Köllner, T. G., & Gershenzon, J. (2009). Monoterpene and sesquiterpene synthases and the origin of terpene skeletal diversity in plants. *Phytochemistry*, *70*, 1621–1637.

[17] Vattekkatte, A., Garms, S., Brandt, W., & Boland, W. (2018). Enhanced structural diversity in terpenoid biosynthesis: enzymes, substrates and cofactors. *Organic & Biomolecular Chemistry*, *16*, 348–362.

[18] O'Maille, P. E., Malone, A., Dellas, N., Andes Hess, B., Smentek, L., Sheehan, I., Greenhagen, B. T., Chappell, J., Manning, G., & Noel, J. P. (2008). Quantitative exploration of the catalytic landscape separating divergent plant sesquiterpene synthases. *Nature Chemical Biology*, *4*, 617–623.

[19] Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. *ArXiv e-prints*. `arXiv:1702.08608`.

[20] Garms, S., Köllner, T. G., & Boland, W. (2010). A multiproduct terpene synthase from *Medicago truncatula* generates cadalane sesquiterpenes via two different mechanisms. *The Journal of Organic Chemistry*, *75*, 5590–5600.

[21] Bülow, N., & König, W. A. (2000). The role of germacrene D as a precursor in sesquiterpene biosynthesis: Investigations of acid catalyzed, photochemically and thermally induced rearrangements. *Phytochemistry*, *55*, 141–168.

[22] McInnes, L., Healy, J., & Melville, J. (2018). UMAP: Uniform manifold approximation and projection for dimension reduction. *ArXiv e-prints*. `arXiv:1802.03426`.

[23] Morcos, F., Pagnani, A., Lunt, B., Bertolino, A., Marks, D. S., Sander, C., Zecchina, R., Onuchic, J. N., Hwa, T., & Weigt, M. (2011). Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proceedings of the National Academy of Sciences*, *108*, E1293–E1301.

[24] Little, D. B., & Croteau, R. B. (2002). Alteration of product formation by directed mutagenesis and truncation of the multiple-product sesquiterpene synthases $\delta$-selinene synthase and $\gamma$-humulene synthase. *Archives of Biochemistry and Biophysics*, *402*, 120–135.

[25] Starks, C. M., Back, K., Chappell, J., & Noel, J. P. (1997). Structural basis for cyclic terpene biosynthesis by tobacco 5-*epi*-aristolochene synthase. *Science*, *277*, 1815–1820.

[26] Salmon, M., Laurendon, C., Vardakou, M., Cheema, J., Defernez, M., Green, S., Faraldos, J. A., & O'Maille, P. E. (2015). Emergence of terpene cyclization in *Artemisia annua*. *Nature Communications*, *6*, 6143.

[27] Li, J.-X., Fang, X., Zhao, Q., Ruan, J.-X., Yang, C.-Q., Wang, L.-J., Miller, D. J., Faraldos, J. A., Allemann, R. K., Chen, X.-Y., & Zhang, P. (2013). Rational engineering of plasticity residues of sesquiterpene synthases from *Artemisia annua*: Product specificity and catalytic efficiency. *The Biochemical Journal*, *451*, 417–426.

[28] Köllner, T. G., O'Maille, P. E., Gatto, N., Boland, W., Gershenzon, J., & Degenhardt, J. (2006). Two pockets in the active site of maize sesquiterpene synthase TPS4 carry out sequential parts of the reaction scheme resulting in multiple products. *Archives of Biochemistry and Biophysics*, *448*, 83–92.

[29] Köllner, T. G., Gershenzon, J., & Degenhardt, J. (2009). Molecular and biochemical evolution of maize terpene synthase 10, an enzyme of indirect defense. *Phytochemistry*, *70*, 1139–1145.

[30] Anishchenko, I., Ovchinnikov, S., Kamisetty, H., & Baker, D. (2017). Origins of coevolution between residues distant in protein 3D structures. *Proceedings of the National Academy of Sciences*, *114*, 9122–9127.

[31] Christianson, D. W. (2006). Structural biology and chemistry of the terpenoid cyclases. *Chemical Reviews*, *106*, 3412–3442.

[32] Fontana, A., de Laureto, P. P., Spolaore, B., & Frare, E. (2012). Identifying disordered regions in proteins by limited proteolysis. In *Intrinsically Disordered Protein Analysis* (pp. 297–318). Springer.

[33] Webb, E. C. et al. (1992). *Enzyme Nomenclature 1992. Recommendations of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology on the Nomenclature and Classification of Enzymes*. Ed. 6. Academic Press.

[34] Finn, R. D., Clements, J., & Eddy, S. R. (2011). HMMER web server: Interactive sequence similarity searching. *Nucleic Acids Research*, *39*, W29–W37.

[35] Bateman, A. et al. (2004). The Pfam protein families database. *Nucleic Acids Research*, *32*, D138–D141.

[36] Sievers, F., Wilm, A., Dineen, D., Gibson, T. J., Karplus, K., Li, W., Lopez, R., McWilliam, H., Remmert, M., Söding, J., Thompson, J. D., & Higgins, D. G. (2011). Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Molecular Systems Biology*, *7*, 539.

[37] Webb, B., & Sali, A. (2014). Comparative protein structure modeling using MODELLER. *Current Protocols in Bioinformatics*, *47*, 5–6.

[38] Bernstein, F. C., Koetzle, T. F., Williams, G. J., Meyer, E. F., Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T., & Tasumi, M. (1977). The Protein Data Bank: A computer-based archival file for macromolecular structures. *Journal of Molecular Biology*, *112*, 535–542.

[39] Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Research*, *25*, 3389–3402.

[40] Pruitt, K. D., Tatusova, T., & Maglott, D. R. (2005). NCBI Reference Sequence (RefSeq): A curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Research*, *33*, D501–D504.

[41] Cheng, J., Randall, A. Z., Sweredoski, M. J., & Baldi, P. (2005). SCRATCH: A protein structure and structural feature prediction server. *Nucleic Acids Research*, *33*, W72–76.

[42] Bakan, A., Meireles, L. M., & Bahar, I. (2011). ProDy: Protein dynamics inferred from theory and experiments. *Bioinformatics*, *27*, 1575–1577.

[43] Unni, S., Huang, Y., Hanson, R. M., Tobias, M., Krishnan, S., Li, W. W., Nielsen, J. E., & Baker, N. A. (2011). Web servers and services for electrostatics calculations with APBS and PDB2PQR. *Journal of Computational Chemistry*, *32*, 1488–1491.

[44] Cock, P. J., Antao, T., Chang, J. T., Chapman, B. A., Cox, C. J., Dalke, A., Friedberg, I., Hamelryck, T., Kauff, F., Wilczynski, B. et al. (2009). BioPython: Freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics*, *25*, 1422–1423.

[45] Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* KDD '16 (pp. 785–794). Association for Computing Machinery.

[46] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V. et al. (2011). Scikit-Learn: Machine learning in Python. *the Journal of Machine Learning Research*, *12*, 2825–2830.

[47] Consortium, U. (2016). UniProt: The universal protein knowledgebase. *Nucleic Acids Research*, *45*, D158–D169.

[48] Matasci, N. et al. (2014). Data access for the 1,000 Plants (1KP) project. *GigaScience*, *3*.

[49] Seemayer, S., Gruber, M., & Söding, J. (2014). CCMpred—fast and precise prediction of protein residue–residue contacts from correlated mutations. *Bioinformatics*, *30*, 3128–3130.

[50] DeLano, W. L. et al. (2002). PyMOL: An open-source molecular graphics tool. *CCP4 Newsletter on Protein Crystallography*, *40*, 82–92.

[51] Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. *Computing in Science & Engineering*, *9*, 90–95.

[52] Beekwilder, J., van Houwelingen, A., Cankar, K., van Dijk, A. D., de Jong, R. M., Stoopen, G., Bouwmeester, H., Achkar, J., Sonke, T., & Bosch, D. (2014). Valencene synthase from the heartwood of Nootka cypress (*Callitropsis nootkatensis*) for biotechnological production of valencene. *Plant Biotechnology Journal*, *12*, 174–182.

CHAPTER 4

# Caretta - A multiple protein structure alignment and feature extraction suite

Mehmet Akdel*, Janani Durairaj*, Dick de Ridder, and Aalt D.J. van Dijk

* authors contributed equally

This chapter has been published as

Akdel, Mehmet, et al. "Caretta–A multiple protein structure alignment and feature extraction suite." *Computational and Structural Biotechnology Journal* 18 (2020): 981-992.

# Abstract

The vast number of protein structures currently available opens exciting opportunities for machine learning on proteins, aimed at predicting and understanding functional properties. In particular, in combination with homology modelling, it is now possible to not only use sequence features as input for machine learning, but also structure features. However, in order to do so, robust multiple structure alignments are imperative.

Here we present Caretta, a multiple structure alignment suite meant for homologous but sequentially divergent protein families which consistently returns accurate alignments with a higher coverage than current state-of-the-art tools. Caretta is available as a GUI and command-line application and additionally outputs an aligned structure feature matrix for a given set of input structures, which can readily be used in downstream steps for supervised or unsupervised machine learning. We show Caretta's performance on two benchmark datasets, and present an example application of Caretta in predicting the conformational state of cyclin-dependent kinases.

Python code available at `https://git.wur.nl/durai001/caretta`

## 4.1 Introduction

Protein structure alignment has recently been gaining attention in the bioinformatics field, becoming almost as popular as its cousin, protein sequence alignment. While sequence alignment aims to use amino acid substitution patterns and physicochemical properties to make a residue-residue correspondence between sequences of related proteins, structure alignment instead usually focuses on making an optimal superposition of the 3D coordinates of backbone atoms to establish such a correspondence. In many cases these two approaches agree with each other, especially in cases where the proteins under consideration share a high sequence homology. However, it has been repeatedly observed[1,2] that some protein families have divergent protein sequences and yet share a high structure, topology, and/or fold similarity, mostly due to the fact that structure tends to evolve slower than sequence[3]. For example, the ubiquitous TIM barrel structural fold is found in over 70 protein families all across nature, and even the most accurate sequence-based techniques cannot find relationships between these diverse sequences with the same structure[4]. In such cases, while sequence alignment may not be successful, structure alignment can still find meaningful residue correspondences.

Structure alignment has had applications in understanding evolutionary conservation and divergence patterns between proteins across different species[5], identifying conserved active site residues involved in catalytic reactions, creating structure-aware sequence profiles[6], structural similarity search against a database[7] and even as a method to design gold standard datasets for evaluating sequence alignment programs[8,9]. One area in which comparing multiple protein structures is only recently becoming popular is machine learning.

Though machine learning is not a new field, its popularity and applicability in bioinformatics has recently grown at a tremendous pace. In the protein and enzyme world, machine learning has successfully been applied to predict protein function, protein-protein interactions, drug-target binding, enzyme substrate specificity, thermostability, catalytic rates, binding affinity, and so on[10–12]. In many of these cases, protein sequences are used due to their widespread availability. However, the increase in both the number of experimentally solved structures, as well as the improvement in structure prediction using homology modelling and co-evolution based approaches, has led to the possibility of incorporating predicted or actual protein structure information (such as residue depth, electrostatic potentials etc.) in such algorithms to better predict and understand outcomes and properties associated with protein families[13,14].

The typical input for a machine learning algorithm has a tabular format, with each row representing an input protein and each column representing a particular feature or attribute extracted across all the proteins considered. Naturally, the construction of such an input table is often performed by means of a multiple protein alignment. Each column then consists of a particular feature value measured across all the residues in a particular alignment position. This then allows the prediction algorithm to look for patterns in these columns which are correlated with the desired response. For example, in an alignment of ten proteins, if one position is a Trp in the five proteins with a high catalytic rate and a Gly in the five with a lower catalytic rate then this residue position may be implicated in the reduction of catalytic activity. The power of machine learning algorithms lies in finding much more complex and interconnected patterns such as this one. Regions in the alignment with many insertions and deletions, however, can be more difficult to handle, as functionally equivalent residues may be split across multiple columns. This makes it harder for a predictor to spot patterns in a single column or link them together. Often, columns with too many gaps without feature information have to be discarded completely from the analysis, with the risk of losing out on predictive and catalytically important residues simply due to an alignment not fit for the task at hand.

Although there are a number of multiple structure alignment tools, different tools excel in different settings. Many existing multiple structure alignment algorithms, such as Matt[15,16], MUSTANG[17] and MultiProt[18], focus on and are optimized for aligning evolutionarily distant proteins, which may be from the same superfamily but only share short stretches of structurally conserved "core" regions. Concentrating on these core regions, typically by aligning short fragments of proteins and then assembling these intermediate alignments, leads to these methods overestimating the number of gaps in the alignment, as observed by Carpentier & Chomilier[19] in their multiple structure alignment benchmark. This is especially a hindrance in evolutionarily conserved families as one would expect long stretches of residue correspondences with only a small number of gaps. Therefore, there is a need for a multiple structure alignment tool aimed at returning accurate alignments with a high coverage for homologous protein families with divergent sequences and conserved structures. Machine learning methods which make use of these high-coverage alignments would then have a larger number of extracted residue features at their disposal, allowing for the pinpointing of under-explored residue positions related to an outcome of interest.

Here we present Caretta, a multiple structure alignment tool that additionally outputs aligned structural feature matrices. Caretta uses a combination of dynamic time warping[20] and progressive pairwise alignment[21] to align structures. The pairwise alignment algorithm makes an initial superposition of the two structures using either a signal-based rotation-invariant approach or secondary structure, and further refines the alignment using a scoring system based on the Euclidean distance between corresponding coordinates. The algorithmic novelty of Caretta is that information about the multiple structure alignment is fed into each progressive pairwise alignment in order to maintain and extend existing aligned blocks without disturbing them with insertions unlikely to be found within the same protein family.

We demonstrate that Caretta covers more residues in its alignments than competing tools while still maintaining accuracy. Testing on the widely used Homstrad dataset[22] shows that Caretta often performs on par with manual curation. Caretta is capable of outputting a matrix of features, such as bond angles and residue fluctuations, extracted from the input structures and aligned according to the multiple structure alignment. We use these feature matrices to demonstrate an example workflow of Caretta in machine learning, for classifying cyclin-dependent kinases (CDK) into active or inactive states[23]. Feature selection allows for pinpointing residues involved in state switching, some of which are confirmed by previous studies. A Caretta GUI application allowing for easy access and visualization of aligned structures and features is provided as well. Taken together, Caretta is a full-featured multiple structure alignment suite which provides tools for creating and exploring accurate structural alignments and for calculating structural features extracted from the proteins aligned, in order to successfully apply machine learning to identify distinguishing characteristics of a family of homologous proteins

## 4.2 Methods

Figure 4.1A depicts the workflow of Caretta for multiple structure alignment: an all *vs.* all pairwise alignment step followed by the construction of a guide tree for progressive alignment, to finally output a multiple alignment. Each intermediate pairwise alignment step uses the dynamic programming approach detailed in Section 4.2.1. These pairwise alignments use a combination of two different approaches (labelled B1 and B2 in Figure 4.1) to construct an initial superposition of structures, described in Section 4.2.2. The progressive alignment step, explained in Section 4.2.3 and Figure 4.1C, combines aligned structures into an alignment intermediate and boosts the weight of well-aligned residue positions, an approach which reduces the chances of unlikely insertions and deletions.

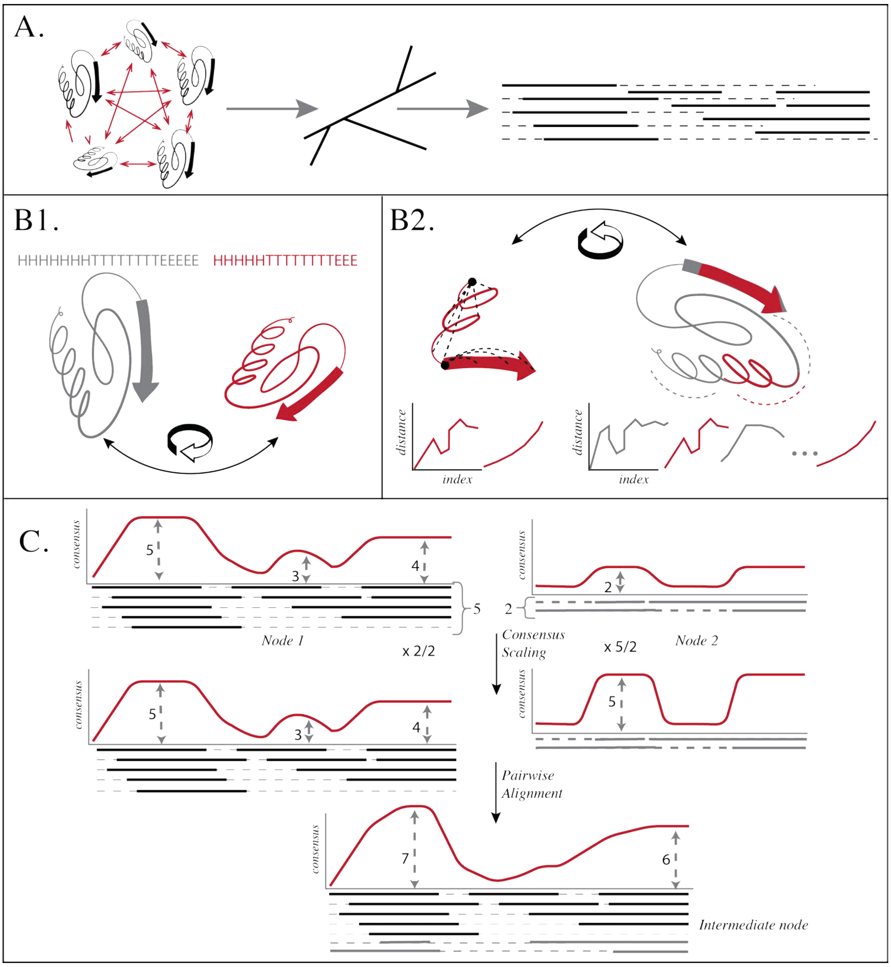


Figure 4.1: **A.** Caretta's multiple structure alignment workflow: an all *vs.* all pairwise alignment step, followed by construction of a guide tree and progressive alignment. **B.** The two approaches for initial rotation and superposition of two structures used in pairwise alignment: 1) aligning secondary structure codes and 2) dynamic time warping on one-dimensional signals of distances from all residues to the first or last residue in a segment. **C.** The guide tree specifies the two neighbours to combine at each progressive alignment step. These two neighbours can either be protein structures or previously combined intermediate nodes. A new intermediate node is created at each step by aligning and combining the two neighbours. The alignment step takes into account the number of times each position has been aligned in each of the two neighbours weighted by the number of structures in each neighbour ($consensus$ row, shown in red). This ensures that when the difference of the two is taken in calculating $Score_C$ (Equation 4.3), positions with fewer gaps get higher scores. After alignment, the $consensus$ row of the new intermediate node keeps track of the number of residues aligned at each position.

### 4.2.1 Dynamic programming based alignment

The algorithm underlying Caretta is dynamic programming alignment with affine gap costs as described by Altschul & Erickson[24]. This algorithm is used in different parts of Caretta with different scoring schemes and different gap open and gap extension penalties, described in the next sections. Supp. Section 4.1 contains pseudocode for all the remaining sections with the dynamic programming alignment algorithm represented as $DPAlign$.

### 4.2.2 Pairwise alignment

Pairwise alignment of two structures depends on the residue-to-residue distance between them. The underlying assumption of a similarity scoring scheme based on residue distance is that the proteins in question are already rotated and centred such that equivalent residues are close to each other. Such a superposition requires a correspondence between residues, *i.e.* an alignment, leading to a chicken and egg problem for pairwise alignment. Caretta solves this by making an initial superposition of two structures using the best out of two coarse alignments: the first based on secondary structure ($SecondarySuperpose$ in Supp. Section 4.1), and the second based on the alignment of one-dimensional rotation-invariant signals derived from overlapping contiguous segments of the two structures ($SignalSuperpose$ in Supp. Section 4.1). These two approaches are represented in Figure 4.1B1 and Figure 4.1B2 respectively, and described below:

1. The first method aligns the residues between two proteins according to their secondary structure elements. The secondary structure score or $Score_S$ is defined as below, where $s_i$ represents the DSSP secondary structure code (Supplementary Table 1) for residue $i$:

$$Score_S(i,j) = \begin{cases} 0 \text{ if } s_i = \text{'-'} \vee s_j = \text{'-'} \\ 1 \text{ if } s_i = s_j \\ -1 \text{ if } s_i \neq s_j \end{cases} \tag{4.1}$$

   This scoring system is used with gap open and gap extend penalties $\sigma_S = 1$ and $\epsilon_S = 0$ (since this scoring scheme works in increments or decrements of 1) to make an initial alignment. The two proteins are then superposed using the Kabsch algorithm[25] to find the rotation and translation matrix that optimally matches the aligning pairs of residues.

2. The second method performs dynamic time warping on rotation-invariant overlapping segments of two structures. Each segment represents each residue $r$ in a thirty-residue stretch of the structure by the Euclidean distances of its $\alpha$-carbon to the $\alpha$-carbon of the first residue in the segment ($\vec{P} = [d_0, d_1, ..., d_n]$). The score between two such segments is given as:

$$Score_P(i,j) = median_d \left( \exp \left( -\frac{(\vec{P}_{i,d} - \vec{P}_{j,d})^2}{10} \right) \right) \tag{4.2}$$

> After determining the alignment of these segments (by using $Score_p$ with zero gap penalties to allow for more leniency as the proteins are not yet in their correct orientation), the optimal rotation and translation of the $\alpha$-carbons of the first residues in each aligning pair of segments are calculated using the Kabsch algorithm[25] and used to superpose the two structures.
>
> This approach is repeated, taking the distances to the last $\alpha$-carbon in each segment instead of the first, to obtain a different superposition.

The superposition from the above two approaches giving the best-scoring alignment is chosen. The scoring method used by Caretta uses an RBF (Gaussian) kernel derived from the Euclidean distance between two (superposed) $\alpha$-carbon coordinates ($\vec{\alpha} = [\alpha_x, \alpha_y, \alpha_z]$), defined below:

$$Score_C(i,j) = \exp\left(-\gamma \sum (\vec{\alpha_i} - \vec{\alpha_j})^2\right) \tag{4.3}$$

Supp. Figure 4.1 shows the distribution of this score for different values of $\gamma$ as a function of Euclidean distance. We chose a $\gamma$ value of 0.03 as this causes a sharp drop to near-zero values at 8 Å while still yielding a score of around 0.6 at the commonly used structural equivalence cutoff of 4 Å, reflecting the belief that residues further away than 8 Å are not likely to be structurally or functionally equivalent.

This score is summed across all paired residues to derive the score of an alignment between two proteins $x$ and $y$:

$$PairwiseAlignmentScore_C(x,y) = \sum_{(i,j)\in\text{aligned residue pairs}} Score_C(x_i, y_i) \tag{4.4}$$

Caretta uses the scoring scheme $Score_C$ and $\sigma_C$ and $\epsilon_C$ as gap open and gap extend penalties (set to 1 and 0.01 for the alignments presented here) on the newly superposed coordinates to find the optimal correspondence between them ($PairwiseAlignment$ in Supp. Section 4.1).

When more than two structures are required to be aligned, pairwise alignments are made for all input structures. This step is essential for the guide tree construction described in the next section.

### 4.2.3 Multiple alignment

The idea behind a progressive alignment approach is to perform step-wise alignments of two stacks of previously aligned structures (or single structures) to result in a final stack of all aligned structures. The order of addition of structures is a crucial factor in the performance of this method. Aligning similar structures first, with a smoother progression towards distantly related structures, increases the chances of a good alignment. We construct a guide tree for determining the order of progression using maximum linkage neighbour joining[26] on the pairwise alignments constructed in Section 4.2.2. The pairwise tree score for two proteins is given by their pairwise alignment score (Equation 4.4) divided by the number of aligning pairs.

With the guide tree in place, the progressive alignment steps start, as illustrated in Figure 4.1C and Supp. Section 4.1 $MultipleAlignment$. While progressive alignment typically consists of independent pairwise alignment steps, the algorithmic novelty of Caretta lies in the introduction of a feedback loop between the state of the multiple structure alignment and each pairwise alignment, explained in detail below. For this purpose, an additional $consensus$ row, of length equal to the protein length, is maintained for each structure, initiated with a consensus weight parameter ($cw$, default=1). This row is concatenated to the coordinates $\vec{\alpha}$ of a protein before $Score_C$ in Equation 4.3 is calculated.

Before two neighbours in the guide tree are aligned, the $consensus$ row of each neighbour is multiplied by half the number of structures represented by the other neighbour. This ensures that when their difference is taken during the calculation of $Score_C$ in Equation 4.3 (with the $consensus$ row attached), the positions with equal $consensus$ values in both receive high scores (as the difference is close to zero), increasing their chances of being aligned. An intermediate node is created with a length equal to the length of the resulting pairwise alignment, representing the aligned stack of structures from the two neighbours. The $x$, $y$, and $z$ coordinates of this intermediate node are calculated by averaging the coordinates of the two initial structures after superposition, across the alignment. At each position in the alignment, the secondary structure code in the intermediate node is taken as the code of the input which does not have a gap, or the code of the first input if both are aligned. The $consensus$ term for each alignment position is the number of aligned residues at that position times the $cw$, i.e. well-aligned positions with fewer gaps have a higher $consensus$ value. This way, Caretta tends to maintain fully aligned core regions by avoiding the insertion of gaps at these locations as progressive alignment proceeds as such gaps are unlikely to happen in conserved protein families.

### 4.2.4 Benchmarking

Data

Caretta takes as input a list of PDB files, along with optional chain identifiers and start and end residue indices. All PDB file parsing is done using ProDy[27] and the secondary structure for each protein is derived using ProDy's execDSSP[28] function.

Caretta was tested on two benchmark datasets, Homstrad[22] and SABmark-Sup[9]. The PDB files for these two datasets were obtained from mTM-align's website[29] and Matt benchmark results[15,30] respectively, in order to directly compare results to the output of these two tools. To this end, the alignments for the Homstrad[22] and SABmark-Sup[9] datasets for Matt[15] and mTM-align[31] were obtained from Mattbench[30] and mTM-align's website[29] respectively. For 35 cases in the SABmark-Sup dataset, mTM-align returned alignments where at least one sequence did not match the corresponding PDB sequence. These cases are not shown in Figure 4.4A. Two protein families, labelled seatoxin and kringle in the Homstrad benchmark set are used to demonstrate and contrast the alignments returned by Caretta, Matt and mTM-align. The structures in these groups are superposed according to the gap-less positions in each alignment and visualized using PyMOL[32].

Metrics

To measure the quality of multiple structure alignments we make use of various metrics. The last two are defined for pairwise alignments, and are calculated for every pair of structures in the multiple structure alignment after superposing all structures to one reference structure, the longest protein. These are then averaged over all pairs to give the final score for a multiple alignment.

- **Gap-less positions** - Positions in an alignment that do not contain any gaps.
- **Homstrad equivalence score** - Percentage of gap-less positions which are present in the corresponding Homstrad reference alignment.
- **RMSD** - The root mean square deviation between two superposed structures in a pairwise alignment is given by:

$$\sqrt{\frac{\sum_{(i,j) \in \text{aligned residue pairs}} (\vec{\alpha_i} - \vec{\alpha_j})^2}{|\text{aligned residue pairs}|}}$$

- **Structurally equivalent residues** - Residues in the same alignment position of a pairwise alignment within 4 Å of each other after superposition.

Measuring Runtime

To estimate Caretta running times, we randomly chose 25 protein structures from the Homstrad dataset with differing lengths as "seeds". Each seed was used to form multiple groups of proteins to be aligned by Caretta. Forming each of these groups involved introducing noise to the seed coordinates to create a given number of members, from 13 to 93 in increments of 30. Caretta was then used to align these groups on a Linux workstation using 20 threads.

### 4.2.5 Feature extraction

In addition to multiple structure alignment, Caretta was designed specifically in order to enable feature extraction for downstream machine learning.

Structural features are extracted for each input protein, aligned according to Caretta's multiple structure alignment. All atom-level features are converted into $\alpha$-carbon, $\beta$-carbon, and mean residue features. For Gly, the $\alpha$-carbon is used for the $\beta$-carbon features as well.

ProDy[27] is used to calculate the 50-mode Gaussian Network Model (GNM) and Anisotropic Network Model (ANM) atom fluctuations using the `calcGNM`/`calcANM` functions followed by the `calcSqFlucts` function.

DSSP features are calculated using ProDy[27] to give hydrogen bond energies, surface accessibility, dihedral angles ($\alpha$), bend angles ($\kappa$), $\phi$, and $\psi$ backbone torsion angles, and tco angles (cosine angle between the C=O of residue $i$ and the C=O of residue $i-1$).

Residue depths are extracted using BioPython[33].

### 4.2.6 Cyclin-dependent kinase classification

Caretta's alignment and feature extraction capabilities are further demonstrated on the task of predicting the functional state of cyclin-dependent kinases (CDKs). PDB IDs of these proteins, along with the corresponding active/inactive labels, were obtained from McSkimming et al.[23]. These proteins were clustered by sequence similarity using an LZW kernel[34], and a single cluster containing 80 CDKs was chosen. A multiple structure alignment was made for these 80 CDKs followed by feature extraction of DSSP features, GNM and ANM fluctuations and residue depths. These features were aligned after discarding positions in the alignment which contained gaps. A logistic regression model with L1 penalty was trained for binary classification of CDK active/inactive state, and tested on 50 random splits of the data, each with 60 training points and 20 test points, using the scikit-learn Python library[35]. The importance of an alignment position was taken to be the sum of the absolute values of the feature coefficients for that position, averaged across all train/test splits. This scoring scheme was used to select the top 15 most informative residue positions, which were visualized on the proline-rich tyrosine kinase 2 (PYK2) CDK structure (PDB ID: 3FZP, chain A) using PyMOL[32].

### 4.2.7 GUI application

The Caretta GUI was built using Dash and Dash-Bio[36]. It takes as input a list of PDB IDs, either from a user-specified folder or from a list of structures associated with a user-inputted Pfam domain, and performs multiple structure alignment on these structures. The results are displayed in three different panels:

- Structure alignment - displays the superposed 3D structures of the input proteins;
- Sequence alignment - displays the multiple sequence alignment, coloured by hydrophobicity;
- Feature alignment - displays aligned structural features. The feature name under consideration can be changed using a drop-down box.

These three panels are interlinked via interactive capabilities. Clicking a protein or a residue position in any of the three panels highlights the corresponding protein or position in the other two. All three panels can also be exported to different file formats for downstream use.

## 4.3 Results

### 4.3.1 Caretta returns accurate alignments with higher coverage

We compare Caretta with two popular multiple structure alignment methods, Matt and mTM-align. Matt is a fragment-based approach, which allows for local flexibility between fragment pairs from two input structures and then uses a dynamic programming algorithm to assemble these intermediate pairs[15]. mTM-align[31] instead performs global alignment and builds upon the pairwise structure alignment algorithm

TM-align[37], which uses the length-independent TM-score as a measure of similarity between two proteins in a dynamic programming approach. mTM-align then progressively assembles these pairwise alignments into a multiple structure alignment.

These two MSA tools were tested along with Caretta on the popular Homstrad and SABmark-Sup datasets. Assessing the quality of multiple structure alignments is a difficult task and, depending on the metric used, different aspects of the alignment come under consideration. While RMSD (root mean square deviation) is often used, it has been observed that fewer aligned residues can easily lead to smaller RMSDs at the expense of a very gap-filled alignment[31], which can easily happen in the case of proteins with conserved cores but flexible regions that are not often aligned. While the conserved core can be responsible for the overall stability and function of the protein, the flexible regions can occur in and around active sites or interaction sites and lead to differences in enzyme specificity towards substrates, products, or interaction partners[38] - making them immensely important for machine learning aimed at predicting determinants of such specificities. Thus, gap-filled alignments focusing on low RMSDs, while accurate and useful for superposition of structures, are suboptimal for machine learning as the features of many potentially relevant residues are discarded due to a lack of data in those positions. In most cases, positions with over a certain percentage of aligned residues are considered, with gaps replaced by zeros or by the average of the feature values in that position[23]. Therefore, when benchmarking Caretta we emphasize the coverage of the alignment along with structural equivalence measures such as RMSD.

The Homstrad dataset is unique in that it provides manually curated and annotated alignments, representing a ground truth. This dataset has examples from various homologous protein families, typical of the kinds of applications where machine learning would be applied. Since these proteins are homologous, a high alignment coverage is expected as many of the residues are functionally equivalent, with few insertions and deletions. In Figure 4.2 we show the percentage of gap-less columns found by each aligner that are the same as the corresponding column in the Homstrad reference (Homstrad equivalence score), against the percentage of all gap-less columns in the alignment. Caretta clearly outperforms the other aligners by regularly finding near-optimal alignments with a high coverage. In the majority of cases (65% for Matt, and 82% for mTM-align), Caretta also finds the same or more structurally equivalent residues within gap-less positions. Taken together, this indicates that the increase in gap-less positions is warranted in that Caretta still finds accurate residue pairings.

Figure 4.3 shows two examples where Caretta does a better job of multiple structure alignment in terms of Homstrad equivalence. The first case shows a family of small, loop-filled structures where the pitfalls of optimizing for RMSD become clear. Matt, in this case, only gap-lessly aligns 8 residues and has a Homstrad equivalence of 3%, while Caretta achieves a Homstrad equivalence score of 58% by correctly aligning areas where structural flexibility makes it difficult to accurately pinpoint equivalent residue pairs. The second case demonstrates a family in which some members structurally deviate from the others in a small region. Such regions are especially relevant for machine learning as they may be responsible for a change in a response variable

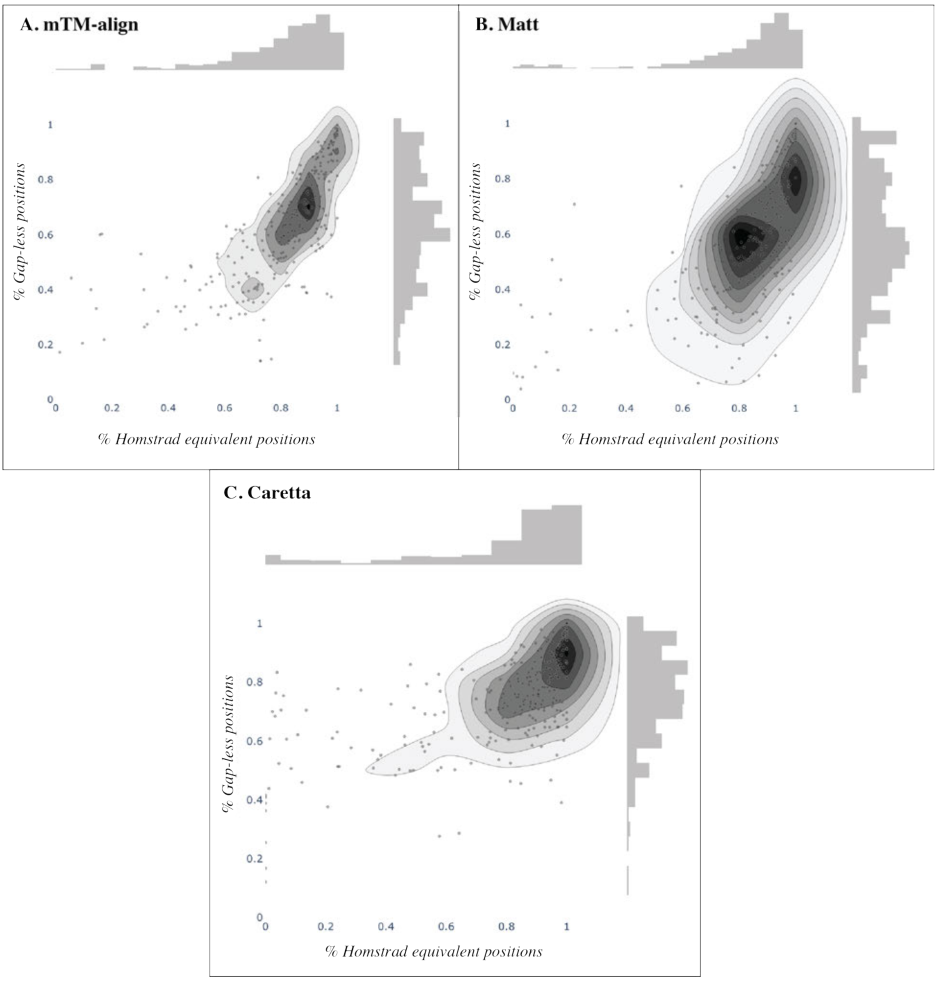


Figure 4.2: Plots showing the percentage of gap-less positions in an alignment which are identical to the corresponding Homstrad reference alignment vs. the percentage of all gap-less positions. **A**, **B**, and **C** show the results for alignments generated by mTM-align, Matt, and Caretta respectively.

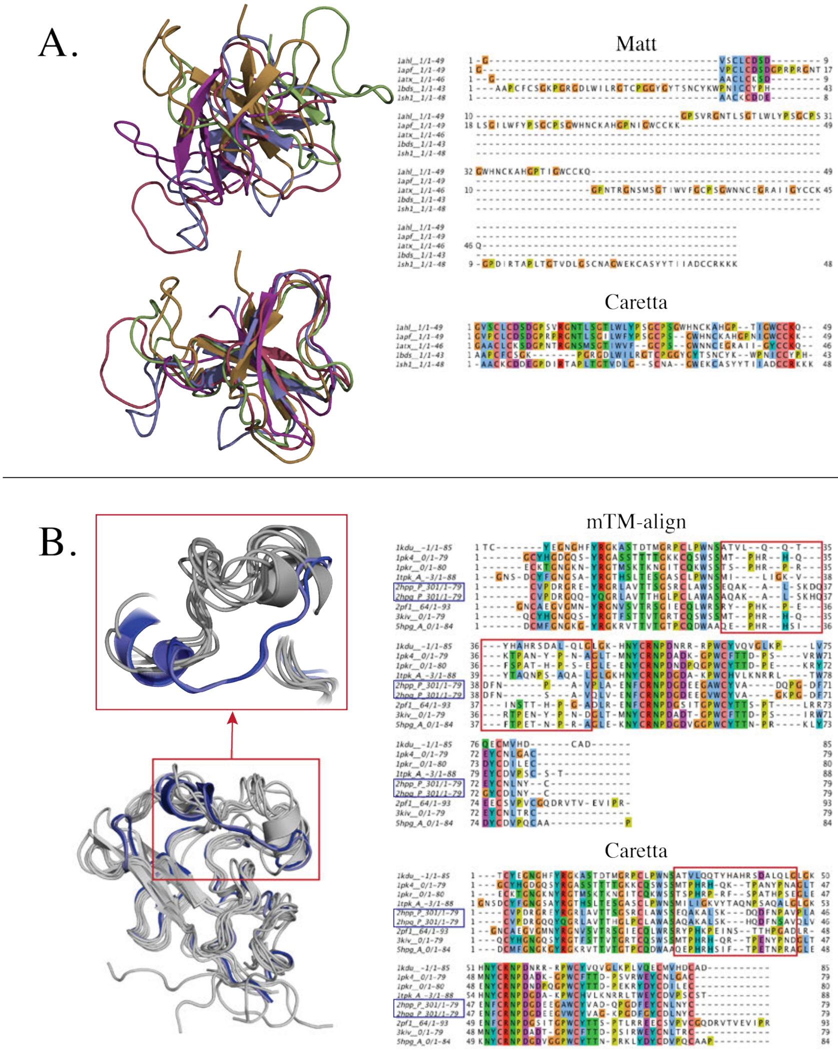


Figure 4.3: Two examples of protein families in which Caretta finds a better alignment than Matt and mTM-align. **A.** structures from the "seatoxin" family (a collection of toxins released by sea anemones), superposed according to alignments made by Matt (top) and Caretta (bottom) respectively, with the alignments shown on the right. **B.** structures from the "Kringle" family (PFAM ID: PF00051) superposed according to Caretta's alignment. Two structurally divergent proteins are highlighted in blue and the region of divergence is highlighted in red, both in the structure superposition and in the corresponding alignments on the right. These two groups of proteins are obtained from the Homstrad benchmark dataset[22].

such as substrate specificity, catalytic rate etc. Both Caretta and mTM-align lead to the same superposition of structures for this family, but mTM-align inserts gaps in the highlighted region such that the two divergent proteins cannot be compared here.

The SABmark-Sup benchmark dataset consists of proteins from the same superfamily with distant homology[9]. These proteins are much harder to align as usually only small fragments have any meaningful correspondence. Though Caretta has adjustable gap penalties that can be useful in such cases to allow for substructure alignment, these are still not the optimal conditions for the algorithm. While Matt is known to yield alignments with low RMSDs in this dataset, this is at the expense of coverage, which is often quite low. Figure 4.4 shows the average RMSD *vs.* average percentage of gap-less columns for Matt, mTM-align and Caretta on the SABmark-Sup dataset. While Matt and mTM-align have a gap-less percentage range typically within 20-60%, this increases to 40-80% for Caretta, often still within the same RMSD range. This indicates that Caretta also performs well at fragmented substructure alignment, though optimizing the gap penalties and consensus weight may improve results further for individual cases.

The time complexity of Caretta's alignment algorithm is $O(n^2l^2)$ where $n$ is the number of proteins and $l$ is the length of the longest protein in the alignment. Figure 4.5 shows the time taken for Caretta alignment for varying numbers and lengths of proteins. These results show that aligning a reasonably large set of protein structures (50-90) with a mid-range residue length (200-300) takes less than 2 hours on a workstation with 20 threads.

### 4.3.2 An application of Caretta in predicting Cyclin-dependent kinase conformation

Apart from $\alpha$-carbon coordinates, residues in a protein carry a wealth of structural information, as a result of the physicochemical differences between amino acids and the many interactions to neighbouring residues. This information can be extracted from structures and used to explore differences and similarities between proteins in the same family performing different functions. To enable such exploration, Caretta calculates and outputs various structural features aligned according to the core columns in the multiple structure alignment. The feature matrices outputted by Caretta can be also used for downstream tasks such as dimensionality reduction and supervised learning. As proteins typically have many hundreds of residues each with tens of features, a feature selection step is recommended for small datasets, to focus on functionally important residues.

An application of Caretta for a classification task is presented using a dataset of cyclin-dependent kinases (CDKs). This family of enzymes is involved in cell cycle regulation and its members share a high degree of structural similarity. Classical kinase inhibitors bind to the ATP site of CDKs and compete for substrate binding[39]. The determination of additional inhibitor binding sites in these enzymes, which would switch their state from active to inactive in terms of substrate binding, is a challenging

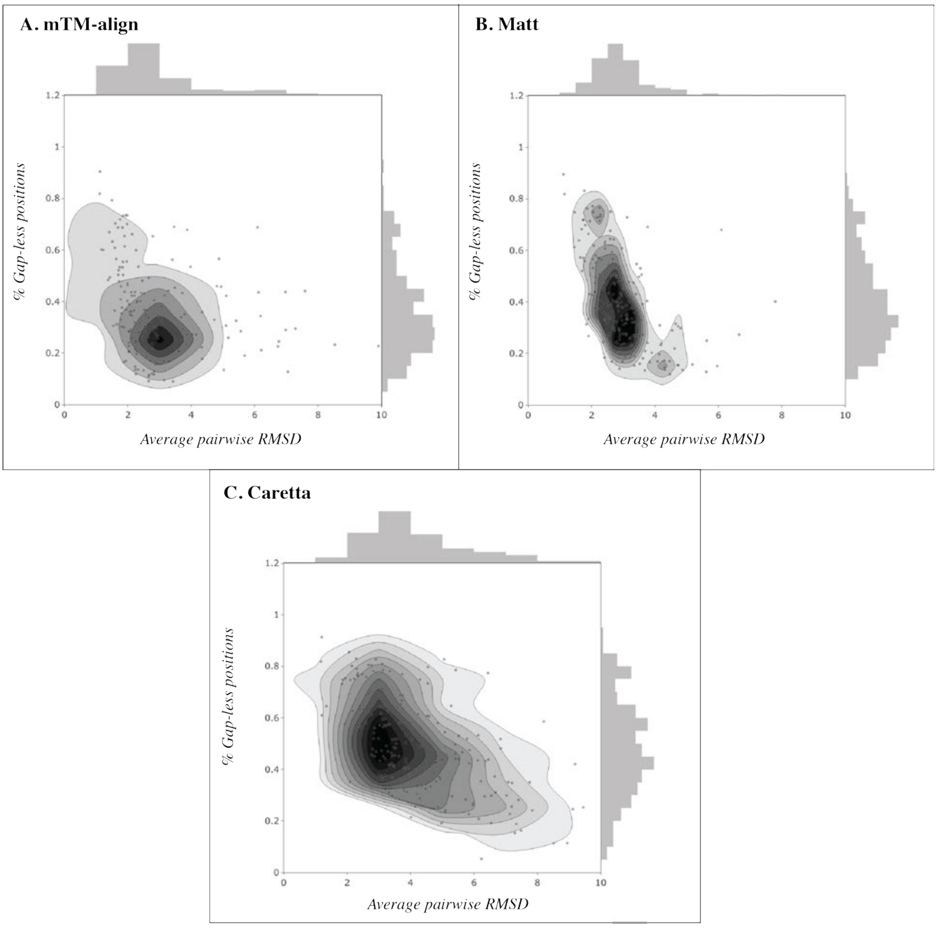


Figure 4.4: Plots showing average pairwise RMSD vs. percentage of gap-less alignment positions across the alignments in the SABmark-sup dataset. **A**, **B**, and **C** show the results for alignments generated by mTM-align, Matt, and Caretta respectively.

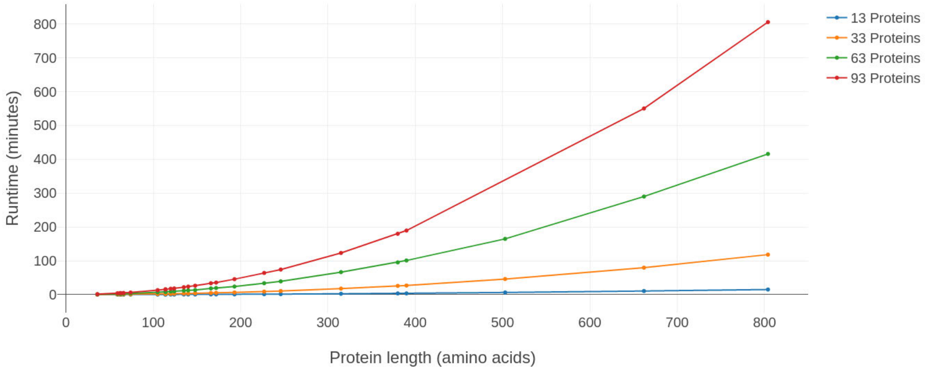


Figure 4.5: Runtime measured in minutes for Caretta alignment using 20 threads on proteins of differing lengths (constructed as described in Section 4.2.4. Each line represents a different number of proteins aligned.

and significant problem in the drug design field. One intriguing aspect of this family is that the same or very similar sequences can switch state depending on their structural conformation which means that sequence similarity cannot be successfully used for classification[23].

Fortunately, a large number of CDK structures have been experimentally solved. We used Caretta to align 80 CDK structures and extracted bond angles, Gaussian and anisotropic network model residue fluctuations, residue depths and solvent accessibility features from these structures. The alignment had a mean pairwise RMSD of 3.08 Å and 128 gap-less positions. We trained a logistic regression model to predict active/inactive states of CDKs using features aligned according to these gap-less positions and attained a mean cross-validation accuracy of 98%, with only 60 structures used for training in each split. The performance is summarized in Figure 4.6A in a Receiver operating characteristic (ROC) curve. Summing the absolute values of the feature coefficients for a residue across all splits allowed us to rank informative residue positions and pinpoint residues relevant to the activation process, as shown in Supp. Figure 4.2. Figure 4.6B labels the fifteen most informative residues on the structure of inactive proline-rich tyrosine kinase 2 (PYK2), co-crystallized along with an ATP-mimetic kinase inhibitor (ATP$\gamma$S). Supp. Figure 4.3 shows a PCA plot of the feature values of these top 15 residues, demonstrating a clear distinction between active and inactive CDKs. Interestingly, a number of the selected residues lie close to the inhibitor, with one falling within the well-studied DFG motif[39], indicated in the figure. The remaining selected residues cluster underneath this motif, indicating flexibility in these regions associated with a conformational change. This simple example demonstrates the power of a robust structural alignment, combined with features describing various aspects of protein structures, in exploring distinguishing characteristics of protein families. Insights gained from such studies can be utilized for mutational studies to engineer enzymes with desired activity, or in inhibitor design. While CDKs are relatively unique in that there are many solved crystal structures, due to the advances in homology modelling as well as the growing size of the PDB, most protein families can be supplemented with accurate structural models, which can then be aligned and analysed in a similar way with Caretta.

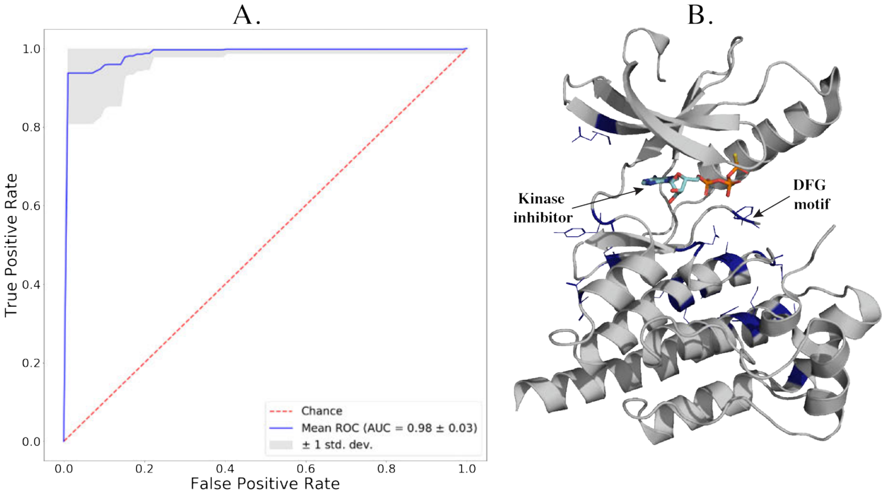


Figure 4.6: **A.** ROC-AUC curve showing the cross-validation performance of the logistic regression model to predict the state (active/inactive) of cyclin-dependent kinases (CDKs). This model is trained on structural residue features (bond angles, residue depths, fluctuations, and solvent accessibilities), and aligned according to Caretta's multiple structure alignment of the CDKs. **B.** Structure of active proline-rich tyrosine kinase 2 (PYK2), co-crystallized with the ATP$\gamma$S ATP-mimetic kinase inhibitor. The catalytically important DFG motif is labelled and the fifteen most predictive residues for active/inactive state determination are coloured in blue and represented as sticks.

### 4.3.3 Caretta can be used to visually explore structure alignments and features.

A Caretta GUI application can be found at `www.bioinformatics.nl/caretta` for aligning selected structures, from either a Pfam domain or a custom folder, and exploring their structural features. Figure 4.7 shows the kind of information that can be obtained. The application is fully interactive, with the sequence and feature alignments linked to the corresponding residues in the structure alignment. Different features such as bond and torsion angles, electrostatics, atom fluctuations etc. can be visualized separately, and the means and standard deviations across all proteins are shown for each position allowing the user to easily pinpoint highly variable or highly conserved residues or residues. While the website only allows for the alignment of up to 40 structures, the application can be installed locally to avoid this restriction. In

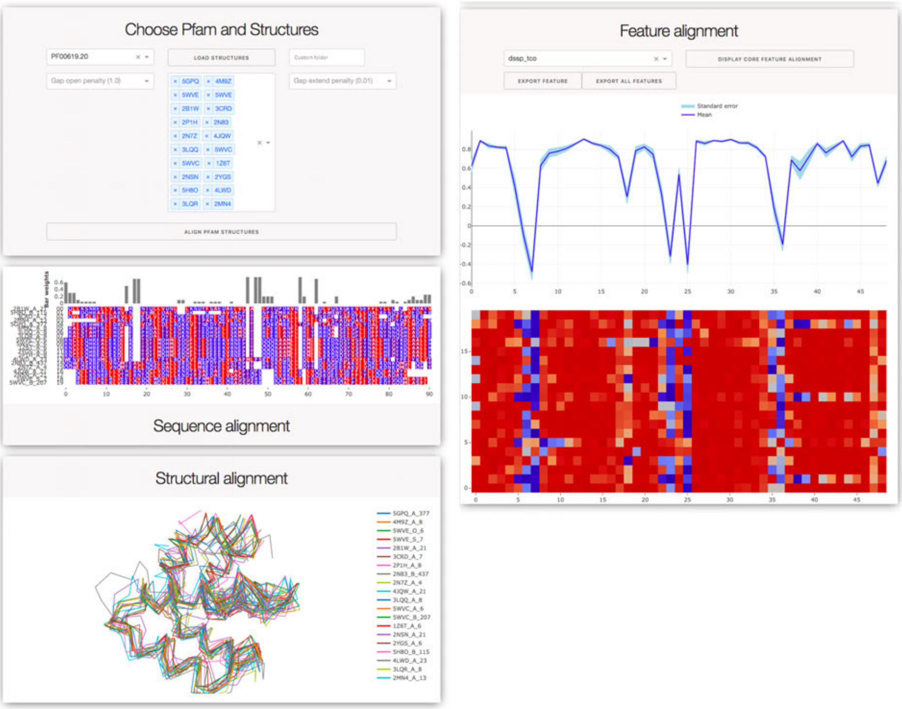


Figure 4.7: An example of Pfam domain alignment possible with the Caretta website (found at `www.bioinformatics.nl/caretta`). The user selects a Pfam domain and is given the list of PDB IDs associated with that domain. The website allows selection of up to 40 PDB IDs to align. Once alignment is complete, three panels are displayed, showing the multiple sequence alignment, the corresponding superposition of the structures, and the alignment of structural features (with a drop-down menu to choose between different features). These three panels are interconnected, allowing the user to select proteins and residue positions across all three views at once.

addition, Caretta can also be installed as a command line application or used as a Python library for easy handling of multiple structures, features, and alignments.

## 4.4 Conclusion

Multiple sequence alignment is an integral part of a broad range of bioinformatics research topics, including phylogenetics, functional domain identification, co-evolution analysis and machine learning to predict functional properties of proteins. Compared to protein sequences, protein structures echo an even deeper evolutionary history that in a more direct way relates to their function. Previously, this kind of analysis was hindered by the scarcity of protein structures available. However, the number of solved protein structures is increasing at a great pace, and structural modelling methods are also improving rapidly, in part due to the use of co-evolutionary information when reliable structural templates are not available. This means it is now possible to analyse patterns correlating with function in a protein family by aligning, comparing and applying machine learning on a large set of solved or modelled structures.

We contribute to this field with Caretta, a multiple structure alignment suite which returns accurate alignments with an increased ratio of aligned positions to make the best use of structural features from functionally comparable residues. Dong et al.[31] noticed that the accuracy of a multiple structure alignment depends heavily on the quality of the individual pairwise alignments, which in turn depends on the initial superposition of two proteins, often accomplished by approximate point cloud registration techniques. Caretta uses signals of distances derived from overlapping contiguous stretches of residues to make this initial superposition, a novel rotation-invariant technique. This, combined with a novel feedback approach to maintain well-aligned blocks of residues in the multiple alignment, works well with protein families where large and numerous stretches of insertions are not expected to be found.

In the Caretta GUI, we coupled structural alignment and feature extraction with a visual interface to pinpoint relevant proteins and residue positions for downstream prediction tasks. This kind of feature selection becomes necessary as proteins typically have many hundreds of residues, each of which is described by a number of structural features. This quickly leads to what is known as the "large **p** small **n**" problem in machine learning, where the number of descriptors far exceeds the number of labelled data points from which to learn. Feature selection in such cases removes noisy and irrelevant features, and can be used to find residue positions correlated with the response variable. We demonstrated this in our application on predicting the conformational state of cyclin-dependent kinases, where we found a small set of predictive residues, some of which lie in previously studied motifs known to be involved in conformational change.

More research into protein families using the approach we present for dealing with structural alignments and residue selection across a large set of structural features will lead to improvements and novel techniques for feature selection, dimensionality reduction, and learning that work well on such large, hierarchically structured data.

Given the prominent role in present-day bioinformatics of both machine learning and homology modelling, this will lead to further breakthroughs in using protein structures to analyse protein function.

## Acknowledgements

We thank Rens Holmer for valuable feedback on early drafts and stimulating discussions on the algorithmic approach, Barbara Terlouw for testing and reporting bugs in Caretta, and Rodrigo Vargas Honorato for his insightful suggestions and comments. This work was supported by the Netherlands Organization for Scientific Research (NWO).

## Supplementary information

All supplementary sections and figures are available online at `https://doi.org/10.1016/j.csbj.2020.03.011`

# References

[1] Flower, D. R., North, A. C., & Sansom, C. E. (2000). The lipocalin protein family: Structural and sequence overview. *Biochimica et Biophysica Acta (BBA)-Protein Structure and Molecular Enzymology*, *1482*, 9–24.

[2] Illergård, K., Ardell, D. H., & Elofsson, A. (2009). Structure is three to ten times more conserved than sequence—A study of structural response in protein cores. *Proteins: Structure, Function, and Bioinformatics*, *77*, 499–508.

[3] Redfern, O. C., Dessailly, B., & Orengo, C. A. (2008). Exploring the structure and function paradigm. *Current Opinion in Structural Biology*, *18*, 394–402.

[4] Nagano, N., Orengo, C. A., & Thornton, J. M. (2002). One fold with many functions: The evolutionary relationships between TIM barrel families based on their sequences, structures and functions. *Journal of Molecular Biology*, *321*, 741–765.

[5] Leibowitz, N., Fligelman, Z. Y., Nussinov, R., & Wolfson, H. J. (2001). Automated multiple structure alignment and detection of a common substructural motif. *Proteins: Structure, Function, and Bioinformatics*, *43*, 235–245.

[6] Rozewicki, J., Li, S., Amada, K. M., Standley, D. M., & Katoh, K. (2019). MAFFT-DASH: Integrated protein sequence and structural alignment. *Nucleic Acids Research*, .

[7] Gibrat, J.-F., Madej, T., & Bryant, S. H. (1996). Surprising similarities in structure comparison. *Current Opinion in Structural Biology*, *6*, 377–385.

[8] Bahr, A., Thompson, J. D., Thierry, J.-C., & Poch, O. (2001). BAliBASE (Benchmark Alignment dataBASE): Enhancements for repeats, transmembrane sequences and circular permutations. *Nucleic Acids Research*, *29*, 323–326.

[9] Van Walle, I., Lasters, I., & Wyns, L. (2004). SABmark—a benchmark for sequence alignment that covers the entire known fold space. *Bioinformatics*, *21*, 1267–1268.

[10] Ding, H., Takigawa, I., Mamitsuka, H., & Zhu, S. (2013). Similarity-based machine learning methods for predicting drug–target interactions: A brief review. *Briefings in bioinformatics*, *15*, 734–747.

[11] Michael, J., Ross, D. et al. (1992). Modelling the structure and function of enzymes by machine learning. *Faraday Discussions*, *93*, 269–280.

[12] Fariselli, P., Pazos, F., Valencia, A., & Casadio, R. (2002). Prediction of protein–protein interaction sites in heterocomplexes with neural networks. *European Journal of Biochemistry*, *269*, 1356–1361.

[13] Berliner, N., Teyra, J., Çolak, R., Lopez, S. G., & Kim, P. M. (2014). Combining structural modeling with ensemble machine learning to accurately predict protein fold stability and binding affinity effects upon mutation. *PLoS One*, *9*, e107353.

[14] Ferraro, E., Via, A., Ausiello, G., & Helmer-Citterich, M. (2006). A novel structure-based encoding for machine-learning applied to the inference of SH3 domain specificity. *Bioinformatics*, *22*, 2333–2339.

[15] Menke, M., Berger, B., & Cowen, L. (2008). Matt: Local flexibility aids protein multiple structure alignment. *PLoS Computational Biology*, *4*, e10.

[16] Shegay, M. V., Suplatov, D. A., Popova, N. N., Švedas, V. K., & Voevodin, V. V. (2019). parMATT: Parallel multiple alignment of protein 3D-structures with translations and twists for distributed-memory systems. *Bioinformatics*, *35*, 4456–4458.

[17] Konagurthu, A. S., Whisstock, J. C., Stuckey, P. J., & Lesk, A. M. (2006). MUSTANG: A multiple structural alignment algorithm. *Proteins: Structure, Function, and Bioinformatics*, *64*, 559–574.

[18] Shatsky, M., Nussinov, R., & Wolfson, H. J. (2002). MultiProt—a multiple protein structural alignment algorithm. In *Guigó r., Gusfield d. (Eds) Algorithms in Bioinformatics. WABI 2002. International Workshop on Algorithms in Bioinformatics* (pp. 235–250). Springer volume 2452.

[19] Carpentier, M., & Chomilier, J. (2019). Protein multiple alignments: Sequence-based versus structure-based programs. *Bioinformatics*, *35*, 3970–3980.

[20] Vintsyuk, T. K. (1968). Speech discrimination by dynamic programming. *Cybernetics and Systems Analysis*, *4*, 52–57.

[21] Hogeweg, P., & Hesper, B. (1984). The alignment of sets of sequences and the construction of phyletic trees: An integrated method. *Journal of Molecular Evolution*, *20*, 175–186.

[22] Mizuguchi, K., Deane, C. M., Blundell, T. L., & Overington, J. P. (1998). HOMSTRAD: A database of protein structure alignments for homologous families. *Protein Science*, *7*, 2469–2471.

[23] McSkimming, D. I., Rasheed, K., & Kannan, N. (2017). Classifying kinase conformations using a machine learning approach. *BMC Bioinformatics*, *18*, 86.

[24] Altschul, S. F., & Erickson, B. W. (1986). Optimal sequence alignment using affine gap costs. *Bulletin of Mathematical Biology*, *48*, 603–616.

[25] Kabsch, W. (1976). A solution for the best rotation to relate two sets of vectors. *Acta Crystallographica Section A: Crystal Physics, Diffraction, Theoretical and General Crystallography*, *32*, 922–923.

[26] Saitou, N., & Nei, M. (1987). The neighbor-joining method: A new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution*, *4*, 406–425.

[27] Bakan, A., Meireles, L. M., & Bahar, I. (2011). ProDy: Protein dynamics inferred from theory and experiments. *Bioinformatics*, *27*, 1575–1577.

[28] Touw, W. G., Baakman, C., Black, J., te Beek, T. A., Krieger, E., Joosten, R. P., & Vriend, G. (2014). A series of PDB-related databanks for everyday needs. *Nucleic Acids Research*, *43*, D364–D368.

[29] Dong, R., Peng, Z., Zhang, Y., & Yang, J. mTM-align benchmark results. URL: `http://yanglab.nankai.edu.cn/mTM-align/benchmark/`.

[30] Menke, M., Berger, B., & Cowen, L. Matt benchmark results. URL: `http://cb.csail.mit.edu/cb/matt/homstrad/`.

[31] Dong, R., Peng, Z., Zhang, Y., & Yang, J. (2017). mTM-align: An algorithm for fast and accurate multiple protein structure alignment. *Bioinformatics*, *34*, 1719–1725.

[32] DeLano, W. L. et al. (2002). PyMOL: An open-source molecular graphics tool. *CCP4 Newsletter on Protein Crystallography*, *40*, 82–92.

[33] Cock, P. J., Antao, T., Chang, J. T., Chapman, B. A., Cox, C. J., Dalke, A., Friedberg, I., Hamelryck, T., Kauff, F., Wilczynski, B. et al. (2009). BioPython: Freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics*, *25*, 1422–1423.

[34] Filatov, G., Bauwens, B., & Kertész-Farkas, A. (2018). LZW-Kernel: Fast kernel utilizing variable length code blocks from LZW compressors for protein sequence classification. *Bioinformatics*, *34*, 3281–3288.

[35] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V. et al. (2011). Scikit-Learn: Machine learning in Python. *the Journal of Machine Learning Research*, *12*, 2825–2830.

[36] Plotly. Dash web-app framework. URL: `https://dash.plot.ly/`.

[37] Zhang, Y., & Skolnick, J. (2005). TM-align: A protein structure alignment algorithm based on the TM-score. *Nucleic Acids Research*, *33*, 2302–2309.

[38] Li, L., Shakhnovich, E. I., & Mirny, L. A. (2003). Amino acids determining enzyme-substrate specificity in prokaryotic and eukaryotic protein kinases. *Proceedings of the National Academy of Sciences*, *100*, 4463–4468.

[39] Endicott, J. A., & Noble, M. E. (2013). Structural characterization of the cyclin-dependent protein kinase family. *Biochemical Society Transactions*, *41*.

# CHAPTER 5

# Geometricus represents protein structures as shape-mers derived from moment invariants

Janani Durairaj, Mehmet Akdel, Dick de Ridder, and Aalt D.J. van Dijk

This chapter has been published as

Durairaj, Janani, et al. "Geometricus represents protein structures as shape-mers derived from moment invariants." *Bioinformatics* 36.Supplement_2 (2020): i718-i725.

# Abstract

**Motivation**: As the number of experimentally solved protein structures rises, it becomes increasingly appealing to use structural information for predictive tasks involving proteins. Due to the large variation in protein sizes, folds, and topologies, an attractive approach is to embed protein structures into fixed-length vectors, which can be used in machine learning algorithms aimed at predicting and understanding functional and physical properties. Many existing embedding approaches are alignment-based, which is both time-consuming and ineffective for distantly related proteins. On the other hand, library- or model-based approaches depend on a small library of fragments or require the use of a trained model, both of which may not generalize well.

**Results**: We present Geometricus, a novel and universally applicable approach to embedding proteins in a fixed-dimensional space. The approach is fast, accurate, and interpretable. Geometricus uses a set of 3D moment invariants to discretize fragments of protein structures into shape-mers, which are then counted to describe the full structure as a vector of counts. We demonstrate the applicability of this approach in various tasks, ranging from fast structure similarity search, unsupervised clustering, and structure classification across proteins from different superfamilies as well as within the same family.

Python code available at `https://github.com/TurtleTools/geometricus`

## 5.1 Introduction

The number of structures added to the Protein Data Bank[1] has been increasing rapidly, with over 10,000 structures deposited in 2019 alone. Meanwhile, major advances have been made in the areas of homology-based and *de novo* protein structure modelling[2]. This increased availability of protein structures has enabled protein biologists and bioinformaticians to start including structural data and information in protein function studies instead of being confined to the sole use of sequence data. These studies address a variety of questions, such as finding remote protein homologs with a similar structural fold, or defining the properties of a single protein family. Protein structures evolve slower than sequences, and encode long-range contact and fold information that are often crucial for protein activity. Hence, our understanding of molecular biology can be greatly enhanced by the inclusion of protein structures.

For both structures and sequences, choosing the right computational method to generate a representation of a protein for comparison and prediction purposes is crucial. This is especially true for machine learning methods, which often require variable-length sequences of amino acids, coordinate sets or other residue descriptors to be transformed into fixed-length representations. These representations can be used as input for supervised and unsupervised machine learning methods or be compared using standard vector distance formulae. As proteins typically cover a wide range of shapes, sizes and topological folds, the choice of representation is not always straightforward and may depend on the scale of the study. For instance, research questions addressing proteins within a single family may opt to use alignment-based representations[3,4]. These have the advantage of easy interpretability, as each residue can be directly mapped to a column in the transformed representation. However, alignment is computationally expensive, and its accuracy decreases with decreasing protein similarity.

To solve this, alignment-free methods were introduced, which learn a reduced and condensed representation of proteins without an explicit alignment. There are many examples of such approaches using machine learning and deep learning methods to learn generic patterns and features of the protein sequence space[5,6]. Structure-based representations also exist[7,8] but are generally more difficult to generate due to the three-dimensional nature of structures compared to the one-dimensional sequences. Some structure-based "alignment-free" methods generate a representation of a protein of the same length as the sequence and then use sequence alignment or calculate sequence similarity to compare these structural sequences in 1D[9,10]. The conserved nature of protein structure circumvents the problem of decreasing accuracy of sequence alignment in these approaches.

Many structure embedding techniques make use of a library of small structural fragments to which fragments of each input structure are compared, usually requiring the calculation of rotations and translations that would orient the input fragment and the library fragment in the same position[7]. To reduce the computational load of these structure-structure comparisons, library sizes are limited. Newer techniques, which make use of deep learning[8], do not need a library but still require a pre-trained model to generate new embeddings. In both cases the size, scale and resolution of the embedding is highly dependent on the initial choice of library fragments or training data used, and thus may not be applicable to research questions about a different protein set. Also, in both these approaches it is difficult to link predictive importance to functionally important residues or structural regions, which is often desired in studies aimed at understanding underlying mechanisms in protein biology.

To address these disadvantages of existing approaches, we introduce Geometricus, a novel structure embedding technique based on 3D rotation-invariant moments. Moment invariants were proposed in the 1960s for 2D images[11] and have been used extensively in the image processing field for object detection[12] and character recognition[13] among other applications. In the 1980s they were subsequently adapted to the 3D field[14], and found applications in the fields of robotics[15], gesture recognition[16], brain morphology[17], and even in structural biology[18]. Rotation-invariant moments yield identical output when performed on any translated and/or rotated version of a set of continuous or discrete three-dimensional coordinates. This implies that coordinate sets can be compared without the need for superposition. Alternatively, any set of coordinates can be represented by a number of these moments.

To generate a Geometricus embedding for a protein structure, we fragment the structure into overlapping $k$-mers based on sequence, as well as into overlapping spheres, calculated for a certain radius, based on 3D coordinate information. Moment invariants are calculated for each of the coordinate sets corresponding to these structural fragments, and then binned into *shape-mers*, each of which represents a set of similar structural fragments. Counting the occurrences of these shape-mers across the protein yields a representation of the whole protein structure as a fixed-length vector of counts, similar to an amino acid $k$-mer count vector describing a protein sequence. As the moment invariant calculation is simple, the entire embedding process runs in the order of tens of milliseconds per protein and is easily parallelized. In addition, each element in the count vector can be mapped back to the residues forming the corresponding shape-mer, allowing for interpretation of predictive residues on par with alignment-based approaches. The shape-mer binning process is easily controllable, allowing for coarse shape-mer definitions for divergent proteins with distinct structures, or a fine-grained resolution for closely related proteins from the same family. This makes Geometricus suitable for a variety of tasks where library-based or model-based embeddings would struggle or require expensive retraining.

We demonstrate the effectiveness and versatility of Geometricus embeddings in a variety of machine learning approaches and other applications applied to datasets of varying structure similarity. Geometricus can be used for very fast structure similarity searches, while maintaining accuracy close to that obtained by alignment-based methods. The innate simplicity of the approach enables flexibility in application, such that embeddings can be optimized for the task at hand, as we demonstrate using datasets with proteins from different superfamilies and within the same family. Geometricus is available as a Python library at `https://github.com/TurtleTools/geometricus`.

## 5.2 Methods

### 5.2.1 Protein Embedding

To generate embeddings for a set of proteins, we define so-called shape-mers which are analogous to sequence $k$-mers. A shape-mer represents a set of similar structural fragments, each a collection of coordinates in 3D space. The following sections describe the process of generating these structural fragments, their subsequent conversion into rotation- and translation-invariant moments, the moment-based grouping of structural fragments into shape-mers, and finally, shape-mer counting to obtain the resulting embedding.

#### Protein Fragmentation

We consider two different ways of dividing a protein with $l$ residues into structural fragments, using its $\alpha$-carbon coordinates, $\boldsymbol{\alpha} = \{\boldsymbol{\alpha_i} | \boldsymbol{\alpha_i} = (\alpha_i^x, \alpha_i^y, \alpha_i^z), i : 1, ..., l\}$.

1. ***k*-mer based** - for a given value of $k$, a protein is divided into $l$ $k$-mer-based structural fragments, $\{C_i^k, i : 1, ..., l\}$ where

$$C_i^k = \{\boldsymbol{\alpha_j} | j \in (\max(1, i - \lfloor k/2 \rfloor), \min(l, i + \lfloor k/2 \rfloor))\}$$

Here $\lfloor\rfloor$ converts a floating point number to the closest integer value below it.

2. **radius based** - for a given radius $r$, a protein is divided into $l$ radius-based structural fragments $\{C_i^r, i : 1, ..., l\}$ where

$$C_i^r = \{\boldsymbol{\alpha_j} | d(\boldsymbol{\alpha_i}, \boldsymbol{\alpha_j}) < r\}$$

with $d(\boldsymbol{\alpha_i}, \boldsymbol{\alpha_j})$ being the Euclidean distance between $\boldsymbol{\alpha_i}$ and $\boldsymbol{\alpha_j}$.

Practically, this is accomplished by constructing a KD-Tree on $\boldsymbol{\alpha}$, using the KD-tree implementation in ProDy v1.10.11[19] and querying by radius with each $\boldsymbol{\alpha_i}$ as the centre.

While the $k$-mer based approach is effective in describing structural fragments that are sequential in nature, such as $\alpha$-helices and loops, the radius-based approach can capture long-range structural contacts as seen in $\beta$-sheets, as well as distinct interaction patterns in space, as found in enzyme active sites. Both fragmentation methods have $\mathcal{O}(l)$ time complexity.

Each resulting structural fragment is then transformed into four moment invariants, described in the next section. In our examples and results section we use a $k$ of 16 and a radius $r$ of 10 Å as a compromise between specificity of the structural fragments and effectiveness of the moment invariants. In principle, optimization of these parameters could lead to further improvements of our approach for specific applications, but we leave this open for future exploration.

Moment Invariants

Three-dimensional moment invariants are computed using the formula of the central moment, defined below for a discrete set of $c$ coordinates, with $(\overline{x}, \overline{y}, \overline{z})$ being the centroid:

$$\mu_{pqr} = \sum_{i=1}^{c} (x_i - \overline{x})^p (y_i - \overline{y})^q (z_i - \overline{z})^r$$

Using this formula, we then compute four moments that were previously used in a structural bioinformatics study to describe enzyme active sites[18]. These include the three second-order rotation invariants ($O_3$, $O_4$, and $O_5$) described by Mamistvalov[20] and a fourth invariant, $F$, described by Flusser et al.[21]. These four moment invariants are defined below:

$$O_3 = \mu_{200} + \mu_{020} + \mu_{002}$$

$$O_4 = \mu_{200} \cdot \mu_{020} + \mu_{200} \cdot \mu_{002} + \mu_{020} \cdot \mu_{002} - \mu_{110}^2 - \mu_{101}^2 - \mu_{011}^2$$

$$O_5 = \mu_{200} \cdot \mu_{020} \cdot \mu_{002} + 2\mu_{110} \cdot \mu_{101} \cdot \mu_{011} - \mu_{002} \cdot \mu_{110}^2 - \mu_{020} \cdot \mu_{101}^2 - \mu_{200} \cdot \mu_{011}^2$$

$$\begin{aligned} F = 15\mu_{111}^2 + \mu_{003}^2 + \mu_{030}^2 + \mu_{300}^2 - 3\mu_{102}.\mu_{120} - 3\mu_{021}.\mu_{201} \\ -3\mu_{030}.\mu_{210} - 3\mu_{102}.\mu_{300} - 3\mu_{120}.\mu_{300} \\ -3\mu_{012}.(\mu_{030} + \mu_{210}) - 3\mu_{003}(\mu_{021} + \mu_{201}) \\ +6\mu_{012}^2 + 6\mu_{120}^2 + 6\mu_{201}^2 + 6\mu_{210}^2 + 6\mu_{021}^2 + 6\mu_{102}^2 \end{aligned}$$

Thus, any structural fragment can be represented by a vector $(O_3, O_4, O_5, F)$. Moment invariant calculation is implemented using Numba v0.48.0[22] and has $\mathcal{O}(c)$ time complexity which is negligible for small values of $c$, as seen for $k$=16 (i.e. a maximum $c$ of 16) and $r$=10 ($c = 18 \pm 6$).

Discretization to Shape-mers

While the moment invariants obtained for each structural fragment can be directly compared, discretizing them enables collecting sets of fragments that resemble each other across multiple proteins. We convert the continuous and real-valued moment invariants to discrete shape-mers as follows:

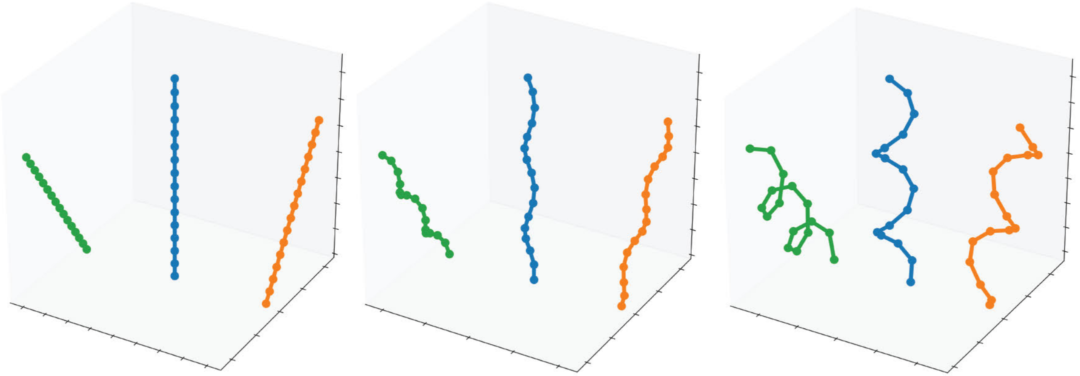
$$\begin{aligned} (O_3', O_4', O_5', F') = (\lfloor m \times \ln(O_3) \rfloor, \lfloor m \times \ln(O_4) \rfloor, \\ \lfloor m \times \ln(O_5) \rfloor, \lfloor m \times \ln(F) \rfloor) \end{aligned}$$

Here $m$ is the resolution parameter, which defines the coarseness of the shape-mers, with higher values leading to more fine-grained separation of structural fragments. Thus, a shape-mer is defined by four discrete numbers and can describe any number of structural fragments. Figure 5.1 gives examples of moment invariant and shape-mer calculations (with $m = 1$) for three synthetic coordinate sets generated with the equation $\{\boldsymbol{\alpha_i} = (R\cos(i), R\sin(i), i), i : 1, ..., 16\}$ for $R = 0$, 0.5, and 2 respectively, each rotated by ±45°, and translated by ±10Å along the $x$-axis.

Counting Shapes

Given a set of $n$ proteins, we generate a collection of shape-mers for each protein. The total number of shape-mers $s$ is then the number of distinct shape-mers observed across all $n$ proteins. A count vector of length $s$ is calculated for each protein, with each element recording the number of times the corresponding shape-mer appears in that protein. This counting is done separately for the $k$-mer and radius based approaches, as they represent different types of structural fragments. The two resulting count vectors are concatenated to form the final protein embedding. The entire embedding process has a time complexity of $\mathcal{O}(nl)$ and takes around 50 milliseconds CPU time for proteins of medium length (400-600 residues). Note that different values for $m$ (the resolution parameter) and different input sets of proteins will lead to different sets of shape-mers and embedding sizes. This allows the user to generate feature spaces tailored to the problem at hand.

A. Structural Fragments



|  |  |  |  |
|---|---|---|---|
| B. Moment Invariants | (340.0, 0.0, 0.0, 0.0) | (343.9, 1,336.6, 1,304.3, 9,596.5) | (402.9, 22,312.2, 333,915.0,164,154.9) |
| C. Natural log | (5.83  0.00, 0.00, 0.00) | (5.84, 7.19, 7.17, 9.16) | (6.00, 10.01, 12.71, 12.00) |
| D. Shape-mers | (5, 0, 0, 0) | (5, 7, 7, 9) | (6, 10, 12, 12) |

Figure 5.1: **A.** Three synthetic structural fragments, with a rotation of ±45° and translation of ±10Å between the middle fragment and the two outer ones. **B.** Moment invariants $(O_3, O_4, O_5, F)$ for each fragment. The three rotated and translated versions have the same moment invariant values. **C.** The natural log-transformed versions of $(O_3, O_4, O_5, F)$ and **D.** shape-mers $(O'_3, O'_4, O'_5, F')$ for each fragment.

### 5.2.2 Datasets

We apply Geometricus to a number of datasets to demonstrate the wide applicability of shape-mer based protein embedding. These are described below. The remaining sections use the acronyms defined here to refer to these datasets.

1. **CASP11** - 87,573 protein structures from the Critical Assessment of protein Structure Prediction XI[23] training set, obtained from the ProteinNet data source[24].
2. **CATH20** - The CATH database of protein structures[25] categorizes proteins hierarchically based on secondary structure class (C), architecture (A), topology (T), and homology (H). From the CATH hierarchy, we selected 3,673 proteins with <20% sequence identity to each other from the top five most populated CAT categories. Table 5.1 shows the number of proteins per CAT category.
3. **SCOP-Lo** - The Structural Classification of Proteins (SCOP) database[26] provides a detailed classification of structures based on their topologies and folds. We adapted the SCOP-Lo dataset from Lo et al.[9]. This dataset comprises 23,912 target proteins from ASTRAL SCOP 1.67 further divided into sets with 10%, 30%, 70% and 100% maximum sequence identity within each group respectively. It also contains a query set of 83 proteins each with at least two proteins from the same SCOP family in the 10% target protein set, and <10% sequence identity to other proteins in the query set.
4. **Pfam10** - The Protein families database (Pfam)[27] collates a large set of protein families. Out of the twenty most populated Pfam domains, the ten accessions with most available structures are considered, resulting in a total of 3,053 structures. Table 5.2 lists the number of proteins for each of these ten Pfam accessions.
5. **CMGC** - 1,822 human protein structures in the CMGC kinase family were collected from the Kinase-Ligand Interaction Fingerprints and Structures (KLIFS) database[28]. These are further divided into 660 cyclin-dependent kinases (CDK), 527 mitogen-activated protein kinases (MAPK), 268 casein kinase 2 (CK2) proteins, 160 dual specificity Tyrosine regulated kinases (DYRK), 122 glycogen synthase kinases (GSK), 61 cdc2-like kinases (CLK), 16 serine/threonine-protein kinases (SRPK), and 8 cyclin-dependent kinase-like kinases (CDKL).
6. **MAPK** - The 527 MAP kinases from the CMGC dataset are considered separately. These comprise 271 p38 MAPK structures (p38), 147 extracellular signal-regulated kinases (ERKs), and 109 c-Jun N-terminal kinases (JNKs).

Note that the low sequence identity between the proteins in many of these datasets clearly underlines the need for structure-based embedding.

### 5.2.3 Visualization of shape-mers and Geometricus embeddings

To visualize two commonly occurring $k$-mer-based shape-mers from the CASP11 dataset, we first randomly selected 1,000 structural fragments described by them. From these 1,000, one fragment was randomly chosen as the base and the remaining were superposed to the base using the Kabsch algorithm[29]. For each fragment, the

| Class | Architecture | Topology | No. of Proteins |
|---|---|---|---:|
| Mainly Alpha | Orthogonal Bundle | Arc Repressor Mutant, subunit A | 465 |
| Mainly Beta | Sandwich | Immunoglobulin-like | 700 |
| Mainly Beta | Sandwich | Jelly Rolls | 401 |
| Alpha Beta | 3-Layer (aba) Sandwich | Rossman Fold | 1,660 |
| Alpha Beta | 2-Layer Sandwich | Alpha-Beta Plaits | 447 |

Table 5.1: Number of proteins in each CAT category in the CATH20 dataset

| Pfam Accession | Short name | Description | No. of Proteins |
|---|---|---|---:|
| PF00005 | ABC_tran | ABC transporter | 187 |
| PF00069 | Pkinase | Protein kinase domain | 438 |
| PF00076 | RRM_1 | RNA recognition motif. (a.k.a. RRM, RBD, or RNP domain) | 269 |
| PF00096 | zf-C2H2 | Zinc finger, C2H2 type | 144 |
| PF00400 | WD40 | WD domain, G-beta repeat | 906 |
| PF00440 | TetR_N | Bacterial regulatory proteins, tetR family | 164 |
| PF02518 | HATPase_c | Histidine kinase-, DNA gyrase B-, and HSP90-like ATPase | 87 |
| PF12796 | Ank_2 | Ankyrin repeats (3 copies) | 263 |
| PF13561 | adh_short_C2 | Enoyl-(Acyl carrier protein) reductase | 273 |
| PF13855 | LRR_8 | Leucine rich repeat | 322 |

Table 5.2: Number of proteins for each Pfam accession in the Pfam10 dataset

best superposition to the base of that fragment and its flipped version, in terms of the minimum Root Mean Square Deviation (RMSD), is taken in order to account for 360° rotations. We also visualized two radius-based shape-mers using two of their structural fragments shown in the context of their respective protein structures.

The Geometricus embeddings of the Pfam10, CMGC, and MAPK datasets, generated for different values of the $m$ parameter, were reduced to two dimensions using the Python implementation (v0.3.10) of the Uniform Manifold Approximation and Projection (UMAP) algorithm by McInnes et al.[30], with the cosine similarity metric and default settings.

### 5.2.4 Structure similarity search

We demonstrated how Geometricus can be used in structure-based similarity searches by applying it to the CATH20 and SCOP-Lo datasets. A pair of proteins is called similar if they share the same CAT category for the CATH20 dataset or the same SCOP category for the SCOP-Lo dataset, and dissimilar otherwise.

Typically, in structure similarity search applications, similarity scores are calculated for a small set of *query* proteins against a larger predefined and preprocessed *target set* of structures. Here, the target set determines which collection of shape-mers will be used in the search. For the CATH20 dataset, 70% of the proteins are randomly chosen as the target set. The SCOP-Lo dataset already has four defined target sets (10%, 30%, 70%, and 100% sequence redundant sets) which are each evaluated separately.

The pairwise similarity measure between two proteins is defined as the cosine similarity of their Geometricus embedding vectors, constructed with a low resolution ($m = 0.25$) to reflect the major structural differences expected between proteins in these two datasets. Proteins with a similarity score above a threshold $t$ are predicted to be similar and those below $t$ are predicted as dissimilar. We calculated similarity scores for all CATH20 proteins against the CATH20 target set, and the 83 SCOP-Lo query proteins against each of the four sequence redundant SCOP-Lo target sets. ROC-AUC curves were constructed by varying $t$ to evaluate the correctness of the similarity search in these five cases.

### 5.2.5 CATH classification

A $k$-nearest neighbour classifier from the scikit-learn python library v0.22.1 (`k=5`, `metric="cosine"`) was trained to predict the CAT category for the proteins in the CATH20 dataset with 50% of the data randomly chosen for training and the remaining for testing. We repeated this five times and report the average accuracy.

### 5.2.6 MAP kinase classification

To demonstrate the applicability of Geometricus for interpretable machine learning on protein structures, we performed classification on the MAPK dataset to predict the type of MAP kinase (namely p38, ERK, or JNK) from protein structure. This was accomplished using the decision tree classifier from the scikit-learn Python library (v0.22.1)[31], with a random 70%-30% split of training and test data. The top two most predictive shape-mers from the trained classifier were then mapped back to the residues that they correspond to and visualized on one p38 structure (PDB ID: 3QUE), one ERK structure (PDB ID: 2OJJ) and one JNK structure (PDB ID 4KKG) using PyMOL[32].

## 5.3 Results

### 5.3.1 Shape-mers capture common structural fragments across protein structures

We performed moment-invariant and shape-mer calculations on over 87,000 proteins in the CASP11 dataset to understand their distributions and patterns found across structurally divergent proteins. Figure 5.2A shows the log-distribution of each of the four moment invariants for the $k$-mer- and radius based structural fragmentation approaches. The radius-based approach shows wider distributions in general, which can be expected: different locations in a protein have different densities of residues leading to differing numbers of coordinates in the radius-based approach, while the $k$-mer based approach largely produces fragments with $k$ coordinates except for some shorter fragments at the N- and C-terminal ends of each protein.

Shape-mers were computed from the moment invariants using a resolution $m$ of 1 (see Methods). The resulting 565 $k$-mer shape-mers and 703 radius shape-mers do not all represent the same number of structural fragments. Figure 5.2B shows the $\log_{10}$ distribution of structural fragment counts represented by each shape-mer. Some shape-mers, at the right end of Figure 5.2B, are found over a million times, and unsurprisingly represent common structural fragments such as short, well-defined $\alpha$-helices. One $k$-mer based and one radius-based example are shown in Figure 5.2C1 and Figure 5.2D1 respectively, both found across most of the proteins in the CASP11 dataset. Conversely, the shape-mers on the very left end of the Figure 5.2B represent only one structural fragment, likely loops or specific folds which are structurally and functionally unique and thus rare. The remaining shape-mers describe anywhere between one and a million fragments and may be specific to certain superfamilies or families of proteins. Figure 5.2C2 shows an extended roll-like shape-mer found in almost 10,000 proteins and Figure 5.2D2 shows a sparse radius shape-mer found on the surfaces and ends of 5,000 proteins.

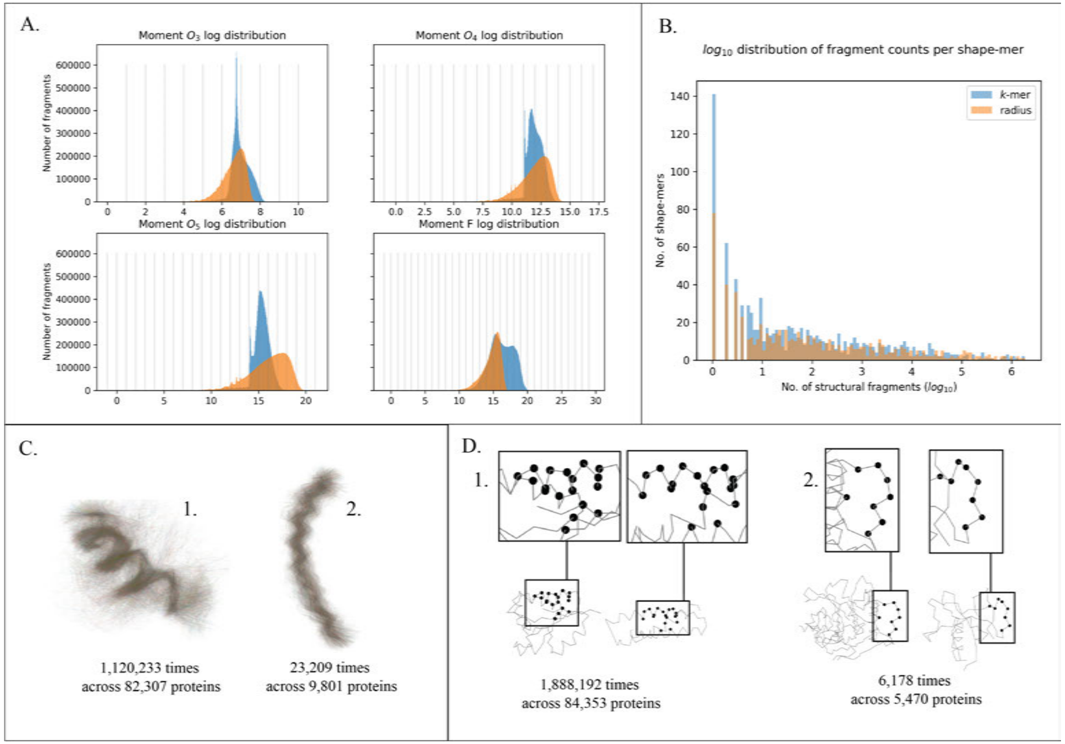


Figure 5.2: **A.** Natural log distribution of each of the four moment invariants across $k$-mer based structural fragments (in blue) and radius-based structural fragments (in orange) from proteins in the CASP11 dataset. **B.** Histogram of ($\log_{10}$ transformed) counts of fragments described by each shape-mer. **C.** Two examples of $k$-mer based shape-mers, shown using a thousand randomly selected fragments, superposed. **D.** Two examples of radius-based shape-mers, magnified, and highlighted as black dots on two protein structures in which they are found.

### 5.3.2 Geometricus can be used for fast and accurate structure similarity search and topology classification

A common application of structure-based embeddings is to perform a fast similarity search for an input structure across a database of structures and return the most similar candidates. We demonstrate the performance of Geometricus on this task using CATH and SCOP classifications as a ground truth measure of protein similarity.

Figure 5.3 shows the Receiver Operating Characteristic (ROC) curves for the CATH20 dataset and for various sequence redundancy levels of the SCOP-Lo dataset, along with their corresponding area under the curve (AUC) values. The all *vs.* all similarity calculation for the 3,673 proteins in the CATH20 dataset took 250 milliseconds. For the SCOP-Lo dataset, query *vs.* target similarity calculation for the 83 query proteins against the 10% target dataset (with 4,332 proteins) took 4 milliseconds and the 100% target set (with 23,912 proteins) took 20 milliseconds. Generating the target dataset embeddings was also fast, taking 2 minutes for the CATH20 dataset and 15 minutes for the 100% SCOP-Lo dataset (excluding file parsing time as this depends on the speed of the disk). Target set embedding time is not as important as search time, as it only has to be run once. Embedding each additional query protein takes 20-60 milliseconds depending on its length. A $k$-nearest neighbours classification of the CATH20 dataset into the five CAT classes showed a high accuracy of 82%.

In both these applications, Geometricus performs favourably compared to results reported by other alignment-free approaches applied to comparable datasets[7,9,10], which typically achieve search AUCs between 0.75 and 0.85 and fold classification accuracy up to 75%. For the structural alphabets defined by Le et al.[10] classification accuracy increases to 80% upon using more sophisticated SVM classifiers with tailored kernels. This approach is not investigated here but would likely improve our fold classification accuracy further. Geometricus comes close to the highly accurate alignment-based methods[10] (with search AUCs exceeding 0.9 and fold classification accuracy exceeding 90%) at a mere fraction of the computational cost.

### 5.3.3 Geometricus can be used across and within protein families

Unlike library-based or deep learning-based structure embedding techniques, Geometricus can be adapted to the type and scale of the problem at hand without sacrificing speed, via the $m$ (resolution) parameter. When comparing proteins from different superfamilies, a coarse discretization of structural fragments is preferred as it is expected that these proteins will have very different structures. However, as the specificity of the problem increases, the proteins under investigation start resembling each other more. In such cases, more specific binning of fragments, *i.e* a higher resolution, is advantageous to better capture their differences. This is demonstrated in Figure 5.4 with the Pfam10, CMGC, and MAPK datasets.

The Pfam10 dataset (Figure 5.4A) consists of proteins that contain one of ten fairly divergent Pfam domains. A low resolution of 0.25 (leading to 26 $k$-mer based shape-mers and 28 radius-based shape-mers) already separates these ten Pfam accessions into well-defined clusters. Some similar accessions, such as the Protein kinase domain

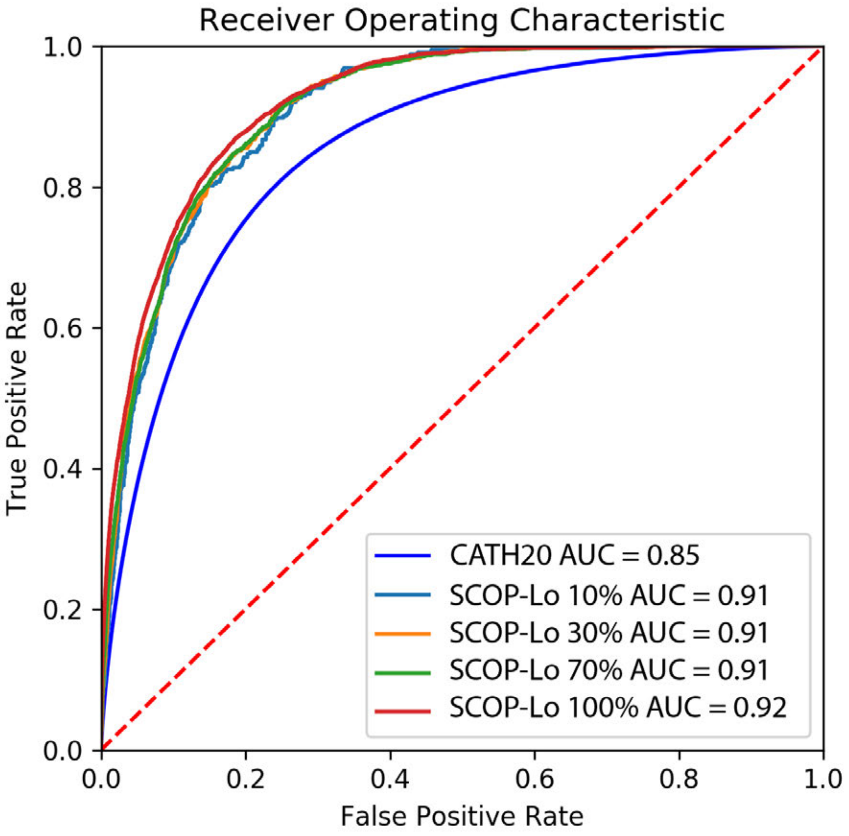


Figure 5.3: ROC curves for an all *vs.* all structure similarity search using Geometricus embeddings on the CATH20 dataset (dark blue) and four similarity searches of 83 query proteins on different sequence redundant target sets from the SCOP-Lo dataset (10% - light blue, 30% - orange, 70% - green, and 100% - red). True positives were determined using CATH and SCOP classifications, as described in the Methods.

(Pkinase) and the Histidine kinase-like ATPase (HATPase_c) cluster closer together as expected. Higher resolutions perform worse on this dataset, as comparable structural fragments are split across different shape-mers.

The CMGC dataset (Figure 5.4B) contains proteins from a group of kinases called the CMGC group (named after the initials of some members). As these proteins are more evolutionarily and functionally related, a higher resolution of 0.5 (resulting in 78 $k$-mer shape-mers and 93 radius shape-mers) is required to achieve a good separation between the individual families within this group.

Finally, the MAPK dataset (Figure 5.4C) consists of MAP kinases, a family of proteins which relay signals from the cell surface to coordinate growth, stress and other responses. This family is divided into subfamilies, here simplified into the p38, ERK, and JNK categories, each of which relay different types of growth and stress signals. A high resolution of 2 (resulting in 1098 $k$-mer shape-mers and 908 radius shape-mers) separates these subfamilies.

Thus, the feature space generated by Geometricus can be altered depending on the structural similarity expected between the proteins under consideration. This is especially advantageous in situations where the proteins under study are from the same family or subfamily and share a common structural fold, or in the case of mutation studies where local structure alterations occur due to single residue changes. In contrast, other embedding techniques are often optimized for divergent structures, and would likely assign the same embedding to each protein in these cases.

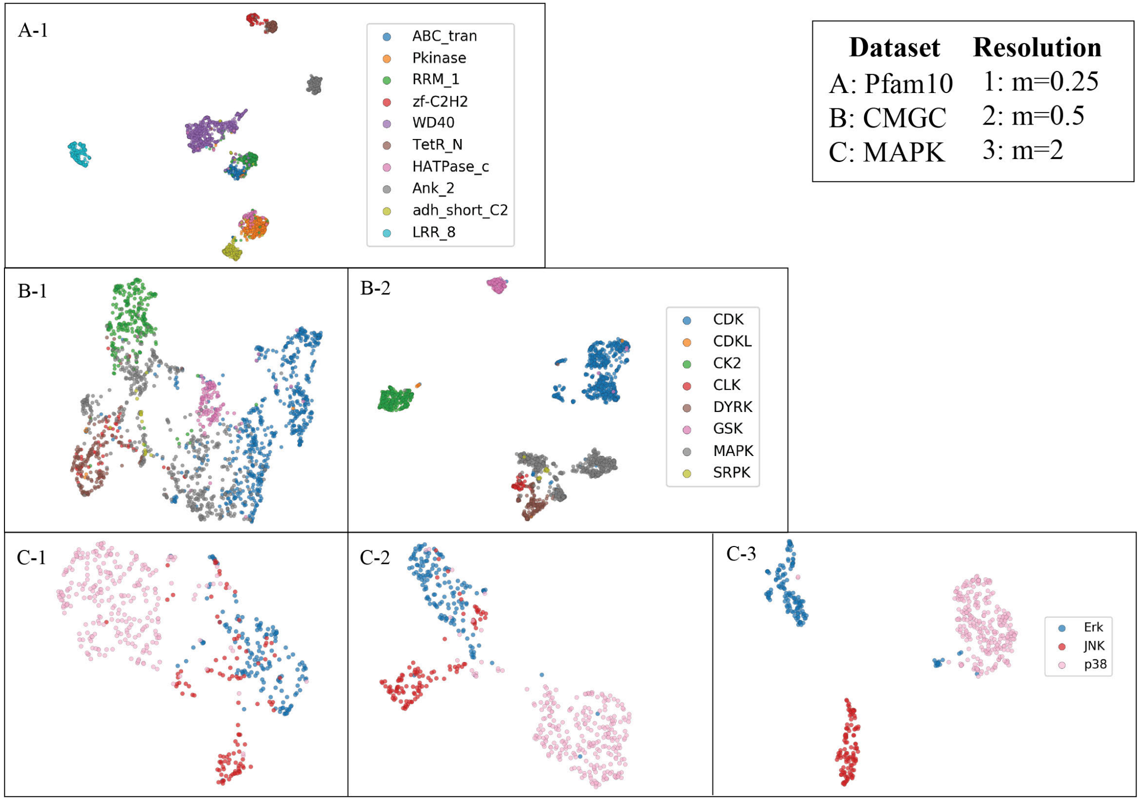
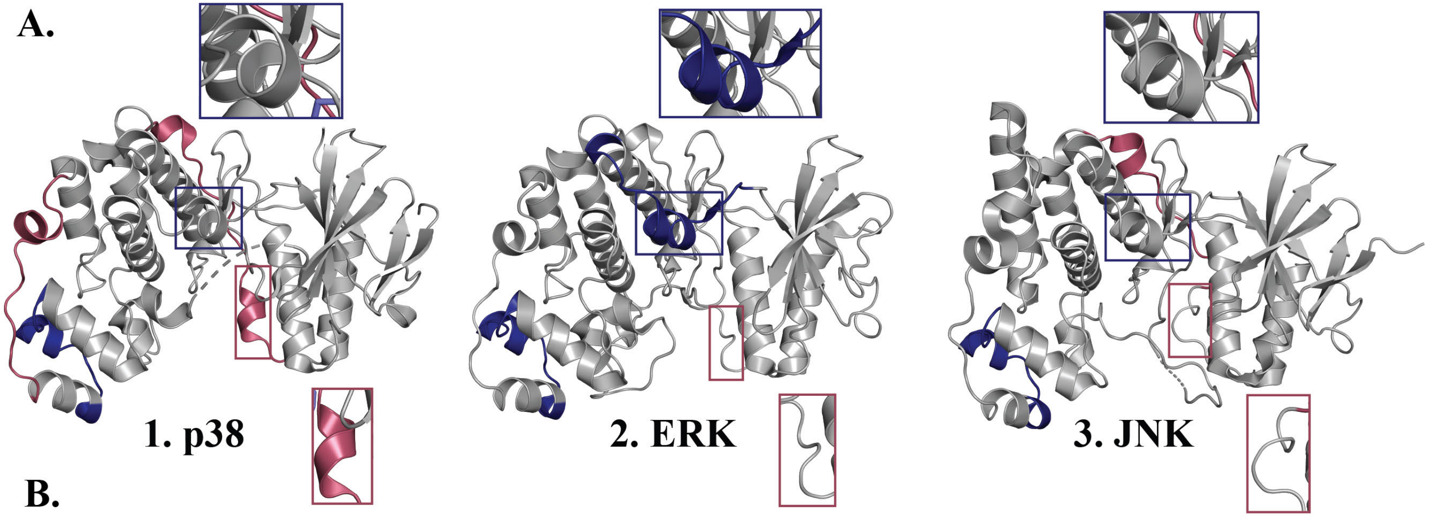


Figure 5.4: The effect of the resolution parameter $m$ on different datasets. A, B, and C represent the Pfam10, CMGC, and MAPK datasets respectively. 1, 2, and 3 represent resolutions of 0.25, 0.5, and 2. As the structural similarity between proteins increases, higher resolutions are needed to achieve a good separation of pre-existing clusters in each dataset.

### 5.3.4 Geometricus can be used as input for interpretable machine learning

Typically, when analysing highly similar proteins as found in the MAPK dataset, one would also be interested in interpreting the results to find functionally important residues or structural regions. Such insights can be directly be applied to select candidate residues for mutational studies or used in directed evolution techniques to engineer proteins and enzymes with desired properties such as substrate specificity[33], drug-target binding affinity[34], interaction specificity[35], or thermostability[36] among others. Geometricus embeddings are well-suited for this kind of learning as each element of an embedding can easily be mapped back to the specific residues of the shape-mer it represents.

We demonstrate this with a classification problem defined for the MAPK dataset, namely to predict the specific subfamily of a MAP kinase. A simple decision tree trained on 70% of the data and tested on the remaining 30% showed an accuracy of 96% for this task. More interestingly, this trained classifier can now be inspected for predictive features. We mapped the top two shape-mers considered the most predictive by the decision tree back to all the residues and locations at which they occur across all the MAPK proteins. These locations are visualized on three example proteins, one from each of the three subfamilies (Figure 5.5A; shape-mer 1 in red, shape-mer 2 in blue). Figure 5.5B details the percentage of proteins from each subfamily which contain each of the two shape-mers, and the average number of times they appear per protein. The first appears more often in p38 kinases at a higher frequency per protein, while the second favours the ERK kinases with over three occurrences per protein on average. Looking at the structures themselves, it becomes clear which particular locations (highlighted and magnified) cause this difference in frequencies, even in such highly similar structures. While this is a simple example, it demonstrates the potential for using Geometricus in interpretable machine learning tasks for protein families.



B.

| 1. | p38 | ERK | JNK |
|---|---|---|---|
| % Proteins | 99% | 41% | 21% |
| No. per protein | 2.7 ± 0.9 | 1.4 ± 0.9 | 1.3 ± 0.5 |

| 2. | p38 | ERK | JNK |
|---|---|---|---|
| % Proteins | 97% | 100% | 89% |
| No. per protein | 2.3 ± 1.2 | 3.0 ± 0.9 | 1.1 ± 0.3 |

Figure 5.5: **A.** The occurrences of two shape-mers (coloured red and blue respectively) most predictive in separating MAPK subfamilies visualized on three MAPK structures: 1. p38 structure (PDB ID: 3QUE), 2. ERK structure (PDB ID: 2OJJ) and 3. JNK structure (PDB ID: 4KKG). For each shape-mer, one location in a structure where it is present is magnified across all three structures and discussed in the text. **B.** The percentage of proteins containing a shape-mer and the average number of times a shape-mer appears per protein across the three MAPK subfamilies, for 1. the first shape-mer (red) and 2. the second shape-mer (blue).

## 5.4 Conclusion

We have presented a novel, fast and accurate approach for protein structure embedding with a wide range of applications. Geometricus uses 3D rotation invariant moments to describe structural fragments such that they can be easily compared across proteins without the need for superposition or alignment. This allows for a blazing fast embedding technique that takes milliseconds to generate an embedding of a protein, and scales linearly with the number of proteins.

The simplicity of this approach also brings with it versatility, as Geometricus does not depend on a fixed library of predefined fragments and can instead grow or shrink depending on the scale of the problem at hand. Therefore, it is readily applied to more specialised prediction tasks focusing on a single protein family with a conserved structural fold where other structure embedders would likely struggle to resolve each protein. The explicit mapping between residues and shape-mers further allows the user to trace back from a predictive model to predictive residues and structural regions, which can broaden our understanding of specific protein and enzyme mechanisms. This makes Geometricus well-suited for machine learning tasks where interpretation is a concern along with accuracy.

While this initial version of Geometricus uses four rotation-invariant moments, more such invariants have been studied[37] and could be added to increase the specificity of a shape-mer. Another possible extension is to include solvent accessibility or amino acid descriptors as rotation-invariant aspects of a residue set. While these additions would likely not be so helpful in tasks spanning diverse proteins, such as structure similarity search, they may be useful in tasks involving enzyme mechanisms[38] or protein/ligand interactions and hotspots[39,40] where the accessibility of a structure fragment as well as its physicochemical and electrostatic properties matter as much as its shape.

Geometricus thus combines a set of highly attractive features that sets it apart from other structure embedding and structure similarity techniques. It is much faster than alignment-based algorithms such as Madej et al.[41] and Ye & Godzik[42], and at the same time highly accurate compared to other alignment-free techniques such as Le et al.[10] and Lo et al.[9]. Unlike most techniques, its independence from a fragment library or predefined training set allows for broad application to generate feature sets for machine learning, even for differentiating mutants - something that has not been explored due to the focus of current techniques on divergent proteins. The shape-mer approach allows for easy interpretability and possible association of specific shapes to function, and its simplicity allows for ease of expansion. Shape-mer similarity could also be utilized to train structure-informed sequence embedding techniques, similar to the approach detailed by[43], or as part of a scoring function to assess protein model quality, a field in which topology has been shown to play a crucial role[44].

Improvements in homology and *de novo* modelling techniques have greatly expanded the number of proteins for which we can accurately model structure. This means that future structure-based machine learning tasks will likely be augmented with structural models to obtain large datasets comparable to those used in sequence-based predictive approaches, where such a fast and versatile structural embedder would be useful. Given the prominent role in present-day bioinformatics of both structural modelling and machine learning, Geometricus embeddings, with possible further embellishments, may lead to breakthroughs in understanding protein function.

## Acknowledgements

We thank Rens Holmer and Barbara Terlouw for valuable feedback on drafts, stimulating discussions and insightful suggestions on the approach and applications presented here.

## Funding

This work was supported by the Netherlands Organization for Scientific Research (NWO), project numbers TTW 15043 (JD) and TTW 14516 (MA).

## References

[1] Bernstein, F. C., Koetzle, T. F., Williams, G. J., Meyer, E. F., Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T., & Tasumi, M. (1977). The Protein Data Bank: A computer-based archival file for macromolecular structures. *Journal of Molecular Biology*, *112*, 535–542.

[2] Senior, A. W., Evans, R., Jumper, J., Kirkpatrick, J., Sifre, L., Green, T., Qin, C., Žídek, A., Nelson, A. W., Bridgland, A. et al. (2020). Improved protein structure prediction using potentials from deep learning. *Nature*, (pp. 1–5).

[3] Simossis, V., Kleinjung, J., & Heringa, J. (2003). An overview of multiple sequence alignment. *Current Protocols in Bioinformatics*, *3*, 3–7.

[4] Ma, J., & Wang, S. (2014). Algorithms, applications, and challenges of protein structure alignment. In *Advances in Protein Chemistry and Structural Biology* (pp. 121–175). Elsevier volume 94.

[5] Alley, E. C., Khimulya, G., Biswas, S., AlQuraishi, M., & Church, G. M. (2019). Unified rational protein engineering with sequence-based deep representation learning. *Nature Methods*, *16*, 1315–1322.

[6] Rao, R., Bhattacharya, N., Thomas, N., Duan, Y., Chen, X., Canny, J., Abbeel, P., & Song, Y. S. (2019). Evaluating protein transfer learning with TAPE. *Advances in Neural Information Processing Systems*, *32*, 9689–9701.

[7] Budowski-Tal, I., Nov, Y., & Kolodny, R. (2010). FragBag, an accurate representation of protein structure, retrieves structural neighbors from the entire PDB quickly and accurately. *Proceedings of the National Academy of Sciences*, *107*, 3481–3486.

[8] Liu, Y., Ye, Q., Wang, L., & Peng, J. (2018). Learning structural motif representations for efficient protein structure search. *Bioinformatics*, *34*, i773–i780.

[9] Lo, W.-C., Huang, P.-J., Chang, C.-H., & Lyu, P.-C. (2007). Protein structural similarity search by Ramachandran codes. *BMC Bioinformatics*, *8*, 307.

[10] Le, Q., Pollastri, G., & Koehl, P. (2009). Structural alphabets for protein structure classification: A comparison study. *Journal of Molecular Biology*, *387*, 431–450.

[11] Hu, M.-K. (1962). Visual pattern recognition by moment invariants. *IRE Transactions on Information Theory*, *8*, 179–187.

[12] Rizon, M., Haniza, Y., Puteh, S., Yeon, A., Shakaff, M., Abdul Rahman, S., Mohd Rozailan, M., Sazali, Y., Hazri, D., & Karthigayan, M. (2006). Object detection using geometric invariant moment. *American Journal of Applied Sciences*, *3(6)*, 1876–1878.

[13] Flusser, J., & Suk, T. (1994). Affine moment invariants: A new tool for character recognition. *Pattern Recognition Letters*, *15*, 433–436.

[14] Sadjadi, F. A., & Hall, E. L. (1980). Three-dimensional moment invariants. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, (pp. 127–136).

[15] Se, S., Lowe, D., & Little, J. (2001). Vision-based mobile robot localization and mapping using scale-invariant features. In *Proceedings 2001 ICRA. IEEE International Conference on Robotics and Automation (Cat. No. 01CH37164)* (pp. 2051–2058). IEEE volume 2.

[16] Kratz, S., & Rohs, M. (2011). Protractor3D: A closed-form solution to rotation-invariant 3D gestures. In *Proceedings of the 16th International Conference on Intelligent User Interfaces* (pp. 371–374).

[17] Mangin, J.-F., Poupon, F., Duchesnay, É., Rivière, D., Cachia, A., Collins, D. L., Evans, A. C., & Régis, J. (2004). Brain morphometry using 3D moment invariants. *Medical Image Analysis*, *8*, 187–196.

[18] Sommer, I., Müller, O., Domingues, F. S., Sander, O., Weickert, J., & Lengauer, T. (2007). Moment invariants as shape recognition technique for comparing protein binding sites. *Bioinformatics*, *23*, 3139–3146.

[19] Bakan, A., Meireles, L. M., & Bahar, I. (2011). ProDy: Protein dynamics inferred from theory and experiments. *Bioinformatics*, *27*, 1575–1577.

[20] Mamistvalov, A. G. (1998). N-dimensional moment invariants and conceptual mathematical theory of recognition n-dimensional solids. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *20*, 819–831.

[21] Flusser, J., Boldys, J., & Zitová, B. (2003). Moment forms invariant to rotation and blur in arbitrary number of dimensions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *25*, 234–246.

[22] Lam, S. K., Pitrou, A., & Seibert, S. (2015). Numba: A LLVM-based python JIT compiler. In *Proceedings of the Second Workshop on the LLVM Compiler Infrastructure in HPC* (pp. 1–6).

[23] Moult, J., Fidelis, K., Kryshtafovych, A., Schwede, T., & Tramontano, A. (2016). Critical assessment of methods of protein structure prediction: Progress and new directions in round XI. *Proteins: Structure, Function, and Bioinformatics*, *84*, 4–14.

[24] AlQuraishi, M. (2019). ProteinNet: A standardized data set for machine learning of protein structure. *BMC Bioinformatics*, *20*, 311.

[25] Pearl, F. M., Bennett, C., Bray, J. E., Harrison, A. P., Martin, N., Shepherd, A., Sillitoe, I., Thornton, J., & Orengo, C. A. (2003). The CATH database: An extended protein family resource for structural and functional genomics. *Nucleic Acids Research*, *31*, 452–455.

[26] Murzin, A. G., Brenner, S. E., Hubbard, T., Chothia, C. et al. (1995). SCOP: A structural classification of proteins database for the investigation of sequences and structures. *Journal of Molecular Biology*, *247*, 536–540.

[27] Bateman, A., Birney, E., Cerruti, L., Durbin, R., Etwiller, L., Eddy, S. R., Griffiths-Jones, S., Howe, K. L., Marshall, M., & Sonnhammer, E. L. (2002). The Pfam protein families database. *Nucleic Acids Research*, *30*, 276–280.

[28] Kooistra, A. J., Kanev, G. K., van Linden, O. P., Leurs, R., de Esch, I. J., & de Graaf, C. (2016). KLIFS: A structural kinase-ligand interaction database. *Nucleic Acids Research*, *44*, D365–D371.

[29] Kabsch, W. (1976). A solution for the best rotation to relate two sets of vectors. *Acta Crystallographica Section A: Crystal Physics, Diffraction, Theoretical and General Crystallography*, *32*, 922–923.

[30] McInnes, L., Healy, J., & Melville, J. (2018). UMAP: Uniform manifold approximation and projection for dimension reduction. *ArXiv e-prints*. `arXiv:1802.03426`.

[31] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V. et al. (2011). Scikit-Learn: Machine learning in Python. *the Journal of Machine Learning Research*, *12*, 2825–2830.

[32] DeLano, W. L. et al. (2002). PyMOL: An open-source molecular graphics tool. *CCP4 Newsletter on Protein Crystallography*, *40*, 82–92.

[33] Ding, H., Takigawa, I., Mamitsuka, H., & Zhu, S. (2014). Similarity-based machine learning methods for predicting drug-target interactions: A brief review. *Briefings in Bioinformatics*, *15*, 734–747.

[34] Michael, J., Ross, D. et al. (1992). Modelling the structure and function of enzymes by machine learning. *Faraday Discussions*, *93*, 269–280.

[35] Fariselli, P., Pazos, F., Valencia, A., & Casadio, R. (2002). Prediction of protein–protein interaction sites in heterocomplexes with neural networks. *European Journal of Biochemistry*, *269*, 1356–1361.

[36] Jia, L., Yarlagadda, R., & Reed, C. C. (2015). Structure based thermostability prediction models for protein single point mutations with machine learning tools. *PLoS One*, *10*.

[37] Žunić, J., Hirota, K., Dukić, D., & Aktaş, M. A. (2016). On a 3D analogue of the first Hu moment invariant and a family of shape ellipsoidness measures. *Machine Vision and Applications*, *27*, 129–144.

[38] Heckmann, D., Lloyd, C. J., Mih, N., Ha, Y., Zielinski, D. C., Haiman, Z. B., Desouki, A. A., Lercher, M. J., & Palsson, B. O. (2018). Machine learning applied to enzyme turnover numbers reveals protein structural correlates and improves metabolic models. *Nature Communications*, *9*, 5252.

[39] Liu, S., Liu, C., & Deng, L. (2018). Machine Learning Approaches for Protein–Protein Interaction Hot Spot Prediction: Progress and Comparative Assessment. *Molecules*, *23*, 2535.

[40] Zheng, N., Wang, K., Zhan, W., & Deng, L. (2019). Targeting virus-host protein interactions: Feature extraction and machine learning approaches. *Current drug metabolism*, *20*, 177–184.

[41] Madej, T., Lanczycki, C. J., Zhang, D., Thiessen, P. A., Geer, R. C., Marchler-Bauer, A., & Bryant, S. H. (2014). MMDB and VAST+: Tracking structural similarities between macromolecular complexes. *Nucleic Acids Research*, *42*, D297–D303.

[42] Ye, Y., & Godzik, A. (2004). FATCAT: A web server for flexible structure comparison and structure similarity searching. *Nucleic Acids Research*, *32*, W582–W585.

[43] Bepler, T., & Berger, B. (2019). Learning protein sequence embeddings using information from structure. *ArXiv e-prints*.

[44] Garg, S., Kakkar, S., & Runthala, A. (2016). Improved protein model ranking through topological assessment. In *Computational Biology and Bioinformatics* (pp. 410–428). CRC Press.

CHAPTER 6

# Fast and adaptive protein structure representations for machine learning

Janani Durairaj*, Mehmet Akdel*, Dick de Ridder, and Aalt D.J. van Dijk

* authors contributed equally

This chapter was presented at the Machine Learning for Structural Biology (MLSB) Workshop, NeurIPS 2020

# Abstract

The growing prevalence and popularity of protein structure data, both experimental and computationally modelled, necessitates fast tools and algorithms to enable exploratory and interpretable structure-based machine learning. Alignment-free approaches have been developed for divergent proteins, but proteins sharing functional and structural similarity are often better understood via structural alignment, which has typically been too computationally expensive for larger datasets. Here, we introduce the concept of rotation-invariant shape-mers to multiple structure alignment, creating a structure aligner that scales well with the number of proteins and allows for aligning over a thousand structures in 20 minutes. We demonstrate how alignment-free shape-mer counts and aligned structural features, when used in machine learning tasks, can adapt to different levels of functional hierarchy in protein kinases, pinpointing residues and structural fragments that play a role in catalytic activity.

## 6.1 Introduction

The dual effect of the ever-growing number of protein structures deposited in the Protein Data Bank[1] and dramatically improved protein structure modelling[2] has led to an increasing number of studies incorporating structure information for predicting and understanding protein function. Structures are essential to our understanding of protein biology as their form dictates function, and they evolve more slowly than sequences. Research questions for which structural data may provide an answer are many and diverse - ranging from searching for remote protein homologs with similar structural folds across the tree of life to exploring the properties of a single protein family in a single species. These two extremes require different approaches, as both the numbers of protein structures involved and the types of insights that can be obtained differ greatly. In the past years, machine learning is proving itself to be crucial in solving these research questions, evident by its meteoric growth in the bioinformatics field. Machine learning algorithms have been applied across divergent protein structures for tasks such as topology classification[3], model quality assessment[4], ligand pocket prediction[5], and mutant stability estimation[6]. For specific protein families, structure-based machine learning has helped with predicting SH2 domain specificity[7], modelling the fitness landscape of cytochrome P450s[8], finding similarities in telomerases[9], and predicting ligand-binding for G-protein coupled receptors[10], among others.

Recently, we published Geometricus[11], a fast alignment-free protein structure embedding approach for describing and comparing divergent proteins. Geometricus defines discrete so-called shape-mers, analogous to sequence $k$-mers, using a set of rotation-invariant moments. The embedding of a protein is then simply the count vector of these shape-mers. This alignment-free approach accurately represents the topological aspects of proteins in machine learning for predicting protein function. The embedding allows for interpretation by mapping predictive shape-mers back to a set of residues in every protein. Given that applications on divergent proteins often encompass tens of thousands of structures, Geometricus provides a good balance between speed and interpretability.

However, for more similar proteins more correspondence between individual residues is expected, and an alignment better captures information about conservation, outliers, and functionally important residue positions. For each residue, a variety of features can be measured according to their relevance to the problem at hand, ranging from amino acid physicochemical properties to electrostatic energies and, in this research, topological properties via rotation-invariant moments. These features when aligned according to a multiple structure alignment generate a matrix that can directly be used as input to a machine learning algorithm. The algorithm looks across the alignment positions for patterns and correlations relevant for predicting the desired response variable. Predictions can be understood in terms of predictive residue positions, which are now easily compared to known catalytic residues or form hypotheses for mutagenesis studies. Gaps in an alignment are considered as missing data, and alignment positions with too many gaps are often discarded, potentially losing out on the predictive power of catalytically important residues split across multiple gap-filled positions. Thus, to generate an alignment-based feature matrix from a set of similar proteins we start with our recently released Caretta multiple structure alignment algorithm, built with the aim of generating high-coverage alignments for use in machine learning[12].

Computational structure modelling, both *de novo* and homology-based, has started to play more of a role in structure-based machine learning research[6,13], leading to datasets with up to thousands of protein structures sharing similar functionality and structural folds. These numbers are difficult to handle with current multiple structure alignment approaches, which generally scale poorly with the number of proteins[14]. In many cases, this can be attributed to the initial all vs. all pairwise alignment step used to generate a guide tree for subsequent progressive alignment steps. Multiple sequence alignment algorithms such as ClustalW[15], Kalign[16] and MUSCLE[17] circumvent this by using alignment-free $k$-tuple similarity, calculated by collecting matching subsequences of length $k$ ($k$-mers) from both input sequences, instead of pairwise alignment. This greatly reduces time complexity and allows for near-linear scaling with the number of proteins. With Geometricus this now becomes possible for structure alignment as well, by defining shape-tuple similarity as the collection of matching shape-mers from each protein's structure, thus completely avoiding the need for all vs. all pairwise structure alignment. The progressive alignment stage still aligns pairs of proteins at a time, and unlike for sequences, the three-dimensional nature of structures necessitates a superposition step in pairwise alignment. This step orients the input protein pair such that distance measures between aligning residues are meaningful. We use rotation-invariant moments to define this initial superposition step. Both of these improvements are incorporated into the Caretta algorithm to give **Caretta-shape**. We demonstrate that Caretta-shape is comparable to other popular structure alignment algorithms in terms of alignment quality and accuracy, while scaling easily to thousands of proteins.

We use the well-studied protein kinase superfamily to demonstrate how Geometricus and Caretta-shape can be used in unison to explore and understand structural similarities and differences between large datasets of protein structures. We showcase both unsupervised and supervised machine learning analyses at the superfamily, family, and subfamily levels, with emphasis on extracting structural insights useful for downstream research.

## 6.2 Methods

### 6.2.1 Caretta-shape

We recently released Caretta[12], a software for multiple structure alignment aimed at generating aligned features for use in machine learning algorithms. The advantage of Caretta in the context of machine learning applications lies in its focus on high coverage alignments using a novel consensus weight mechanism, which improves the information content of the aligned features. Here we detail the modifications made to the Caretta algorithm in Caretta-shape. Python code for Caretta-shape is available at `https://github.com/TurtleTools/caretta`

#### Shape-tuple similarity for fast guide tree construction

An all vs. all similarity matrix is constructed for input proteins by calculating the Bray-Curtis similarity between each protein pairs' Geometricus count vectors (with $k$-mer size $k = 20$ and resolution $m = 2$). The guide tree for determining the order of progressive alignments is constructed using maximum linkage neighbour joining[18] on this similarity matrix.

#### Rotation-invariant moment-based superposition

Caretta-shape replaces the signal- and secondary structure-based superposition scheme of Caretta by moment-based superposition. For each of the two structures to be aligned, four moment invariants are calculated for each residue with a fixed $k$-mer size (set to $k = 20$), $\vec{M} = [O_3, O_4, O_5, F]$ (named as in Durairaj et al.[11]). To ensure that the four moments contribute equally to the distance measure, each moment is normalized across both structures by subtracting the mean and dividing by the standard deviation to form $\vec{M'}$. The two series of normalized moment invariants are then aligned by dynamic programming using the Gaussian Caretta score:

$$Score_M(i,j) = \exp\left(-\gamma_m \sum (\vec{M'_i} - \vec{M'_j})^2\right) \tag{6.1}$$

with $\gamma_m = 0.6$. The aligning residues are used to calculate the optimal superposition using the Kabsch algorithm[19], after which coordinate-based superposition is performed as in the Caretta algorithm with default parameters ($\gamma = 0.03$, gap open penalty = 1, gap extend penalty = 0.01, consensus weight = 1). Parameter optimization for specific tasks could improve the results presented here, but we leave this open for future exploration.

#### Benchmarking

Caretta-shape was tested on two benchmark datasets, Homstrad[20] and SABmark-superfamily[21]. The PDB files for these two datasets (390 sets with 3–27 proteins each

| **Kinase group** | CMGC | TK | CAMK | AGC | STE | TJL | CK1 | Atypical | Other |
|---|---|---|---|---|---|---|---|---|---|
| **No. of proteins** | 2,049 | 2,057 | 811 | 517 | 455 | 654 | 209 | 346 | 648 |

Table 6.1: Number of proteins in the KinaseAll dataset across the eight major kinase groups (the rest are labelled as "Other").

from Homstrad and 425 sets with 3-42 proteins each from SABmark-superfamily) were obtained from mTM-align's website[22] and Matt benchmark results[23] respectively, in order to directly compare results to the output of these two tools. To this end, the Matt[23] and mTM-align[22] alignments for the Homstrad[20] and SABmark-Sup[21] datasets were obtained from their respective websites. For 17 cases in the Homstrad dataset, mTM-align returned alignments where at least one sequence did not match the corresponding PDB sequence. These cases were not considered.

We report various quality metrics of multiple structure alignments obtained from the different tools benchmarked. For both benchmark datasets we report the average (median) TM-score of the alignment, a measure that takes into account both the structural equivalence of corresponding residues and the overall coverage of the alignment[24]. We also report the median percentage of positions in the alignment without gaps (gapless positions), an aspect important to consider when using aligned features as input to machine learning algorithms, as gaps are seen as missing data and may cause loss of information about the residue positions in which they occur. In addition, the Homstrad dataset provides a set of manually curated reference alignments, for which we define an accuracy score (Acc.) that measures the number of correct gapless positions found, *i.e* gapless positions which are equivalent to positions in the corresponding reference alignment, divided by the total number of gapless positions in the reference alignment.

To estimate Caretta-shape running times, we chose four proteins from the SABMark dataset as "seeds", with lengths 100, 300, 504, and 714 respectively. Each seed was used to form multiple groups of proteins by introducing noise of up to 5 Å to each of the seed coordinates, to create a given number of members, from 10 to 1010 in increments of 200. Each noisy structure was further rotated by a random angle (between 0°and 360°) along a randomly selected axis. Caretta-shape was then used to align these groups on a Linux workstation using a single thread.

### 6.2.2 Protein kinases

Data

Protein kinase PDB files with group and family annotations were collected from the kinase–ligand interaction fingerprints and structure database (KLIFS)[25], resulting in 7,746 monomeric structures collectively named the KinaseAll dataset. Table 6.1 lists the number of proteins in this dataset in each kinase group.

Data about active and inactive states of kinase structures was taken from work by McSkimming et al.[26] yielding 1,773 kinases marked as having active conformations and 1,592 structures in inactive conformations. This dataset is referred to as the KinaseActive dataset in the text. A subset of 514 cyclin-dependent kinases from this set was further analysed and referred to as the CDKActive dataset.

### Shape-mers

Geometricus count vectors were calculated for kinase structures using a $k$-mer size `k = 20` and a resolution `m = 2`. These were visualized using a t-SNE embedding calculated using the scikit-learn Python library[27] with `perplexity = 30` and default parameters.

Shape-mers distinguishing a kinase group $G$ were found as those shape-mers which are present in $> 95\%$ of the proteins within $G$ and whose mean count value within $G$ is at least one more than the mean count outside $G$. We visualized distinguishing shape-mers for the STE, AGC, and TK kinase groups using three representative structures, one from each group, with PyMOL[28]. Each shape-mer can have multiple occurrences across a protein, some of which are shared across groups and do not contribute to the distinguishing nature of the shape-mer. To overcome this, we only visualize occurrences of a group's shape-mer in the group's structure which are absent in similar positions across the two structures from the other groups.

Agglomerative clustering was performed on the count vectors, again using the scikit-learn library[27], with the Bray-Curtis affinity metric and a distance threshold of 0.63. This threshold was decided using an all vs. all pairwise alignment for 232 kinase structures, up to 40 from each kinase family, from which we took the mean Geometricus similarity score of pairs with an alignment TM-score $> 0.95$.

### Alignment

Subsets of kinase structures were aligned using Caretta-shape with the same parameters as used in benchmarking. For the CK2 alignment, we superposed 292 structures according to the aligning positions and depict each structure as grey lines passing through the $\alpha$-carbon coordinates using Matplotlib[29]. The mean and standard deviation of all coordinates were depicted as a black line and coloured circles respectively.

### Machine Learning

Gradient Boosting trees were used for machine learning tasks due to their high generalization potential and capability to include missing features as a separate category. These were implemented using the XGBoost Python library[30] with a maximum depth of 5 and remaining default parameters. Kinase active vs. inactive state classification was performed on the KinaseActive and CDKActive datasets with five-fold cross validation. For the KinaseActive dataset consisting of divergent proteins, Geometricus count vectors were used as features. For the CDKActive dataset consisting of the structurally similar cyclin-dependent kinases, aligned moment invariant values were used.

In both cases, feature importance values of each predictor were averaged across cross-validation folds and summed across features. The top 2 predictive shape-mers from the KinaseActive classifier and top 10 predictive residues from the CDKActive classifier are considered in the text. Predictive shape-mer occurrences were mapped back to their corresponding residues. Predictive shape-mer residues and predictive residues from the alignment-based approaches were visualized on representative structures using PyMOL[28].

## 6.3 Results

### 6.3.1 Fast and accurate multiple structure alignment with rotation-invariant moments

Most machine learning algorithms accept a tabular, fixed dimensional matrix as input, with rows representing individual data points and columns representing features measured across all data points. For proteins sharing high structural similarity, this can be accomplished by organizing residue-level features in the order dictated by a multiple structure alignment. Desired properties of this alignment would be high accuracy in terms of structural equivalence of residues, high coverage in order to include as many relevant residue positions as possible instead of just highly conserved positions, and high speed to be able to align and re-align large datasets of proteins in typical parameter selection and validation pipelines. Here we demonstrate that Caretta-shape possesses all three of these properties.

We benchmarked Caretta-shape with the Homstrad[20] and SABMark-superfamily[21] alignment datasets, and compared against two popular structure aligners, Matt[23] and mTM-align[31]. Table 6.2 shows average quality metrics across these datasets and demonstrates that Caretta-shape returns high quality, accurate alignments with high coverage. The pairwise alignment step in Matt and mTM-align is prohibitive, with runtime complexities of $O(n^2l^3log(l))$ and $O(n^2l^2)$ respectively (where $n$ is the number of proteins and $l$ is the length of the longest protein). mTM-align's authors mention that 80-90% of their running time is spent in this step[31]. Shape-tuple similarity reduces this step to $O(n^2)$ in Caretta-shape. The entire Homstrad dataset takes only 4 minutes to align with Caretta-shape, compared to half an hour using the old Caretta algorithm and mTM-align, both of which are 10-15 times faster than Matt[31].

Figure 6.1 shows the runtime of Caretta-shape on a single thread across synthetic datasets with differing lengths and numbers of proteins. Over a thousand medium-length proteins can be aligned in as little as 20 minutes on a personal computer with a single thread. Further speed improvements such as those employed by multiple sequence alignment algorithms[32] or by the use of graphical processing units (GPUs) could extend Caretta-shape to aligning hundreds of thousands of protein structures in hours; these approaches are left for further exploration.

| Aligner | Homstrad | | | SABMark-superfamily | |
|---|---|---|---|---|---|
| | TM-score | % Gapless | Acc. | TM-score | % Gapless |
| mTM-align | 0.88 | 0.61 | 0.84 | 0.77 | 0.32 |
| Matt | 0.85 | 0.56 | **0.87** | 0.68 | 0.25 |
| Caretta | **0.92** | 0.73 | **0.87** | **0.82** | **0.46** |
| Caretta-shape | **0.92** | **0.74** | **0.87** | 0.81 | 0.45 |

Table 6.2: Average TM-score and percentage of gapless columns across Homstrad and SABmark-superfamily datasets. As the Homstrad dataset also provides reference alignments, "Acc." shows the number of gapless columns present in the corresponding reference alignment divided by the total number of gapless columns in the reference alignment.

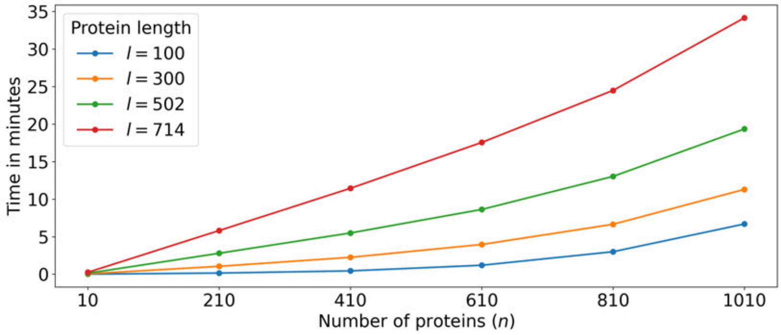


Figure 6.1: Running time in minutes of Caretta-shape on synthetic datasets with differing number of proteins and proteins with different lengths.

### 6.3.2 Structure-based exploration of protein kinases

In the past decades, protein kinases have become an alluring target for drug discovery due to the important role they play in key signal transduction pathways. These phosphotransferase enzymes mediate the transfer of the phosphate moeity from high-energy molecules such as ATP to their substrates, and are classified into broad groups based on the substrates they act on. Their popularity has led to a boom in the number of experimentally solved kinase structures with different ligands and inhibitors bound. The kinase–ligand interaction fingerprints and structure database (KLIFS)[25] now contains 7,746 monomeric structures covering 308 kinases across 8 groups and 3,341 unique ligands.

This superfamily as a whole has divergent protein structures for which only a small 85-residue catalytic segment can successfully be aligned[33]. However, individual kinase families, each consisting of up to a thousand structures, share common structural folds that lend well to alignment. With a combination of Caretta-shape and Geometricus we are able to pinpoint differences between kinase groups, align kinase families, and predict conformational change across and within kinase families all in a matter of an hour.

#### Divergent groups of proteins

Figure 6.2A shows a t-SNE embedding of Geometricus shape-mer count vectors of all 7,746 kinase monomers in the KinaseAll dataset, coloured by the group to which they belong. Clear separation is seen between groups, with smaller clusters visible within each group. These mostly correspond to the kinase families, some of which are labelled in the figure. In Figure 6.2B, for three kinase groups, we look at some shape-mers present in the members of that group and absent in the others. Many of these regions in the structure do not lie in the alignable catalytic stretch and thus would not have been found using alignment-based methods.

#### Similar families of proteins

By clustering Geometricus count vectors, we arrive at clusters of proteins displaying high structural similarity which are better suited to alignment. Table 6.3 shows clusters with $> 100$ proteins obtained after performing agglomerative clustering with a distance threshold derived from comparison of Geometricus similarity scores with pairwise alignment TM-scores (described in Methods). Each cluster only contains proteins from a single kinase group and many are dominated by a single kinase family (labelled in Figure 6.2A), demonstrating that Geometricus similarity scores can be used to assign proteins to functional groups when annotations are lacking. We used Caretta-shape to align the proteins within each cluster. The average TM-score of each cluster alignment (reported in Table 6.3) is very high, confirming their structural similarity. Figure 6.3 shows the coordinate standard deviations for the CK2 alignment, demonstrating how alignments can be used to assess residue conservation and pinpoint outliers or sub-groups.

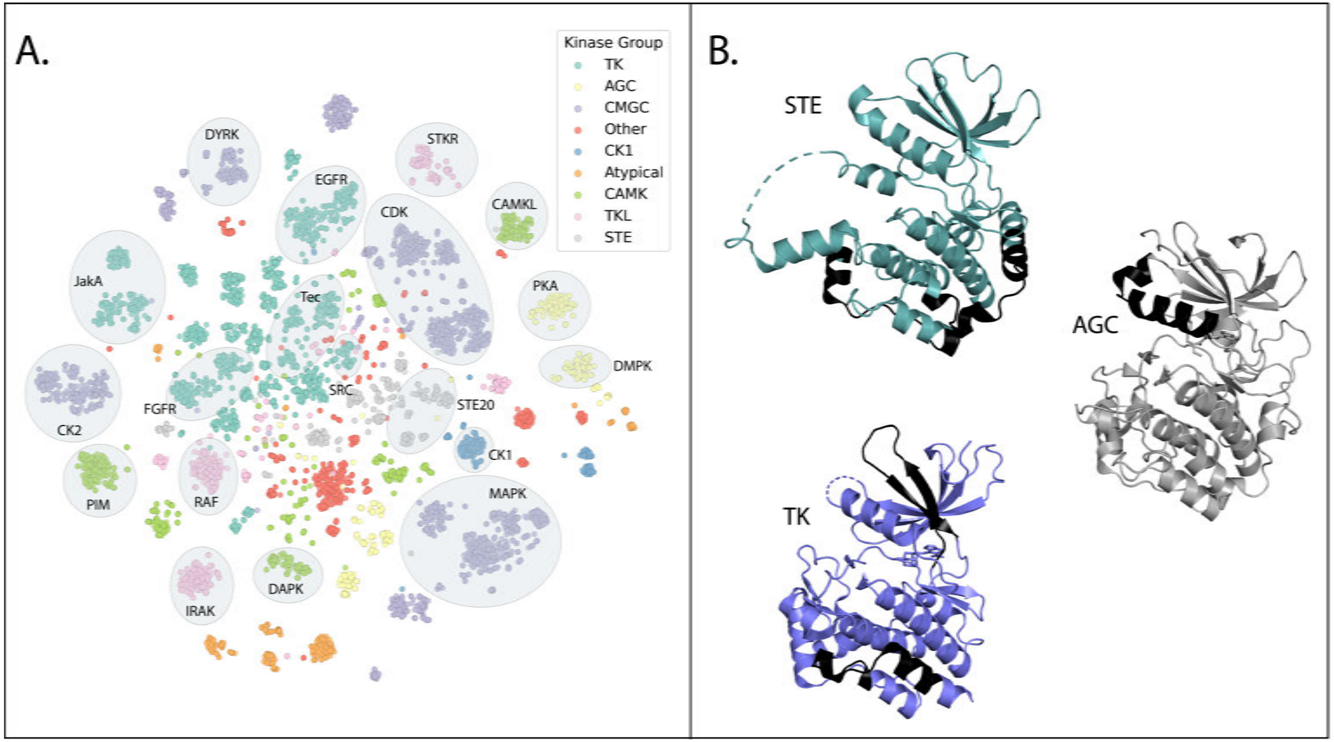


Figure 6.2: **A.** t-SNE dimensionality reduction of Geometricus shape-mer count vectors for 7,746 kinase monomers, colored by kinase group. 19 kinase families, corresponding to the clusters in Table 6.3, are labelled. **B.** Shape-mers found in one group and absent in the others, colored black for representative structures of the STE (PDB ID: 4USE), AGC (PDB ID: 3OCB), and TK (PDB ID: 6AAH) kinase groups.

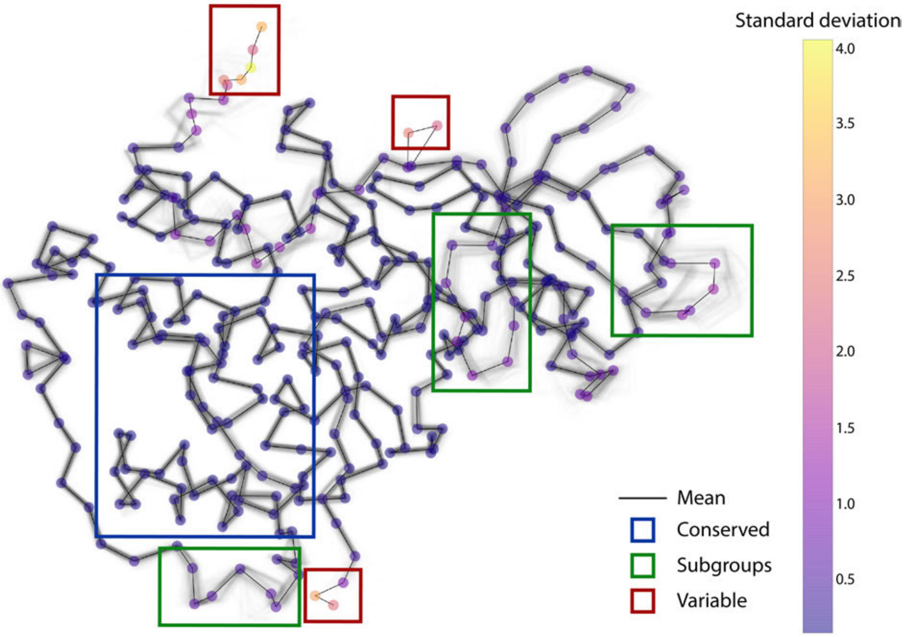


Figure 6.3: 292 CK2 kinase structures superposed according to their Caretta-shape alignment, each depicted as grey lines passing through the $\alpha$ carbon coordinates. The mean is depicted as a black line with each residue position coloured according to its standard deviation. The blue box marks a well-conserved region, green boxes mark regions showing subgroups where different structures follow different distinct paths (visible in lighter grey), and red boxes mark outlier regions where each protein has highly differing coordinates.

### 6.3.3 Kinase activity from different perspectives

The protein kinase domain can undergo dramatic conformational changes when reacting to regulatory signals in signalling pathways. These changes are controlled by protein-protein interactions, phosphorylation, and ligand binding[34]. Drug discovery efforts often aim to target specific kinase conformations and thus benefit from an understanding of conformational activation across kinases and how this activation differs across the different kinase groups and families.

Using a dataset of 3,365 kinase structures labelled as being in active or inactive conformations[26], we aim to classify a structure as belonging to one of these two states as well as pinpoint structural elements responsible for the change. We demonstrate how the alignment-free Geometricus is well-suited to tackle such classification problems across diverse proteins, such as those belonging to the different kinase groups, and Caretta-shape alignment allows for zooming into the idiosyncrasies of a single family.

| No. of proteins | Avg. TM-score | Group | Top family | Purity (%) |
|---:|---:|---|---|---:|
| 741 | 0.96 | CMGC | CDK | 82 |
| 661 | 0.96 | CMGC | MAPK | 100 |
| 570 | 0.94 | TK | Tec | 28 |
| 454 | 0.95 | TK | JakA | 59 |
| 346 | 0.95 | TK | FGFR | 61 |
| 292 | 0.99 | CMGC | CK2 | 100 |
| 288 | 0.94 | TK | EGFR | 100 |
| 248 | 0.95 | CMGC | DYRK | 64 |
| 230 | 0.96 | AGC | PKA | 62 |
| 176 | 0.99 | CAMK | PIM | 100 |
| 163 | 0.96 | CAMK | DAPK | 73 |
| 157 | 0.98 | TKL | RAF | 100 |
| 157 | 0.98 | TKL | IRAK | 100 |
| 141 | 0.98 | TKL | STKR | 100 |
| 139 | 0.98 | CAMK | CAMKL | 100 |
| 137 | 0.96 | STE | STE20 | 100 |
| 125 | 0.98 | CK1 | CK1 | 91 |
| 117 | 0.98 | AGC | DMPK | 100 |
| 108 | 0.96 | TK | Src | 66 |

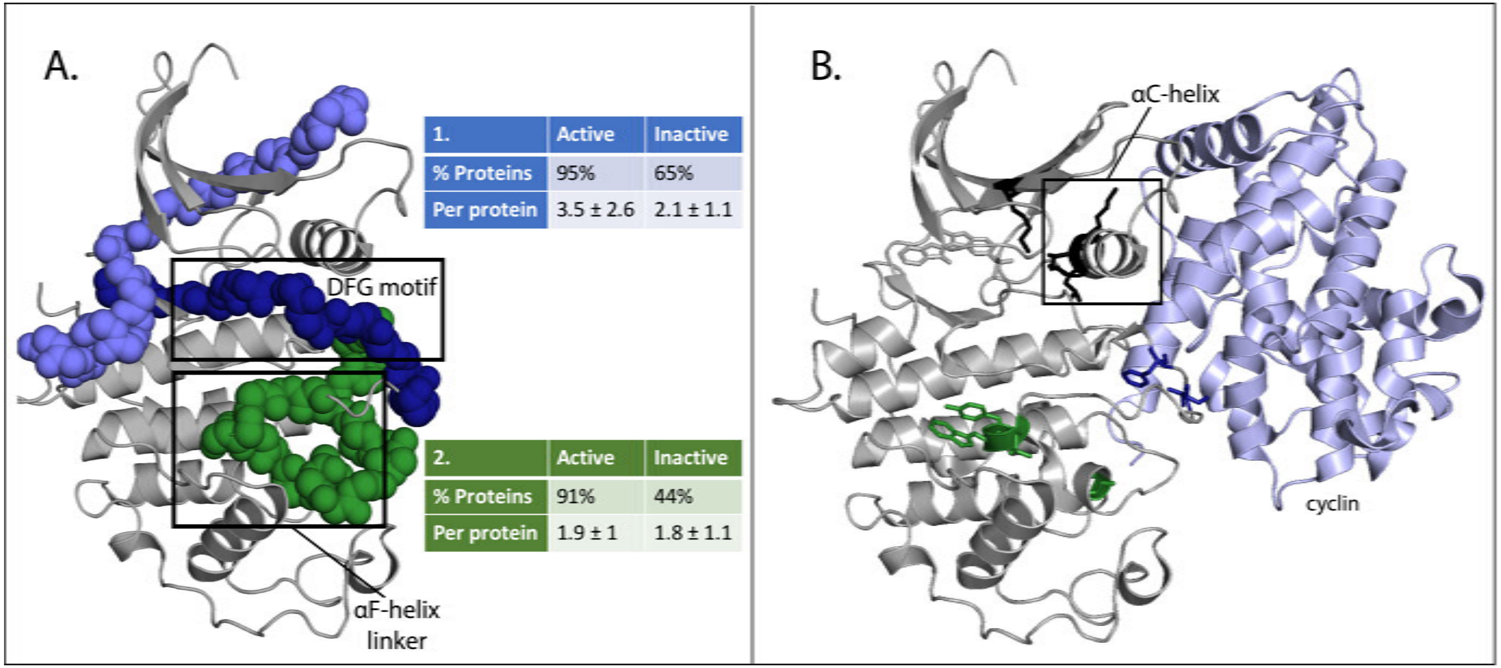
Table 6.3: Clusters of kinases obtained from Geometricus count vector clustering, also labelled in Figure 6.2A. For each cluster we report the average TM-score of its Caretta-shape alignment, the group in which its members belong, and the most frequent family along with the percentage of the family's occurrence in the cluster (purity).

### Activation across divergent kinases

To inspect activation across all kinases, we trained a Gradient Boosting classifier on Geometricus embeddings of the KinaseActive dataset. The five-fold cross validation accuracy of this classifier was $96 \pm 0.01\%$. Figure 4A shows the top two predictive shape-mers and their prevalence across active and inactive kinases. These shape-mers are also depicted on an example kinase structure (PDB ID 1E9H). One shape-mer, in dark blue, is localized in the DFG motif which lies in the well-established activation segment[26]. Another, in green, lies in the linker region connecting the activation segment to the $\alpha$F-helix which acts as an organizing hub in the activation process[35]. The DFG motif shape-mer is repeated (in light blue) but since Geometricus works with counts we cannot distinguish the true predictive motif using a single structure. More clarity is obtained when looking across multiple structures, as the dark blue occurrence is present in many structures in the active conformational state while the light blue occurrence is not.

### Activation in the cyclin-dependent kinase family

While the Geometricus approach gives us good prediction performance and pinpoints critical structural regions, it misses some structural regions that are specific to certain families. For instance, the cyclin-dependent kinase (CDK) family is dependent on the formation of a CDK-cyclin complex. Upon binding, cyclin induces conformational changes in the kinase domain that allow for autophosphorylation of the activation segment to produce a fully active kinase[36]. Thus, CDKs are further allosterically regulated through cyclin-binding, an aspect not seen in our coarse-grained classifier trained across all kinases. To analyse a specific subfamily such as CDKs, an alignment based approach can be beneficial due to the high structural similarity between proteins and expected residue correspondences. We create a Caretta-shape alignment across 514 CDKs, resulting in an alignment of 399 residues with an average TM-score of 0.96. A Gradient Boosting classifier is trained on the aligned moment invariant values of each CDK, resulting in a very high accuracy of $99\%$. Figure 4B depicts the top 10 predictive residues. While some residues are again found in the DFG motif and $\alpha$F-helix linker regions, residues in the $\alpha$C-helix which forms part of the cyclin-CDK interface are also found as predictive, indicating that this predictor picks up CDK-specific patterns relevant for kinase conformational change.



| 1. | Active | Inactive |
|---|---|---|
| % Proteins | 95% | 65% |
| Per protein | 3.5 ± 2.6 | 2.1 ± 1.1 |

| 2. | Active | Inactive |
|---|---|---|
| % Proteins | 91% | 44% |
| Per protein | 1.9 ± 1 | 1.8 ± 1.1 |

Figure 6.4: **A.** Occurrences of two shape-mers (shape-mer 1 in dark and light blue, shape-mer 2 in green) found predictive for conformational state classification on a representative kinase structure (PDB ID: 1E9H). The DFG motif and $\alpha$F-helix linker are labelled. The percentage of proteins containing each shape-mer and the average number of times the shape-mer appears per protein across the active and inactive kinases is shown. **B.** Top ten predictive residues for cyclin-dependent kinase conformational state classification shown along with the position of cyclin (light blue). Residues near the DFG motif and $\alpha$F-helix linker are coloured blue and green respectively, while the remaining predictive residues are coloured black. The $\alpha$C-helix in the CDK-cyclin interface is labelled.

## 6.4 Conclusion

With growing numbers of experimentally solved protein structures and proteins capable of being computationally modelled, structures are seeing increasing use in machine learning applications. To that end we present Caretta-shape, a very fast and accurate multiple structure alignment algorithm based on the concept of rotation-invariant moments, aimed at generating aligned structural features for machine learning.

Depending on the similarity between proteins under study, an alignment-free or alignment-based approach is preferred and each presents its own advantages and insights. We adapt these two approaches to the protein kinase superfamily, which consists of structurally divergent protein groups as well as more similar protein families. We use machine learning to tackle active/inactive conformational state prediction across all kinase families with Geometricus and across the cyclin-dependent kinase family with Caretta-shape alignments. These two approaches lead to the exploration of different aspects of catalytic mechanisms: one aspect explains commonalities within all proteins in this diverse superfamily, and the other zooms in on peculiarities displayed by a single family.

Computational structure modelling is capable of expanding datasets of proteins into the thousands. Once the expensive but automated modelling steps are complete, analyses similar to the ones presented here, both unsupervised and supervised, can be carried out with comparable ease allowing for fast iteration and adaptive exploration of protein biology.

## Broader Impact

The research presented here includes a novel multiple structure alignment algorithm and a demonstration of recently developed algorithms for analysing protein structures with machine learning. Researchers in structural bioinformatics and enzymology may benefit from this work for obtaining structural insights from their data. The ideas discussed also form a fertile basis for more complex algorithms that leverage the increasing amounts of data and recent advances in machine learning and deep learning techniques aimed at such structured, high-dimensional data.

# References

[1] Bernstein, F. C., Koetzle, T. F., Williams, G. J., Meyer, E. F., Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T., & Tasumi, M. (1977). The Protein Data Bank: A computer-based archival file for macromolecular structures. *Journal of Molecular Biology*, *112*, 535–542.

[2] Senior, A. W., Evans, R., Jumper, J., Kirkpatrick, J., Sifre, L., Green, T., Qin, C., Žídek, A., Nelson, A. W., Bridgland, A. et al. (2020). Improved protein structure prediction using potentials from deep learning. *Nature*, (pp. 1–5).

[3] Jain, P., Garibaldi, J. M., & Hirst, J. D. (2009). Supervised machine learning algorithms for protein structure classification. *Computational Biology and Chemistry*, *33*, 216–223.

[4] Cao, R., Adhikari, B., Bhattacharya, D., Sun, M., Hou, J., & Cheng, J. (2017). QAcon: Single model quality assessment using protein structural and contact information with machine learning techniques. *Bioinformatics*, *33*, 586–588.

[5] Krivák, R., & Hoksza, D. (2018). P2Rank: Machine learning based tool for rapid and accurate prediction of ligand binding sites from protein structure. *Journal of cheminformatics*, *10*, 39.

[6] Berliner, N., Teyra, J., Çolak, R., Lopez, S. G., & Kim, P. M. (2014). Combining structural modeling with ensemble machine learning to accurately predict protein fold stability and binding affinity effects upon mutation. *PLoS One*, *9*, e107353.

[7] Ferraro, E., Via, A., Ausiello, G., & Helmer-Citterich, M. (2006). A novel structure-based encoding for machine-learning applied to the inference of SH3 domain specificity. *Bioinformatics*, *22*, 2333–2339.

[8] Romero, P. A., Krause, A., & Arnold, F. H. (2013). Navigating the protein fitness landscape with Gaussian processes, . *110*, E193–E201.

[9] Lee, J.-H., Hamilton, M., Gleeson, C., Caragea, C., Zaback, P., Sander, J. D., Li, X., Wu, F., Terribilini, M., Honavar, V. et al. (2008). Striking similarities in diverse telomerase proteins revealed by combining structure prediction and machine learning approaches. In *Biocomputing 2008* (pp. 501–512). World Scientific.

[10] Vass, M., Kooistra, A. J., Ritschel, T., Leurs, R., de Esch, I. J., & de Graaf, C. (2016). Molecular interaction fingerprint approaches for GPCR drug discovery. *Current Opinion in Pharmacology*, *30*, 59–68.

[11] Durairaj, J., Akdel, M., de Ridder, D., & van Dijk, A. D. J. (2020). Geometricus represents protein structures as shape-mers derived from moment invariants. *Bioinformatics*, *36*, i718–i725.

[12] Akdel, M., Durairaj, J., de Ridder, D., & van Dijk, A. D. J. (2020). Caretta-a multiple protein structure alignment and feature extraction suite. *Computational and Structural Biotechnology Journal*, *18*, 981–992.

[13] Cavasotto, C. N., & Palomba, D. (2015). Expanding the horizons of G protein-coupled receptor structure-based ligand discovery and optimization using homology models. *Chemical Communications*, *51*, 13576–13594.

[14] Ma, J., & Wang, S. (2014). Algorithms, applications, and challenges of protein structure alignment. In *Advances in Protein Chemistry and Structural Biology* (pp. 121–175). Elsevier volume 94.

[15] Thompson, J. D., Gibson, T. J., & Higgins, D. G. (2003). Multiple sequence alignment using ClustalW and ClustalX. *Current Protocols in Bioinformatics*, (pp. 2–3).

[16] Lassmann, T., & Sonnhammer, E. L. (2005). Kalign–an accurate and fast multiple sequence alignment algorithm. *BMC bioinformatics*, *6*, 1–9.

[17] Edgar, R. C. (2004). MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, *32*, 1792–1797.

[18] Saitou, N., & Nei, M. (1987). The neighbor-joining method: A new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution*, *4*, 406–425.

[19] Kabsch, W. (1976). A solution for the best rotation to relate two sets of vectors. *Acta Crystallographica Section A: Crystal Physics, Diffraction, Theoretical and General Crystallography*, *32*, 922–923.

[20] Mizuguchi, K., Deane, C. M., Blundell, T. L., & Overington, J. P. (1998). HOMSTRAD: A database of protein structure alignments for homologous families. *Protein Science*, *7*, 2469–2471.

[21] Van Walle, I., Lasters, I., & Wyns, L. (2004). SABmark—a benchmark for sequence alignment that covers the entire known fold space. *Bioinformatics*, *21*, 1267–1268.

[22] Dong, R., Peng, Z., Zhang, Y., & Yang, J. mTM-align benchmark results. URL: `http://yanglab.nankai.edu.cn/mTM-align/benchmark/`.

[23] Menke, M., Berger, B., & Cowen, L. (2008). Matt: Local flexibility aids protein multiple structure alignment. *PLoS Computational Biology*, *4*, e10.

[24] Zhang, Y., & Skolnick, J. (2004). Scoring function for automated assessment of protein structure template quality. *Proteins: Structure, Function, and Bioinformatics*, *57*, 702–710.

[25] Kooistra, A. J., Kanev, G. K., van Linden, O. P., Leurs, R., de Esch, I. J., & de Graaf, C. (2016). KLIFS: A structural kinase-ligand interaction database. *Nucleic Acids Research*, *44*, D365–D371.

[26] McSkimming, D. I., Rasheed, K., & Kannan, N. (2017). Classifying kinase conformations using a machine learning approach. *BMC Bioinformatics*, *18*, 86.

[27] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V. et al. (2011). Scikit-Learn: Machine learning in Python. *the Journal of Machine Learning Research*, *12*, 2825–2830.

[28] DeLano, W. L. et al. (2002). PyMOL: An open-source molecular graphics tool. *CCP4 Newsletter on Protein Crystallography*, *40*, 82–92.

[29] Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. *Computing in Science & Engineering*, *9*, 90–95.

[30] Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* KDD '16 (pp. 785–794). Association for Computing Machinery.

[31] Dong, R., Peng, Z., Zhang, Y., & Yang, J. (2017). mTM-align: An algorithm for fast and accurate multiple protein structure alignment. *Bioinformatics*, *34*, 1719–1725.

[32] Sievers, F., Wilm, A., Dineen, D., Gibson, T. J., Karplus, K., Li, W., Lopez, R., McWilliam, H., Remmert, M., Söding, J., Thompson, J. D., & Higgins, D. G. (2011). Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Molecular Systems Biology*, *7*, 539.

[33] Van Linden, O. P., Kooistra, A. J., Leurs, R., De Esch, I. J., & De Graaf, C. (2014). KLIFS: A knowledge-based structural database to navigate kinase–ligand interaction space. *Journal of Medicinal Chemistry*, *57*, 249–277.

[34] Johnson, L. N., Noble, M. E., & Owen, D. J. (1996). Active and inactive protein kinases: Structural basis for regulation. *Cell*, *85*, 149–158.

[35] Kornev, A. P., Taylor, S. S., & Ten Eyck, L. F. (2008). A helix scaffold for the assembly of active protein kinases. *Proceedings of the National Academy of Sciences*, *105*, 14377–14382.

[36] Russo, A. A., Jeffrey, P. D., & Pavletich, N. P. (1996). Structural basis of cyclin-dependent kinase activation by phosphorylation. *Nature Structural Biology*, *3*, 696–700.

CHAPTER 7

# Matchmaker - A joint structure-molecule framework to predict protein-compound specificity

Janani Durairaj, Thamara Hesselink, Harro J. Bouwmeester, Dick de Ridder, Jules Beekwilder, and Aalt D.J. van Dijk

# Abstract

Plant sesquiterpene synthases (STSs) catalyse the formation of a large number of volatile sesquiterpenes. Predicting the exact sesquiterpene produced by a given STS is a challenging task, yet crucial for a number of biotechnological applications. Here, we present a novel machine learning framework, Matchmaker, that combines protein structural features derived from STS homology models with chemical features derived from sesquiterpenes to successfully predict product specificity of STSs. We experimentally characterize 64 novel STSs across STS sequence and chemical space, on which we demonstrate Matchmaker's performance. Matchmaker also enables pinpointing residues involved in the formation of specific sesquiterpenes. These insights will be useful for future studies looking to engineer enzymes with desired specificity.

## 7.1 Introduction

A number of protein families are capable of collectively modifying or producing one or many of a wide range of possible compounds. This is especially the case for families involved in the production of natural products or specialized metabolites – compounds not required for an organisms primary biochemical pathways of cell growth and reproduction, yet crucial for environmental adaptation[1]. The intense evolutionary pressure over nearly half a billion years to keep up with surrounding predators, competitors, and allies alike combined with limits on the amount of available energy and material to spend on this ongoing battle has resulted in a relatively small number of proteins capable of producing and modifying a vast array of natural product compounds[2]. Thus, these proteins are very diverse in terms of sequence, across families and even within the same family across different species.

Biotechnological applications involving specialized metabolites often require computational methods to predict substrate or product specificity. These can be useful to screen for proteins with desired specificity, to predict specificity in proteins from a specific species or containing other desired characteristics such as high thermostability, to optimize production in different host systems, and to understand the residues involved with the aim of engineering mutants and novel proteins to possess required specificity or inhibit/enhance protein activity. The immense sequence diversity of these enzyme families has led to approaches tailored to specific proteins, for example those combining molecular docking and molecular dynamics to screen possible compounds and intermediates against a single protein[3–8]. Such techniques often have a significant computational cost, require extensive fine-tuning and focus entirely on a single protein, thus hindering the interpretation of results from a family perspective. Another common technique for screening is simple sequence identity to proteins with known products. While this has some success for proteins from previously characterized species or genera, it does not generalize well to proteins from rare species or proteins producing rare compounds. More importantly, such an approach gives no indication of the residues or regions in the protein responsible for transferring a particular specificity. This knowledge is crucial for successful engineering of novel mutants with desired properties.

The family of plant sesquiterpene synthases (STSs), responsible for the production of compounds called sesquiterpenes involved in plant fragrances, is a perfect example of a diverse natural product enzyme family associated with a diverse range of specialized metabolites. STSs collectively produce hundreds of different sesquiterpene compounds, with many STSs being promiscuous in nature and capable of individually producing multiple compounds at varying levels. Figure 7.1 shows the STS reaction cascade, starting from the substrate farnesyl diphosphate (FPP) which undergoes a series of cyclizations, hydride shifts, methyl shifts, rearrangements, re- and deprotonations to produce the final enzyme products. In previous research, we demonstrated that all the products of a multi-product STS tend to arise from the same reaction pathway, indicating that chemical similarity between sesquiterpene compounds is important to consider when predicting product specificity[9]. In addition, we showed that sequence identity in this family did not explain product specificity, [10], but protein structure carries more information about reaction specificity[9]. Thus, predicting product specificity in STSs requires an approach that takes into account STS structure information and sesquiterpene chemical information.

Recently, machine learning has become a popular choice for a variety of general protein-compound specificity predictors. Deep learning has been used to predict interaction specificity on large datasets consisting of tens to hundreds of thousands of protein-ligand pairs. However, while these models tend to have superior performance when applied to broad and diverse datasets consisting of proteins spanning multiple superfamilies, their performance drastically decreases when applied to small datasets of individual families[11]. Thus, a variety of shallow machine learning algorithms relying on expert-based descriptors have been developed and demonstrated on independent families of just hundreds of proteins. These approaches can be categorized according to the kinds of input they take – some use only the protein as input, some only the compound, and some are proteochemometric and use both protein and molecule descriptors. Classification algorithms need a sufficient number of samples per class in order to successfully learn generalizable patterns for each class. This presents a problem for approaches that take only the protein as input and attempt to classify the specificity of the protein, as the number of such classes (of compounds) they can successfully predict is limited by the number of labelled proteins available in each class. Thus, to ensure sufficient data per class, compounds are typically grouped according to some shared criteria such as the presence of a particular chemical skeleton, shared parent compound etc., as was done in our previous research on classifying STSs into two categories based on the parent cation of their products[9]. This unfortunately often means that these predictors only narrow down compound specificity but cannot resolve the final compound, still leaving researchers with numerous possible answers. In addition, the chemical similarity between compounds in the different classes contains relevant information for prediction but is left unused. A similar situation occurs in ligand-centric approaches. These typically use a large set of molecules as input and predict specificity of these molecules to a single protein and thus cannot be used to make claims about other proteins. Proteochemometric approaches, on the other hand, do not have these limitations and thus are ideal for predicting compound specificity in individual protein families such

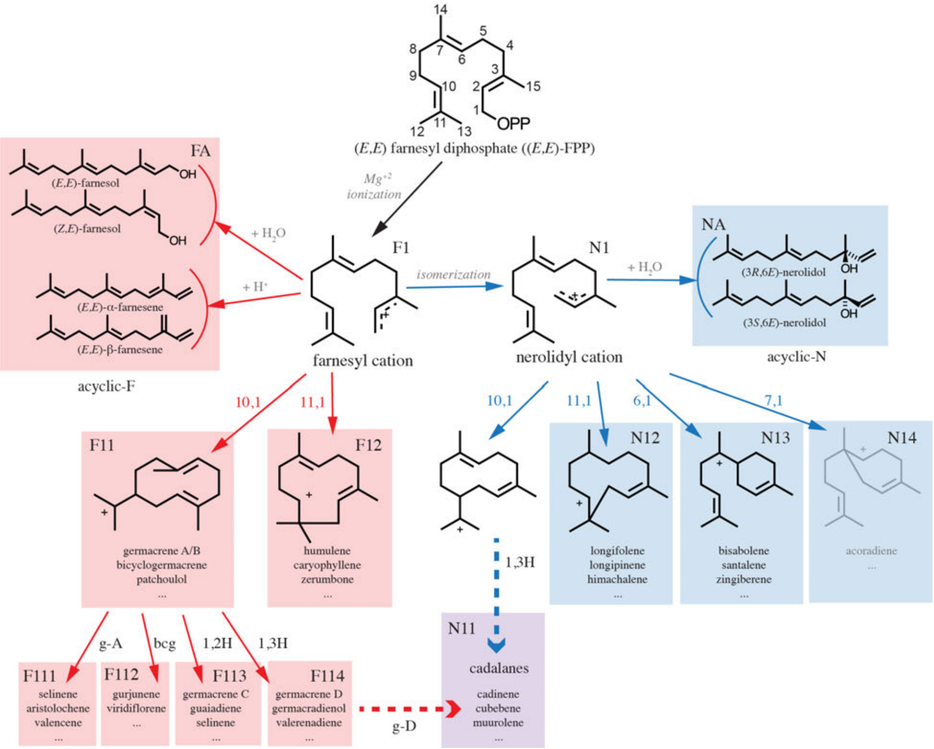


Figure 7.1: The reaction mechanism of sesquiterpene production starts with farnesyl diphosphate ((*E*,*E*)-FPP). Loss of the diphosphate moiety (OPP) leads to farnesyl cation formation. The farnesyl cation can subsequently be converted to the nerolidyl cation. Acyclic sesquiterpenes (acyclic-F and acyclic-N) are formed from these two cations by proton loss or reaction with water molecules. Possible cyclizations for both cations are indicated in the figure. The subsequently formed cyclic cations undergo modifications and rearrangements to form cyclic sesquiterpenes. Some of these sesquiterpenes (g-A and bcg) themselves act as neutral intermediates which can be re-protonated and undergo further reactions to form more products. Products are also formed from specific charged intermediates such as a 1,2- or 1,3-hydride shift of the 10,1-cyclized farnesyl cation (1,2H, 1,3H) and the cadalane skeleton (cadalanes), which can be formed via either of the two precursor cations, or via acid-induced rearrangement of germacrene D. The 7,1-cyclization of the nerolidyl cation, shown in grey, is not found in plant-derived sesquiterpenes. g-A = germacrene A, g-D = germacrene D, bcg = bicyclogermacrene

as the STSs. However, previous proteochemometric approaches either used protein sequence derived features[12] or protein-ligand complex derived features[13,14]. While the former misses out on useful structure information, the latter requires structures of complexes or expensive docking studies to generate these. In addition, a number of these predictors use kernel methods, such as support vector machines, to generate protein-compound similarity matrices for learning. These are difficult to interpret, in that it is not straightforward to learn from a trained predictor what residues and regions are predictive for specificity.

Here we present a novel joint framework for STS product specificity prediction, Matchmaker, which considers sequence and structure features extracted from modelled 3D structures of unbound proteins along with chemical features extracted from the free state of individual compounds. The framework is trained on protein-compound pairs and returns a compatibility score. This score, when sorted across all possible compound combinations for a given protein, reveals the most probable product. In addition, the framework enables inspection of predictive features and residues for each individual compound, allowing researchers to drill down into compound-specific structural regions and design mutant studies tailored towards desired compounds.

We evaluate two settings: predicting the compound produced by a specific STS, i.e. *protein-centric prediction*, and screening for STSs producing desired compounds, i.e. *compound-centric prediction*, demonstrating the ease of adaptation of a trained Matchmaker framework to different biotechnological applications. We also consider the inclusion of minor or side products of multi-product STSs, and the effect this has on performance. To comprehensively evaluate our framework on STSs from different species with varying levels of sequence identities to known enzymes, we experimentally characterize 64 novel STSs to use as an independent test set. Compound-specific predictive residues returned by our framework overlap with results of previous theoretical research and mutation studies specific to the compound under consideration, implying that the advanced interpretation capabilities of Matchmaker can be used to customize experimental studies to compounds and proteins of choice. Matchmaker thus provides a unique, structure-based and highly interpretable proteochemometric compound specificity prediction approach, likely to be useful for other protein families with similar properties where pinpointing residues important in reaction mechanisms is as important as predicting specificity.

## 7.2 Methods

### 7.2.1 Overview

Figure 7.2A depicts the usage and workflow of Matchmaker, starting from a set of protein sequences and a set of compound SMILES strings. Homology modelling is used to generate structural models for each STS sequence, followed by the extraction of sequence- and structure-derived features and the alignment of these features across the set of proteins based on structural alignment. Chemical fingerprints are generated for each compound from its SMILES representation. A gradient boosting classifier is then trained on all protein-compound pairs, with pairs consisting of compounds produced by a protein labelled as positive and the remaining as negative. For test proteins, compatibility scores are generated across all possible compounds, producing a protein-compound compatibility score matrix (Figure 7.2A). Figure 7.2B depicts the two settings for obtaining predictions from this matrix:

1. **Protein-centric prediction setting**: to predict product specificity for a protein, compatibility scores are sorted across compounds for that protein, and the compound with the highest score is its predicted product.
2. **Compound-centric prediction setting**: to predict proteins containing desired specificity from a set of possible candidates, compatibility scores are sorted across proteins in the column corresponding to the desired compound. The proteins with the highest scores in this column are most likely to produce the compound of interest.

The subsections below delve into the details of applying the above prediction approach to plant STSs.

### 7.2.2 Proteins

For training Matchmaker we used the proteins in the plant STS database (`https://www.bioinformatics.nl/sesquiterpene/synthasedb/`) for which product specificity has been experimentally determined via GCMS and NMR studies[10]. This dataset consists of 302 enzymes, covering 106 species in 68 genera, and is henceforth referred to as the training set.

#### Test set selection and characterization

As an independent test set, we selected and experimentally characterized novel putative STSs from sequenced plant genomes and transcriptomes. We started with a set of 18,667 putative STS candidates from the TrEMBL database[15] and the 1000 plant transcriptome project[16], defined as proteins containing the Terpene_synth_C Pfam domain (Pfam ID: PF03936). These were extracted using HMMER (version) with default settings. We used the approach detailed by Terzyme[17] to create hidden Markov models (HMMs) for characterized STSs, monoterpene synthases (MTSs), and diterpene synthases (DTSs) and further restricted candidates to those having an STS HMM score over 500 (which was the average STS HMM score of proteins in the STS database) and greater than the corresponding MTS and DTS scores. To remove non-functional enzymes, we discarded sequences which didn't start with Methionine, which had sequence lengths more than one standard deviation away from characterized enzymes (both considering full sequence length and C-terminal domain sequence length), and which did not contain the conserved RXR, DDXXD, and NSE/DTE motifs[10,18]. From this resulting set we used three different strategies to select 160 candidates to characterize:

1. **Filling sequence space**: We iteratively selected 50 sequences with the lowest pairwise sequence identity to their closest sequence from both the training set

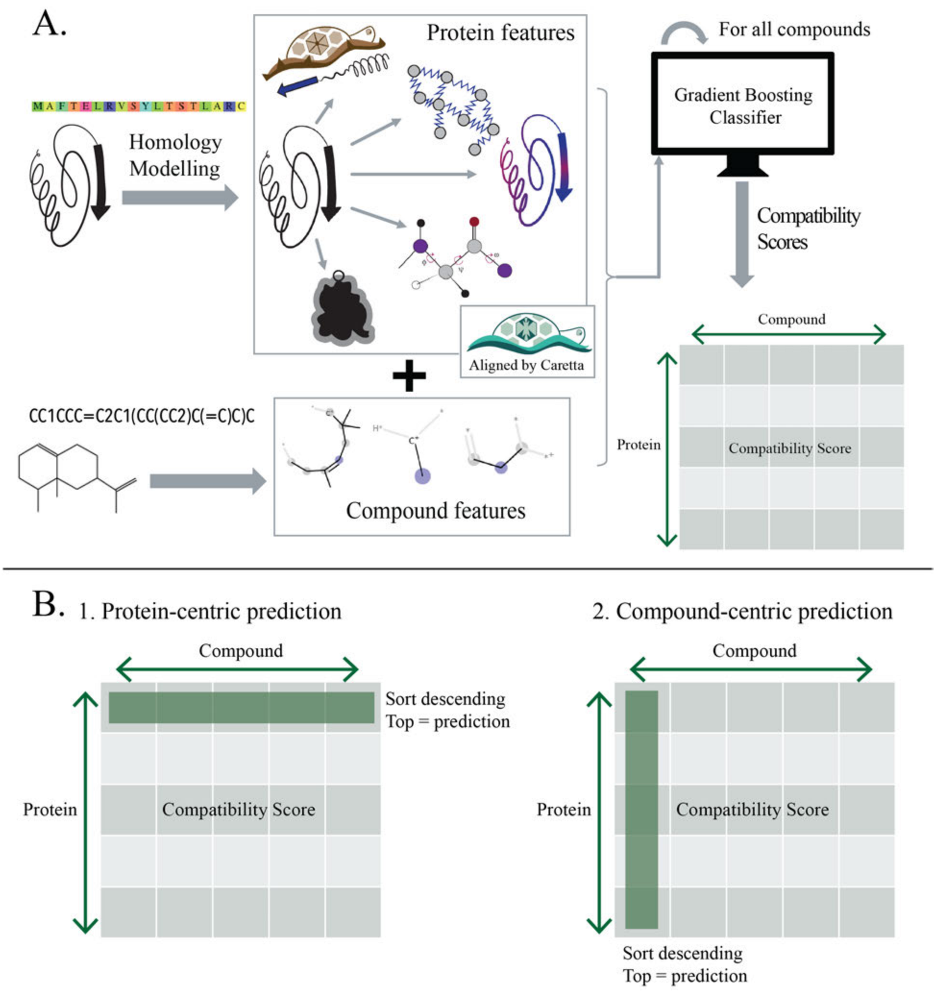


Figure 7.2: **A** Matchmaker workflow, starting from a set of protein sequences and a set of molecule SMILES strings, and resulting in a protein-compound compatibility score matrix. **B** The two settings for generating predictions from the compatibility score matrix.

and previously chosen set. As this experiment was performed before the second version of the plant STS database with updated literature studies was released[9], it used only the STSs from the first version[10] in the training set.

2. **Filling mutation space**: We obtained 26 cyclization-specific residues (according to the cyclizations depicted in Figure 7.1) by inspecting the sequence alignment of already characterized enzymes. These enzymes were divided into cyclization groups based on the cyclization of their major product. Residue positions where the most common amino acid was not the same across all groups, and at least one group had a >50% conservation at that position were considered as cyclization-specific. These positions were used to create a graph of characterized and uncharacterised STSs, with edges representing the number of mutations within the cyclization-specific residues. From this graph, we selected 30 as yet uncharacterised enzyme sequences with the highest load centrality, i.e. those acting as a bridge between characterized enzymes.
3. **Increasing nerolidyl cation representation**: In order to deal with the imbalance in our database, with over 70% of the characterized enzymes producing compounds derived from the farnesyl cation, we used our previous cation specificity predictor[9] to select 80 sequences at least 10 residues away from characterized enzymes, from species with <2 characterized enzymes, with <90% identity to each other and with >5% predicted nerolidyl specificity. We also ensured that none of these selected enzymes had protein features close to those of nerolidol synthases, as described in Durairaj et al.[9].

Codon-optimized versions of the selected candidates were produced using JCat[19]. The resulting synthetic genes were ordered from BaseClear (Leiden), and cloned into MCS1 (BamHI, NotI) cloning sites of the expression vector pACYCDuet-1 (Novagen, CmR). The resulting constructs, and an empty pACYCDuet, were transformed into chemically competent *Escherichia coli* BL21 DE3 carrying the plasmid pMEV, which is a kanamycin-resistant version of a mevalonate pathway-expressing plasmid[20]. Transformants were plated on LB plates supplemented with kanamycin (50 µg/ml), chloramphenicol (50 µg/ml) and 1% glucose. A starter culture of all clones was grown overnight in 5 ml LB medium supplemented with kanamycin (50 µg/ml), chloramphenicol (50 µg/ml) and 1% glucose at 37 °C, 250 rpm. 20 ml YT supplemented with kanamycin (50 µg/ml) and chloramphenicol (50 µg/ml) in 100 ml Erlenmeyer flasks were inoculated using the starter cultures at an OD600 of 0.1. The culture was incubated at 37 °C, 250 rpm until the OD600 reached 0.45-0.55. At this point, 20 µl 1 M IPTG was added, and 2 ml n-dodecane, and cultures were further incubated for 24h at 30 °C and 250 rpm. Subsequently, cultures were centrifuged (15 min 3,600×g), and the dodecane fraction was collected. Dodecane was diluted 1:100 in hexane, and used for GCMS analysis according to Di Girolamo et al.[21]. Eluting peaks were identified by their mass spectrum using the NIST V8 library[22].

Table 7.1: **The six structures used for multi-template modelling**

| Name | PDB ID | Resolution | Species | Product | Cation |
|---|---|---|---|---|---|
| GACS | 3G4F | 2.65Å | *Gossypium arboreum* | (+)-$\delta$-cadinene | cadalane |
| AGBS | 3SDU | 1.89Å | *Abies grandis* | $\alpha$-bisabolene | nerolidyl |
| AABS | 4FJQ | 2.00Å | *Artemisia annua* | $\alpha$-bisabolol | nerolidyl |
| AAHS | 4GAX | 1.99Å | *Artemisia annua* | $\gamma$-humulene | farnesyl |
| HMVS | 5JO7 | 2.15Å | *Hyoscyamus muticus* | vetispiradiene | farnesyl |
| TEAS | 5EAU | 2.15Å | *Nicotiana tabacum* | 5-*epi*-aristolochene | farnesyl |

Sequence extraction

The C-terminal domain sequence, that contains the active site, was extracted from each STS using HMMER[23] and the Pfam[24] domain PF03936. The C-terminal domain sequence is defined from the starting position of the HMMER hit to the end of the protein.

Measuring sequence similarities

A distance matrix of the C-terminal domain sequences of all 160 novel STS candidate enzymes from Section 7.2.2 and 302 training set STSs was constructed using the pairwise sequence k-tuple measure described by Wilbur & Lipman[25], implemented in Clustal Omega (version 1.2.4)[26]. This distance matrix was then used to construct a multi-dimensional scaling (MDS) plot using scikit-learn (version 0.19.1)[27].

Homology modelling

For each STS, 500 multi-template homology models were created of the C-terminal domain sequence using MODELLER[28], with six STS structures from the PDB[29] as templates, as listed in Table 7.1. These were aligned to each sequence using the C-terminal PF03936 Pfam domain[24] as a guide, using Clustal Omega[26]. Three $Mg^{2+}$ ions were also included while modelling. The model with the lowest normalized DOPE score was selected for feature extraction.

Feature extraction and alignment

The same protein features were extracted for each STS from both sequence and modelled structure as described in our previous work[9]. This includes residue-level features derived from sequence conservation, flexibilities based on normal mode analysis, Coulomb and Born electrostatics, bond angles, residue depths, and surface accessibility. In addition, our recently released topological feature extraction tool, Geometricus[30], was used to extract rotation invariant moments for each residue of each modelled structure. 16 moments were calculated considering the $C\alpha$ coordinates of residues from 6 positions upstream and downstream of each residue. These include the 4 moments described in our manuscript[30]($O_3$, $O_4$, and $O_5$ and $F$) and 12 independent third order moments from Flusser et al.[31] ($phi_{2-13}$). Features were aligned according to a multiple structure alignment generated by Caretta-shape[32] using default parameters. Gaps in the alignment were represented as NaNs.

### 7.2.3 Compounds

We collected all sesquiterpene compounds produced by STSs in our dataset and obtained their SMILES strings from the PubChem database[33] using the Python wrapper for the PubChem REST API[34], PubChemPy (version 1.0.4). Since in many cases the exact chirality of the compound was not determined during experimental characterization, we use canonical SMILES without chirality information. This resulted in 64 unique compounds of which 4, produced by 11 STSs, are unique to the test set.

#### Feature extraction

Morgan circular fingerprints[35] with a radius of 3 were calculated from the SMILES strings of each compound using RDKit (version 2020.09)[36]. Only bit indices present in at least one sesquiterpene from the training set were retained, resulting in 816 features per compound.

### 7.2.4 Matchmaker prediction framework

Matchmaker uses gradient boosting trees trained on concatenated protein and compound features ($X$). The training set consists of all possible pairs of the 302 training STSs and 64 training compounds. The response variable ($y$) for pairs consisting of STSs and their major products are marked as 1 and the rest as 0. Matchmaker-Multi on the other hand includes all products of promiscuous STSs, i.e. $y = 1$ for pairs consisting of STSs and any of their products.

The classifier is constructed using XGBoost[37]. We use gradient boosting parameters consistent with the low number of proteins and high number of features in our dataset in order to avoid overfitting - a high number of trees (`n_trees = 1000`), a low learning rate (`learning_rate = 0.01`), subsampling data points for each tree (`subsample = 0.8`), and subsampling features for each tree and level (`colsample_bytree = 0.5`, `colsample_bylevel = 0.5`). To handle the high rate of imbalance between positive ($y = 1$) and negative ($y = 0$) samples, we also control the balance of positive and negative weights (`scale_pos_weight` = $\frac{\text{number of positive pairs}}{\text{number of negative pairs}}$). The posterior predicted probabilities of the trained gradient boosting model are termed as the compatibility scores of protein-compound pairs.

### 7.2.5 Prediction and evaluation

We evaluate Matchmaker on STSs across the two different settings depicted in Figure 7.2B, and compare it to prediction using sequence identity, defined as the Clustal Omega k-tuple similarity measure[26] between C-terminal domain sequences.

### Protein-centric prediction setting

The goal in this setting is to correctly predict the product of a given protein. Our previous research[9,10] established that overall protein similarity in STSs is more associated with phylogeny than product specificity - i.e. STSs from the same species are similar irrespective of the compounds they produce. This necessitates careful evaluation that takes this phylogeny into account. Hence, instead of random cross-validation, we use a genus-based split with each fold consisting of STSs from six genera at a time. For each protein in each validation set, we generate compatibility scores for each possible product. After sorting these scores for each protein in descending order, we inspect the topmost, top-3, and top-5 predicted compounds and count the prediction as correct if the major product is present in these sets (and, in Section 7.3.3, if any of the products is present). These same scores are used across the remaining settings as well. For sequence-based prediction, we use a 1-nearest neighbour classifier on sequence identity. As a baseline, we also calculated results for random prediction, assigning labels based on the frequencies of compounds in our dataset. This setting is also used for our independent test set of newly characterized enzymes using a Matchmaker predictor trained on the entirety of the training set.

### Compound-centric prediction setting

In this setting, the goal is to predict proteins which produce a particular compound, aiming to reduce the experimental cost of screening for desired specificity. To evaluate this setting, we attempt to predict proteins producing sesquiterpenes with more than 10 examples in the training set. This is considering the compatibility scores across columns corresponding to the selected compounds. We report the score obtained by sorting the compatibility scores in descending order (Figure 7.2B2) and counting the number of proteins in the top 10 which truly produce the corresponding compound for each compound. To evaluate sequence identity based screening, for each selected compound, the top-10 proteins with the highest sequence identity to any training protein producing the compound are considered. To obtain a percentage, we divide these counts by the count obtained when assuming perfect recall. For the three compounds most common in the training set, the Receiver Operating Characteristic (ROC) curve is calculated for Matchmaker and sequence-identity based scoring.

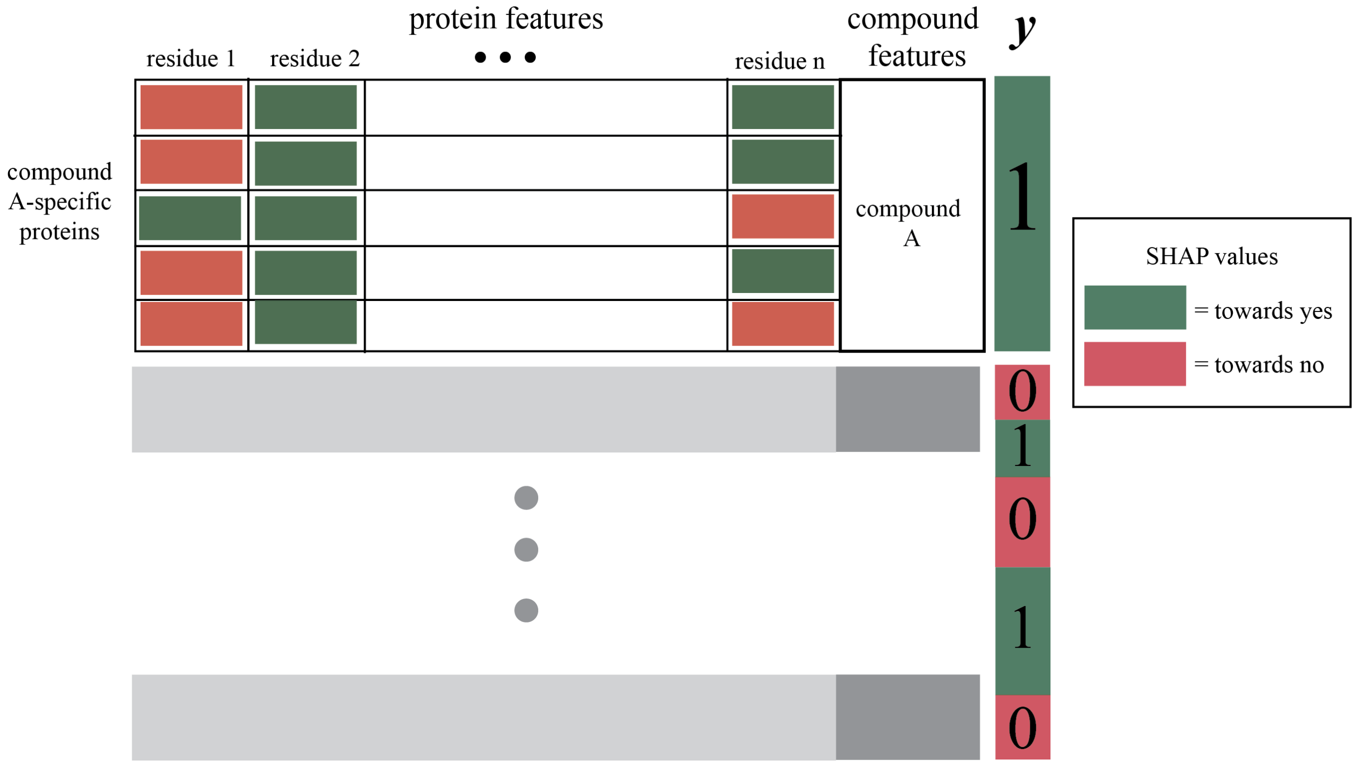


Figure 7.3: Obtaining predictive residue scores for a specific compound (compound A) using SHAP scores from a trained Matchmaker model. Only SHAP scores of proteins producing compound A are considered, i.e. $y = 1$ and compound features correspond to compound A. The remaining SHAP scores (in grey) are ignored for compound A.

### 7.2.6 Interpretation and visualization

The trained classifier can give insights about residues involved in determining compound specificity. The overall residue predictiveness score is calculated by averaging Gradient boosting importance scores from the trained model across protein features for each residue.

To assess the predictive value of residues for each individual compound we use SHAP[38] which assigns a positive or negative value to each feature of each data point, determining each feature's influence on the prediction result for that data point. For a given compound, say compound A, we collect all protein-A pairs consisting of proteins producing compound A (i.e. $y = 1$) and add the SHAP scores across features for each residue, obtaining a score for each residue that indicates its predictive value in the context of compound A specificity. This is illustrated in Figure 7.3. These values are then further averaged across the collected pairs. Positive scores in the resulting array correspond to residues which push the prediction towards a higher compatibility score, which are thus most likely to be involved in the production of that compound.

Overall residue scores and compound-specific scores for four compounds are visualized on the tobacco aristolochene synthase structure (TEAS, PDB ID: 5EAT)[39] using PyMOL[40]. The size of the residue is proportional to the predictive score.

## 7.3 Results

### 7.3.1 Characterization of novel STSs

Out of the 160 enzymes selected as described in the Methods, we experimentally characterized 64 novel STSs across the STS sequence space. Of the remaining 96 genes, 10 could not be cloned or transformed into *Escherichia coli* and 86 did not produce detectable amounts of product. We use the 64 productive STSs as an independent test set. Figure 7.4 shows the distribution of sesquiterpenes across the STSs in the training set and test set.

Figure 7.5 depicts the sequence identities of the 302 STSs in the training set with the 160 selected genes using multi-dimensional scaling (MDS), a technique used to visualize the level of similarity of individual objects in a dataset such that the between-object distances are preserved as well as possible. Therefore, STSs appearing close in Figure 7.5 have a high sequence identity, while those further away have lower identity. STSs from the current database, novel STSs with successfully detected products, and STS candidates which could not be cloned or did not produce detectable amounts of products are labelled with different colours, with markers differentiating dicot, monocot and conifer STSs.

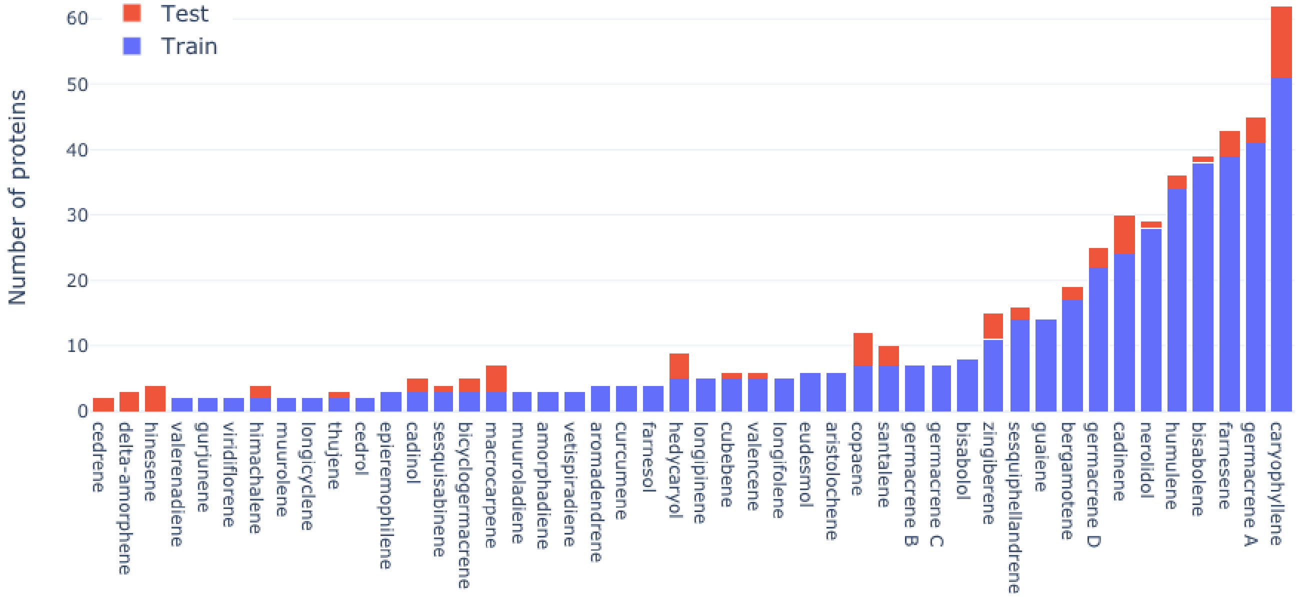


Figure 7.4: The number of STSs from the plant STS database training set and novel characterized STS test set producing certain sesquiterpenes. Sesquiterpenes produced by a single enzyme are not shown.

Since STSs producing the acyclic sesquiterpene nerolidol are numerous in our dataset, and many were shown in previous research to have easily detectable sequence and structural patterns[9,10], these would not be as useful in enhancing our knowledge of product specificity. Thus, we discarded sequences with protein features close to nerolidol STSs during our selection process. This is reflected back in the characterization results - in Figure 7.5 none of the novel enzymes are close to the cluster of nerolidol synthases (in green), and only one novel nerolidol synthase was detected (Figure 7.4). The test set also does not consist of any STSs from coniferous species, these were likely discarded from the selection process due to their longer than average sequence lengths and lower sequence identities compared to the remaining STSs, which lowers their corresponding HMM scores. After considering these exceptions, Figures 7.4 and 7.5 show that the novel STSs are evenly distributed across STS sequence and product space, indicating that the experimental characterization study successfully broadens our coverage of STSs and presents an independent and representative test set for evaluating product prediction.

More than half of the candidates (42/80) found by the first two selection techniques, based on sequence similarity to existing STSs irrespective of product, were productive enzymes but produced predominantly farnesyl cation-derived sesquiterpenes, with only two examples of nerolidyl cation-derived STSs. However, of the 80 candidates obtained from the third selection technique, which made use of our STS cation predictor[9] to select STSs with a chance of producing sesquiterpenes derived from the nerolidyl cation, 15 out of 22 productive enzymes were nerolidyl cation-derived STSs. This demonstrates that our cation prediction approach is accurate on productive enzymes and can be used for screening of enzymes from a desired cation, but it cannot differentiate between productive and unproductive STSs. It is difficult to pinpoint the exact reason for genes producing unproductive enzymes and no discernable clustering can be seen in Figure 7.5 differentiating productive from unproductive enzymes. One possibility is that these genes encode for monoterpene synthases (MTSs) instead of STSs. While we attempt to discard MTSs from our selection using the approach detailed by Terzyme[17], this relies on HMMs covering the entire sequence, and evidence of an MTS and STS sharing 97% sequence identity[21] demonstrates that this is not a foolproof strategy. Another possibility is the presence of sequencing and assembly errors, which are prevalent in such large-scale sequencing datasets and are not easy to correct. Thus, differentiating between productive and unproductive STSs is a prediction problem of its own right and could be attempted in future studies using data from experiments such as ours.

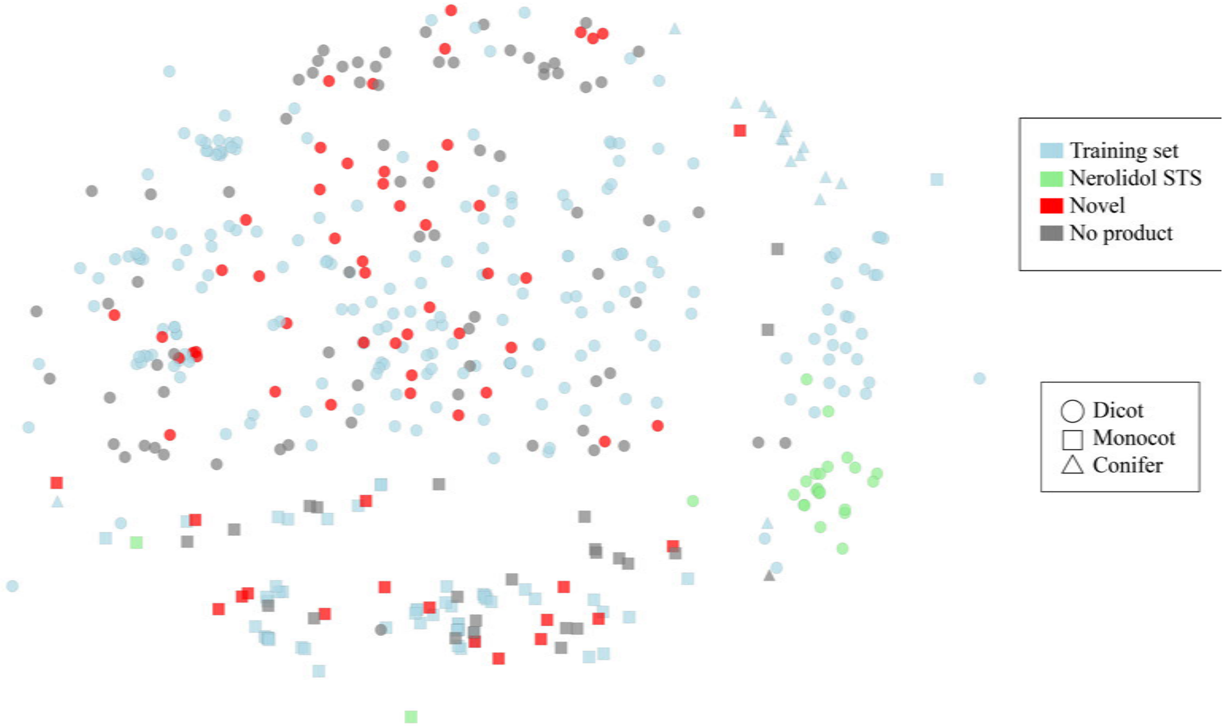


Figure 7.5: Multi-dimensional scaling plot of training STSs (in light blue) and novel characterized enzymes based on sequence identity. Training set STSs producing nerolidol are shown in light green instead. Proteins which could not be cloned or whose products could not be detected are labelled as "No product" (in grey), novel productive enzymes are in red. Markers differentiate dicot STSs (circles), monocot STSs (squares) and STSs from coniferous species (triangles).

| Predictor | Top | Top-3 | Top-5 |
|---|---|---|---|
| **Random** | 7% | 19% | 27% |
| **Sequence** | 30% | 38% | 48% |
| **Matchmaker** | 37% | 57% | 62% |

Table 7.2: Prediction results of random prediction, sequence identity-based prediction and Matchmaker. The columns indicate the percentage of STSs with correctly predicted major products when taking into account the topmost, top-3 and top-5 compounds with the highest compatibility scores.

### 7.3.2 Matchmaker predicts STS product specificity in different settings

Table 7.2 shows validation results for the protein-centric prediction setting, where the goal is to predict the correct product for a given enzyme, using random prediction (which selects products randomly based on their frequencies in the training data), sequence-identity based prediction, and Matchmaker prediction when considering the topmost, top-3 and top-5 predicted products. Matchmaker improves over sequence based prediction and achieves over 60% accuracy when considering the top 5 products. Given the sheer number of possible sesquiterpenes, Matchmaker prediction can greatly narrow down the possible products produced by any novel STS of interest.

In contrast to the prediction setting, some applications such as screening assays and directed evolution studies may require a selection of proteins that are likely to have a desired compound specificity. We assessed performance in such a compound-centric setting by checking if enzymes with high compatibility scores for a given compound are truly specific for that compound (see Methods for details). Matchmaker found such enzymes with a 40% probability in the top-10, compared to just 14% using sequence identity. Figure 7.6 shows the ROC curve and area under the curve measure for the three most populous sesquiterpenes in our training set, again demonstrating the superiority of Matchmaker prediction to sequence identity based screening. In this setting, Matchmaker can greatly reduce the number of characterization experiments needed to locate STSs producing a desired product, an application which could be particularly useful for diversifying the number of STSs making industrially valuable sesquiterpenes.

Both settings demonstrate that Matchmaker has a significant advantage over sequence identity-based approaches in biotechnological applications.

### 7.3.3 Incorporating minor products improves prediction

STSs, like many other natural product enzyme families involved in producing the wide array of specialized metabolites, can be highly promiscuous and often produce a number of products at once. In previous research, we observed that all the products of a promiscuous STS predominantly arise from the same chemical pathway[9]. This indicates that chemical similarity could hold information valuable for predicting specificity, possibly linked to active site shape and size constraints and compound

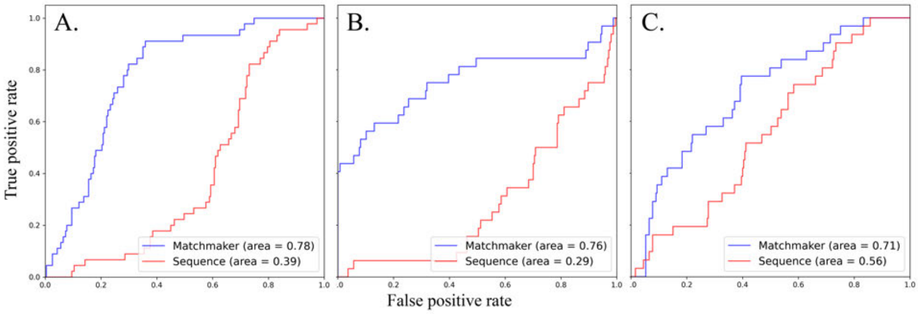


Figure 7.6: The Receiver Operating Characteristic (ROC) curves comparing Matchmaker compatibility scores and sequence identity for the three most common products in our training set - **A.** caryophyllene, **B.** germacrene A, **C.** farnesene

.

| **Predictor** | **Top** | **Top 3** | **Top 5** |
|---|---|---|---|
| **Sequence-Multi** | 34% | 48% | 59% |
| **Matchmaker** | 41% | 61% | 70% |
| **Matchmaker-Multi** | 43% | 63% | 71% |

Table 7.3: Prediction results of Matchmaker, Matchmaker-Multi and sequence identity-based prediction, with the latter two considering all products of a promiscuous STS as positives in the training process. The columns show the percentage of STSs with correctly predicted products when considering all products of a promiscuous STS as correct, looking at the topmost, top-3, and top-5 compounds with the highest compatibility scores.

positioning for enzymatic reactions. Such situations are easily handled by the Matchmaker framework by marking pairs between a protein and all its products as 1. Table 7.3 shows the performance obtained by adding minor or side products to our training process (labelled "-Multi") and evaluating any product produced by an STS as correct. When considering the top-5 predicted products, Matchmaker-Multi correctly predicts partial product specificity for over 71% of STSs. Thus, Matchmaker easily adapts to multi-compound specificity, and can be applied in situations where even trace amounts of sesquiterpene are of interest.

Such protein-centric multi-product prediction would also be useful in settings where all STSs from a specific plant species are under study. Since sesquiterpenes are volatile metabolites, volatile studies on plant parts such as flowers and stems give an indication of which compounds are produced by that species. This information, combined with Matchmaker prediction, can help produce highly accurate and comprehensive prediction reports linking each putative STS from a species to their respective products in the species' volatile profile.

| Predictor | Top | Top 3 | Top 5 |
|---|---|---|---|
| **Sequence** | 8 | 20 | 24 |
| **Matchmaker** | 16 | 22 | 27 |
| **Sequence-Multi** | 15 | 29 | 32 |
| **Matchmaker-Multi** | 19 | 34 | 39 |

Table 7.4: Prediction results on the test set of novel characterized enzymes. Sequence and Matchmaker only consider major products during training and evaluation, while Sequence-Multi and Matchmaker-Multi include all products of promiscuous STSs. The columns show the number of STSs with correctly predicted products (out of 64 novel STSs) when looking at the topmost, top-3, and top-5 compounds with the highest compatibility scores.

### 7.3.4 Matchmaker predicts specificity for novel STSs

Table 7.4 shows the prediction results for the 64 novel STSs producing sesquiterpenes present in the training set. Matchmaker shows clear improvement over sequence identity-based prediction, despite the fact that the distribution of sesquiterpenes in the test set differs from that in the training set. This indicates that Matchmaker is an accurate and useful tool for STS screening in biotechnological applications.

In future work, we intend to integrate the novel enzymes characterized in this study into our database of characterized plant STSs, greatly expanding our coverage of STS sequence space and allowing for more accurate predictors that generalize across diverse sequences and products. Similar to the manner in which we used the cation predictor to increase our representation of nerolidyl cation-derived products, the Matchmaker predictor could be used in an active learning setting[41], to select candidates producing relatively rarer sesquiterpenes or for which prediction scores are not as robust. Such an expanded database could further improve product specificity prediction while simultaneously reducing the number of experiments needed to be performed to obtain STSs with desired specificity.

### 7.3.5 Combinations of residues are predictive for different products

A major advantage of the framework presented here, where residue-level protein features are considered explicitly, lies in its interpretability. Firstly, the underlying Gradient Boosting classifier provides a score for each feature, indicating how useful or valuable it was in the construction of the boosted decision trees within the model. The more often a feature plays a role in making key decisions, the higher its relative importance score. These feature importances are depicted in Figure 7.7 with higher values indicated in red. In previous research, we performed cation specificity classification in STSs using a protein-centric approach, dividing sesquiterpenes into two groups based on their precursor cation[9]. As the cation specificity classifier also depended on gradient boosting trees, we performed a similar analysis of feature importance, finding that many of the 30 predictive residues found overlapped with residues that changed cation specificity in multiple mutation studies. Unsurprisingly,

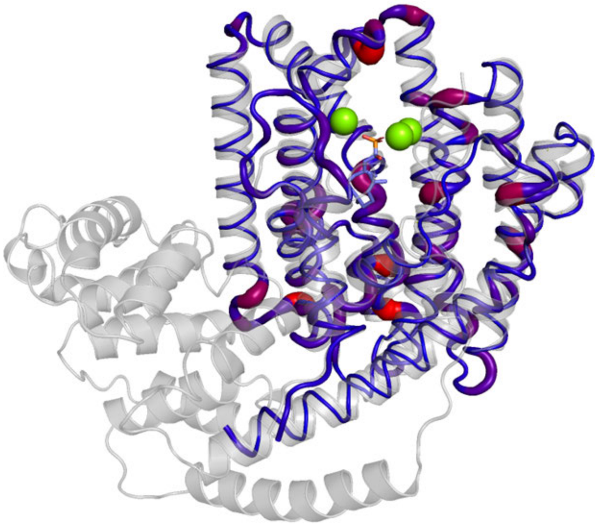


Figure 7.7: Gradient Boosting predictive scores per residue, as depicted on the tobacco aristolochene synthase structure. The size of the residue is proportional to its predictive score. The score is also depicted on a blue-red colour scale, with blue representing lower scores and red representing higher scores.

we now find 12 of these residues again as predictive for product specificity. In addition, product-predictive residues cluster near the sides and bottom of the active site cavity, where the bulk of rearrangement and modification reactions take place.

However, while inspecting Gradient Boosting importances is useful to obtain an overall picture of residues likely involved in compound specificity, the proteochemometric approach of Matchmaker calls for a more compound-specific interpretation. For this we turn to SHAP (SHapley Additive exPlanations) analysis on the trained model, based on the game theoretical concept of Shapley values[38]. The goal of SHAP is to explain a single prediction by computing the contribution of each feature to the prediction. In short, this is done with coalitional game theory, assuming each feature is a "player" in a coalition game with the prediction as a payout, and calculating how to fairly distribute the payout among the players. This gives us a straightforward approach to calculate predictive features per compound, simply by aggregating SHAP scores across positive pairs for that compound.

In Figure 7.8 we depict these aggregated SHAP scores for four prevalent compounds in our dataset - germacrene D (farnesyl cation, 10,1 cyclization in Figure 7.1), humulene (farnesyl cation 11,1 cyclization), bisabolene (nerolidyl cation, 6,1 cyclization), and nerolidol (nerolidyl cation acyclic), all deriving from distinct precursor cations. It is clear that combinations of different residue positions are involved in predictions, and that some positions are similar across compounds, but many are distinct. For example, the highly flexible H3-$\alpha$1 loop (in teal, Figure 7.8A, B, C, D), close to the catalytic NSE/DTE motif[42], is a common predictive feature across compounds. Given the inability to crystallize this region in three crystal structures, it is likely that structural models, and hence residue features, differ drastically here in concordance with compound specificity. The G2 helix kink (in yellow, Figure 7.8A, C, D) is predictive for all three cyclic compounds but not for the acyclic nerolidol, which matches our theoretical understanding of the underlying mechanism where carbonyl oxygens of the residues in this kink direct the cationic end of the substrate for further cyclization[39]. Residues in the A-C loop, which contains the RXR motif, are predictive only for nerolidol (in red, Figure 7.8B). Previous research has indeed shown that changes in conserved amino acids in this loop are observed predominantly in nerolidol synthases[10]. Similarly, the J-K loop has previously been implicated in 11,1 cyclization to form humulene[43], explaining its relative importance in our predictor (in brown, Figure 7.8C). The residues found predictive for bisabolene synthases (in dark blue, Figure 7.8D) match what we know from maize mutants about the importance of active site-adjacent residues in bisabolyl-cation derived products[44].

The interpretation of predictive residues could form the basis for designing mutation studies to change compound specificity, reduce promiscuity, or increase activity. While some residues have been mutated in the past with demonstrable effects, as mentioned above, inspecting SHAP scores for a prediction can help decide whether the same residues may be as crucial in STSs of a different species or specificity. In addition, residues which have not yet been mutated in any STS, perhaps due to lack of direct proximity to the substrate, form excellent candidates for future studies. Product specificity, activity, and promiscuity in these enzymes is a complex cocktail of far-reaching effects from residues both in and around the active site[45], something that is likely true for many other protein and enzyme families as well.

### 7.3.6 Matchmaker as a software library for proteochemometric specificity prediction

Our results on the STS enzyme family demonstrate the applicability of the Matchmaker framework to protein-compound compatibility prediction in small datasets without expensive procedures, and with advanced interpretation capabilities. The same approach can easily be adapted to other protein families and prediction tasks such as substrate specificity, drug or inhibitor binding etc. In many cases, compound specificity is not an exclusive feature of a protein and demonstrable specificity for one protein-compound pair does not preclude specificity of that protein for another compound. This is especially true in drug discovery applications, where proteins often successfully bind to multiple inhibitors or drug-like molecules; predicting off-target effects is an ongoing challenge in the field. We provide Matchmaker as a Python library at `https://git.wur.nl/durai001/matchmaker`, with modules for generating structural models and extracting features, downloading compounds as SMILES strings and generating molecular fingerprints, training and evaluating the prediction framework, and inspecting predictive residues per compound which can then be visualized in any molecular visualization software of choice.

## 7.4 Conclusion

We have presented a novel, interpretable machine learning framework for STS product specificity prediction, combining insights, concepts, and algorithms from previous research on STSs and structural bioinformatics. Matchmaker improves upon sequence

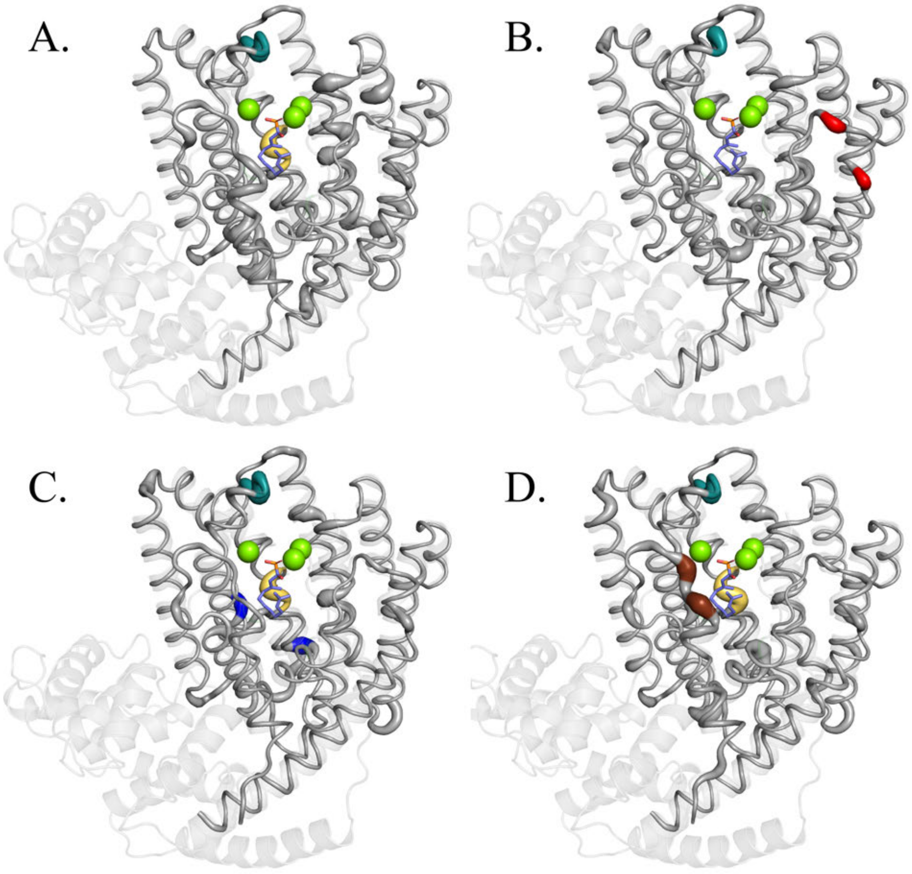


Figure 7.8: SHAP-based predictive scores per residue for four different sesquiterpenes, **A** germacrene D, **B** nerolidol, **C** bisabolene, and **D** humulene. The size of the residue is proportional to its predictive score. Coloured residues are discussed in the text - teal residues in A, B, C and D are in the H3-$\alpha$1 loop; yellow residues in A, C and D correspond to the G2 helix kink; red residues in B are in the A-C loop; dark blue residues in C were mutated in maize STSs by Köllner et al.[44]; and brown residues in D are in the J-K loop.

identity based screening, adapts to different prediction settings, and provides an attractive alternative for biotechnological applications. We experimentally characterized 64 novel STSs, and were able to accurately predict product specificity for 70% of these. We plan to use this data set in future work to improve prediction performance across diverse sequences and compounds. Inspecting the trained Matchmaker allows to locate residues responsible for the formation of specific sesquiterpenes. This level of advanced interpretation sets Matchmaker apart from other prediction methods, enabling the design of mutation studies and protein engineering experiments with input derived from across the entire STS family. Matchmaker is also available as a Python library, and could be applied to other protein families to obtain predictive insights on residues driving compound specificity. The universe of natural products comprises a huge diversity of enzymes and compounds, in plants and also in bacteria and fungi. Many of these are highly sought after for industrial bioproduction and engineering. Interpretable machine learning on sparsely characterized proteochemometric data, as espoused by Matchmaker, enables guided and efficient experimental studies and design.

## References

[1] Osbourn, A. E., & Lanzotti, V. (2009). *Plant-Derived Natural Products*. Springer.

[2] Firn, R. D., & Jones, C. G. (2003). Natural products–a simple model to explain chemical diversity. *Natural Product Reports*, *20*, 382–391.

[3] Kobe, B., Kampmann, T., Forwood, J. K., Listwan, P., & Brinkworth, R. I. (2005). Substrate specificity of protein kinases and computational prediction of substrates. *Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics*, *1754*, 200–209.

[4] Yao, J., Chu, Y., An, R., & Guo, H. (2012). Understanding product specificity of protein lysine methyltransferases from QM/MM molecular dynamics and free energy simulations: The effects of mutation on SET7/9 beyond the Tyr/Phe switch. *Journal of Chemical Information and Modeling*, *52*, 449–456.

[5] Tu, Y., Deshmukh, R., Sivaneri, M., & Szklarz, G. D. (2008). Application of molecular modeling for prediction of substrate specificity in cytochrome P450 1A2 mutants. *Drug Metabolism and Disposition*, *36*, 2371–2380.

[6] Brookes, J. C., Galigniana, M. D., Harker, A. H., Stoneham, A. M., & Vinson, G. P. (2012). System among the corticosteroids: Specificity and molecular dynamics. *Journal of The Royal Society Interface*, *9*, 43–53.

[7] Tian, B.-X., Wallrapp, F. H., Holiday, G. L., Chow, J.-Y., Babbitt, P. C., Poulter, C. D., & Jacobson, M. P. (2014). Predicting the functions and specificity of triterpenoid synthases: A mechanism-based multi-intermediate docking approach. *PLoS Computational Biology*, *10*, e1003874.

[8] O'Brien, T. E., Bertolani, S. J., Zhang, Y., Siegel, J. B., & Tantillo, D. J. (2018). Predicting productive binding modes for substrates and carbocation intermediates in terpene synthases —bornyl diphosphate synthase as a representative case. *ACS Catalysis*, *8*, 3322–3330.

[9] Durairaj, J., Melillo, E., Bouwmeester, H. J., Beekwilder, J., de Ridder, D., & van Dijk, A. D. J. (2021). Integrating structure-based machine learning and co-evolution to investigate specificity in plant sesquiterpene synthases. *PLoS Computational Biology*, *17*, e1008197.

[10] Durairaj, J., Di Girolamo, A., Bouwmeester, H. J., de Ridder, D., Beekwilder, J., & van Dijk, A. D. J. (2019). An analysis of characterized plant sesquiterpene synthases. *Phytochemistry*, *158*, 157–165.

[11] Playe, B., & Stoven, V. (2020). Evaluation of deep and shallow learning methods in chemogenomics for the prediction of drugs specificity. *Journal of Cheminformatics*, *12*, 11.

[12] van Westen, G. J., Swier, R. F., Cortes-Ciriano, I., Wegner, J. K., Overington, J. P., IJzerman, A. P., van Vlijmen, H. W., & Bender, A. (2013). Benchmarking of protein descriptor sets in

proteochemometric modeling (part 2): Modeling performance of 13 amino acid descriptor sets. *Journal of Cheminformatics*, *5*, 42.

[13] Boyles, F., Deane, C. M., & Morris, G. (2021). Learning from docked ligands: Ligand-based features rescue structure-based scoring functions when trained on docked poses, .

[14] Wójcikowski, M., Kukiełka, M., Stepniewska-Dziubinska, M. M., & Siedlecki, P. (2019). Development of a protein–ligand extended connectivity (PLEC) fingerprint and its application for binding affinity predictions. *Bioinformatics*, *35*, 1334–1341.

[15] Boeckmann, B., Bairoch, A., Apweiler, R., Blatter, M.-C., Estreicher, A., Gasteiger, E., Martin, M. J., Michoud, K., O'Donovan, C., Phan, I., Pilbout, S., & Schneider, M. (2003). The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Research*, *31*, 365–370.

[16] Matasci, N. et al. (2014). Data access for the 1,000 Plants (1KP) project. *GigaScience*, *3*.

[17] Priya, P., Yadav, A., Chand, J., & Yadav, G. (2018). Terzyme: A tool for identification and analysis of the plant terpenome. *Plant Methods*, *14*, 4.

[18] Degenhardt, J., Köllner, T. G., & Gershenzon, J. (2009). Monoterpene and sesquiterpene synthases and the origin of terpene skeletal diversity in plants. *Phytochemistry*, *70*, 1621–1637.

[19] Grote, A., Hiller, K., Scheer, M., Münch, R., Nörtemann, B., Hempel, D. C., & Jahn, D. (2005). JCat: A novel tool to adapt codon usage of a target gene to its potential expression host. *Nucleic Acids Research*, *33*, W526–W531.

[20] (). Fungal volatile compounds induce production of the secondary metabolite Sodorifen in *Serratia plymuthica*, .

[21] Di Girolamo, A., Durairaj, J., van Houwelingen, A., Verstappen, F., Bosch, D., Cankar, K., Bouwmeester, H., de Ridder, D., van Dijk, A. D. J., & Beekwilder, J. (2020). The santalene synthase from *Cinnamomum camphora*: Reconstruction of a sesquiterpene synthase from a monoterpene synthase. *Archives of Biochemistry and Biophysics*, *695*, 108647.

[22] Lemmon, E. W., Huber, M. L., & McLinden, M. O. (2007). NIST standard reference database 23: Reference fluid thermodynamic and transport properties-REFPROP, version 8.0, .

[23] Potter, S. C., Luciani, A., Eddy, S. R., Park, Y., Lopez, R., & Finn, R. D. (2018). HMMER web server: 2018 update. *Nucleic Acids Research*, *46*, W200–W204.

[24] Finn, R. D., Bateman, A., Clements, J., Coggill, P., Eberhardt, R. Y., Eddy, S. R., Heger, A., Hetherington, K., Holm, L., Mistry, J. et al. (2013). Pfam: The protein families database. *Nucleic Acids Research*, *42*, D222–D230.

[25] Wilbur, W. J., & Lipman, D. J. (1983). Rapid similarity searches of nucleic acid and protein data banks. *Proceedings of the National Academy of Sciences*, *80*, 726–730.

[26] Sievers, F., Wilm, A., Dineen, D., Gibson, T. J., Karplus, K., Li, W., Lopez, R., McWilliam, H., Remmert, M., Söding, J., Thompson, J. D., & Higgins, D. G. (2011). Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Molecular Systems Biology*, *7*, 539.

[27] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V. et al. (2011). Scikit-Learn: Machine learning in Python. *the Journal of Machine Learning Research*, *12*, 2825–2830.

[28] Webb, B., & Sali, A. (2014). Comparative protein structure modeling using MODELLER. *Current Protocols in Bioinformatics*, *47*, 5–6.

[29] Bernstein, F. C., Koetzle, T. F., Williams, G. J., Meyer, E. F., Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T., & Tasumi, M. (1977). The Protein Data Bank: A computer-based archival file for macromolecular structures. *Journal of Molecular Biology*, *112*, 535–542.

[30] Durairaj, J., Akdel, M., de Ridder, D., & van Dijk, A. D. J. (2020). Geometricus represents protein structures as shape-mers derived from moment invariants. *Bioinformatics*, *36*, i718–i725.

[31] Flusser, J., Suk, T., & Zitova, B. (2016). *2D and 3D Image Analysis by Moments*. John Wiley & Sons.

[32] Durairaj, J., Akdel, M., de Ridder, D., & van Dijk, A. D. (2021). Fast and adaptive protein structure representations for machine learning. *bioRxiv*.

[33] Kim, S. et al. (2016). PubChem substance and compound databases. *Nucleic Acids Research*, *44*, D1202–D1213.

[34] Kim, S., Thiessen, P. A., Bolton, E. E., & Bryant, S. H. (2015). PUG-SOAP and PUG-REST: Web services for programmatic access to chemical information in PubChem. *Nucleic Acids Research*, *43*, W605–W611.

[35] Rogers, D., & Hahn, M. (2010). Extended-connectivity fingerprints. *Journal of Chemical Information and Modeling*, *50*, 742–754.

[36] Landrum, G. et al. (2006). RDKit: Open-source cheminformatics, . URL: `http://www.rdkit.org/`.

[37] Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* KDD '16 (pp. 785–794). Association for Computing Machinery.

[38] Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems 30* (pp. 4765–4774). Curran Associates, Inc.

[39] Starks, C. M., Back, K., Chappell, J., & Noel, J. P. (1997). Structural basis for cyclic terpene biosynthesis by tobacco 5-*epi*-aristolochene synthase. *Science*, *277*, 1815–1820.

[40] DeLano, W. L. et al. (2002). PyMOL: An open-source molecular graphics tool. *CCP4 Newsletter on Protein Crystallography*, *40*, 82–92.

[41] Settles, B. (2009). *Active Learning Literature Survey*. Technical Report, University of Wisconsin-Madison Department of Computer Sciences.

[42] Christianson, D. W. (2006). Structural biology and chemistry of the terpenoid cyclases. *Chemical Reviews*, *106*, 3412–3442.

[43] Davis, E. M., & Croteau, R. (). Cyclization enzymes in the biosynthesis of monoterpenes, sesquiterpenes, and diterpenes. In *Biosynthesis: Aromatic Polyketides, Isoprenoids, Alkaloids*.

[44] Köllner, T. G., Degenhardt, J., & Gershenzon, J. (2020). The product specificities of maize terpene synthases TPS4 and TPS10 are determined both by active site amino acids and residues adjacent to the active site. *Plants*, *9*, 552.

[45] Koo, H. J., Vickery, C. R., Xu, Y., Louie, G. V., O'Maille, P. E., Bowman, M., Nartey, C. M., Burkart, M. D., & Noel, J. P. (2016). Biosynthetic potential of sesquiterpene synthases: Product profiles of Egyptian Henbane premnaspirodiene synthase and related mutants. *The Journal of Antibiotics*, *69*, 524–533.

CHAPTER 8

# The *Turterra* web portal for protein family data visualization and analysis

Janani Durairaj*, Barbara Terlouw*, Nico Louwen*, Dick de Ridder, Marnix Medema, and Aalt D.J. van Dijk

* authors contributed equally

# Abstract

Scientists who study individual protein families often perform a wide range of bioinformatics analyses using an assortment of different tools. There is mostly limited compatibility between tools, and without programming skills it is difficult and time-consuming to integrate and comprehensively investigate the different outputs, especially for large datasets. Here, we present Turterra: an accessible and comprehensive analysis portal for protein families. Turterra automatically constructs a web-portal from user-provided data, interactively visualising multiple sequence alignments, phylogenetic trees, protein structures and chemical substrates/products side by side. In this portal, data can be filtered by user-defined categories such as accession, species or compound specificity, so that the user can easily visualise relevant subsets of data. Turterra also provides the option to build multiple sequence and structure alignments, phylogenetic trees and homology models from scratch. Once the portal has been built, new sequences can be uploaded by end-users and compared to existing datasets, making Turterra the perfect engine for quick analysis of multifaceted data and for rapid deployment of protein portals. This will accelerate protein family research and facilitate collaboration between researchers working on the same families.

Turterra code and documentation are available at `https://git.wur.nl/durai001/turterra`.

## 8.1 Introduction

With a wealth of biological data published each day, bioinformatics analyses have become standard practice for researchers in the fields of biology and biochemistry. Novel protein sequences can be instantly compared to large databases such as Genbank[1], SWISS-PROT[2], Pfam[3] and PDB[4], giving scientists direct insight into potential biological functions of their sequences. Once sufficient related proteins have been collected, a wide variety of analyses are routinely performed, including the construction of multiple sequence alignments, phylogenetic trees and structural homology models to approximate 3D architecture. Such family-level analyses are particularly important for enzymes, which often share conserved folds as well as variable parts linked to their activity and specificity. Understanding and visualising patterns of conservation and variability is essential for grasping how different components relate to the various aspects of an enzyme's function.

Resources such as the tools published by EMBL-EBI[5] have definitely made large-scale sequence analysis more accessible to non-bioinformaticians. Still, life scientists have to employ a small army of different tools to analyse their protein data, including tree builders[6,7], homology modellers[8,9], and sequence aligners[10,11]. It can be an arduous and time-consuming task to integrate different, not always compatible outputs of independent tools into comprehensive analyses that capture the essence of a research question. Also, it is often useful to look only at subsets of a dataset at a given time, or to add a new data point to an existing analysis without having to redo the entire analysis for the existing data entries. This is especially true for large datasets, which due to sheer scale are usually difficult to visualise in a digestible manner. Even for experienced bioinformaticians it can take considerable time to integrate and summarise different tool outputs into an intelligible format.

To decrease the time spent by bioinformaticians and molecular biologists on integrating different data outputs, several integrative analysis packages for protein families have already been developed. These include among others JalView 2[12], a tool that specialises in multiple sequence alignment analysis which can link to external web services like PDBe[13] for 3D structure visualisation and ViennaRNA[14] for RNA secondary structure prediction; and Zebra3D[15], which focuses on 3D analysis of protein homologs by 3D similarity search and structure alignment. While these tools are great at integrating and visualising the data types they were designed to handle, each has their shortcomings. For instance, Jalview 2 does not provide the option to build or analyse computationally modelled structures alongside structures in the PDB database, which would be a valuable addition given the increased accuracy of computational modelling techniques. While Zebra3D does allow for the analysis of modelled structures, it lacks the interactivity that tools such as JalView 2 possess. Also, the analyses produced by these tools are not easily filterable on metadata. The inclusion of such an option would enable the targeted study or direct comparison of protein subsets based on auxiliary information such as mutation and variant studies, substrate specificity, inhibitor binding, and phenotypes at various temperature and pH conditions. Finally, the analyses performed by existing tools are not easily shareable, as most tools were designed for individual use and not targeted at communities of researchers who might work on the same protein (sub-)families. A multifaceted data analysis platform that allows users to easily publish their data to the web would truly elevate integrative analyses to the next level: researchers would not have to waste time performing computationally expensive analyses that have already been done, and could easily contribute to existing efforts that aim to understand their protein family of interest. Consequently, protein family information could be coherently stored in a single place for the benefit of research communities.

To provide the scientific community with a shareable platform for state-of-the-art multifaceted data analysis, we have developed Turterra: an accessible and comprehensive analysis portal for protein families. Here, we present a detailed overview of Turterra's features and demonstrate its versatility through two examples: the family of sesquiterpene synthase (STS) enzymes, which catalyse a single substrate into hundreds of 15-carbon sesquiterpene molecules, which give many plants and their fruits their smell[16]; and the family of non-ribosomal peptide synthetase (NRPS) adenylation domains, which govern the composition of microbial non-ribosomal peptides by specifically binding amino acid substrates in their active site[17].

## 8.2 Methods and implementation

Turterra contains two main executable scripts: Turterra-build and Turterra. Both scripts and their dependencies were written in Python (v3.9)[18].

### 8.2.1 Turterra-build

Turterra-build creates the data used for visualisation and analysis by Turterra. If Turterra-build is run without arguments, it will only create the required folder architecture, which can then be manually populated by the user with the appropriate files. However, the user can also specify if they want Turterra-build to create homology modelled structures, a multiple sequence alignment, multiple structure alignment, profile HMM, phylogenetic tree, or any subset of the above. Multiple sequence alignments are created with MUSCLE (v3.8, default settings)[11], structure alignments with Caretta-shape (v1.0, default settings)[19], profile HMMs with HMMER (v3.3.2, default settings)[20], phylogenetic trees with FastTree (v2.1.10, default settings)[21], and homology models with MODELLER (v10.0, default: 250 models per sequence, no loop refinement)[9]. The arguments and settings for each tool can be configured separately with a YAML-formatted configuration file. From each set of generated homology models, the model with the lowest normalised DOPE score is selected as the best model, and used for analysis and visualisation purposes.

### 8.2.2 Turterra

Turterra constructs and runs the layout of the main web portal with the Python package Dash (v1.15.0)[22]. Apart from the built-in Dash components used for buttons, dropdown lists, and uploading files, the web portal contains five Dash components that handle data analysis: SequenceViewer, AlignmentChart, Molecule3DViewer, and Molecule2DViewer from dash-bio (v0.4.8)[23], and Cytoscape from dash-cytoscape (v0.2.0)[24]. These visualise protein sequences; multiple sequence alignments (sequence-based or structure-based); (homology modelled) protein structures in 3D; substrate and/or product structures in 2D; and phylogenetic trees, respectively. A tab-separated file is provided by the user with Accession, Species, and Compounds as required columns defining each protein's accession, the species it is obtained from, and the compound (ligand/product) specificity. All other columns in the file are considered as extra data and are used in the portal to filter data. The dash-extensions package[25] is used to allow downloading filtered subsets of data in each panel via its Download component, and to facilitate server-side filesystem storage of the data used by the application via the ServersideOutput component - this allows storing large datasets of proteins, alignments, models etc. without burdening the user's browser with intensive data transfer operations. Compound chemical structures are parsed from SMILES format using the cheminformatics kit PIKAChU (unpublished software). ProDy (v2.0)[26] is used to parse protein structures from PDB files, combined with the dash-bio-utils (v0.0.6)[27] package to convert each structure into the format required for visualisation by Molecule3DViewer.

The portal also includes a component that allows the user to upload new sequences and/or associated structures. New sequences are appended to the existing sequence-based sequence alignment using MUSCLE (v3.8)[11]. Similarly, Caretta (v1.0)[28] rapidly aligns new structures to the consensus structure computed from the original set of 3D (homology modelled) structures, to then generate a new consensus structure. Uploaded sequences are appended to the existing phylogenetic tree using the phylogenetic placement tool epa-ng (v0.3.8, default settings, model: JTT)[29], and the resulting JPLACE output is parsed and converted to Newick format.

The various components and panels are interconnected by means of a series of callback functions, with different buttons and selections acting as triggers to activate changes in other entities, resulting in a highly networked application integrating the different views of data. Figure 8.1 depicts this as a flowchart, with arrows representing the relationship between triggers and their corresponding outputs.

### 8.2.3 Extending Turterra

To enable developers to easily add and interlink new custom panels, we have built Turterra in an extensible and modular fashion and provide tools and resources to quickly on-board developers to the code base. One such tool is a script to generate flowcharts using Mermaid diagram syntax (described at `https://mermaid-js.github.io`), depicting the flow of information from each component to the other. Figure 8.1 shows this chart generated for the current portal. This can be used as a reference when drafting a new component or panel, to pinpoint other components that may act as inputs, triggers, or outputs to the new one. With the predefined variable naming scheme described in our documentation, developers can get a birds-eye view of their components labelled and styled according to their component type and the panel they are placed in, allowing for easy debugging of highly interconnected code. In addition, we have extensive documentation aimed at Python developers new to the Dash library or to GUI programming in general.

### 8.2.4 Data preparation

For this paper, we assembled two datasets to visualise in Turterra: a dataset of 302 STS enzymes from the characterized plant STS database (`https://bioinformatics.nl/sesquiterpene/synthasedb`)[30], and a dataset of 1,093 NRPS adenylation domains (in-house data). For each accession, the datasets included an amino acid sequence in FASTA format, a structural (homology) model in PDB format, and the chemical structure of the enzyme's or domain's product or substrate respectively in SMILES format. Adenylation domain sequences were trimmed to only contain the N-terminal domain, as the C-terminal domain was too flexible and variable to obtain high-quality homology models. The STS homology models cover only the C-terminal domains of their sequences.

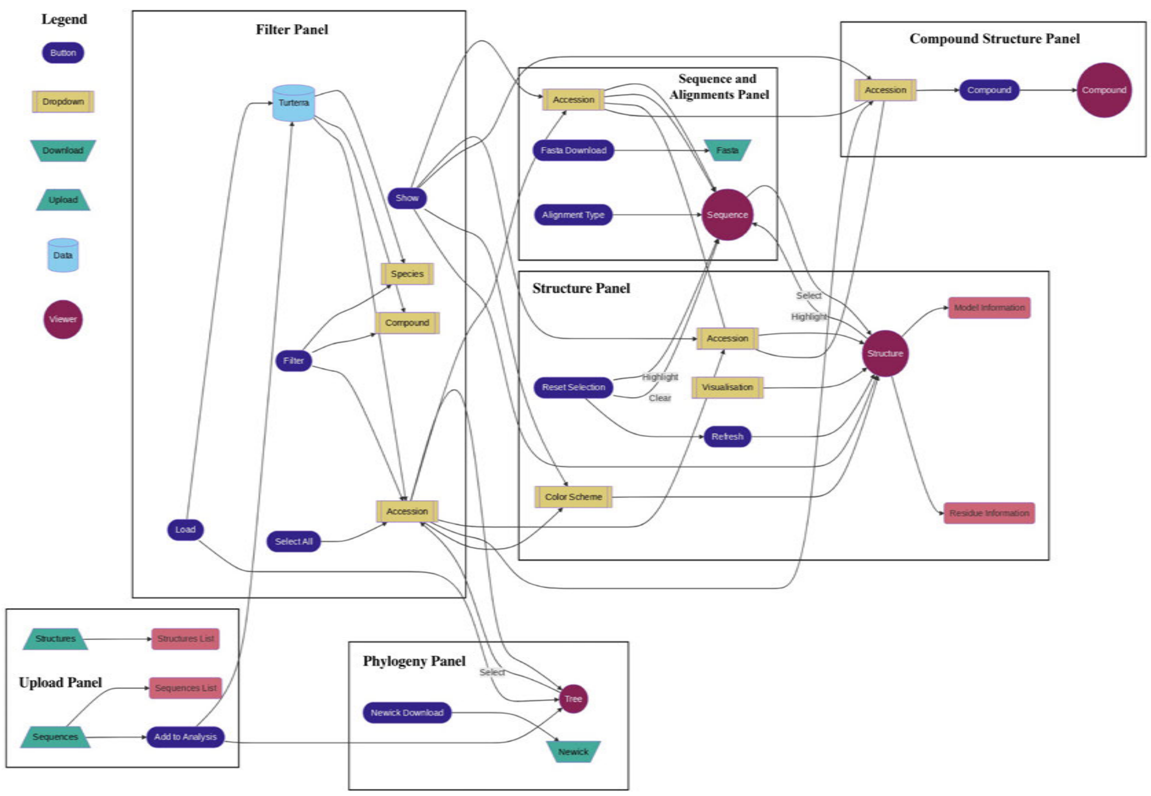


Figure 8.1: The auto-generated flowchart of callback functions between various buttons, dropdowns, viewers and other components on the Turterra website. An arrow from one component to another implies that a change in the first component acts as a trigger to change the second component.

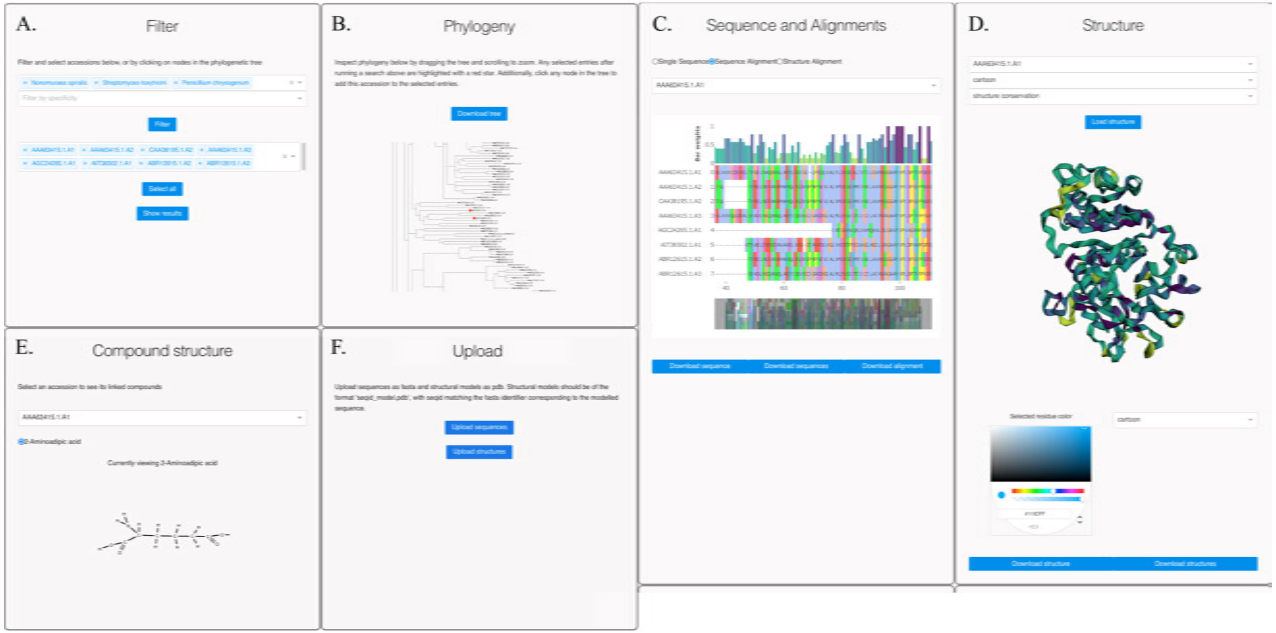
## 8.3 Results and Discussion

### 8.3.1 Turterra: an easy-to-use portal for protein family analysis

For researchers interested in a quick initial assessment of their dataset and molecular biologists less experienced with bioinformatics tools, Turterra-build provides the option to build a multiple sequence alignment, structure alignment, phylogenetic tree and homology models, or a subset of these, from scratch. The user can provide as little as a FASTA file containing the sequences of interest, and a folder of PDB files to be used as templates for homology modelling, from which turterra-build will create all the files that are required for the construction of the web portal. Later, any individual files can be replaced by user-provided versions. Required formats are described in-depth in the Turterra manual, making it straightforward for researchers of any discipline to provide the necessary files.

From this user-provided or Turterra-generated data, Turterra builds a comprehensive web-portal providing different views into the data. Datasets can be easily filtered according to user-defined categories and phylogeny, allowing for straightforward analysis and visualisation of relevant protein subsets. After construction of the portal, new sequences and corresponding (modelled) structures can be uploaded and compared to the pre-existing dataset. In settings where a researchers' Turterra portal is shared with others, the upload panel allows each individual viewer to independently compare their own data to the dataset shared by the portal. This unique combination of features makes Turterra the perfect engine for quick analysis of multifaceted biological data and rapid portal building for publication to the web.

To showcase Turterra's functionalities, we constructed Turterra web portals for two example enzyme families: a dataset of 302 STS enzymes, and a dataset of 1,093 NRPS adenylation domains, at `https://bioinformatics.nl/turterra`. STSs are a large family of plant enzymes responsible for the synthesis of a large variety of sesquiterpenes: plant natural products that help give plants and fruits their distinctive smell[16]. Previous research has shown that sequence similarity in these enzymes is explained more by phylogeny than similarity in product specificity[30]. However, structural information has been successfully used to group these enzymes by precursor cation specificity[31], thus indicating that researchers studying STSs would benefit from an integrated appraisal of sequence, phylogeny, and structure, enabled by Turterra. Like STSs, NRPS adenylation domains are also involved in the biosynthesis of natural products. They are found in both bacteria and fungi, and are core components of the much larger modular macro-enzymes called NRPSs[32]. These enzymes produce peptide scaffolds, the composition of which is determined by the specificity of NRPS adenylation domains for certain amino acids. Subtle differences in sequence and structure of these otherwise highly similar domains result in the recognition of over 100 different amino acid substrates[33,34], making NRPS adenylation domains very suitable for the multifaceted analysis Turterra provides. We use these two portals to describe Turterra's functionalities and performance below.



Filter and select accessions below, or by clicking on nodes in the phylogenetic tree

Inspect phylogeny below by dragging the tree and scrolling to zoom. Any selected entries after running a search above are highlighted with a red star. Additionally, click any node in the tree to add this accession to the selected entries.

Select an accession to see its linked compounds

Currently viewing 2-Aminoadipic acid

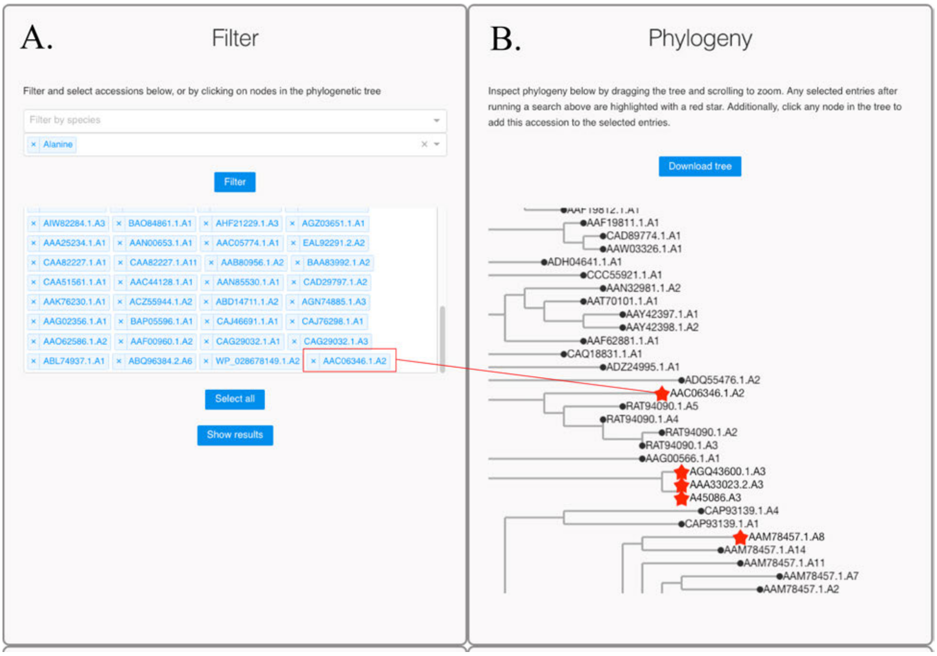
Upload sequences as fasta and structural models as pdb. Structural models should be of the format 'seqid_model.pdb', with seqid matching the fasta identifier corresponding to the modelled sequence.

Figure 8.2: The panels of Turterra web-portal built for NRPS adenylation domains. **A.** The Filter panel displays options to select proteins by various user-defined criteria. The resulting set of accessions are then available throughout the remaining panels. **B.** The Phylogeny panel displays a zoomable phylogenetic tree of all proteins, with selected accessions highlighted. Clicking on an unselected node in this tree adds the corresponding accession to the selection in the Filter panel. **C.** The Sequence and Alignments panel can show an individual searchable protein sequence, a sequence-based sequence alignment (as shown), or a structure-based sequence alignment, based on the radio button selected. The alignments are interactive, allow zooming into specific regions, and display the level of conservation at each position on top. **D.** The Structure panel displays a zoomable, interactive, and customizable view of the 3D protein structure. **E.** The Compound Structure panel displays substrate/product chemical structures of selected accessions. **F** The Upload panel allows the end-user to integrate and compare new sequences and structures with the existing dataset.

### 8.3.2 Turterra in action

Figure 8.2 depicts the six interconnected analysis panels in Turterra, described in detail below.

After loading the user-provided dataset, Turterra has two panels useful for filtering subsets of proteins - the Filter panel (Figure 8.2A) and the Phylogeny panel (Figure 8.2B). The former provides filtering options based on user-defined categories such as species and compound specificity. The latter depicts a phylogenetic tree of all proteins - protein subsets can be defined, expanded or narrowed by selecting or de-selecting clades or single enzymes in the phylogeny panel, allowing for inspecting similarities and differences between phylogenetically related proteins. These two panels are linked: selecting accessions via either panel updates the other as shown in Figure 8.3. For example, the accession outlined in red in Figure 8.3A was selected by clicking on the corresponding node in the Phylogeny panel in Figure 8.3B.



Filter and select accessions below, or by clicking on nodes in the phylogenetic tree

Filter by species

× Alanine

× AIW82284.1.A3 × BAO84861.1.A1 × AHF21229.1.A3 × AGZ03651.1.A1
× AAA25234.1.A1 × AAN00653.1.A1 × AAC05774.1.A1 × EAL92291.2.A2
× CAA82227.1.A1 × CAA82227.1.A11 × AAB80956.1.A2 × BAA83992.1.A2
× CAA51561.1.A1 × AAC44128.1.A1 × AAN85530.1.A1 × CAD29797.1.A2
× AAK76230.1.A1 × ACZ55944.1.A2 × ABD14711.1.A2 × AGN74885.1.A3
× AAG02356.1.A1 × BAP05596.1.A1 × CAJ46691.1.A1 × CAJ76298.1.A1
× AAO62586.1.A2 × AAF00960.1.A2 × CAG29032.1.A1 × CAG29032.1.A3
× ABL74937.1.A1 × ABQ96384.2.A6 × WP_028678149.1.A2 × AAC06346.1.A2

Inspect phylogeny below by dragging the tree and scrolling to zoom. Any selected entries after running a search above are highlighted with a red star. Additionally, click any node in the tree to add this accession to the selected entries.

Figure 8.3: **A** the Filter panel and **B** the Phylogeny panel are interlinked - accessions selected in one reflect in the other, as names in the Filter panel and with red stars in the Phylogeny panel. The accession outlined in red in A has been added by clicking on the corresponding node in B. The phylogeny tree in B can be exported as a Newick file using the Download button.

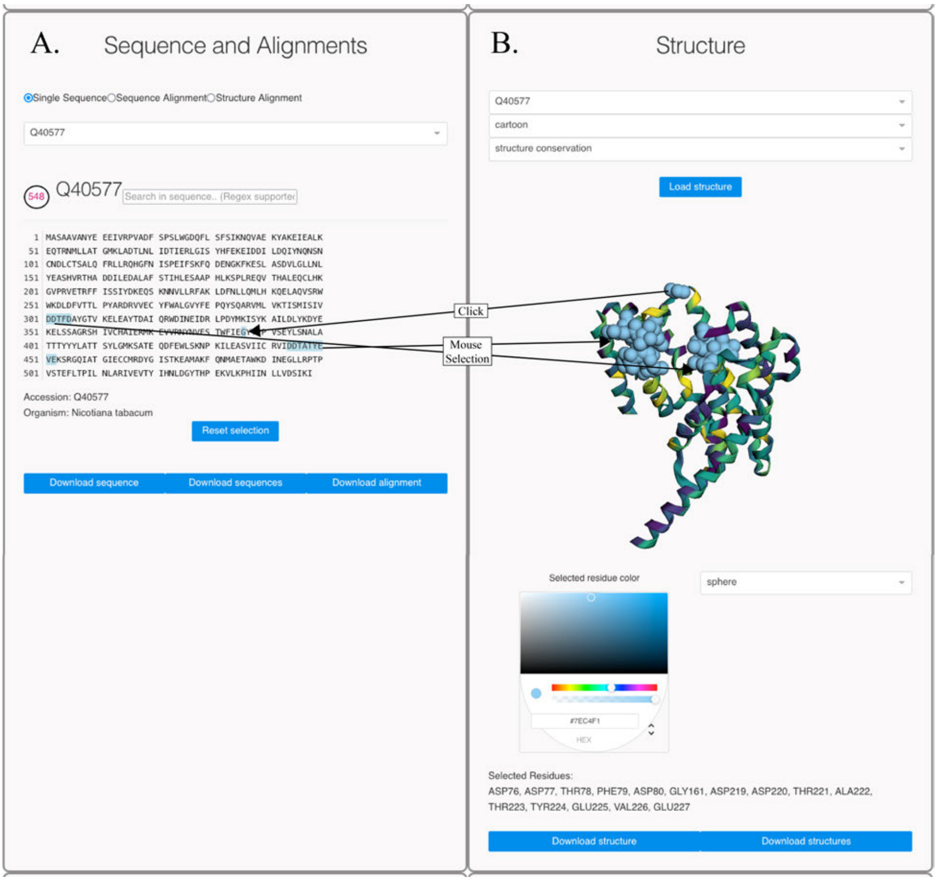
Once a set of accessions are selected, the next three panels allow visualisation of their sequences, 3D structures, and the compounds that they produce or modify. These are, respectively, the Sequence and Alignments panel (Figure 8.2C), the Structure panel (Figure 8.2D), and the Compound Structure panel (Figure 8.2E).

The Sequence and Alignments panel can display an individual sequence (single sequence mode), a sequence-based sequence alignment (sequence alignment mode) or a structure-based sequence alignment (structure alignment mode), controlled by a set of radio buttons. The alignments are interactive, zoomable, and scrollable, with conservation at each position represented as a bar on top of the alignment. The Structure panel displays an interactive 3D protein structure, and has multiple visualisation styles and colour schemes to choose from, described in the online documentation. In addition, the Sequence panel (in single sequence mode) and the Structure panel are interlinked as shown in Figure 8.4: the residues highlighted in the sequence correspond to those highlighted and labelled in the Structure panel. These highlights can be defined in either panel through mouse selection and interactive clicking respectively, and selected residues can be visualised differently. This is especially useful to inspect the structural context of residues deemed as important from literature or from conservation analysis, and to map the sequence context of influential residues in the structure, determined by mutational studies or by analysis of crystal structures. For example, Figure 8.4 depicts the STS from tobacco producing the sesquiterpene 5-*epi*-aristolochene[35]. The two stretches of sequence selected with the mouse in Figure 8.4A correspond to two known STS catalytic motifs[16] which are seen to surround the active site cavity in the structure in Figure 8.4B. The residue selected by clicking in the structure in Figure 8.4B caps this active site cavity, and thus is interesting to map back to its sequence and alignment context.

The Compound panel displays the two-dimensional structure of compounds associated with a protein, with an option to select one if multiple such compounds exist. The protein visualised across the Sequence and Alignments, Structure, and Compound Structure panels is the same and can be chosen using the accession selection dropdowns in any of these panels.

Finally, the Upload panel (Figure 8.2F) allows end-users to compare their novel proteins to the proteins in the loaded dataset. Users can upload sequences and optionally structures of their proteins and have them integrated into the existing phylogenetic tree, sequence alignment, and structure alignment without recalculation for the whole dataset. This enables comparison of putative or newly characterized proteins, and mutants or variant sequences with previously characterized proteins and their existing literature - especially useful in collaborative studies with researchers working on different subsets or variants of the same protein family.

Turterra gracefully handles missing or incomplete data: proteins without associated structures or compounds can still be analysed from a sequence and phylogeny perspective, and incomplete structural models with missing residues are still correctly mapped to the corresponding sequence. In addition, both filtered and expanded data can be exported and downloaded by end-users: the phylogeny tree in Newick format, sequences and alignments as FASTA files, superposed structures as PDB files, and chemical compounds as SMILES strings.



●Single Sequence ○Sequence Alignment ○Structure Alignment

Q40577

548 Q40577 Search in sequence.. (Regex supported

```
  1 MASAAVANYE EEIVRPVADF SPSLWGDQFL SFSIKNQVAE KYAKEIEALK
 51 EQTRNMLLAT GMKLADTLNL IDTIERLGIS YHFEKEIDDI LDQIYNQNSN
101 CNDLCTSALQ FRLLRQHGFN ISPEIFSKFQ DENGKFKESL ASDVLGLLNL
151 YEASHVRTHA DDILEDALAF STIHLESAAP HLKSPLREQV THALEQCLHK
201 GVPRVETRFF ISSIYDKEQS KNNVLLRFAK LDFNLLQMLH KQELAQVSRW
251 WKDLDFVTTL PYARDRVVEC YFWALGVYFE PQYSQARVML VKTISMISIV
301 DDTFDAYGTV KELEAYTDAI QRWDINEIDR LPDYMKISYK AILDLYKDYE
351 KELSSAGRSH IVCHAIERMK EVVRNYNVES TWFIEGY[illegible]P VSEYLSNALA
401 TTTYYYLATT SYLGMKSATE QDFEWLSKNP KILEASVIIC RVIDDTATYE
451 VEKSRGQIAT GIECCMRDYG ISTKEAMAKF QNMAETAWKD INEGLLRPTP
501 VSTEFLTPIL NLARIVEVTY IHNLDGYTHP EKVLKPHIIN LLVDSIKI
```

Accession: Q40577

Organism: Nicotiana tabacum

Q40577

cartoon

structure conservation

sphere

#7EC4F1 HEX

Selected Residues:
ASP76, ASP77, THR78, PHE79, ASP80, GLY161, ASP219, ASP220, THR221, ALA222, THR223, TYR224, GLU225, VAL226, GLU227

Figure 8.4: **A** the Sequence and Alignments panel in single sequence mode and **B** the Structure panel are interlinked - residues selected with the mouse in A ("Mouse Selection") are highlighted in B with residue numbers labelled in the list at the bottom, and residues clicked in B ("Click") are highlighted in A. The colour scheme and visualisation style of the structure can be changed using the provided dropdowns in B. The available colour schemes and styles are described in the documentation. Selected residues can be visualised differently from the rest of the structure using the provided colour picker and style selection dropdown - in the figure they are shown as light blue spheres. The three Download buttons in A allow exporting the sequence of the selected accession, the sequences of all filtered accessions, and the aligned sequences of filtered accessions respectively as FASTA files (with the radio buttons controlling whether the sequence alignment or structure alignment is exported). The two download buttons in B export the structure of the selected accession and the superposed structures of filtered accessions respectively as PDB files.

| Dataset | No. proteins | Avg. sequence length | Avg. model length | Creation | Loading | Response | Upload |
|---|---|---|---|---|---|---|---|
| **STS** | 302 | 551 ± 62 | 265 ± 3 | 7 min | 2s | ~0.1s | 4.2s |
| **NRPS** | 1093 | 475 ± 54 | 370 ± 22 | 25 min | 6s | ~0.3s | 37s |

Table 8.1: Portal creation, loading, response, and upload times for the STS and NRPS datasets, for which the average protein sequence length and structural model lengths are given. "Creation" is the time taken for generating sequence and structure alignments on a single thread. We don't include the model and tree generation times here as this is highly dependent on the program and settings used and hence would differ significantly depending on the user and use case. "Loading" is the amount of time taken upon pressing the "Load Data" button in the portal. "Response" is the time taken to register mouse click and selection events. "Upload" is the time taken for integrating a new sequence and structure into each portal using the Upload panel.

### 8.3.3 Performance and distribution

Table 8.1 depicts the data creation, loading, response and upload times for the STS and NRPS datasets. Since the bulk of data transfer for data shared across components is performed on the server side, the Turterra portal easily scales to datasets with thousands to tens of thousands of proteins. Portals which are published and made available to users through the web can be accessed simply via a URL and do not consume any space on the end-user's filesystem unless data is explicitly downloaded. Each user receives their own session key on the portal maintainer's side which is used to store their filtering and analysis options. These keys can be used to keep track of the time since a user's analysis to inform them of the results or to eventually clear their data once enough time has passed.

### 8.3.4 Opportunities for extension

The Turterra source code provides a good starting point for developing more customized panels holding information specific to certain proteins, use cases, or studies. For example, a mutation panel could connect experimentally solved or modelled structure mutants with their phenotypes and allow users to visualise these mutated residues in the structure and sequence panels. Specialized predictors of compound specificity could generate their own panels giving a detailed prediction report for each protein, and linking to predictive residues in other panels as well as predicted compounds in the compound panel. For multiple researchers working on a shared project, a Turterra portal could contain annotations and notes linked to the sequence, structure, compound, or phylogeny, allowing productive collaboration and easy sharing of results. With thorough documentation and the availability of helper scripts to auto-generate flowcharts of inputs and outputs, adding new panels and interlinking these with existing panels is straightforward for novice programmers as well. As usage becomes more widespread, we envision an open-source community of developers designing plug-and-play panels for Turterra users.

## 8.4 Conclusion

Turterra provides a comprehensive solution for highly interactive, multifaceted biological data analysis, allowing even researchers with limited bioinformatics experience to quickly analyse and filter their protein data in a single place. It is the ideal framework for in-house or web-based server publication, facilitating collaboration between researchers working on the same protein families. For more experienced users, Turterra's accessible source code makes the portal easily customisable and extendable to better fit specific protein families or to allow for integration into existing tools and databases. We expect that Turterra will drastically cut time spent on data analysis by researchers in all fields of protein biology, and will organically grow to suit the needs of the scientific community.

## References

[1] Benson, D. A., Cavanaugh, M., Clark, K., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., & Sayers, E. W. (2012). GenBank. *Nucleic Acids Research*, *41*, D36–D42.

[2] Boeckmann, B., Bairoch, A., Apweiler, R., Blatter, M.-C., Estreicher, A., Gasteiger, E., Martin, M. J., Michoud, K., O'Donovan, C., Phan, I., Pilbout, S., & Schneider, M. (2003). The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Research*, *31*, 365–370.

[3] Bateman, A. et al. (2004). The Pfam protein families database. *Nucleic Acids Research*, *32*, D138–D141.

[4] Burley, S. K., Berman, H. M., Kleywegt, G. J., Markley, J. L., Nakamura, H., & Velankar, S. (2017). Protein Data Bank (PDB): the single global macromolecular structure archive. *Protein Crystallography*, (pp. 627–641).

[5] Madeira, F., Park, Y. M., Lee, J., Buso, N., Gur, T., Madhusoodanan, N., Basutkar, P., Tivey, A. R., Potter, S. C., Finn, R. D. et al. (2019). The EMBL-EBI search and sequence analysis tools APIs in 2019. *Nucleic Acids Research*, *47*, W636–W641.

[6] Stamatakis, A. (2014). RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, *30*, 1312–1313.

[7] Huerta-Cepas, J., Serra, F., & Bork, P. (2016). ETE 3: Reconstruction, analysis, and visualization of phylogenomic data. *Molecular Biology and Evolution*, *33*, 1635–1638.

[8] Schwede, T., Kopp, J., Guex, N., & Peitsch, M. C. (2003). SWISS-MODEL: An automated protein homology-modeling server. *Nucleic Acids Research*, *31*, 3381–3385.

[9] Webb, B., & Sali, A. (2016). Comparative protein structure modeling using MODELLER. *Current Protocols in Bioinformatics*, *54*, 5–6.

[10] Sievers, F., Wilm, A., Dineen, D., Gibson, T. J., Karplus, K., Li, W., Lopez, R., McWilliam, H., Remmert, M., Söding, J., Thompson, J. D., & Higgins, D. G. (2011). Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Molecular Systems Biology*, *7*, 539.

[11] Edgar, R. C. (2004). MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, *32*, 1792–1797.

[12] Waterhouse, A. M., Procter, J. B., Martin, D. M., Clamp, M., & Barton, G. J. (2009). Jalview Version 2—a multiple sequence alignment editor and analysis workbench. *Bioinformatics*, *25*, 1189–1191.

[13] Velankar, S., Best, C., Beuth, B., Boutselakis, C., Cobley, N., Sousa Da Silva, A., Dimitropoulos, D., Golovin, A., Hirshberg, M., John, M. et al. (2010). PDBe: Protein data bank in Europe. *Nucleic Acids Research*, *38*, D308–D317.

[14] Lorenz, R., Bernhart, S. H., Zu Siederdissen, C. H., Tafer, H., Flamm, C., Stadler, P. F., & Hofacker, I. L. (2011). ViennaRNA package 2.0. *Algorithms for Molecular Biology*, *6*, 1–14.

[15] Timonina, D., Sharapova, Y., Švedas, V., & Suplatov, D. (2021). Bioinformatic analysis of subfamily-specific regions in 3D-structures of homologs to study functional diversity and

conformational plasticity in protein superfamilies. *Computational and Structural Biotechnology Journal*, *19*, 1302–1311.

[16] Degenhardt, J., Köllner, T. G., & Gershenzon, J. (2009). Monoterpene and sesquiterpene synthases and the origin of terpene skeletal diversity in plants. *Phytochemistry*, *70*, 1621–1637.

[17] Schwarzer, D., Finking, R., & Marahiel, M. A. (2003). Nonribosomal peptides: From genes to products. *Natural Product Reports*, *20*, 275–287.

[18] Van Rossum, G. et al. (2007). Python programming language. In *USENIX Annual Technical Conference* (p. 36). volume 41.

[19] Durairaj, J., Akdel, M., de Ridder, D., & van Dijk, A. D. (2021). Fast and adaptive protein structure representations for machine learning. *bioRxiv*.

[20] Finn, R. D., Clements, J., & Eddy, S. R. (2011). HMMER web server: Interactive sequence similarity searching. *Nucleic Acids Research*, *39*, W29–W37.

[21] Price, M. N., Dehal, P. S., & Arkin, A. P. (2010). FastTree 2–approximately maximum-likelihood trees for large alignments. *PLoS One*, *5*, e9490.

[22] Hossain, S., Calloway, C., Lippa, D., Niederhut, D., & Shupe, D. (2019). Visualization of bioinformatics data with Dash Bio. In *Proceedings of the 18th Python in Science Conference* (pp. 126–133).

[23] Dash-Bio: Dash components for bioinformatics. URL: `https://dash.plotly.com/dash-bio`.

[24] Dash-Cytoscape: A component library for Dash aimed at facilitating network visualization in Python, wrapped around Cytoscape.js. URL: `https://dash.plotly.com/cytoscape`.

[25] Dash-extensions: Extensions for Plotly Dash. URL: `https://github.com/thedirtyfew/dash-extensions/`.

[26] Bakan, A., Meireles, L. M., & Bahar, I. (2011). ProDy: Protein dynamics inferred from theory and experiments. *Bioinformatics*, *27*, 1575–1577.

[27] Dash-bio-utils: Simple parsing tools that supplement dash-bio. URL: `http://github.com/plotly/dash-bio-utils`.

[28] Akdel, M., Durairaj, J., de Ridder, D., & van Dijk, A. D. J. (2020). Caretta-a multiple protein structure alignment and feature extraction suite. *Computational and Structural Biotechnology Journal*, *18*, 981–992.

[29] Barbera, P., Kozlov, A. M., Czech, L., Morel, B., Darriba, D., Flouri, T., & Stamatakis, A. (2019). EPA-ng: Massively parallel evolutionary placement of genetic sequences. *Systematic Biology*, *68*, 365–369.

[30] Durairaj, J., Di Girolamo, A., Bouwmeester, H. J., de Ridder, D., Beekwilder, J., & van Dijk, A. D. J. (2019). An analysis of characterized plant sesquiterpene synthases. *Phytochemistry*, *158*, 157–165.

[31] Durairaj, J., Melillo, E., Bouwmeester, H. J., Beekwilder, J., de Ridder, D., & van Dijk, A. D. J. (2021). Integrating structure-based machine learning and co-evolution to investigate specificity in plant sesquiterpene synthases. *PLoS Computational Biology*, *17*, e1008197.

[32] Norn, C., Wicky, B. I., Juergens, D., Liu, S., Kim, D., Koepnick, B., Anishchenko, I., Baker, D., & Ovchinnikov, S. (2020). Protein sequence design by explicit energy landscape optimization. *bioRxiv*.

[33] Challis, G. L., Ravel, J., & Townsend, C. A. (2000). Predictive, structure-based model of amino acid recognition by nonribosomal peptide synthetase adenylation domains. *Chemistry & Biology*, *7*, 211–224.

[34] Stachelhaus, T., Mootz, H. D., & Marahiel, M. A. (1999). The specificity-conferring code of adenylation domains in nonribosomal peptide synthetases. *Chemistry & Biology*, *6*, 493–505.

[35] Starks, C. M., Back, K., Chappell, J., & Noel, J. P. (1997). Structural basis for cyclic terpene biosynthesis by tobacco 5-*epi*-aristolochene synthase. *Science*, *277*, 1815–1820.

# CHAPTER 9

# General discussion

The introduction to this thesis establishes the biotechnological opportunities presented by secondary metabolite (SM) enzyme families and puts forth a combination of structural bioinformatics and machine learning (ML) to take advantage of these opportunities. During the research described in this thesis, I developed several strategies along these lines. Here, I expand on further challenges posed by these enigmatic families and how these may be overcome in the future with innovation in ML techniques. Given recent advances in the field, I contemplate the future of structural bioinformatics and the new frontiers likely to define the next decade of computational efforts. Finally, I discuss aspects of protein structures that go beyond their rigid coordinates and aspects of proteins that go beyond their structures, both of which play a role in most biological phenomena.

## 9.1 The tangled web of natural product enzymes

The numerous approaches designed and explored in this thesis for product prediction in the single enzyme family of plant sesquiterpene synthases (STSs) already point to the sheer complexity and diversity shown by such families specialized for producing SMs. This complexity is further compounded by three factors, discussed below, which to various extents also translate to the broader field of protein bioinformatics.

### 9.1.1 Generalizing to novel protein and compound space

Many SM enzymes are promiscuous: they can accept multiple substrates, catalyse multiple kinds of reactions, or produce multiple SMs as products[1,2]. Promiscuous product specificity in the STS family is explored in **Chapter 3** and **Chapter 7**. Due to the size, shape, and properties of the active site, which confers substrate and reaction specificity, promiscuous product enzymes often have high chemical similarity among their products with the minor products typically derived from intermediates along the same reaction pathway as the major product. Conversely, however, there could be multiple different active site conformations resulting in similar product specificity. Both these aspects are crucial to consider given that the enzymes responsible for the majority of already known SMs have not been characterized, and no doubt the number of SMs and their biosynthetic enzymes yet to be identified far exceeds the number of known ones. This indicates that prediction approaches trained on sparse datasets of characterized SM enzymes to predict known specificity will miss out on, or mischaracterize, the huge untapped potential of novel SM enzymes and their products. We observed this in **Chapter 7** when our attempts to search for novel nerolidyl cation-derived enzymes (which were a minority in our dataset) led to many more non-productive enzymes than those from the majority class - indicating that predictions beyond the space of enzymes similar to those in our characterized dataset are not robust. Similarly, the predicted enzymes that we characterized mostly produced commonly occurring sesquiterpenes as products, suggesting that the identification of rare or novel sesquiterpene-producing enzymes is not yet solved.

One avenue to alleviate some of these issues is semi-supervised learning, which is the branch of ML concerned with using both labelled and unlabelled data for learning tasks[3]. Typically, semi-supervised learning algorithms attempt to improve performance in a classification task by utilizing information from unlabelled data points associated with labelled data points, thus increasing the robustness of the classifier to the data distribution. Though semi-supervised learning is not a new concept, recent approaches have successfully applied these algorithms to millions of unlabelled data points across a variety of protein learning tasks[4–10]. In the case of STSs, TPSs, and other SM enzymes, this unlabelled data could be in the form of the tens of thousands of uncharacterised proteins from sequenced genomes and transcriptomes annotated as belonging to the same enzyme family. It could also be in the form of naturally occurring SM compounds without associated enzymes, and even putative SM molecules with computationally predicted production paths consisting of reactions known to be catalysed by the SM enzyme family, but as yet unobserved in nature[11–14]. Such data could be additional input to predictors such as those in **Chapters 3 and 7** to both increase their robustness towards prediction of uncharacterised SM enzymes and their potential in predicting rare or novel SMs.

In addition, the field of ML in SM enzyme families is usually closely interlinked with experimental characterization studies. This facilitates opportunities where the results and inspection of ML models can be used to design or suggest novel experiments, enabling active learning. Active learning is a paradigm which iteratively selects the most significant unlabelled data points for experimental labelling, allowing for the greatest improvement in predictive performance at the lowest experimental cost[15]. Some common strategies for this selection include density-based selection, where data points from dense regions are selected, uncertainty-based selection which picks points that the current classifier struggles with most, representative-based selection which tries to find points most representative of the dataset, estimated-error reduction based selection which aims to reduce the classifier's maximal estimated error, and ensemble-based selection which combines multiple of the above criteria[16]. In the field of drug development, researchers have used existing knowledge (e.g., about signalling pathways), investigator insight and intuition to guide paths through experimental space, a process often hindered by incomplete or incorrect pathway information and the difficulty of making predictions about complex pathway interactions. Active learning has provided an attractive alternative to this, and has been used to characterize structurally diverse hits for virtual screening[17], to pinpoint compounds binding a target molecule in a minimum number of iterations of biochemical testing[18], and to predict off-target and unexplored biological responses[19] (see Reker & Schneider[20] for a review of active learning approaches in the drug discovery field). To aid our understanding of protein fitness, active learning via Gaussian processes has been used to select and experimentally characterize small sets of diverse proteins with differing thermostability, again with the aim of improving generalization of the stability predictor[21,22]. Also in applications such as rescue mutant prediction[23,24], protein-protein interaction prediction[25], and inhibitor structure prediction[26], active learning was shown to significantly reduce experimental cost for improved performance. We explored a simplified version of uncertainty-based active learning in this thesis by using the cation predictor in **Chapter 3** to select enzymes for characterization which were predicted to be in the minority class, in order to increase our representation of that class. These new training points will likely improve the generalization of the product predictor in **Chapter 7** to products derived from this class. More advanced active learning paradigms could similarly be used to select which STSs to characterize in order to improve prediction of STSs producing rare or novel sesquiterpenes, or to select relevant mutants to characterize as input to models predicting catalytic activity for STSs producing a specific product. While this aspect is not explored in this thesis, it presents a logical next step to computational STS research, as understanding determinants of activity could be crucial in engineering STSs to produce industrial levels of sesquiterpenes in demand. Thus, active learning could provide the path forward to further extend ML methods into the field of plant specialized metabolism without the need for huge datasets and large-scale experiments.

### 9.1.2 Generalizing across domains and distributions

The diversity of proteins across different species, conditions, and functional groups may present a significant challenge for prediction algorithms. In this thesis we observed the distinct separation of STSs from coniferous species compared to those from other plants. ML tasks involving data with such distinct subgroups require careful design of evaluation strategies, to ensure that the patterns learned generalize across the subgroups. **Chapter 3** demonstrates the need for this; we show that testing our predictor on a randomly selected set of STSs overestimates its performance on STSs from species rare in the dataset. This separation is even larger in the case of bacterial and fungal STSs, and other terpene synthases (TPSs) using different substrates.

From an evolutionary perspective, phylogenetic analyses can help locate shared ancestry and hypothesize evolutionary mechanisms leading to the diversity we see. For example, TPSs have been divided into seven clades on the basis of phylogeny, with two clades including members distantly related to primary metabolism and the others, showing greater diversification, involved in specialized metabolism[27]. Some lineages have members specific to certain substrates and even to certain products, thus attempts to predict this lineage specificity[28] are also beneficial to the kind of compound specificity screening explored in this thesis, though much of this specificity is still unexplained by phylogenetic analysis alone. Examining evolutionary relatedness between TPSs both within individual species such as tomato[29], grape[30] and *Cannabis sativa*[31] as well as across species[32–34], has pinpointed a number of orthologous genes, as well as species-specific genes which deviate from overall TPS sequence patterns.

From an ML perspective, the presence of distinct subgroups in data could lead to a situation where ML algorithms learn different patterns for each subgroup. Such patterns are often not meaningful in some subgroups due to insufficient amounts of data, but also not desired as many insights about biological mechanisms can be obtained by looking into similarities across these proteins rather than fixating on where they diverged, unlike in the phylogenetic analysis where this divergence was the focus. ML models trained on a single subgroup, on the other hand, may not perform well on a broader set, limiting their usefulness and compelling researchers to go through the process of data collection, preprocessing, training, and evaluation for different subgroups individually. Domain adaptation is one ML strategy that can be used to alleviate this issue in the case of distinct domains such as bacterial, fungal, and plant STSs, or STSs and other kinds of TPSs such as mono- and di-TPSs[35]. Domain adaptation is concerned with transferring knowledge to a target domain using information from different but related source domains. One approach to do this is using large amounts of unlabelled data from different source domains to obtain implicit information about the target domain[36–38]. In **Chapter 3** we utilized a similar concept, where uncharacterised putative TPS sequences (covering STSs but also mono-, di- and other TPSs) were used to spot possible catalytically co-evolving residues linked with STS cation specificity. Another approach attempts to normalize features obtained from labelled source and target domain instances - either via weight transfer[39], or by attempting to extract domain-invariant features[40]. Bacterial and fungal STSs share a number of products with plant STSs but contain distinct sequence and structural features - thus, such domain adaptation approaches may be useful in learning STS compound specificity across the tree of life. STSs from different plant clades, producing sesquiterpenes from different reaction paths, or characterized under different experimental conditions do not constitute distinct domains - these instead represent data which is not independent and identically distributed (i.i.d). While some domain adaptation concepts may also translate to these settings, recent approaches to deal with non-i.i.d data can also be found in the field of federated or decentralized learning, where ML needs to be performed over datasets generated at different devices and locations. Algorithms for this purpose have been created which attempt to deal with both skewed data distributions[41] and labels[42], but often require more labelled data than is currently available for STSs, though this is changing fast with the rate of novel STS characterization studies such as the one in **Chapter 7**.

Additionally, some aspects of individual protein families are likely to be shared across the protein universe as a whole - including phylogenetic, fluorescence, pair-wise contact, structural, and subcellular localization properties, many of which may not actually be well-defined for the family under study due to lack of experimental data (especially true for proteins in the so-called "dark proteome" [43]). Two associated ML strategies, transfer learning and self-supervised learning, can take advantage of this implicit universe-level protein data. Both fall under the category of "pretraining", a technique to learn more effective representations of data to use as input for ML algorithms, but whereas transfer learning uses labelled data from a larger target domain to learn a good representation of the source domain, self-supervised learning does not require annotated labels for target domain data[44]. These techniques would enable learning global protein representations which may emphasize characteristics of proteins relevant to a particular task, followed by individual retraining in one family to learn idiosyncrasies specific to these proteins for that task. Transfer learning has been used in protein modelling[45] and model quality assessment[46], in predicting the effects of mutation[47] and post-translational modifications[48]. Self-supervised learning has been implemented to generate embeddings for proteins mainly using techniques from natural language processing[49–52]; these embeddings were found to capture various global properties of proteins such as amino acid characteristics, topological folds, and other physiological properties. Endeavours such as the recently released Tasks

Assessing Protein Embeddings (TAPE) datasets will be critical to compare these and future embeddings across a variety of protein ML tasks such as secondary and tertiary structure prediction, remote homology detection, fluorescence and stability landscape prediction, function prediction and more[53]. However, these approaches often suffer from a lack of interpretability, as it is often not straightforward or even possible to go from a learned embedding to features and regions of the original input proteins which are relevant to a predictive task. In **Chapter 5** we described a protein structure embedding that captures structural characteristics across diverse and homologous protein families while still being able to locate and interpret predictive structural regions in individual proteins. As the field of ML moves more towards "opening the black box" [54], I expect to see more advanced versions of such flexible and interpretable protein embeddings.

### 9.1.3 Specialized portals to integrate data from different sources

One of the first challenges in the work in this thesis was the collection of characterized STS enzyme data. In **Chapter 2** we performed an extensive literature survey to gather characterization studies of these enzymes and their products, and this was further expanded upon in **Chapters 3**. Despite the presence of curated aggregators such as SwissProt which aim to collect information across the protein universe, such individual literature surveys and databases are widely prevalent for multiple protein families[55–57]. This is partly because the speed at which new proteins are experimentally characterized is too high for universal curated databases to keep up with, as we saw in **Chapter 2** where our literature survey returned nearly twice as many characterized STSs as those annotated on SwissProt. However, an additional major advantage of specialized databases and portals is their capability to include auxiliary information only relevant for the protein family under consideration, enabling easy comparison across different species, annotated properties, experimental conditions and more.

This is especially true for SM enzyme families, as the intricate and highly complex metabolic networks present in cells involve numerous external factors and a lot of research has been done that does not always fit into the standardized forms and fields of universal databases. For instance, apart from protein sequence, plant SM production via SM enzyme families can be influenced by transcription factors, the development stage of the plant, morphogenetic factors such as the tissues involved and concentrations of various molecules within these tissues, and environmental factors such as pH, temperature, and biotic stresses[58]. Some of these aspects, such as concentrations of certain molecules and ions, pH, and temperature also influence experimental studies and may alter the detected profiles of enzymes in terms of the molecules produced and their abundance[59]. In some cases, experiments have been performed to adapt enzymatic production to novel microbial systems, or to modify a product profile by changing the substrates provided. In addition, there are a multitude of mutation studies, such as those described in Chapter 1, that affect various aspects of enzyme activity and are also performed under varying conditions. It can be worthwhile to document this information in a way that allows for comparison and exploration across all annotated enzymes and experiments.

Standardization of experimental characterization studies is difficult due to varying conditions in labs across the world and differing hypotheses and expectations. This underlying level of noise and experimental bias can negatively impact computational studies. However, thorough documentation of these conditions in the proposed specialized portals may allow for better handling of different experimental settings explicitly within the model. In **Chapter 8** we present a starting point for a protein family portal, that includes information on sequence, structure, substrate, product, and conservation, and allows for comparison across proteins based on any number of these axes. This could be extended to include information on catalytic activity, mutations and their effect, and various other kinds of experiments too specific to be included in universal databases yet crucial for the understanding of individual families. They would also be the ideal centralized setting to incorporate various predictors such as the ones described in this thesis. Cross-talk between portals could be designed to aid analysis of interactions between proteins from different families. While initially the issue of keeping a portal continuously updated with novel experiments lies on the shoulders of the authors, increasing usage across the research community leads to submission of novel data becoming convention. This is often seen in databases and portals for model organisms[60], which serve as a focal point for the research community.

## 9.2 The promising future of structural bioinformatics

One landmark in 2020 (from a structural bioinformatics perspective) was the Critical Assessment of Structure Prediction (CASP14) result. DeepMind's AlphaFold2 algorithm prevailed over its competition, so much so that the CASP organizers released a press statement declaring the protein structure problem for single protein chains solved. Despite contrasting opinions from researchers in the field on whether or not this is a hyperbolic statement, it clearly represents an important and historic breakthrough in the field, demarcating the start of a new era in structural bioinformatics. Given this leap forward in structure prediction combined with various advances in experimental structure determination such as deep mutation scanning and cryo-electron microscopy, and the uninterrupted pace of traditional experimental structure resolution, it is not far-fetched to foresee an age where protein structure information is as prevalent and ubiquitous as sequence. Judging by the velocity of these advancements, this age is right around the corner.

This opens up a number of new opportunities in structural bioinformatics, and also places some urgency on a few long-standing open challenges. A number of these can be anticipated from the AlphaFold2 results themselves. Firstly, the press release emphasized "single protein chains" for a reason - complex structures are yet to be successfully predicted at the same breakthrough levels. Thus, using ML and deep learning (DL) for protein-protein interaction and interface prediction could be the next frontier. Many, if not most, interaction prediction algorithms rely on sequence data as input simply due to their wider availability[61]. However, structure-based prediction is more accurate and will likely become more sought after, with Wass et al.[62], Zhang et al.[63], Fout et al.[64], Townshend et al.[65], Sanchez-Garcia et al.[66], Gainza et al.[67]

representing some early entrants into the field. In addition, the fields of cryo-EM and cryo-electron tomography (cryo-ET), driven by improvements in the underlying technology and in algorithms for image processing, have transformed drastically in the past decade into high-throughput, high-resolution structural biology techniques for describing macromolecular complexes[68,69]. By capturing millions of snapshots of the molecule of interest, each carrying a unique molecule in its own conformational state, cryo-EM holds promise to reveal the conformational landscape of very large dynamic macromolecular complexes and molecular machines. While some existing areas for improvement lie in sample preparation and experimental design, computational algorithms for on-the-fly image processing and 3D reconstruction are also in high demand. Fast and accessible machine learning based tools to analyse and condense cryo-EM data have started to appear[70,71] and will likely become crucial tools in protein complex determination. Recently, a study combined incomplete cryo-EM data with parts of AlphaFold2-predicted domains to obtain a full-length atomic model for a SARS-CoV-2 protein, allowing for hypotheses linking specific residues to RNA binding[72]

A second area ripe with opportunities lies in protein design, also referred to as the inverse protein-folding problem. The typical goal of protein design is to identify an amino acid sequence that will stabilize a desired protein conformation or binding interaction, in order to raise thermostability, control binding specificities, increase binding affinity by scaffolding binding sites, design novel interfaces to disease-related signalling proteins, introduce novel ligand binding sites, and more[73,74]. While a number of computational approaches are being used in protein design[75] there is still a great deal of room for improvement. For instance, current prediction techniques have difficulty distinguishing between functional and non-functional proteins which have a high degree of similarity, and the space of possible proteins is too large and too functionally sparse to search exhaustively naturally, in the laboratory, and even computationally. Directed evolution is one technique used to alleviate this issue, as it sidesteps the need for accurate function prediction. Inspired by natural evolution, directed evolution starts from a (set of) known functional proteins and accumulates beneficial mutations via an iterative protocol of mutation and selection. This generates a library of modified sequences followed by screening to identify mutants and variants with improved properties, with further rounds of diversification until fitness goals are achieved. While this approach implicitly imposes limits in the lab – there are an enormous number of ways to mutate any given protein and even the most high-throughput screening or selection methods can sample only a fraction of these sequences – ML methods can intelligently select new variants to screen, thereby reaching higher fitness levels than are possible through screening alone[76]. One way to do this is by using information from unimproved sequences from previous rounds to learn functional relationships from experimental data even when the underlying biophysical mechanisms are not well understood. Additionally, specificity predictors such as those detailed in **Chapters 3 and 7** can be inspected for predictive or relevant residues to improve or modify compound specificity which can then form the starting points for mutation, thus reducing the sequence space to experimentally explore.

Another area with scope with improvement is *de novo* protein design of rare or unnatural folds. Approaches similar to the first AlphaFold have been used successfully in design tasks[77–79], indicating that the AlphaFold2 breakthrough may also cause a leap in protein design prediction. However, since ML-based structure prediction algorithms typically rely on natural protein structures during their learning phase, it is expected that the patterns they learn apply more to naturally occurring proteins than artificially designed ones, implying that synthetic protein design may still be out of reach. One major downside of existing DL-based structure prediction techniques is that prediction acts merely as an alternative to an experimental technique - it does not provide us with any more understanding of the processes behind the folding of proteins. However, the process of constructing idealized folds during protein design can reveal new information about the physical and structural constraints that dictate which conformations a protein can adopt[80,81]. Such insights could be of vital importance to solving fundamental biological questions behind the evolution of proteins, as well as critical to further improvement of protein engineering and design[82].

Finally, it seems to be time for a new sub-field of structural bioinformatics, which Mohammed AlQuraishi aptly dubbed "comparative structuromics". This sub-field would be concerned with tools, algorithms, and techniques to compare and contrast assorted datasets of protein structures to answer a variety of biological questions - the evolutionary relationships between structural orthologs, interaction networks and how they are affected by structural changes, folding and changes within different cellular contexts and organisms, and how structure and folding is coupled with different functional characteristics. Just as there exists a wide variety of tools for answering analogous questions from a sequence perspective, there need to be tools in structural bioinformatics, such as those described in **Chapters 4, 5, and 6**, that are as easy to use, as intuitive to interpret, as optimized, and as feature-rich as these sequence-based counterparts. I envision that in the coming decade people will reach for structure-based tools and algorithms nearly as often as they reach for BLAST, and that machine learning will play a major role in many of these tools, just as it has with sequence.

## 9.3 Structure as one part of the puzzle

The three-dimensional coordinates of a protein structure or computationally generated model provides information about the kinds of folds and structural motifs within a protein, the interaction networks between non-sequential residues, and the positioning of amino acid sidechains, many of which are relevant to protein activity. However, proteins are much more than these coordinates, just as they are much more than their one-dimensional sequence representation.

Since proteins are inherently dynamic in nature, their true "structure" is an ensemble of possible conformations, with some areas of the protein displaying more flexibility than others. This is further influenced by the constant interaction of proteins with the surrounding solvent, small molecules, nucleic acids, peptides and of course other proteins, all of which drive conformational changes within the protein. Protein biological activity often involves adopting specific conformations, contributions from local fluctuations, and even large-scale structural transitions between different conformations. Correspondingly, structural flexibility ranges from small sidechain fluctuations to fragments of highly disordered or unstructured regions becoming ordered to large rearrangement of the entire backbone. In fact, the old paradigm that sequence encodes structure, and structure determines function can now be rephrased as sequence encodes structure, structure determines dynamics, and dynamics encodes function[83].

Molecular dynamics (MD) is a commonly used technique to model protein flexibility. Given the positions of all atoms in a protein system (i.e. including water or other solvent molecules, and sometimes a lipid bilayer), MD calculates the force exerted on each atom by all other atoms as a function of time, using a molecular mechanics force field fit to the results of quantum mechanical calculations and, typically, to certain experimental measurements[84]. MD simulations typically involve millions or billions of time steps and calculations of millions of interatomic interactions during each time step, causing them to be computationally extremely expensive. Moreover, MD does not address covalent bond formation or breakage, both crucial in a number of enzyme families, leading to the need for the even more expensive and challenging set up of Quantum mechanics/molecular mechanics (QM/MM) simulations, in which a small part of the system is modelled using quantum mechanical calculations and the remainder by MD simulation[85]. Since MD is often too time-consuming, computational-resource hungry, or difficult to set up to be applied effectively to large-scale protein systems, coarse-grained (CG) modelling with Monte Carlo (MC) simulations and elastic network models (ENM) with normal mode analysis (NMA) both provide simplified protein representations that still allow for quite accurate understanding of protein flexibility. CG protein representations reduce amino acid residues or even whole fragments of secondary structural elements to so-called united atoms, thus greatly reducing the number of degrees of freedom. For example, the CABS CG model[86] represents each amino acid as four united atoms corresponding to the $C\alpha$, $C\beta$, the centre of mass of the sidechain atoms, and the centre of the $C\alpha - C\alpha$ pseudobond to the next amino acid. The SURPASS CG model[87] goes even further and condenses four consecutive amino acids into a united atom at the centre of mass of their $C\alpha$s. CG models are typically used with MC simulation, a technique that maps the distribution of possible protein conformations through a very long random sequence of small local moves controlled via knowledge-based force fields[83]. In contrast to CG models, the ENM representation of a protein consists of nodes corresponding to residues with identical harmonic springs connected residues closer than a set distance threshold[88]. The simple harmonic oscillations of these interconnected springs around an energy minimum define the normal modes, with different modes accounting for low frequency large-scale movements up to high frequency small local fluctuations.

Together, these computational techniques can provide information about globular protein flexibility and mutations[89,90], large-scale structural transitions (e.g. from active to inactive conformations)[91–94], and conformations involved in the formation of protein complexes[95]. They have also been used to assess and refine 3D models[96–98], improve ligand positioning[99,100], and to create receptor ensembles for ensemble docking[101,102]. The faster and cruder CG-MC and ENM-NMA approaches can be combined with atomistic-level MD, providing efficient strategies and starting points for multiscale simulations of proteins and complexes[103]. While ML is becoming more prevalent in the MD and CG-MC fields, to construct force field models, energy surfaces, sampling etc.[104,105], future efforts will likely also utilize the flexibility information obtained from these techniques to use as input in ML-based predictors of protein function, as we did in **Chapters 3 and 7**. This was also demonstrated in previous research to improve over static structure-based prediction[106]. The topological featurization described in **Chapter 5** could be a starting point for such advances.

Sometimes, computational prediction methods produce important information as a by-product of the prediction process. For example, the model scores produced during homology modelling can be useful to select the correctly folded model produced by *ab initio* modelling[107]. Scores produced during computational docking can be used to predict interacting partners for proteins[62]. Similarly, it has been noticed that comparing accessible surface area predicted from sequence alone to that obtained from structure can give an idea of areas of the structure which are buried during binding, due to the difference in the sources of information used for these two predictions[108]. The recent Essential Site Scanning Analysis algorithm mimics the crowding induced upon substrate binding by adding the heavy atoms of each residue as additional network nodes into the $\alpha$-carbon-based ENM, and measuring the effect this increased density has on the resulting ENM normal modes - residues that cause significant changes can play key roles in altering the global dynamics of proteins on ligand-binding[109]. Such normal mode perturbation analyses have also been used to detect allosteric interactions and pockets, where one site on a molecule is perturbed by an effector and causes a functional change at another, possibly distant, site[109,110].

Eventually protein ML will need to explicitly or implicitly handle all of these different sources of features, also including the underlying physicochemical properties of the amino acids involved, as they all play crucial roles in determining protein activity, interaction, and function. End-to-end approaches may help select and utilize the combination of features relevant to the task at hand.

Furthermore, biological function is only partly determined by an individual protein – its genomic and cellular contexts also play a big role. Each protein is determined by an underlying gene sequence, but the mapping from gene to protein is not so straightforward, complicated by the existence of alternatively spliced transcript variants[111], pre-protein sequences in need of further processing[112], and moonlighting pseudoenzymes[113]. In addition, transcription factors[114], post-translational modifications[115], the developmental stage of an organism's life, their subcellular localization and environment in the cell, and even the extra-cellular conditions all have an effect on protein expression and function. More often than not, proteins also work in concert with a wide variety of other entities, ranging from metal ions and cofactors[116], water and other solvent molecules[117], small molecule ligands[118], peptides, nucleic acids, and other proteins.

One area of study focused on integrating these different contexts of proteins and their complex interactions is network biology. This field is crucial for the accurate modelling of biological systems, and given the influx of data from high-throughput interaction assays and large-scale multi-omics studies, a great target for ML and DL methods. The future holds an increasing number of opportunities for this combination of network biology and ML[119] – in understanding and fighting diseases by inspecting protein and gene interaction networks, in locating off-target effects of drugs and concocting valuable drug combination therapies based on chemical networks and multi-omics data from drug treatments[120], in understanding microbial interactions through metabolic networks, in finding biosynthetic gene clusters through gene neighbourhoods, transcriptomics, and expression profiling, and in designing synthetic gene circuits combining interconnected genes, promoters, and ribosome binding sites. Each of the individual entities involved have their own set of features and representations and there is a long way to go to harness the networked intricacies and complexities of these systems.

## 9.4 Closing remarks

The field of plant specialized metabolism is exciting to work in, mainly due to the growing prevalence and diversity of computational methodologies capable of addressing a number of research questions and limitations of previous approaches. Structural bioinformatics, with its capabilities of inspecting the relationships and mechanisms of the proteins and compounds involved, and machine learning, with its capacity to identify general patterns across entire protein families and superfamilies, both present numerous opportunities and applications to specialized metabolism and protein biology. With this thesis I have contributed to the field by developing novel computational methods for structural bioinformatics, and providing insights into the puzzling and sought-after family of terpene synthase enzymes. With the boundaries of experimental biology and bioinformatics starting to blur, computational approaches guiding experimental ones and experiments providing data to create predictors nearly as accurate, the ability to understand biological function is closer than ever.

# References

[1] Kreis, W., & Munkert, J. (2019). Exploiting enzyme promiscuity to shape plant specialized metabolism. *Journal of Experimental Botany*, *70*, 1435–1445.

[2] Nobeli, I., Favia, A. D., & Thornton, J. M. (2009). Protein promiscuity and its implications for biotechnology. *Nature Biotechnology*, *27*, 157–167.

[3] Chapelle, O., Schölkopf, B., & Zien, A. (2006). *Semi-Supervised Learning (Adaptive Computation and Machine Learning)*. The MIT Press.

[4] van Engelen, J. E., & Hoos, H. H. (2020). A survey on semi-supervised learning. *Machine Learning*, *109*, 373–440.

[5] Xia, Z., Wu, L.-Y., Zhou, X., & Wong, S. T. (2010). Semi-supervised drug-protein interaction prediction from heterogeneous biological spaces. *BMC Systems Biology*, *4*, S6.

[6] Jiang, J. Q., & McQuay, L. J. (2012). Predicting protein function by multi-label correlated semi-supervised learning. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, *9*, 1059–1069.

[7] Nguyen, T.-P., & Ho, T.-B. (2012). Detecting disease genes based on semi-supervised learning and protein–protein interaction networks. *Artificial Intelligence in Medicine*, *54*, 63–71.

[8] Shi, L., Lei, X., & Zhang, A. (2011). Protein complex detection with semi-supervised learning in protein interaction networks. *Proteome Science*, *9*, S5.

[9] Bahi, M., & Batouche, M. (2018). Drug-target interaction prediction in drug repositioning based on deep semi-supervised learning. In *IFIP International Conference on Computational Intelligence and Its Applications* (pp. 302–313). Springer.

[10] Dong, Y., Sun, Y., & Qin, C. (2018). Predicting protein complexes using a supervised learning method combined with local structural information. *PLoS One*, *13*, e0194124.

[11] da Silva, W. M. C., Andersen, J. L., Holanda, M. T., Walter, M. E. M. T., Brigido, M. M., Stadler, P. F., & Flamm, C. (2019). Exploring plant sesquiterpene diversity by generating chemical networks. *Processes*, *7*, 240.

[12] Tian, B., Poulter, C. D., & Jacobson, M. P. (2016). Defining the product chemical space of monoterpenoid synthases. *PLoS Computational Biology*, *12*, e1005053.

[13] da Silva, W. M. C., de Andrade, D. P., Andersen, J. L., Walter, M. E. M. T., Brigido, M., Stadler, P. F., & Flamm, C. (2020). Computational simulations for cyclizations catalyzed by plant monoterpene synthases. In J. C. Setubal, & W. M. Silva (Eds.), *Advances in Bioinformatics and Computational Biology* (pp. 247–258). Springer.

[14] Duigou, T., du Lac, M., Carbonell, P., & Faulon, J.-L. (2019). RetroRules: A database of reaction rules for engineering biology. *Nucleic Acids Research*, *47*, D1229–D1235.

[15] Yu, H., Yang, X., Zheng, S., & Sun, C. (2019). Active learning from imbalanced data: A solution of online weighted extreme learning machine. *IEEE Transactions on Neural Networks and Learning Systems*, *30*, 1088–1103.

[16] Settles, B. (2009). *Active Learning Literature Survey*. Technical Report, University of Wisconsin-Madison Department of Computer Sciences.

[17] Fujiwara, Y., Yamashita, Y., Osoda, T., Asogawa, M., Fukushima, C., Asao, M., Shimadzu, H., Nakao, K., & Shimizu, R. (2008). Virtual screening system for finding structurally diverse hits by active learning. *Journal of Chemical Information and Modeling*, *48*, 930–940.

[18] Warmuth, M. K., Liao, J., Rätsch, G., Mathieson, M., Putta, S., & Lemmen, C. (2003 Mar-Apr). Active learning with support vector machines in the drug discovery process. *Journal of Chemical Information and Computer Sciences*, *43*, 667–673.

[19] Naik, A. W., Kangas, J. D., Langmead, C. J., & Murphy, R. F. (2013). Efficient modeling and active learning discovery of biological responses. *PLoS One*, *8*, e83996.

[20] Reker, D., & Schneider, G. (2015). Active-learning strategies in computer-assisted drug discovery. *Drug Discovery Today*, *20*, 458–465.

[21] Romero, P. A., Krause, A., & Arnold, F. H. (2013). Navigating the protein fitness landscape with Gaussian processes, . *110*, E193–E201.

[22] Stegle, O., Payet, L., Mergny, J.-L., MacKay, D. J. C., & Leon, J. H. (2009). Predicting and understanding the stability of G-quadruplexes. *Bioinformatics*, *25*, i374–i1382.

[23] Danziger, S. A., Baronio, R., Ho, L., Hall, L., Salmon, K., Hatfield, G. W., Kaiser, P., & Lathrop, R. H. (2009). Predicting positive p53 cancer rescue regions using Most Informative Positive (MIP) active learning. *PLoS Computational Biology*, *5*, e1000498.

[24] Danziger, S. A., Zeng, J., Wang, Y., Brachmann, R. K., & Lathrop, R. H. (2007). Choosing where to look next in a mutation sequence space: Active Learning of informative p53 cancer rescue mutants. *Bioinformatics*, *23*, i104–114.

[25] Mohamed, T. P., Carbonell, J. G., & Ganapathiraju, M. K. (2010). Active learning for human protein-protein interaction prediction. *BMC Bioinformatics*, *11*, S57.

[26] Reker, D., Schneider, P., & Schneider, G. (2016). Multi-objective active machine learning rapidly improves structure–activity models and reveals new protein–protein interaction inhibitors. *Chemical Science*, *7*, 3919–3927.

[27] Chen, F., Tholl, D., Bohlmann, J., & Pichersky, E. (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*, *66*, 212–229.

[28] Priya, P., Yadav, A., Chand, J., & Yadav, G. (2018). Terzyme: A tool for identification and analysis of the plant terpenome. *Plant Methods*, *14*, 4.

[29] Zhou, F., & Pichersky, E. (2020). The complete functional characterisation of the terpene synthase family in tomato. *New Phytologist*, *226*, 1341–1360.

[30] Martin, D. M., Aubourg, S., Schouwey, M. B., Daviet, L., Schalk, M., Toub, O., Lund, S. T., & Bohlmann, J. (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*, 226.

[31] (). Terpene synthases and terpene variation in *Cannabis sativa*, .

[32] Hillwig, M. L., Xu, M., Toyomasu, T., Tiernan, M. S., Wei, G., Cui, G., Huang, L., & Peters, R. J. (2011). Domain loss has independently occurred multiple times in plant terpene synthase evolution. *The Plant Journal*, *68*, 1051–1060.

[33] Singh, B., & Sharma, R. A. (2015). Plant terpenes: Defense responses, phylogenetic analysis, regulation and clinical applications. *3 Biotech*, *5*, 129–151.

[34] Jia, Q., Chen, X., Köllner, T. G., Rinkel, J., Fu, J., Labbé, J., Xiong, W., Dickschat, J. S., Gershenzon, J., & Chen, F. (2019). Terpene synthase genes originated from bacteria through horizontal gene transfer contribute to terpenoid diversity in fungi. *Scientific Reports*, *9*, 9223.

[35] Kouw, W. M. (). An introduction to domain adaptation and transfer learning. *ArXiv e-prints*.

[36] Herndon, N., & Caragea, D. (2014). Predicting protein localization using a domain adaptation approach. In M. Fernández-Chimeno, P. L. Fernandes, S. Alvarez, D. Stacey, J. Solé-Casals, A. Fred, & H. Gamboa (Eds.), *Biomedical Engineering Systems and Technologies* Communications in Computer and Information Science (pp. 191–206). Springer.

[37] Yu, L., Li, R., Zeng, X., Wang, H., Jin, J., Ge, Y., Jiang, R., & Xu, M. (2021). Few shot domain adaptation for in situ macromolecule structural classification in cryoelectron tomograms. *Bioinformatics*, *37*, 185–191.

[38] Rios, A., Kavuluru, R., & Lu, Z. (2018). Generalizing biomedical relation classification with neural adversarial domain adaptation. *Bioinformatics*, *34*, 2973–2981.

[39] Xu, Y., Min, H., Wu, Q., Song, H., & Ye, B. (2017). Multi-instance metric transfer learning for genome-wide protein function prediction. *Scientific Reports*, *7*, 41831.

[40] Shaw, D., Chen, H., & Jiang, T. (2019). DeepIsoFun: A deep domain adaptation approach to predict isoform functions. *Bioinformatics*, *35*, 2535–2544.

[41] Briggs, C., Fan, Z., & Andras, P. (2020). Federated learning with hierarchical clustering of local updates to improve training on non-IID data. In *2020 International Joint Conference on Neural Networks (IJCNN)* (pp. 1–9).

[42] Hsieh, K., Phanishayee, A., Mutlu, O., & Gibbons, P. (2020). The non-IID data quagmire of decentralized machine learning. In H. D. III, & A. Singh (Eds.), *Proceedings of the 37th International Conference on Machine Learning* (pp. 4387–4398). PMLR volume 119 of *Proceedings of Machine Learning Research*.

[43] Perdigão, N., Heinrich, J., Stolte, C., Sabir, K. S., Buckley, M. J., Tabor, B., Signal, B., Gloss, B. S., Hammang, C. J., Rost, B., Schafferhans, A., & O'Donoghue, S. I. (2015). Unexpected features of the dark proteome. *Proceedings of the National Academy of Sciences*, *112*, 15898–15903.

[44] Yang, X., He, X., Liang, Y., Yang, Y., Zhang, S., & Xie, P. (). Transfer learning or self-supervised learning? A tale of two pretraining paradigms. *ArXiv e-prints*.

[45] Wang, S., Li, Z., Yu, Y., & Xu, J. (2017). Folding membrane proteins by deep transfer learning. *Cell Systems*, *5*, 202–211.e3.

[46] Hurtado, D. M., Uziela, K., & Elofsson, A. (2018). Deep transfer learning in the assessment of the quality of protein models. *ArXiv e-prints*. `arXiv:1804.06281`.

[47] Shamsi, Z., Chan, M., & Shukla, D. (2020). TLmutation: Predicting the effects of mutations using transfer learning. *The Journal of Physical Chemistry B*, *124*, 3845–3854.

[48] He, F., Wang, R., Gao, Y., Wang, D., Yu, Y., Xu, D., & Zhao, X. (2019). Protein ubiquitylation and sumoylation site prediction based on ensemble and transfer learning. In *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)* (pp. 117–123).

[49] Alley, E. C., Khimulya, G., Biswas, S., AlQuraishi, M., & Church, G. M. (2019). Unified rational protein engineering with sequence-based deep representation learning. *Nature Methods*, *16*, 1315–1322.

[50] Bepler, T., & Berger, B. (2019). Learning protein sequence embeddings using information from structure. *ArXiv e-prints*.

[51] Heinzinger, M., Elnaggar, A., Wang, Y., Dallago, C., Nechaev, D., Matthes, F., & Rost, B. (2019). Modeling aspects of the language of life through transfer-learning protein sequences. *BMC Bioinformatics*, *20*, 723.

[52] Rives, A., Meier, J., Sercu, T., Goyal, S., Lin, Z., Liu, J., Guo, D., Ott, M., Zitnick, C. L., Ma, J., & Fergus, R. (2021). Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences*, *118*.

[53] Rao, R., Bhattacharya, N., Thomas, N., Duan, Y., Chen, X., Canny, J., Abbeel, P., & Song, Y. S. (2019). Evaluating protein transfer learning with TAPE. *Advances in Neural Information Processing Systems*, *32*, 9689–9701.

[54] Azodi, C. B., Tang, J., & Shiu, S.-H. (2020). Opening the black box: Interpretable machine learning for geneticists. *Trends in Genetics*, .

[55] Kooistra, A. J., Kanev, G. K., van Linden, O. P., Leurs, R., de Esch, I. J., & de Graaf, C. (2016). KLIFS: A structural kinase-ligand interaction database. *Nucleic Acids Research*, *44*, D365–D371.

[56] Munk, C., Isberg, V., Mordalski, S., Harpsøe, K., Rataj, K., Hauser, A., Kolb, P., Bojarski, A., Vriend, G., & Gloriam, D. (2016). GPCRdb: The G protein-coupled receptor database–an introduction. *British Journal of Pharmacology*, *173*, 2195–2207.

[57] Caboche, S., Pupin, M., Leclere, V., Fontaine, A., Jacques, P., & Kucherov, G. (2007). NORINE: A database of nonribosomal peptides. *Nucleic Acids Research*, *36*, D326–D331.

[58] Verma, N., & Shukla, S. (2015). Impact of various factors responsible for fluctuation in plant secondary metabolites. *Journal of Applied Research on Medicinal and Aromatic Plants*, *2*, 105–113.

[59] Koo, H. J., Vickery, C. R., Xu, Y., Louie, G. V., O'Maille, P. E., Bowman, M., Nartey, C. M., Burkart, M. D., & Noel, J. P. (2016). Biosynthetic potential of sesquiterpene synthases: Product profiles of Egyptian Henbane premnaspirodiene synthase and related mutants. *The Journal of Antibiotics*, *69*, 524–533.

[60] Leonelli, S., & Ankeny, R. A. (2012). Re-thinking organisms: The impact of databases on model organism biology. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences*, *43*, 29–36.

[61] Khatun, M. S., Shoombuatong, W., Hasan, M. M., & Kurata, H. (2020). Evolution of sequence-based bioinformatics tools for protein-protein interaction prediction. *Current Genomics*, *21*, 454–463.

[62] Wass, M. N., Fuentes, G., Pons, C., Pazos, F., & Valencia, A. (2011). Towards the prediction of protein interaction partners using physical docking. *Molecular Systems Biology*, *7*, 469.

[63] Zhang, Q. C. et al. (2012). Structure-based prediction of protein–protein interactions on a genome-wide scale. *Nature*, *490*, 556–560.

[64] Fout, A., Byrd, J., Shariat, B., & Ben-Hur, A. (2017). Protein interface prediction using graph convolutional networks. In *Proceedings of the 31st International Conference on Neural Information Processing Systems* NIPS'17 (pp. 6533–6542). Red Hook, NY, USA: Curran Associates Inc.

[65] Townshend, R., Bedi, R., Suriana, P., & Dror, R. (2019). End-to-end learning on 3D protein structure for interface prediction. *Advances in Neural Information Processing Systems*, *32*.

[66] Sanchez-Garcia, R., Sorzano, C. O. S., Carazo, J. M., & Segura, J. (2019). BIPSPI: A method for the prediction of partner-specific protein-protein interfaces. *Bioinformatics*, *35*, 470–477.

[67] Gainza, P., Sverrisson, F., Monti, F., Rodolà, E., Boscaini, D., Bronstein, M. M., & Correia, B. E. (2020). Deciphering interaction fingerprints from protein molecular surfaces using geometric deep learning. *Nature Methods*, *17*, 184–192.

[68] Cheng, Y. (2018). Single-particle cryo-EM—How did it get here and where will it go. *Science*, *361*, 876–880.

[69] Danev, R., Yanagisawa, H., & Kikkawa, M. (2019). Cryo-electron microscopy methodology: Current aspects and future directions. *Trends in Biochemical Sciences*, *44*, 837–848.

[70] Alnabati, E., & Kihara, D. (2020). Advances in structure modeling methods for cryo-electron microscopy maps. *Molecules*, *25*, 82.

[71] Zhong, E. D., Bepler, T., Berger, B., & Davis, J. H. (2021). CryoDRGN: Reconstruction of heterogeneous cryo-EM structures using neural networks. *Nature Methods*, *18*, 176–185.

[72] Gupta, M. et al. (2021). CryoEM and AI reveal a structure of SARS-CoV-2 Nsp2, a multifunctional protein involved in key host processes.. .

[73] Kuhlman, B., & Bradley, P. (2019). Advances in protein structure prediction and design. *Nature Reviews Molecular Cell Biology*, *20*, 681–697.

[74] Huang, P.-S., Boyken, S. E., & Baker, D. (2016). The coming of age of de novo protein design. *Nature*, *537*, 320–327.

[75] Pan, X., & Kortemme, T. (2021). Recent advances in de novo protein design: Principles, methods, and applications. *Journal of Biological Chemistry*, *296*, 100558.

[76] Yang, K. K., Wu, Z., & Arnold, F. H. (2019). Machine-learning-guided directed evolution for protein engineering. *Nature Methods*, *16*, 687–694.

[77] Anishchenko, I., Chidyausiku, T. M., Ovchinnikov, S., Pellock, S. J., & Baker, D. (2020). De novo protein design by deep network hallucination. .

[78] Norn, C., Wicky, B. I., Juergens, D., Liu, S., Kim, D., Koepnick, B., Anishchenko, I., Baker, D., & Ovchinnikov, S. (2020). Protein sequence design by explicit energy landscape optimization. *bioRxiv*.

[79] Tischer, D., Lisanza, S., Wang, J., Dong, R., Anishchenko, I., Milles, L. F., Ovchinnikov, S., & Baker, D. (2020). Design of proteins presenting discontinuous functional sites using deep learning. .

[80] Lin, Y.-R., Koga, N., Tatsumi-Koga, R., Liu, G., Clouser, A. F., Montelione, G. T., & Baker, D. (2015). Control over overall shape and size in de novo designed proteins. *Proceedings of the National Academy of Sciences of the United States of America*, *112*, E5478–5485.

[81] Marcos, E., Chidyausiku, T. M., McShan, A. C., Evangelidis, T., Nerli, S., Carter, L., Nivón, L. G., Davis, A., Oberdorfer, G., Tripsianes, K., Sgourakis, N. G., & Baker, D. (2018). De novo design of a non-local $\beta$-sheet protein with high stability and accuracy. *Nature Structural & Molecular Biology*, *25*, 1028–1034.

[82] Baker, D. (2019). What has de novo protein design taught us about protein folding and biophysics? *Protein Science*, *28*, 678–683.

[83] Kmiecik, S., Kouza, M., Badaczewska-Dawid, A. E., Kloczkowski, A., & Kolinski, A. (2018). Modeling of protein structural flexibility and large-scale dynamics: Coarse-grained simulations and elastic network models. *International Journal of Molecular Sciences*, *19*, 3496.

[84] Hollingsworth, S. A., & Dror, R. O. (2018). Molecular dynamics simulation for all. *Neuron*, *99*, 1129–1143.

[85] Quesne, M. G., Borowski, T., & de Visser, S. P. (2016). Quantum mechanics/molecular mechanics modeling of enzymatic processes: Caveats and breakthroughs. *Chemistry – A European Journal*, *22*, 2562–2581.

[86] Koliński, A. et al. (2004). Protein modeling and structure prediction with a reduced representation. *Acta Biochimica Polonica*, *51*.

[87] Dawid, A. E., Gront, D., & Kolinski, A. (2017). SURPASS low-resolution coarse-grained protein modeling. *Journal of Chemical Theory and Computation*, *13*, 5766–5779.

[88] Atilgan, A. R., Durell, S. R., Jernigan, R. L., Demirel, M. C., Keskin, O., & Bahar, I. (2001). Anisotropy of fluctuation dynamics of proteins with an elastic network model. *Biophysical Journal*, *80*, 505–515.

[89] Jamroz, M., Orozco, M., Kolinski, A., & Kmiecik, S. (2013). Consistent view of protein fluctuations from all-atom molecular dynamics and coarse-grained dynamics with knowledge-based force-field. *Journal of Chemical Theory and Computation*, *9*, 119–125.

[90] Frappier, V., & Najmanovich, R. J. (2014). A coarse-grained elastic network atom contact model and its use in the simulation of protein dynamics and the prediction of the effect of mutations. *PLoS Computational Biology*, *10*, e1003569.

[91] Tekpinar, M., & Zheng, W. (2010). Predicting order of conformational changes during protein conformational transitions using an interpolated elastic network model. *Proteins*, *78*, 2469–2481.

[92] Kmiecik, S., Gront, D., Kouza, M., & Kolinski, A. (2012). From coarse-grained to atomic-level characterization of protein dynamics: Transition state for the folding of B domain of protein A. *The Journal of Physical Chemistry. B*, *116*, 7026–7032.

[93] Mahajan, S., & Sanejouand, Y.-H. (2015). On the relationship between low-frequency normal modes and the large-scale conformational changes of proteins. *Archives of Biochemistry and Biophysics*, *567*, 59–65.

[94] Yang, L., Song, G., & Jernigan, R. L. (2007). How well can we understand large-scale protein motions using normal modes of elastic network models? *Biophysical Journal*, *93*, 920–929.

[95] Takada, S., Kanada, R., Tan, C., Terakawa, T., Li, W., & Kenzaki, H. (2015). Modeling structural dynamics of biomolecular complexes by coarse-grained molecular simulations. *Accounts of Chemical Research*, *48*, 3026–3035.

[96] Singharoy, A., Teo, I., McGreevy, R., Stone, J. E., Zhao, J., & Schulten, K. (2016). Molecular dynamics-based refinement and validation for sub-5 Å cryo-electron microscopy maps. *eLife*, *5*.

[97] Mirjalili, V., Noyes, K., & Feig, M. (2014). Physics-based protein structure refinement through multiple molecular dynamics trajectories and structure averaging. *Proteins*, *82 Suppl 2*, 196–207.

[98] Gniewek, P., Kolinski, A., Jernigan, R. L., & Kloczkowski, A. (2012). Elastic network normal modes provide a basis for protein structure refinement. *The Journal of Chemical Physics*, *136*, 195101.

[99] Schneider, J., Korshunova, K., Si Chaib, Z., Giorgetti, A., Alfonso-Prieto, M., & Carloni, P. (2020). Ligand pose predictions for human G protein-coupled receptors: Insights from the Amber-based hybrid molecular mechanics/coarse-grained approach. *Journal of Chemical Information and Modeling*, *60*, 5103–5116.

[100] Wang, A., Zhang, Y., Chu, H., Liao, C., Zhang, Z., & Li, G. (2020). Higher accuracy achieved for protein-ligand binding pose prediction by elastic network model-based ensemble docking. *Journal of Chemical Information and Modeling*, *60*, 2939–2950.

[101] Cavasotto, C. N. (2012). Normal mode-based approaches in receptor ensemble docking. In R. Baron (Ed.), *Computational Drug Discovery and Design* Methods in Molecular Biology (pp. 157–168). New York, NY: Springer.

[102] Evangelista Falcon, W., Ellingson, S. R., Smith, J. C., & Baudry, J. (2019). Ensemble docking in drug discovery: How many protein configurations from molecular dynamics simulations are needed to reproduce known ligand binding? *The Journal of Physical Chemistry B*, *123*, 5189–5195.

[103] Stansfeld, P. J., & Sansom, M. S. P. (2011). From coarse grained to atomistic: A serial multiscale approach to membrane protein simulations. *Journal of Chemical Theory and Computation*, *7*, 1157–1166.

[104] Noé, F., Tkatchenko, A., Müller, K.-R., & Clementi, C. (2020). Machine learning for molecular simulation. *Annual Review of Physical Chemistry*, *71*, 361–390.

[105] Noé, F., De Fabritiis, G., & Clementi, C. (2020). Machine learning for protein folding and dynamics. *Current Opinion in Structural Biology*, *60*, 77–84.

[106] Glazer, D. S., Radmer, R. J., & Altman, R. B. (2008). Combining molecular dynamics and machine learning to improve protein function recognition. In *Biocomputing 2008* (pp. 332–343). World Scientific.

[107] Khare, S., Bhasin, M., Sahoo, A., & Varadarajan, R. (2019). Protein model discrimination attempts using mutational sensitivity, predicted secondary structure, and model quality information. *Proteins: Structure, Function, and Bioinformatics*, *87*, 326–336.

[108] Porollo, A., & Meller, J. (2007). Prediction-based fingerprints of protein–protein interactions. *Proteins: Structure, Function, and Bioinformatics*, *66*, 630–645.

[109] Kaynak, B. T., Bahar, I., & Doruker, P. (2020). Essential site scanning analysis: A new approach for detecting sites that modulate the dispersion of protein global motions. *Computational and Structural Biotechnology Journal*, *18*, 1577–1586.

[110] Greener, J. G., & Sternberg, M. J. (2015). AlloPred: Prediction of allosteric pockets on proteins using normal mode perturbation analysis. *BMC Bioinformatics*, *16*, 335.

[111] Kim, E., Goren, A., & Ast, G. (2008). Alternative splicing: Current perspectives. *BioEssays*, *30*, 38–47.

[112] Owji, H., Nezafat, N., Negahdaripour, M., Hajiebrahimi, A., & Ghasemi, Y. (2018). A comprehensive review of signal peptides: Structure, roles, and applications. *European Journal of Cell Biology*, *97*, 422–441.

[113] Ribeiro, A. J. M., Das, S., Dawson, N., Zaru, R., Orchard, S., Thornton, J. M., Orengo, C., Zeqiraj, E., Murphy, J. M., & Eyers, P. A. (2019). Emerging concepts in pseudoenzyme classification, evolution, and signaling. *Science Signaling*, *12*.

[114] Latchman, D. S. (1997). Transcription factors: An overview. *The International Journal of Biochemistry & Cell Biology*, *29*, 1305–1312.

[115] Audagnotto, M., & Dal Peraro, M. (2017). Protein post-translational modifications: In silico prediction tools and molecular modeling. *Computational and Structural Biotechnology Journal*, *15*, 307–319.

[116] Fischer, J. D., Holliday, G. L., Rahman, S. A., & Thornton, J. M. (2010). The structures and physicochemical properties of organic cofactors in biocatalysis. *Journal of Molecular Biology*, *403*, 803–824.

[117] Prabhu, N., & Sharp, K. (2006). Protein-solvent interactions. *Chemical Reviews*, *106*, 1616–1623.

[118] Li, X., Wang, X., & Snyder, M. (2013). Systematic investigation of protein–small molecule interactions. *IUBMB Life*, *65*, 2–8.

[119] Camacho, D. M., Collins, K. M., Powers, R. K., Costello, J. C., & Collins, J. J. (2018). Next-generation machine learning for biological networks. *Cell*, *173*, 1581–1592.

[120] Fuentealba, M., Dönertaş, H. M., Williams, R., Labbadia, J., Thornton, J. M., & Partridge, L. (2019). Using the drug-protein interactome to identify anti-ageing compounds for humans. *PLoS Computational Biology*, *15*, e1006639.

# Summary

Plant specialized metabolites (SMs) are crucial to plants and to humanity, with numerous applications in food, healthcare, agriculture, and cosmetics. The enzyme families involved in producing SMs, such as the terpene synthases, are very diverse, both across and within families. Understanding and predicting compound specificity of these enzymes is critical for biotechnological applications and protein engineering. The growing availability of structure data and improved computational modelling techniques puts us in the position to use structural bioinformatics and machine learning (ML) techniques to learn patterns across all enzymes in an SM family, instead of focusing on a few structures or mutants. In this thesis I explore new algorithms and approaches to analyse datasets of SM families and take advantage of their complex structural data.

In **Chapter 1** I introduce the terpene synthases and place them in context among the wider field of plant specialized metabolism. Their importance in both the plant and human worlds is discussed along with a history of the elucidation of their catalytic mechanisms via structural and mutational studies. I explore the various opportunities and challenges offered by computational techniques, found in the structural bioinformatics and ML fields, to better understand such elusive SM enzyme families. In **Chapter 2** I describe the creation of a database of experimentally characterized plant sesquiterpene synthases (STSs), collected from literature studies, covering over 250 enzymes collectively responsible for the production of over a hundred sesquiterpene compounds. These proteins are analysed from a sequence perspective leading to interesting results on previously studied motifs, as well as the conclusion that phylogeny plays a larger role in STS sequence similarity than product specificity. This further expedited the need for protein structure information, extracted using homology modelling. In **Chapter 3** I put forth an analysis of STS major and minor products, demonstrating that sesquiterpenes produced by an STS tend to be derived from the same reaction path. This enabled us to simplify the idea of product prediction to parent cation prediction, where I show that ML on the modelled STS structures out-performs sequence-based approaches.

To make further use of this structural information, in Chapters 4, 5 and 6 I developed structural bioinformatics embeddings for ML applications, resulting in an embedding allowing alignment-free comparison of the topologies and shapes contained in a structure, and a multiple structure alignment algorithm for structural features. The former, termed Geometricus and presented in **Chapters 5 and 6**, uses a concept from computer vision called rotation invariant moments to extract and count "shape-mers", structural analogues to sequence $k$-mers. The latter, Caretta, presented in **Chapters 4 and 6** is a multiple structure aligner that incorporates Geometricus shape-mer counting to scale to many thousands of proteins, and includes a feedback loop between single proteins and the progressively created alignment to return accurate and high-coverage alignments. To enable downstream ML analyses, Caretta also extracts and outputs aligned feature matrices, including the moment invariants used by Geometricus as a novel feature source describing protein shape and topology.

This novel feature extraction and alignment approach is applied in **Chapter 7** to the task of predicting STS product specificity. To increase our coverage of STS sequence and compound space we use what we learned in **Chapters 2 and 3** to select and experimentally characterize over 60 new STSs. As the number of possible products precludes the classification approach in **Chapter 3**, I create a joint protein-compound framework combining aligned protein structural features with chemical compound features to both successfully predict product specificity, and pinpoint residues involved in the formation of each sesquiterpene.

Many of the analyses and techniques used in this thesis are common across protein biology and bioinformatics. To allow life scientists to explore the interconnected properties of their protein family of interest from a variety of different perspectives, and share these findings across the web, in **Chapter 8** I present Turterra, an interactive data visualization portal.

**Chapter 9** concludes this thesis by describing ongoing challenges in studying SM enzyme families and their potential solutions from an ML perspective. I expand the discussion to the broader field of protein structure bioinformatics and the many opportunities it holds for enhancing our understanding of biological function.

# List of publications

**Durairaj, J.**; Di Girolamo, A.; Bouwmeester, H. J.; de Ridder, D.; Beekwilder, J.; van Dijk, A. D. J. An analysis of characterized plant sesquiterpene synthases. *Phytochemistry 2019, 158, 157–165.*

Akdel, M.; **Durairaj, J.**; de Ridder, D.; van Dijk, A. D. J. Caretta - A multiple protein structure alignment and feature extraction suite. *Computational and Structural Biotechnology Journal 2020.*

Di Girolamo, A.; **Durairaj, J.**; van Houwelingen, A.; Verstappen, F.; Bosch, D.; Cankar, K.; Bouwmeester, H.; de Ridder, D.; van Dijk, A. D. J.; Beekwilder, J. The santalene synthase from *Cinnamomum camphora*: Reconstruction of a sesquiterpene synthase from a monoterpene synthase. *Archives of Biochemistry and Biophysics 2020, 695, 108647.*

**Durairaj, J.**; Akdel, M.; de Ridder, D.; van Dijk, A. D. J. Geometricus represents protein structures as shape-mers derived from moment invariants. *Bioinformatics 2020, 36 (Supplement_2), i718–i725.*

**Durairaj, J.**; Melillo, E.; Bouwmeester, H. J.; Beekwilder, J.; de Ridder, D; van Dijk, A. D. J. Integrating structure-based machine learning and co-evolution to investigate specificity in plant sesquiterpene synthases. *PLOS Computational Biology 2021, 17 (3), e1008197.*

# Acknowledgements

The past years of scientific research have been both thrilling and fulfilling, culminating in this thesis. I have many people to thank for their contributions and support in achieving this milestone.

Aalt-Jan and Dick for the opportunity to come all the way to Wageningen for my Master thesis and then again to continue on to my PhD. Aalt-Jan, your attention to detail, astute comments, far-sighted suggestions and ever-sunny disposition were critical in shaping this thesis and in keeping up my motivation. You provided me with the perfect role model for effective teaching and supervision, and I hope to be able to emulate some of your qualities in my career. Dick, I'm grateful for your guidance and constant support throughout these years and for being so willing to take risks on new ideas. We've had some wonderful conversations about science, literature, and cats, and you've given me plenty of great advice for the future - I hope these will continue. Both of you had so much faith in me, and it has helped me become a more confident independent researcher. Your belief in my abilities and the occasional push to reach higher have been instrumental, this thesis would not have been possible without you.

I really appreciated being part of such an interdisciplinary project. Jules, Alice, Dirk, thanks for all our meetings, collaborations, great conversations, and practical insights that broke through my tunnel vision. Thamara and Jules, thanks of course for all the experimental support, and also for showing me around the lab and explaining the techniques involved. I'm very grateful to the larger TTW consortium, our half-yearly meetings were always delightful and your support and interest in my ideas and work were a constant source of motivation. Elena, Alice, Matthew, Harro, Frank, our sojourn in Germany remains one of my favourite conference memories.

This thesis and much of the research I've worked on in this time were fueled by collaboration. Special thanks to my partner Mehmet - our crazy side projects worked out great and here's to a future of many more. Giusi and Katja, Harro, Changsheng, Yanting and Lemeng - I really enjoyed working together and thanks to you I've gained experience in a number of different problems and techniques. Eric, thanks for being such a friendly external supervisor and for connecting me to Cloe and others. Barbara, thanks for being a great collaborator and wonderful friend.

I owe a lot to the close-knit family that is the Bioinformatics group and its past and honorary members (once a bioinformatician, always a bioinformatician). Raúl,

Miguel, Vittorio, Carlos, you always made me feel I had people to rely on. Lucas and Rens, our endless conversations about life, science, programming, and everything in between were often the highlight of my day. Sahar and Ben (and Dukesy), you helped me gain my confidence in a completely new country. Everyone who joined the lunches, the borrels, or got together for visiting museums, playing games and watching movies, Elvira, Hernando, Siavash, Jorge, Judith, Chiara, Eef, Serina, Margi, Sevgin, Ping, Brent, Sina, Jelle, I learned about so many corners of the world from all of you. I look forward to future collaborations and trips with many of you. Marie-Jose and Maria, thanks for all the administrative support. The staff, Harm, Marnix, Sandra, Justin, Anne, thanks for all your efforts to make the group such a warm and welcome place, as well as all the informative sessions you prepared for us. It was an honour to be a part of this group.

I wouldn't be here without my family; they have supported me throughout my education. I'm grateful to everyone in India and around the world for all the love and encouragement. Thanks, Amma and Appa for patiently listening every week as I listed all the projects, meetings and conferences on my mind. Thanks, Patti, for being such a beacon of joy and support always. Thanks, Mehmet, for rekindling my curiosity and enthusiasm whenever it started to wane. And a special thanks to my favourite writing critic, Geetha Mama, all the way from first encouraging me to write about anything at all in my teens, to always being ready to lend an ear to the latest writing challenge in my PhD. I had to do this one without your help, I hope you like it anyway!

# Education Statement of the Graduate School
# Experimental Plant Sciences

| | |
|---|---|
| **Issued to:** | **Janani Durairaj** |
| **Date:** | **15 September 2021** |
| **Group:** | **Bioinformatics** |
| **University:** | **Wageningen University & Research** |

| | **1) Start-Up Phase** | *date* | *cp* |
|---|---|---|---|
| ► | **First presentation of your project** | | |
| | Title: Novel Enzymes for Fragrance and Flavour | 9 Nov 2017 | 1.5 |
| ► | **Writing or rewriting a project proposal** | | |
| ► | **MSc courses** | | |
| | *Subtotal Start-Up Phase* | | 1.5 |

| | **2) Scientific Exposure** | *date* | *cp* |
|---|---|---|---|
| ► | **EPS PhD student days** | | |
| | EPS Get2Gether, Soest, NL | 11-12 Feb 2019 | 0.6 |
| | EPS Get2Gether, Soest, NL | 10-11 Feb 2020 | 0.6 |
| ► | **EPS theme symposia** | | |
| | EPS Theme 4 symposium 'Genome Biology', Amsterdam, NL | 25 Sep 2018 | 0.3 |
| ► | **Lunteren Days and other national platforms** | | |
| | Annual Meeting Experimental Plant Sciences, Lunteren, NL | 9-10 Apr 2018 | 0.6 |
| | Dutch Bioinformatics & Systems Biology conference (BioSB) conference, Lunteren, NL | 15-16 May 2018 | 0.6 |
| | Dutch Bioinformatics & Systems Biology conference (BioSB) conference, Lunteren, NL | 2-3 Apr 2019 | 0.6 |
| | Dutch Bioinformatics & Systems Biology conference (BioSB) conference, Online | 27-28 Oct 2020 | 0.6 |
| | Netherlands Society on Biomolecular Modelling (NSBM 2017) Fall Meeting, Utrecht, NL | 29 Nov 2017 | 0.3 |
| | Netherlands Society on Biomolecular Modelling (NSBM 2018) Fall Meeting, Utrecht, NL | 16 Nov 2018 | 0.3 |
| ► | **Seminars (series), workshops and symposia** | | |
| | Seminar: B-Wise Jens Allmer & Jesse van Dam | 5 Sep 2017 | 0.2 |
| | Seminar: B-Wise Pulva Kulkarni & Twan America | 7 Nov 2017 | 0.2 |
| | Seminar: B-Wise Mathijs Nieuwenhuis & Jorge Navarro Muñoz | 5 Dec 2017 | 0.2 |
| | Seminar: B-Wise Anton Feenstra & Ehsan Motazedi | 9 Jan 2018 | 0.2 |
| | Seminar: B-Wise Justin van der Hooft & Victor Carrion | 6 Feb 2018 | 0.2 |
| | Seminar: B-Wise Martijn Derks & Rik Kooke | 6 Mar 2018 | 0.2 |
| | Seminar: B-Wise Jeroen de Ridder | 3 Apr 2018 | 0.1 |
| | Seminar: B-Wise Sumanth Mutte & Hernando Suarez Duran | 1 May 2018 | 0.2 |
| | Seminar: B-Wise Joana Gonçalves & Jasper Depotter | 5 Jun 2018 | 0.2 |
| | Seminar: B-Wise Gurnoor Singh and Janani Durairaj (as speaker) | 4 Sept 2018 | 0.1 |
| | Seminar: B-Wise Christian Gilissen and Mohammad Alanjary | 2 Oct 2018 | 0.2 |
| | Seminar: B-Wise Erik van den Bergh & Willem Kruijer | 6 Nov 2018 | 0.2 |
| | Seminar: B-Wise Rachel Cavill & Mehmet Akdel | 4 Dec 2018 | 0.2 |
| | Seminar: B-Wise Rik van Rosmalen & Sevgin Demirci | 8 Jan 2019 | 0.2 |
| | Seminar: B-Wise Gerben Hermes & Pariya Berouhzi | 5 Feb 2019 | 0.2 |
| | Seminar: B-Wise Martin Huijnen & Mark Sterken | 5 Mar 2019 | 0.2 |
| | Seminar: B-Wise Simon van Heeringen & Chiara Bortoluzzi | 7 May 2019 | 0.2 |
| | Seminar: B-Wise Jorge Roel Touris & Victoria Pascal Andreu | 4 Jun 2019 | 0.2 |
| | Seminar: B-Wise Veronika Laine & Raúl Wijfjes | 3 Sep 2019 | 0.2 |
| | Seminar: B-Wise Eliana Papoutsoglou & Roeland Voorrips | 1 Oct 2019 | 0.2 |
| | Symposium: Wageningen 100 Years WUR Science Week: Wilhelm Huck, Philip Ball, Philip Brey | 12-14 Mar 2018 | 0.3 |
| | Symposium: Celebrating five years of bioinformatics collaboration @EPS | 10 Jul 2019 | 0.2 |
| | Workshop: BioExcel 2nd SIG Meeting: Advanced Simulations for Biomolecular Research, Athens, GR | 8 Sep 2018 | 0.3 |
| | Workshop: ELIXIR \| 3D-Bioinfo: Integrating structural and functional data to support in silico predictions in drug design, ECCB 2020, Online | 3 Sep 2020 | 0.3 |
| ► | **Seminar plus** | | |
| ► | **International symposia and congresses** | | |
| | European Conference on Computational Biology (ECCB 2018), Athens, GR | 8-12 Sep 2018 | 1.1 |
| | Intelligent Systems for Molecular Biology/European Conference on Computational Biology (ISMB/ECCB 2019), Basel, CH | 21-25 Jul 2019 | 1.2 |
| | International Meeting on the Biosynthesis, Function and Synthetic Biology of Isoprenoids (TERPNET 2019), Halle (Saale), DE | 26-30 Aug 2019 | 1.2 |
| | European Conference on Computational Biology (ECCB 2020), Online | 7-8 Sep 2020 | 0.6 |
| | Neural Information Processing Systems Conference (NeurIPS 2020), Online | 7-12 Dec 2020 | 0.9 |
| | From Information to Function: a systems biology view of the processes of life (infotofunc) - A tribute to Anna Tramontano, Online | 20 Apr 2021 | 0.3 |
| ► | **Presentations** | | |
| | Poster: 'Product Specificity in Plant Sesquiterpene Synthases', Annual meeting EPS | 9-10 Apr 2018 | 1.0 |
| | Poster: "Structure and Sequence-Based Prediction of Sesquiterpene Synthase Product Specificity", ECCB | 12-13 Sep 2018 | 1.0 |
| | Poster: "Structure-based Prediction of Terpene Synthase Product Specificity", ECCB 2019 | 23-25 Jul 2019 | 1.0 |
| | Poster: "Fast and adaptive protein structure representations for machine learning", MLSB | 12 Dec 2020 | 1.0 |
| | Poster and Flash Talk: "Fast and adaptive protein structure representations for machine learning", infotofunc | 20 Apr 2021 | 1.0 |
| | Talk: "Exploration of sequences, structures, and mechanisms in fragrance producing enzymes", B-Wise | 4 Sep 2018 | 1.0 |
| | Talk: "Exploration of fragrance producing enzymes in plants", EPS Theme 4 | 25 Sep 2018 | 1.0 |
| | Talk: "Novel enzymes for fragrance and flavour", NSBM 2018 | 16 Nov 2018 | 1.0 |
| | Talk: "Predicting sesquiterpene synthase product specificity", TERPNET 2019 | 27 Aug 2019 | 1.0 |

| | | | |
|---|---|---|---|
| | Talk: "Structure-based prediction of sesquiterpene synthase product specificity", BioSB 2019 | 2 Apr 2019 | 1.0 |
| | Talk: "Predicting Sesquiterpene Synthase Product Specificity", Applied Metabolic Systems (AMS) Clustermeeting, WUR | 27 May 2020 | 1.0 |
| | Talk: “Geometricus represents protein structures as shape-mers derived from moment invariants”, ECCB 2020 | 8 Sep 2020 | 1.0 |
| | Talk: "Comparing protein structures with and without alignment", BioSB 2020 | 27 Oct 2020 | 1.0 |
| ► | **3rd year interview** | | |
| ► | **Excursions** | | |
| | | *Subtotal Scientific Exposure* | 27.7 |

| 3) In-Depth Studies | | *date* | *cp* |
|---|---|---|---|
| ► | **Advanced scientific courses & workshops** | | |
| | BioSB: Protein structures: production, prowess, power, promises, and problems, Nijmegen, NL | 30 Oct - 3 Nov 2017 | 3.0 |
| | BioSB: Optimisation techniques in Bioinformatics and Systems Biology, Wageningen, NL | 12-16 Feb 2018 | 1.5 |
| | Summer School in Gaussian Processes and Uncertainty Quantification, Sheffield, UK | 9-12 Sep 2019 | 1.2 |
| ► | **Journal club** | | |
| | Literature Discussion, Bioinformatics Group | 2017-2021 | 3.0 |
| ► | **Individual research training** | | |
| | | *Subtotal In-Depth Studies* | 8.7 |

| 4) Personal Development | | *date* | *cp* |
|---|---|---|---|
| ► | **General skill training courses** | | |
| | EPS Introduction Course, Wageningen, NL | 26 Sep 2017 | 0.3 |
| | Wageningen Graduate Schools course: Working on your PhD in times of crisis, Online | 1-11 Mar 2021 | 0.5 |
| | Wageningen Graduate Schools course: Career Perspectives, Online | Jun-Jul 2021 | 1.6 |
| ► | **Organisation of meetings, PhD courses or outreach activities** | | |
| | Course: BioSB Algorithms for Biological Networks, Wageningen, NL | 25-29 Jun 2018 | 1.5 |
| | Course: Wageningen Data Science Week, Machine Learning Crash Course, Wageningen, NL (https://research.wur.nl/en/publications/crash-course-machine-learning) | 3 Feb 2020 | 1.5 |
| | Workshop: Code-Discussion Session, Wageningen, NL | 21 Feb 2018 | 0.0 |
| | Workshop: Numba Code-Along Session, Wageningen, NL | 4 Apr 2018 | 0.0 |
| | Workshop: Interactive Visualization, Gelselaar, NL | 16 Jul 2019 | 0.0 |
| | Seminar: Updates in Scientific Python, Wageningen + Gelselaar, NL | 24 Apr 2018, 16 Jul 2019 | 0.0 |
| ► | **Membership of EPS PhD Council** | | |
| | | *Subtotal Personal Development* | 5.4 |

| 5) Teaching & Supervision Duties | | *date* | *cp* |
|---|---|---|---|
| ► | **Courses** | | |
| | Machine Learning: Practical assistant | Feb - Mar 2018 | 3.0 |
| | Machine Learning: Practical assistant, Grading | 18 Feb - 8 Mar 2019 | |
| | Machine Learning: Practical assistant, Grading | 17 Feb - 6 Mar 2020 | |
| | Machine Learning: Course material preparation, Practical assistant, Grading | 15 Feb - 5 Mar 2021 | |
| | Data Science Concepts: Course material preparation | Feb 2020 | |
| | Data Analysis and Visualization: Guest lecture | 22 Nov 2019, 27 Nov 2020 | |
| ► | **Supervision of BSc/MSc students** | | |
| | MSc student, Dàmi Rebergen, "Prediction of the enzymatic reactions of sesquiterpene synthases: comparing sequence and structural information in machine learning" | 19 Feb - 27 Jun 2018 | 3.0 |
| | BSc student, Chris Congleton, "Predicting product category of monoterpene synthases based on protein sequence data" | 20 Mar - 3 Jul 2019 | |
| | MSc student internship, Kenneth Rivadeneira Guadamud, "Bioinformatic approach for predicting novel natural product enzymes" | 22 Feb - 19 Sep 2019 | |
| | BSc student, Nico Louwen, “The Turterra web portal for natural product enzyme family data visualization and analysis” | 7 May – 3 Jul 2020 | |
| | | *Subtotal Teaching & Supervision Duties* | 6.0 |

| **TOTAL NUMBER OF CREDIT POINTS*** | **49.3** |
|---|---|

Herewith the Graduate School declares that the PhD candidate has complied with the educational requirements set by the Educational Committee of EPS with a minimum total of 30 ECTS credits.

** A credit represents a normative study load of 28 hours of study.*

The research described in this thesis was financially supported by the Dutch Research Council NWO (TTW 15043) and the Graduate School Experimental Plant Sciences.

Cover design by Janani Durairaj

Printed by Proefschriften, Ede