

Exploring Enantiomeric Determinants in Natural Product Biosynthesis via an Integrated Molecular Modeling and Artificial Intelligence Approach

1. Scientific Background

1.1. Highlight

Earthly and marine plants, animals, fungi, and bacteria, among other organisms, produce a vast array of natural products, and these have been studied extensively in the past century. Enantiomeric pairs of these natural products frequently occur in nature, but the factors that determine which antipode is produced are not well understood. **Considering the fundamental role of enantiomers in molecular recognition by proteins across all of nature, this field demands more intensive investigation and dedicated research. In this proposal we aim to explore enantiomeric determinants in natural product biosynthesis via an integrated molecular modeling and artificial intelligence approach.**

1.2. Natural Products and Enantiomers

Earthly and marine plants, animals, fungi, and bacteria, among other organisms, produce a vast array of secondary metabolites, commonly called natural products.² Unlike primary metabolites, which are essential for survival, secondary metabolites are not required for basic life functions; however, they often support reproductive or defensive roles within the species that generate them.³⁻⁴ From a medicinal perspective, natural products are an invaluable source of bioactive compounds, including antitumor, antibacterial, anti-inflammatory, insecticidal, and immunosuppressive agents, among others. These bioactivities have been extensively harnessed in drug discovery and development projects.⁵⁻⁶

In many cases, chiral natural products are biosynthesized in nature as single, optically pure enantiomers, with only one specific form produced by the organism.^{2, 7} For instance, the diterpene Taxol is synthesized by the *Taxus* species (yew trees) with a distinct stereochemistry, which is crucial to its function as a potent anticancer agent.⁸⁻⁹ However, enantiomeric pairs of natural products do frequently occur in nature (**Figure 1**). These mirror-image compounds are often found in different genera or species, where one enantiomer is isolated from one species and its opposite from another. Occasionally, a single species may produce both enantiomers, which can be isolated either as racemic or scalemic mixtures.^{1, 7, 10}

Bioactive natural products have been studied extensively in the past century, yet the enzymatic synthesis of enantiomeric natural products is not well understood. This results from multiple factors, including the natural predominance of one enantiomer over its counterpart, which can leave the less common enantiomer undetected or unknown. Additionally, limited information on the sequence and structure of the enzymes involved in enantiomer formation adds to this challenge, and the enantiomeric characterization of natural products, which can be challenging,¹¹ is not always reported in the literature. **Considering the**

fundamental role of enantiomers in molecular recognition by proteins across all of nature, this field demands more intensive exploration and dedicated research. Several excellent studies have explored how to predict enantiomeric specificity in relatively simple enzyme reactions,¹²⁻¹³ but natural product biosynthesis presents a significant challenge due to the complex chemistry.

In our research group we have dedicated much attention in recent years to the biosynthesis of terpenes via terpene synthases (TPS).¹⁴⁻¹⁹ TPS catalyze the first step in the formation of

terpenoids, which comprise the largest class of natural products in nature with well over 80,000 known compounds.²⁰ TPS also form building blocks for other natural products, like steroids, saponins, carotenoids, meroterpenoids, and alkaloids.^{7, 21} The intricate structures generated by TPS are the result of substrate binding and folding in the active site, enzyme-controlled carbocation reaction cascades, and final reaction quenching.^{16, 20, 22} The chemical reactions taking place in TPS can be extremely complex, involving highly specific ring formations, proton and hydride shifts, and Wagner–Meerwein rearrangements, spanning up to a dozen discrete chemical steps involving carbocations.²³⁻²⁵ The universal substrates for TPS are relatively simple C_{5n} isoprenoid diphosphates (*n*=1, 2, 3, ...) precursors; the most common being monoterpenes (*n*=2), sesquiterpenes (*n*=3) and diterpenes (*n*=4) (Figure 2).^{20, 22, 26-28} The corresponding substrates are called geranyl diphosphate (GPP), farnesyl diphosphate (FPP), and geranyl geranyl diphosphate (GGPP), respectively (Figure 2A). Subsequent functionalizing enzymes, like P450 monooxygenases, acyltransferases, and glycosyltransferases, generate functionalized terpenes, i.e., terpenoids. Enantiomeric terpenes and terpenoids are relatively common, though they are typically

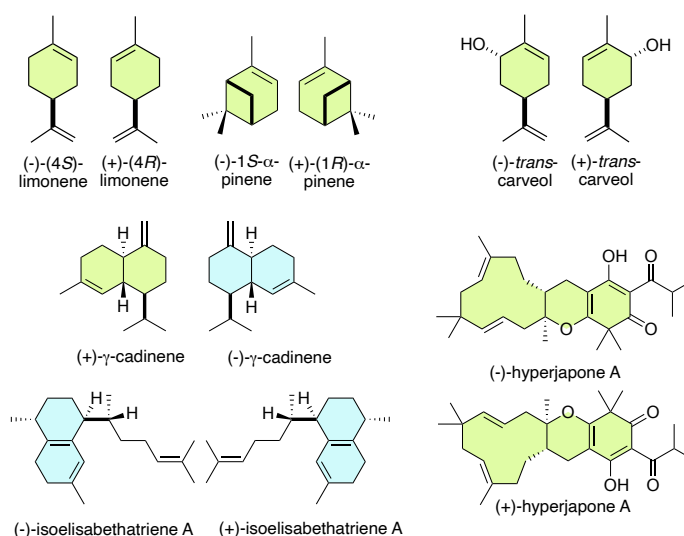


Figure 1. Examples of natural products with known enantiomers. (±)-limonene and (±)-α-pinene are monoterpenes, (±)-γ-cadinene are sesquiterpenes, (±)-isoelisabethatriene A are diterpenes, (±)-trans-carveol are monoterpeneoids, and (±)-hyperjapone A are hypothesized to be formed from the reaction between the achiral sesquiterpene humulene and a phloroglucinol intermediate via a Diels-alderase.¹ Green colored rings encode molecules from plant sources, while blue color encodes molecules from microbial sources.

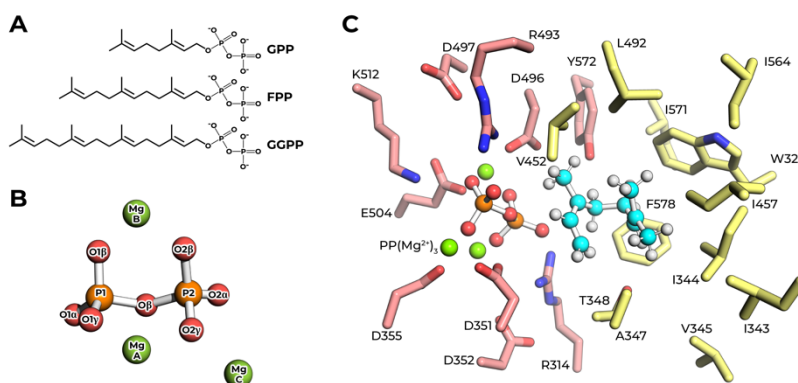


Figure 2. (A) Different TPS substrates: GPP, FPP, and GGPP. (B) Naming convention for pyrophosphate and magnesium atoms. (C) Biphasic active site of class I TPS. Polar residues with carbons are shown in pink, hydrophobic residues with carbons are shown in yellow, carbocation carbons are colored cyan, and magnesium ions are colored green.

restricted to mono-, sesqui-, and, rarely, di-terpenes and terpenoids.^{7, 22, 29-31} The enantiomeric determinants in TPS are not well understood, although many terpenes have been fully characterized and in the case of limonene (**Figure 1**) both crystal structures have been resolved.³²⁻³³ This is partly due to the intricate chemistry involved in TPS, where the enzyme's specific role in the detailed reaction mechanism often remains unclear.

Recently, we employed an integrated approach combining structural, bioinformatics, and EnzyDock mechanistic docking tools to address the ligand binding in class I TPS.¹⁵ In class I TPS the reaction is initiated via heterolytic C-O bond cleavage. We brought initial data suggesting that TPS bind their ligand in a binary mode: connecting the isoprenoid moiety to

either O1 α or O2 α of the diphosphate (PP) moiety (**Figure 2, 3**). We were also able to show that our docking approach (EnzyDock^{14,34}) was able to correctly predict this binary binding mode preference in all cases tested.¹⁵ This new ligand binding rule is rooted in evolutionary differences between TPS,³⁵ and we brought evidence that this alteration in binding, and subsequent chemistry, is due to TPS originating from plants (*p*TPS) or microorganisms (*m*TPS).

Importantly, we further suggested that this difference can cast light on the frequent observation that the chiral TPS products or intermediates of plant³⁶ and bacterial³⁷ terpene synthases represent opposite enantiomers.^{22, 37} For instance, isolation of enantiomeric sesquiterpenes has revealed that terrestrial and marine plant sources sometimes produce opposite enantiomers.³⁸⁻³⁹ A fascinating example involving a monoterpene is the biosynthesis of 1,8-cineol by Cineol Synthase (**Figure 4**).^{31, 40} Although the final product is achiral, the reaction mechanism proceeds via the chiral terpinyl cation intermediate, as shown by isotope labeling experiments. In this case, the (*R*)-terpinyl cation is formed in *Salvia officinalis* (plant),⁴⁰ while the bacterial enzyme (*Streptomyces clavuligerus*, *sc*) proceeds via the (*S*)-terpinyl cation.³¹ Analysis of the active site from crystal structures and EnzyDock docking studies, suggested that the active site architectures in these plant and microbial enzymes have evolved to accommodate different enantiomers (**Figure 4**).

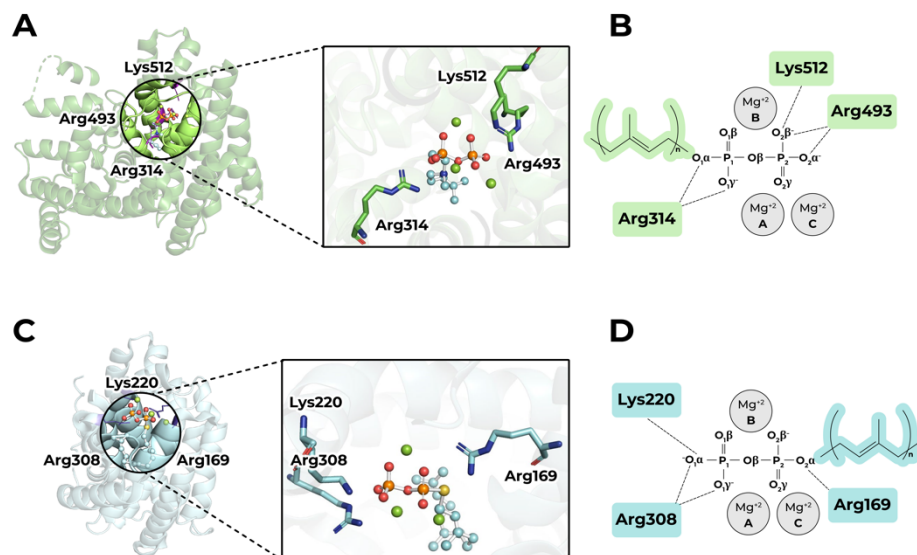


Figure 3. Top: (A) *Salvia officinalis* bornyl diphosphate synthase (BPPS), an example of a **plant TPS**, where the isoprenoid group is connected via O1 α (PDB ID: 1n21). **(B)** Naming convention for positively charged residues binding the diphosphate group for plant TPS follows the residue numbers in *Salvia officinalis* BPPS. **Bottom: (C)** *Aspergillus terreus* aristolochene synthase (AS), an example of a **microbial TPS**, where the isoprenoid group is connected via O2 α (PDB ID: 4kux). **(D)** Naming convention for positively charged residues binding the diphosphate group for microbial TPS follows the residue numbers in *Aspergillus terreus* AS.

Yet, many examples exist of opposite enantiomers both hailing from the same source, like the monoterpenes (+)-limonene (*Citrus sinensis*) and (-)-limonene (*Mentha spicata*) both hail from plant sources and both bind to the O1 α position of the pyrophosphate (Figure 2).^{32-33,41} Additionally, two α -pinene synthases forming opposing enantiomers were isolated from *Pinus taeda*,⁴² and two germacrene D-synthases from *Solidago canadensis* form both (+)- and (-)-germacrene D.⁴³ Hence, the same species often form different enantiomers, and the factors determining chirality in TPS are likely complex and multifaceted. It is the goal of this proposed project to explore the enantiomeric determinants in terpene biosynthesis in TPS and in subsequent terpene binding and functionalization in P450 enzymes.

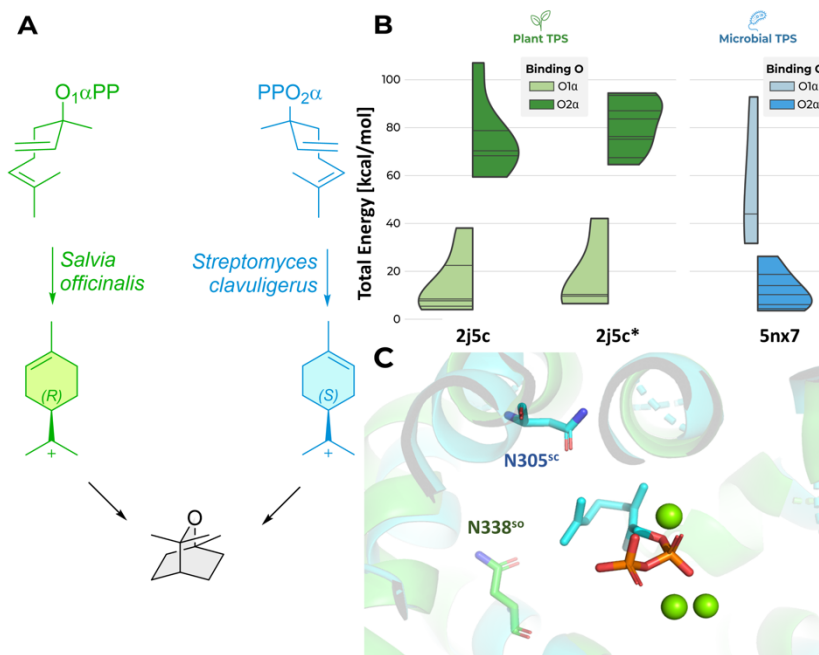


Figure 4. (A) The biosynthesis of 1,8-cineol by Cineol Synthase. Different enantiomeric intermediates yield an achiral product. (B) EnzyDock docking reveals preference for O1 α binding in pTPS (with and without(*) an active site water) and O2 α binding in mTPS. (C) Active site of Cineol Synthases pTPS (PDB ID: 2j5c, green) and mTPS (PDB ID: 5nx7, cyan). The presumably catalytically important Asn residue is in flanking positions in pTPS (N338) and mTPS (N305).

2. Research Objectives and Expected Significance

Little is known about the intricate genetic, biocatalytic, and mechanistic details of how enantiomeric natural products are formed.⁴⁴ Modifications of chemoselectivity in TPS, P450, and other natural product enzymes require the change of only a few amino acids, but the sequence differences between antipodal synthases show that the alteration of substrate enantioselectivity is the result of significant sequence differences.⁴² The objective of this project is to explore the enantiomeric determinants in terpene biosynthesis (in class I TPS) and subsequent terpene binding and functionalization (in P450). To this end, we plan to utilize an integrated molecular modeling and artificial intelligence (AI) approach to explore key differences between enzymes forming natural product enantiomers. We will construct annotated databases from collected data of known terpenes and terpenoids, the protein sequences of known TPS and relevant P450s, and available crystal structures from databases and the literature. We will apply computational biology and chemistry modeling and data mining approaches to augment and extract knowledge from these databases. This knowledge will include correlations between enzyme chemistry (e.g., mechanisms, role of enzyme in catalysis, and formation of enantiomers) and enzyme taxonomy and physical attributes (e.g., active

site structure, conserved 3D motifs). The databases and knowledge created during this project will be made available via a cloud-based web site.

Understanding the determinants of the biosynthesis of enantiomeric natural products is critical for many areas of science, like drug development, enzyme catalysis and design, biodiversity and evolution, molecular sensors, and synthetic biology and biotechnology. Hence, we expect the knowledge regarding the determinants of terpene and terpenoid biosynthesis and in particular enantiomeric specificity, will be important for a broad scientific audience.

3. Detailed Description of the Proposed Research

3.1. Working Hypothesis

The chirality of secondary metabolites formed in nature is determined by the enzymes that synthesize them, and specifically the active site architecture and dynamics, substrate positioning, and the detailed reaction mechanism. These, in turn, are shaped by factors ranging from evolutionary pressure on the organism to form biologically important enantiomers to the inherent chemistry of the reaction being catalyzed. Hence, to appreciate the determinants of enantiomer formation, it is necessary to understand both the taxonomy of the enzymes and the enzyme reactions in atomic detail. As described in the Scientific Background (1.2), we recently discovered a deep connection between the substrate binding mode in TPS and the enzyme's taxonomy, which might be connected to enantiomeric prevalence in plants and microorganisms.¹⁵ Additional factors that might play a role in enantiomeric preferences in TPS enzymes are substrate fold in the active site; *R/S* preferences for the chiral substrates linalyl-PP (mono-TPS), nerolidyl-PP (sesqui-TPS), or geranyl-linalyl-PP (di-TPS) which are required substrate intermediates in many mechanisms; and *Re/Si* approach of cations to double bonds.

In the following, we will detail a research program designed to further our understanding of the biological formation of enantiomers of terpenes and terpenoids, which together constitute the largest family of natural products.

3.2. Project Design and Methods

In this research project we will construct annotated databases of natural products, natural product enzymes, and their catalytic reactions. We will apply computational biology and chemistry modeling and data mining approaches to augment and extract knowledge regarding catalysis, and in particular enantiomer formation, from these databases. *We stress that our analysis will cover all terpenes and terpenoids; not just those with identified enantiomers.* The overall project design is presented in Figure 5 and a detailed description of the methods appears below.

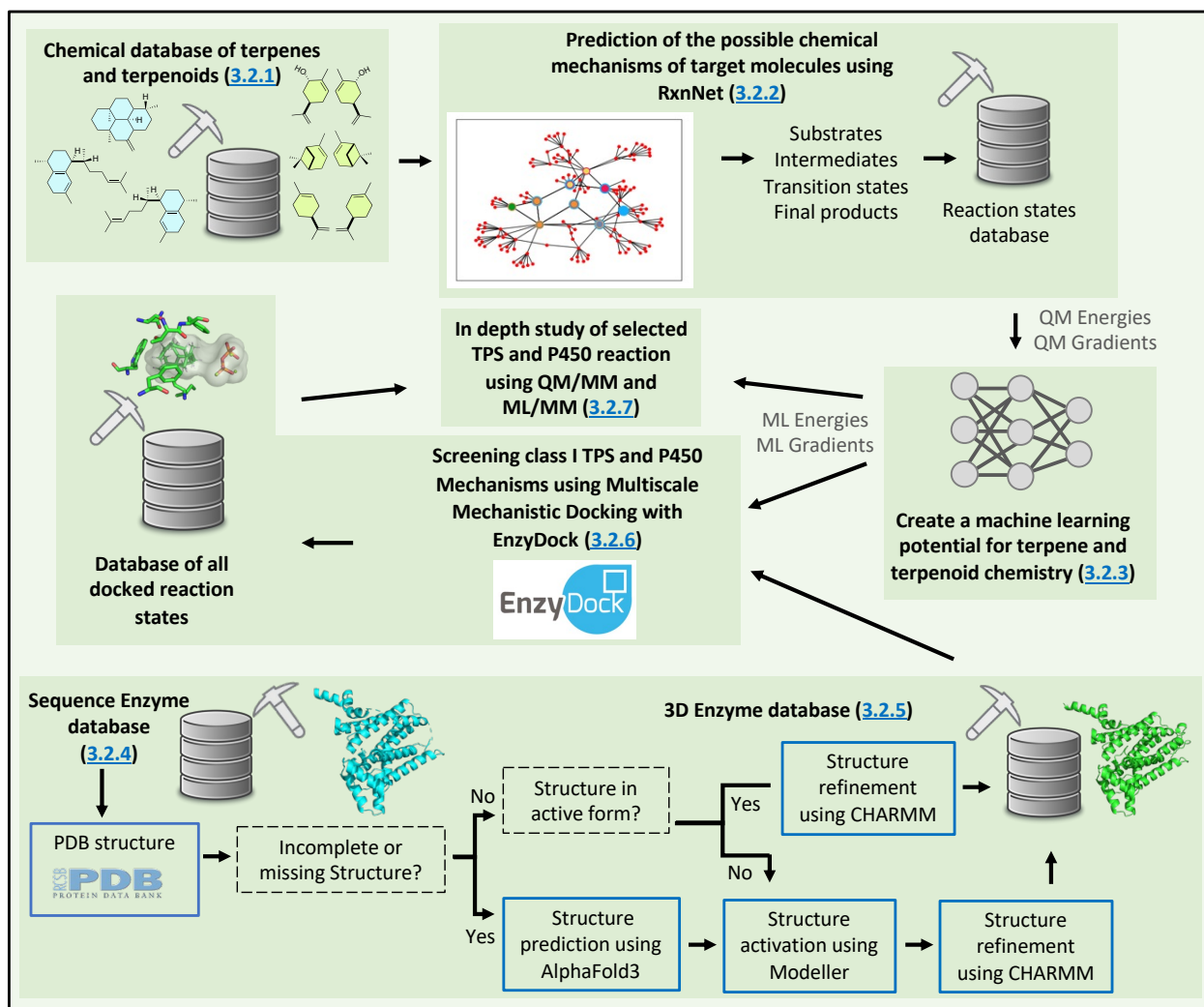


Figure 5. Workflow showing relation between project parts and databases generated. Databases are shown as stacks of grey disks and the adjacent pickaxe represents data mining. **(3.2.1)** Collect all known mono-, sesqui-, and di-terpenes and terpenoids from available databases and literature, and create an annotated database. Both enantiomeric antipodes will be deposited if available. **(3.2.2)** Create all possible mechanisms for the terpenes (e.g., carbocation reactions) and terpenoids (e.g., oxidation reactions) using RxnNet and compute energies for all stationary points using quantum chemistry methods (substrates, intermediates, transition states, products), and create an annotated database and link to products database. **(3.2.3)** Create machine learning (ML) potential energy models for terpene and terpenoid chemistry using the energies from the reaction states database and augmented by energy calculations of additional non-stationary points. **(3.2.4)** Create an annotated local sequence database for TPS and P450 enzymes and link to products and possible mechanisms in case these are known. **(3.2.5)** Create 3D structure database for the TPS and P450 enzymes in the sequence database. A structure that is not an active form (e.g., apo-state or not fully closed holo-state), will be “activated” using specialized modeling techniques described in the text. If a crystal structure exists in active structural form, it will be completed using Modeller (if missing residues) and refined using CHARMM and added to the 3D enzyme database. If a crystal structure exists in inactive structural form,

it will be modelled into an active form using Modeller and CHARMM. If no crystal structure exists, it will be modelled using AlphaFold3 and modelled into an active form using Modeller and CHARMM as needed. (3.2.6) Using the reaction mechanism states and 3D enzyme structures databases, in conjunction with the ML potential, we will perform mechanistic docking using EnzyDock. The docking results will be curated and the most likely docked mechanisms will be deposited to the docked reaction states database. (3.2.7) Detailed quantum mechanics-molecular mechanics (QM/MM) and ML/MM free energy simulations of TPS and P450 enzyme mechanisms where both enantiomers are known.

3.2.1. Create and mine a chemical database of known terpenes and terpenoids. We will collect all known mono-, sesqui-, and di-terpenes and terpenoids originating from TPS and P450 enzymes from available databases and literature, and create an annotated database. Both enantiomeric antipodes will be deposited if known. As a starting point, we will use the excellent Natural Product Atlas,⁴⁵ MARTS-DB (www.marts-db.org), and TeroKit databases as starting points, which have > 150,000 terpenoids.⁴⁶ Additionally, we will use a range of standard search tools like PubChem, Reaxys, SciFinder, ChemSpider, and Web of Science to identify additional terpenes and terpenoids. The dataset will be annotated with information like molecule identification (e.g., name, molecular formula, SMILES, etc.), classification (e.g., mono-, sesqui-, di-terpene), structural data (e.g., 2D, 3D structures, chiral centers and enantiomeric characterization), origin (plants, fungi, or bacteria), biosynthetic pathway (substrate, pathway, enzymes involved), biological/ecological/pharmaceutical activity, and literature references. The database will also include links to other databases in this project, like the reaction mechanism database. Links to external publicly available databases, like PubChem, will be included where much additional information is available.

Based on this information we will apply machine learning (ML) approaches to obtain insights into the distribution of terpene and terpenoid enantiomers in nature. The ML methods will include clustering algorithms, which can reveal patterns of enantiomer distributions across different taxonomic groups (e.g., k-means and hierarchical clustering), classification algorithms (e.g., Support Vector Machines (SVM), Random Forests (RF), k-Nearest Neighbors (k-NN), and dimensionality reduction (e.g., Principal Component Analysis (PCA) or t-SNE). If sufficient data is available, deep learning (DL) tools will be applied. These tools are commonly used in our group.

3.2.2. Create and mine a chemical database for reaction mechanisms leading to terpenes and terpenoids using literature, RxnNet, and quantum mechanical calculations. We will create plausible mechanisms for the terpenes (carbocation reactions) and terpenoids (oxidation reactions) and compute energies for all stationary points using our RxnNet approach (see below) and quantum chemistry methods in the gas-phase⁴⁷⁻⁴⁸ and chloroform solvent, which has been used in nano-capsules terpene synthesis.⁴⁹⁻⁵⁰ The reaction mechanism information will be used to create an annotated reaction database, which will be used by our docking approach EnzyDock (see 3.2.6 below). The RxnNet mechanisms will be carefully compared

with the vast number of mechanisms proposed in earlier literature (experimental and computational), e.g. ref. ^{22, 48, 51-52} and in the MARTS-DB database (www.marts-db.org) which has collected > 1,500 mechanisms from the literature. We stress that extensive earlier work has shown that computational modeling of gas-phase reactions for these systems can provide crucial insights into the enzymatic process.^{17, 48}

Inspired by earlier work,⁵³⁻⁵⁶ we recently developed a carbocation reaction tree generator Python code called RxnNet (Figure 6), which can automatically generate any carbocation intermediate based on the SMILES string of a substrate,⁵⁷ using predefined, known chemical steps encoded as SMARTS strings (e.g., cyclizations, migrations, rearrangements, proton and hydride transfers) (several manuscript describing RxnNet are in preparation).

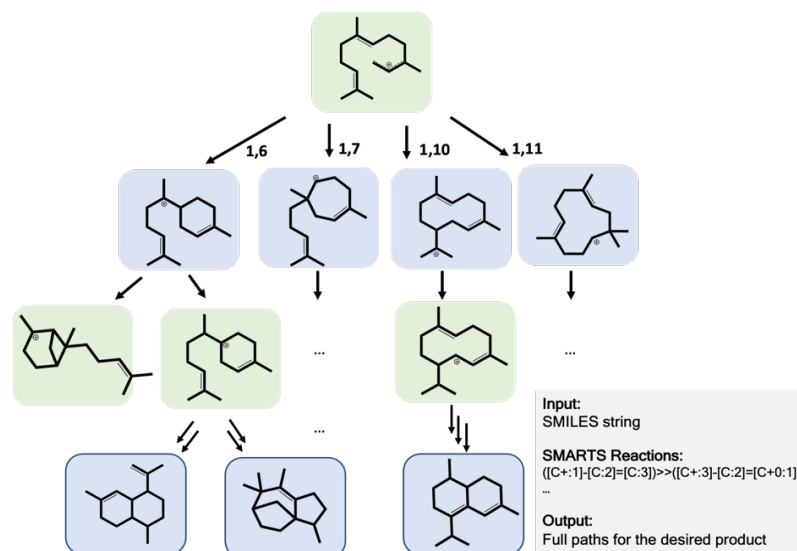


Figure 6. Example of reaction tree generated for a sesquiterpenyl cation, which is formed from farnesyl diphosphate. The first reaction step shows all possible cyclizations, which allows distinction between different TPS according to sequence.

These steps are combined to generate distinct mechanisms, including substrates, intermediates, transition states, and products. Special care is taken to allow for chirality and isotope labeling⁵⁸ in RxnNet, so that all fine details of the chemical reactions are accounted for. RxnNet is written in Python and makes extensive use of the RDKit library (www.rdkit.org/), and it includes automated dispatching of quantum chemical calculations using the xTB, Gaussian 16 and ORCA programs. RxnNet performs conformational search using classical force fields, followed by semi-empirical (SE) (GFNn-xTB) or ML (e.g., AIMNet2) calculations on the lowest energy conformers followed by final density functional theory (DFT) scoring using M062X or ω B97M-V with suitable basis sets (e.g., def2-TZVPP), as these are robust methods (e.g., see Table 1 in ⁵⁹) and can reliably treat terpene chemistry as we have shown in a recent study.⁶⁰ This protocol identifies both intermediates using standard geometry optimization and transition states using a novel combination of a geodesic initial reaction pathway⁶¹ guess followed by the climbing image nudged elastic band (NEB) algorithm.⁶² In the current project, the RxnNet program will be expanded to include P450 enzymatic reactions that modify terpenes (e.g., hydroxylation, epoxidation, desaturation, oxidative ring formation and rearrangement, and C-C bond cleavage). We will employ the RxnNet program to map the reaction mechanisms for the reactions in the chemical database of known terpenes and terpenoids that are generated by TPS and P450 enzymes (section 3.2.1).

The database will include 1D, 2D, and 3D information about all reaction states (substrates, intermediates, transition states, products) of the TPS and P450 products. The database will be augmented by

comparison and references to the extensive reaction information that already exists in the literature (see e.g.,^{22, 48, 51}). The three-dimensional structures (e.g., Cartesian coordinates) for intermediates and transition states, together with models for the substrates and the products, will be used for docking (**3.2.6**) in enzymes from the three-dimensional protein structure database (**3.2.5**), as described below.

We will apply ML/DL approaches (e.g., see list of methods in **3.2.1**) to identify possible correlations between details of the most plausible reaction mechanism in the gas-phase and in chloroform with information from the chemical database of terpenes and terpenoids (e.g., origin, biosynthetic pathway, biological/ecological/pharmaceutical activity). For instance, previous work has identified correlation between the initial ring closure step in sesquiterpene biosynthesis and product origin.⁶³⁻⁶⁴

3.2.3. Create a ML potential for efficient and accurate large-scale docking using EnzyDock and free energy simulations.

To be able to rapidly screen many reaction pathways in EnzyDock (see **3.2.6** below) and perform free energy simulations (see **3.2.7** below), it will be advantageous to employ fast, yet accurate, potentials. We will create ML potential energy models for terpene and terpenoid chemistry using the energies from the reaction states database (see section **3.2.2**) and augmented by energy calculations of additional non-stationary points (e.g., using random coordinate perturbations and molecular dynamics (MD) simulations). This will be extended to also include ML/MM potentials based on data from QM/MM calculations performed for these systems in the past,^{17, 19, 65-66} as well as calculations we will perform herein. Our strategy will follow the ML and ML/MM approach presented in ref.⁶⁷, which has been used in enzyme calculations. In the ML/MM implementation, each ML atom is represented by its local environment matrix to capture the internal interactions within the ML region (learned from the QM interactions), while the MM atoms are represented as point charges without atomic identities for the ML–MM electrostatic interactions and the ML–MM van der Waals interactions are treated at the MM level. The PI (D. T. M.) spent four weeks this past summer with the authors of⁶⁷ and received detailed training in developing and using this ML and ML/MM approach. Considering the complexity of the chemistry involved in TPS and P450 chemistry, we will not attempt to learn the potential energy surface from scratch. Rather, we will adopt a Δ -ML approach using a SE method (e.g., DFTB⁶⁸⁻⁶⁹) as the basis for energy calculations as suggested in⁶⁷, and ML will be used to learn the difference between the SE method and the high level DFT used (e.g., ω B97M-V from section **3.2.2**). All methods described in this section, including the Δ -ML/MM methods,⁶⁷ have been implemented in the CHARMM simulation program,⁷⁰ which is the platform used by EnzyDock.^{14, 34} We will also explore the use of equivariant graph neural network (EGNN) approaches initially proposed by Welling⁷¹ and co-workers and Kozinsky and co-workers (NequIP),⁷² which are data efficient and showed excellent performance on the QM9 and MD-17 databases, respectively.⁷¹ We have recently used EGNN in a ML study.⁷³

3.2.4. Create and mine a sequence database of known TPS and P450 enzymes.

The goal of this database and accompanying analysis is to point to residues likely to be involved in stabilizing specific

carbocation skeletons and co-evolved active site residues, as well as form the basis for construction of 3D models (see **3.2.5** below). For instance, it is well established that TPS active sites are rich in aromatic residues,²⁰ which stabilize cations via π -cation interactions, but many other interactions are possible.¹⁶ We will gather sequence data and functionally annotate all TPS and P450 enzymes with known products (for sesquiterpene synthases this information exists in part⁶³ and this data will be downloaded and updated as needed).⁷⁴ Enzyme sequences (e.g., from all UniProt proteins, the OneKP transcriptome dataset and the microbial genome database) will be identified using hidden Markov models, multiple sequence alignment, followed by phylogenetic analysis, feature extraction (e.g., position-specific scoring matrix and position-specific frequency matrix), and clustering. This will classify the enzymes, e.g., distinguish between mono-, sesqui-, and diterpene synthases using methods like Terzyme⁷⁵ or the convolutional neural network-based DeepEC,⁷⁶ which is freely available Python code. The mono-, sesqui- and di-TPS will be grouped separately according to the known initial carbocation cyclization intermediates, as the initial cyclization determines the initial branching of the reaction mechanism, using our mechanistic database (**3.2.2**). Subsequent clustering of sequences according to ensuing carbocation reaction steps will be attempted, with input from the results from RxnNet (**3.2.2**). The P450 synthases will be annotated according to the kind of chemical reaction they perform. As a starting point, we will analyze existing plant⁷⁷ and bacterial⁷⁸ P450 databases.

We will link the sequence database with the chemical database of known terpenes and terpenoids (**3.2.1**) and the chemical database for reaction mechanisms (**3.2.2**). The current approach expands on excellent previous work on sesquiterpenes, where the focus was correlation between the initial cyclization step in sesquiterpene formation and taxonomy.⁶³⁻⁶⁴

3.2.5. Create and mine a 3D structure database for known TPS and P450 using available crystal structures and refined AlphaFold3 models. We will create a 3D structure database for the TPS and P450 enzymes in the sequence database (see previous section, **3.2.4**). 3D structural enzyme information for the TPS and P450 families is important both for understanding the chemical mechanism of how terpenes and terpenoids are formed, and in particular enantiomer formation, and for potential engineering of the enzymes to design new product portfolios.¹⁶⁻¹⁷ We have previously generated an initial dataset of selected class I mono-, sesqui-, and di-TPSs in their catalytically fully closed, i.e., “active” form (i.e., fully closed holo state with key conserved hydrogen bonds intact).¹⁵ Structure “activation” was achieved by enforcing a set of conserved hydrogen bonds via restrained MD simulations and minimization.¹⁵ This dataset will be expanded to include the TPS and P450 enzymes in the sequence database (see previous section, **3.2.4**). The enzymes will be modeled in their catalytically competent, “active” state. This will be achieved by using a workflow combining available crystal structures, AlphaFold3⁷⁹, and Modeller⁸⁰ (**Figure 5**). Whenever possible, crystal structures will be employed. If these experimental structures are either of the apo enzyme or the holo enzyme not in a fully catalytically competent state, the experimental structure will be modeled (i.e., “activated”) to generate the fully active enzyme, as we have shown in the past for, e.g., taxadiene synthase^{65, 81} and other TPS.¹⁵

AlphaFold3 will be used to generate initial models in the many cases where no experimental crystal structure exists. However, these structures might not be accurate enough for TPS modeling at the atomic level, as we have recently indicated.⁸² Hence, subsequent refinement, or “activation”, of the model is necessary. This will be achieved using Modeller, relying on existing TPS structures resolved in their active state, e.g., for TPS the *Salvia officinalis* bornyl diphosphate synthase (PDB ID: 1n21⁸³) and *Aspergillus terreus* aristolochene synthase (PDB ID: 4kux⁸⁴) are good template structures. For P450 enzymes, the structures of P450cam (PDB: 1DZ4,⁸⁵ 2CPP,⁸⁶ 4WJS⁸⁷ for monoterpenes) and CYP76AH1 (PDB: 5YM3 for diterpenes⁸⁸) are relevant and will be used for modeling. Specific enzyme regions will be remodeled as needed to place highly conserved residues in positions required for catalysis, as identified by high quality crystal structures and prior modeling studies in our group. The final structure models will be relaxed using MD simulations and minimization using the CHARMM program,⁷⁰ which allows simulations with many restraints that will keep important conserved protein-protein and protein-ligand interactions⁸⁹ intact during the simulations. Tasks automation will be achieved by writing scripts in Python, Linux, and CHARMM, as we have done previously.¹⁵

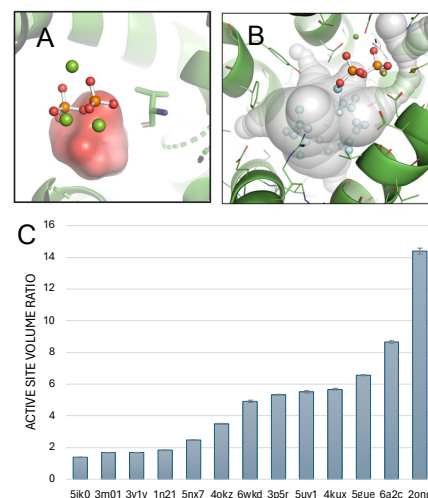


Figure 7. (A) Electrostatic potential in a TPS. **(B)** Active site volume in a TPS. **(C)** Asymmetry of active sites in several TPS, defined as the ratio between active site volume near the bound O (O1 α for pTPS, O2 α for mTPS) and the other O (O1 α for mTPS, O2 α for pTPS) (see **Figure 3**).

We will apply ML techniques along with carefully crafted descriptors to analyze the active sites in TPS and P450 enzymes. The enzymes in these families will be aligned in 3D according to CATH fold classification (<https://www.cathdb.info/>) to obtain meaningful spatial alignments. 3D descriptors include active site volume, active site asymmetry (**Figure 7**),¹⁵ accessible surface area, active site depth, active site curvature, hydrophobicity/hydrophilicity patterns, electrostatic potential (**Figure 7**),¹⁵ and rigidity/flexibility. Amino acid physicochemical features, like ability to stabilize carbocations (as encoded in descriptors like zScales or VHRS),⁹⁰ will be extracted and clustered in 3D. This will define carbocation stabilizing residues within the active site 3D space according to TPS sub-families and binding of terpenes in P450 enzymes. Using such descriptors, we will apply ML/DL techniques⁹¹ to obtain correlation between conserved 3D features in TPS and P450 enzymes and their taxonomy and chemical function (i.e., chemistry, see **3.2.2**). We will carefully analyze and compare the differences between pairs of enzymes forming enantiomers.

3.2.6. Create and mine a 3D enzyme mechanism database of TPS and P450 enzymes with docked states (bound substrates, intermediates, transition states, and products) and database mining. We will perform mechanistic docking of the compounds in the chemical database of known terpenes and terpenoids (**3.2.1**). To this end, we will employ our EnzyDock program in conjunction with the chemical database for reaction mechanisms leading to terpenes and terpenoids (**3.2.2**) and the 3D structure database

for known TPS and P450 (3.2.5). Here we will provide a brief description of EnzyDock and how it can be an important tool for understanding the reactions in TPS and P450 enzymes.

We recently developed EnzyDock, which is a CHARMM-based docking program,⁷⁰ and has conceptual similarities with docking tools targeting multistep reactions in TPS.⁹²⁻⁹⁴ Since EnzyDock relies on CHARMM functionalities, it benefits from decades of development by the CHARMM development team.⁷⁰ EnzyDock includes a series of protocols to predict the chemically relevant orientations, conformations, and energies of reaction coordinate states (Figure 8). The main feature incorporated into EnzyDock is mechanism-based multi-state consensus docking that allows the docking of multiple states (reaction substrate, intermediates, transition states, products) in a mechanistically consistent (i.e., consensus or similar poses) and induced-fit manner. For instance, this assures that the substrate (GPP, FPP, GGPP) in TPS folds correctly if one performs docking with the product as a template (or “seed”) for the docking of all states. We note that EnzyDock is a *docking-tool*; it does not compute free energy profiles, which can be obtained from e.g. umbrella sampling,⁹⁵ metadynamics,⁹⁶ or transition-path sampling⁹⁷ and this will be addressed below in section 3.2.7. Consensus docking is achieved by applying geometric restraints on reaction states relative to a pre-determined “seed” state, such that all states are docked with similar poses (within a given user-defined threshold), and a reaction Pathfinder module identifies all geometrically matching poses along a reaction path (Figure 8).¹⁷ MD and MC simulated annealing sampling of bound states is performed on a grid representing the enzyme, and poses are scored using the CHARMM36⁹⁸ and CGenFF⁹⁹ force fields (FF) and optional refinement using QM/MM³⁴ with a range of QM methods (e.g., SE, DFT). Solvation is modeled using implicit solvation and explicit waters may be included. We have applied EnzyDock to diverse enzyme systems, such as terpene synthases,^{14-17, 25, 34} racemases, Diels-Alderase, covalently bound ligands,^{17, 34} and the main protease in SARS-CoV-2.¹⁰⁰

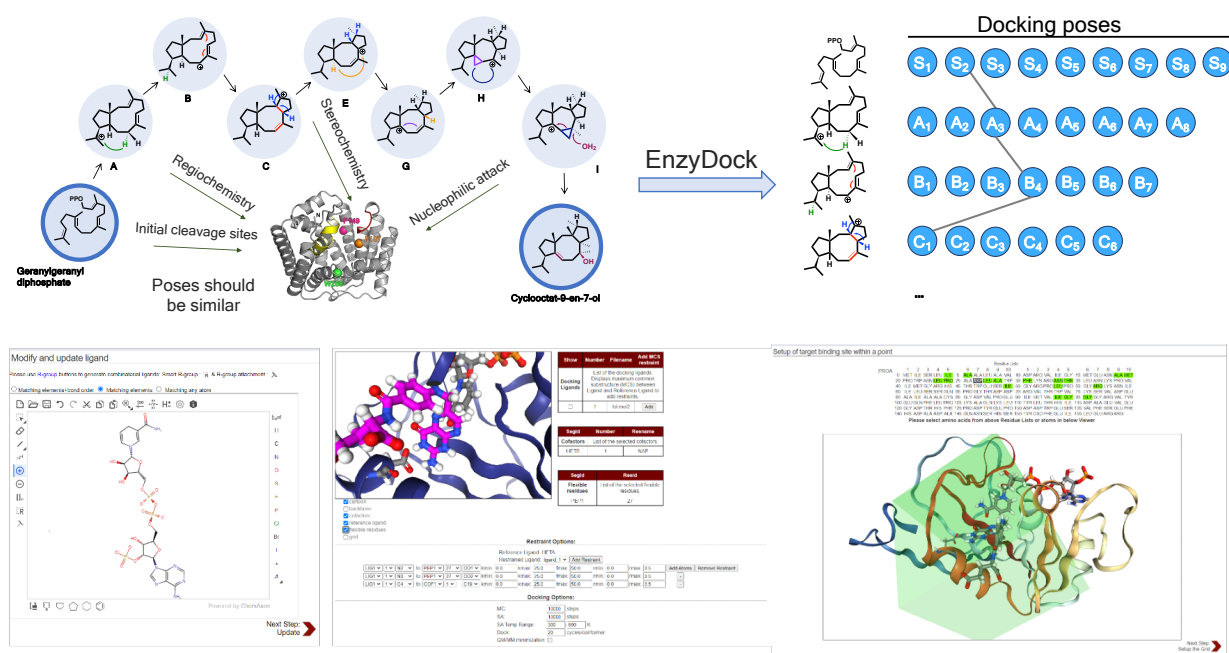


Figure 8. Main concepts of EnzyDock. **Top Left:** Dock multiple states into enzyme (substrate, **S**, intermediates **A-I**, product) with chemical information like regio- and stereochemistry encoded via restraints. **Top Right:**

Large sets of poses are obtained, and geometrically matching poses are found via the Pathfinder program (grey line). If a “seed” state is defined (e.g., here the substrate, **S**), restraints are applied on subsequent states to enforce poses similar to **S**. **Bottom:** EnzyDock implementation in CHARMM-GUI.¹⁰¹ **Left:** Ligand building. **Middle:** Defining restraints. **Right:** Defining docking grid and flexible active site residues.

In the current project we will score poses using the Δ -ML/MM potential developed in section **3.2.3**. This will allow us to rapidly and accurately score binding modes of substrate, intermediates, transition states, and products for probable mechanisms for each enzyme in the 3D structure database for known TPS and P450 at a low cost. To allow rapid screening of the many ligand states involved in mechanistic docking, we have already implemented an interface between RxnNet and EnzyDock, which allows providing EnzyDock with Python Pickle files of mechanisms from RxnNet. The docking results will be curated and compared with available mutational data in the literature for cases where this exists. The most probable docked mechanisms will be deposited to the docked reaction database, which will be mined (see **3.2.8** below). We will analyze and compare the docking results for pairs of enzymes forming enantiomers. In TPS enzymes, this includes substrate fold in the active site and oxygen binding preference ($O1\alpha$ vs $O2\alpha$, see **Figures 2, 3**); *R/S* preferences for the chiral substrates linalyl-PP (mono-TPS), nerolidyl-PP (sesqui-TPS), or geranyl-linalyl-PP (di-TPS) where relevant for mechanism; and *Re/Si* approach of cations to double bonds. In P450s we will look for differential binding of enantiomeric terpenes and functionalization creating new chiral positions leading to enantiomers.

3.2.7. Complete mechanistic studies. We will perform in depth study of selected TPS and P450 reactions where both enantiomers are known using QM/MM and Δ -ML/MM potentials and free energy simulations. We will compute the potential of mean force profiles using multidimensional umbrella sampling,⁹⁵ which we have used extensively for TPS reactions and is a well-established approach. We will compare the QM(ω B97M-V)/MM and Δ -ML/MM free energy profiles for selected reactions and if necessary, improve the Δ -ML/MM by further training the ML model (transfer learning). A suitable system for this purpose is limonene synthase, for which crystal structures exist for both (-)-4S- and (+)-4R-limonene synthases in their holo-form.³²⁻³³ Once the accuracy of the Δ -ML/MM has been validated, we will use this potential exclusively due to its efficiency. The accuracy of the simulations will be validated against experimentally observed product distributions. Based on the simulation data, we will identify patterns in the enantio-selectivity role played by active site residues and cofactors (PP, Mg^{2+} ions) obtained from the complete mechanistic studies and docking (**3.2.6**).

3.2.8. Data integration and mining. In this project, we will produce annotated databases of terpenes and terpenoids and their enzymes (sequences and 3D structures), biosynthesis reactions, and structural models of all reaction states for all reactions studied. **This is a wealth of information that will be mined together to obtain deeper knowledge of TPS and P450 enzyme catalysis in general and enantiomeric specificity in particular.** To this end, we will adopt the novel approach in ref. ⁹¹, termed EzMechanism, which can automatically infer mechanistic paths for a given 3D active site and enzyme reaction, based on a set of

catalytic rules compiled from the Mechanism and Catalytic Site Atlas, a database of enzyme mechanisms. Currently this atlas only contains five TPS reactions, which is far too few to allow efficient learning of complex TPS and P450 reactions. Additionally, the proposed TPS mechanisms in the atlas are not curated and are not always in line with the consensus view in the literature (e.g., for trichodiene synthase⁵¹). Here we propose to combine the extensive mechanistic information generated in this proposal based on the catalytic rules in RxnNet (3.2.2), together with 3D enzyme models (3.2.5) and docked reaction states (3.2.6), to allow greatly enhanced generation of TPS and P450 reaction rules and hypotheses. Our extensive experience in modeling TPS reactions will allow us to generate highly specific TPS rules, both chemical rules like in RxnNet and rules of the role of the enzyme in guiding reaction cascades, and particularly enantioselectivity. Once applied to TPS reactions, we will apply the same approach to P450 reactions. We expect that the rules and hypotheses generated by this model will provide deep insight into differences between enantiomer catalysis and will also clarify areas where we have insufficient knowledge.

3.2.9. Data Management. The data collected will initially be arranged in local databases that will be hosted on our servers at Bar-Ilan University and once a database is complete will be hosted in the cloud for public access (e.g., Amazon Web Service or Google Cloud). The databases will be written in Python using cloud-ready databases like PostgreSQL, MySQL, or MongoDB. The database curation will include data collection and selection, data cleaning, annotation and enrichment, verification and validation, and documentation. Links between databases will be implemented using Python libraries like Foreign Data Wrappers (for PostgreSQL and MySQL) or Database References (for MongoDB). All databases will be hosted on a single website; since the focus of this project is basic science, the initial website will be simple with limited features.

4. Potential caveats

(1) The ML potential might not be accurate and general enough to treat the complex carbocation and oxidation reactions encountered in this project. We will address this by trying different ML architectures, as well as increasing the amount of training data. However, if the ML potentials accuracy is insufficient, we will employ traditional EnzyDock docking scoring and QM(ω B97M-V)/MM for simulations. This means a smaller number of systems can be studied using free energy simulations, but this is not likely to impact the general conclusions of the project. (2) We might not identify new rules for biosynthesis of enantiomers. Indeed, this is a possibility. However, this in itself is an important finding, as it would underscore the complexity of the problem. Furthermore, the wealth of annotated data, mechanistic and structural data, and deep knowledge generated regarding TPS and P450 biosynthesis will be highly valuable for the scientific community.

5. Expected Outcome and Dissemination

This project will generate searchable databases with a basic, easily accessible web-interface that will serve the scientific community. The databases will include:

- (1) Annotated terpenes and terpenoids, which includes enantiomeric information and all relevant information regarding the enzyme responsible for its synthesis (section 3.2.1).
- (2) Terpene and terpenoid reactions, which includes all possible reaction mechanisms towards products in the database, including the presumed dominant mechanism. This will include accurate DFT energies for the gas-phase and chloroform reactions towards all products (section 3.2.2).
- (3) Enzyme structure database of TPS and P450 enzymes in closed catalytically active form (section 3.2.5).
- (4) Enzyme mechanism database with all enzymes in closed catalytically active form with all docked states (substrates, intermediates, transition states, and products) (section 3.2.6).

Additionally, we will further develop the freely available EnzyDock and RxnNet programs. Finally, we will generate a wealth of knowledge and identify rules for enantiomeric determinants in TPS and P450 enzymes.

6. Preliminary Results

The **EnzyDock** program developed in our group has been tested on relevant systems.^{17, 34, 100} Pipelines for streamlined work with protein preparation, structural alignment, modeling and docking of large data sets already exist in our group (Figure

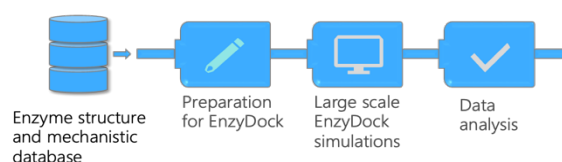


Figure 9. Workflow for automated EnzyDock simulations of database of enzymes and mechanisms for TPS and P450.

9).^{17, 100} The **RxnNet** program is a fully functional code and is ready for use in this project.

A key question in this project is how many terpenes and terpenoids have *known enantiomers and known enzymes*. We performed a preliminary, non-exhaustive search of databases and the literature, and identified significant number of such cases: monoterpenes (6), sesquiterpenes (12), and diterpenes (3). This is sufficient to initiate this project.

7. Group details

PI experience. We have the required expertise and experience for the proposed study. We have extensive experience in enzyme modeling using enhanced sampling techniques, like umbrella sampling, and we have extensive experience in multiscale simulations.^{17, 25, 102-103} Our research has focused intensely on TPS modeling (e.g.,^{14-17, 25, 34}). We have the required experience in ML modeling and ML potentials.^{15, 73, 100} **Available**

Resources: Hardware. Each researcher has a desktop or laptop computer. The group has several Linux clusters. In total, the group possesses over 50 compute nodes with a total of ca. 2,000 cores and 5 GPU nodes.

Software. CHARMM, Schrodinger suite of modeling programs, Gaussian, Q-Chem, ORCA, and additional software packages are available. We have licenses to the search engines mentioned in section 3.2.1.

Personnel. 9 PhD students, 3 MSc students, and 2 post-doctoral fellows. In the current project we will dedicate personnel as follows: 1 post-doc (P450, sections 3.2.1 - 3.2.9), 1 PhD student (TPS, sections 3.2.1 – 3.2.9), 1 MSc student (TPS, 3.2.6 - 3.2.7).

References:

1. Bitchagno, G. T. M.; Nchiozem-Ngnitedem, V.-A.; Melchert, D.; Fobofou, S. A. Demystifying racemic natural products in the homochiral world. *Nat. Rev. Chem.* **2022**, *6*, 806-822.
2. Miller, K. A.; Tsukamoto, S.; Williams, R. M. Asymmetric total syntheses of (+)- and (–)-versicolamide B and biosynthetic implications. *Nat. Chem.* **2009**, *1*, 63-68.
3. Challis, G. L.; Hopwood, D. A. Synergy and contingency as driving forces for the evolution of multiple secondary metabolite production by *Streptomyces* species. *Proc. Natl. Acad. Sci. U.S.A.* **2003**, *100 Suppl 2*, 14555-61.
4. Herbert, R. B., *The Biosynthesis of Secondary Metabolites*. 2nd ed. ed.; Chapman and Hall: New York, 1989.
5. Clardy, J.; Walsh, C. Lessons from natural molecules. *Nature* **2004**, *432*, 829-837.
6. Gunatilaka, A. A. L. Natural Products from Plant-Associated Microorganisms: Distribution, Structural Diversity, Bioactivity, and Implications of Their Occurrence. *J. Nat. Prod.* **2006**, *69*, 509-526.
7. Finefield, J. M.; Sherman, D. H.; Kreitman, M.; Williams, R. M. Enantiomeric Natural Products: Occurrence and Biogenesis. *Angew. Chem. Int. Ed.* **2012**, *51*, 4802-4836.
8. Appendino, G. The phytochemistry of the yew tree. *Nat. Prod. Rep.* **1995**, *12*, 349-360.
9. Köksal, M.; Jin, Y.; Coates, R. M.; Croteau, R.; Christianson, D. W. Taxadiene synthase structure and evolution of modular architecture in terpene biosynthesis. *Nature* **2011**, *469*, 116-120.
10. Mazzotta, S.; Rositano, V.; Senaldi, L.; Bernardi, A.; Allegrini, P.; Appendino, G. Scalemic natural products. *Nat. Prod. Rep.* **2023**, *40*, 1647-1671.
11. De los Santos, Z. A.; Wolf, C. Optical Terpene and Terpenoid Sensing: Chiral Recognition, Determination of Enantiomeric Composition and Total Concentration Analysis with Late Transition Metal Complexes. *J. Am. Chem. Soc.* **2020**, *142*, 4121-4125.
12. Ran, X.; Jiang, Y.; Shao, Q.; Yang, Z. J. EnzyKR: a chirality-aware deep learning model for predicting the outcomes of the hydrolase-catalyzed kinetic resolution. *Chem. Sci.* **2023**, *14*, 12073-12082.
13. Li, Z.-L.; Pei, S.; Chen, Z.; Huang, T.-Y.; Wang, X.-D.; Shen, L.; Chen, X.; Wang, Q.-Q.; Wang, D.-X., . . . Ao, Y.-F. Machine learning-assisted amidase-catalytic enantioselectivity prediction and rational design of variants for improving enantioselectivity. *Nat. Commun* **2024**, *15*, 8778.
14. Schwartz, R.; Zev, S.; Major, D. T., Chapter Ten - Mechanistic docking in terpene synthases using EnzyDock. In *Methods Enzymol.*, Rudolf, J., Ed. Academic Press: 2024; Vol. 699, pp 265-292.
15. Schwartz, R.; Zev, S.; Major, D. T. Differential Substrate Sensing in Terpene Synthases from Plants and Microorganisms: Insight from Structural, Bioinformatic, and EnzyDock Analyses. *Angew. Chem. Int. Ed.* **2024**, *63*, e202400743.
16. Raz, K.; Levi, S.; Gupta, P. K.; Major, D. T. Enzymatic control of product distribution in terpene synthases: insights from multiscale simulations. *Curr. Opin. Biotechnol.* **2020**, *65*, 248-258.
17. Raz, K.; Driller, R.; Dimos, N.; Ringel, M.; Brück, T.; Loll, B.; Major, D. T. The Impression of a Nonexisting Catalytic Effect: The Role of CotB2 in Guiding the Complex Biosynthesis of Cyclooctat-9-en-7-ol. *J. Am. Chem. Soc.* **2020**, *142*, 21562-21574.
18. Major, D. T.; Freud, Y.; Weitman, M. Catalytic control in terpenoid cyclases: multiscale modeling of thermodynamic, kinetic, and dynamic effects. *Curr. Opin. Chem. Biol.* **2014**, *21*, 25-33.
19. Major, D. T.; Weitman, M. Electrostatically Guided Dynamics—The Root of Fidelity in a Promiscuous Terpene Synthase? *J. Am. Chem. Soc.* **2012**, *134*, 19454-19462.
20. Christianson, D. W. Structural and Chemical Biology of Terpenoid Cyclases. *Chem. Rev.* **2017**, *117*, 11570-11648.
21. Jozwiak, A.; Sonawane, P. D.; Panda, S.; Garagounis, C.; Papadopoulou, K. K.; Abebie, B.; Massalha, H.; Almekias-Siegl, E.; Scherf, T., . . . Aharoni, A. Plant terpenoid metabolism co-opts a component of the cell wall biosynthesis machinery. *Nature Chemical Biology* **2020**, *16*, 740-748.
22. Dickschat, J. S. Bacterial terpene cyclases. *Nat. Prod. Rep.* **2016**, *33*, 87-110.
23. Meguro, A.; Motoyoshi, Y.; Teramoto, K.; Ueda, S.; Totsuka, Y.; Ando, Y.; Tomita, T.; Kim, S.-Y.; Kimura, T., . . . Kuzuyama, T. An Unusual Terpene Cyclization Mechanism Involving a Carbon–Carbon Bond Rearrangement. *Angew. Chem. Int. Ed.* **2015**, *54*, 4353-4356.
24. Hong, Y. J.; Tantillo, D. J. The energetic viability of an unexpected skeletal rearrangement in cyclooctatin biosynthesis. *Org. Biomol. Chem.* **2015**, *13*, 10273-10278.

25. Driller, R.; Janke, S.; Fuchs, M.; Warner, E.; Mhashal, A. R.; Major, D. T.; Christmann, M.; Brück, T.; Loll, B. Towards a comprehensive understanding of the structural dynamics of a bacterial diterpene synthase during catalysis. *Nat. Commun* **2018**, *9*, 3971.
26. Croteau, R. Biosynthesis and catabolism of monoterpenoids. *Chem. Rev.* **1987**, *87*, 929-954.
27. Cane, D. E., Sesquiterpene Biosynthesis: Cyclization Mechanisms. In *Comprehensive Natural Products Chemistry: Isoprenoids Including Carotenoids and Stereoids*, Cane, D. E., Ed. Pergamon Press: Oxford, 1999; Vol. 2, pp 155-200.
28. Christianson, D. W. Unearthing the roots of the terpenome. *Curr. Opin. Chem. Biol.* **2008**, *12*, 141-150.
29. Lauterbach, L.; Rinkel, J.; Dickschat, J. S. Two Bacterial Diterpene Synthases from *Allokutzneria albata* Produce Bonnadiene, Phomopsene, and Allokutznerene. *Angew. Chem. Int. Ed.* **2018**, *57*, 8280-8283.
30. Lauterbach, L.; Goldfuss, B.; Dickschat, J. S. Two Diterpene Synthases from *Chryseobacterium*: Chryseodiene Synthase and Wanjudiene Synthase. *Angew. Chem. Int. Ed.* **2020**, *59*, 11943-11947.
31. Rinkel, J.; Rabe, P.; zur Horst, L.; Dickschat, J. S. A detailed view on 1,8-cineol biosynthesis by *Streptomyces clavuligerus*. *Beilstein J. Org. Chem.* **2016**, *12*, 2317-2324.
32. Hyatt, D. C.; Youn, B.; Zhao, Y.; Santhamma, B.; Coates, R. M.; Croteau, R. B.; Kang, C. Structure of limonene synthase, a simple model for terpenoid cyclase catalysis. *Proc. Natl. Acad. Sci. U.S.A.* **2007**, *104*, 5360-5365.
33. Kumar, R. P.; Morehouse, B. R.; Matos, J. O.; Malik, K.; Lin, H.; Krauss, I. J.; Oprian, D. D. Structural Characterization of Early Michaelis Complexes in the Reaction Catalyzed by (+)-Limonene Synthase from *Citrus sinensis* Using Fluorinated Substrate Analogues. *Biochemistry* **2017**, *56*, 1716-1725.
34. Das, S.; Shimshi, M.; Raz, K.; Nitoker Eliaz, N.; Mhashal, A. R.; Ansbacher, T.; Major, D. T. EnzyDock: Protein–Ligand Docking of Multiple Reactive States along a Reaction Coordinate in Enzymes. *J. Chem. Theory Comput.* **2019**, *15*, 5116-5134.
35. Syrén, P.-O., Chapter Twelve - Ancestral terpene cyclases: From fundamental science to applications in biosynthesis. In *Methods Enzymol.*, Rudolf, J., Ed. Academic Press: 2024; Vol. 699, pp 311-341.
36. Wang, Z.; Nelson, D. R.; Zhang, J.; Wan, X.; Peters, R. J. Plant (di)terpenoid evolution: from pigments to hormones and beyond. *Nat. Prod. Rep.* **2023**, *40*, 452-469.
37. Rudolf, J. D.; Alsup, T. A.; Xu, B.; Li, Z. Bacterial terpenome. *Nat. Prod. Rep.* **2021**, *38*, 905-980.
38. Melching, S.; Bülow, N.; Wihstutz, K.; Jung, S.; König, W. A. Natural occurrence of both enantiomers of cadina-3,5-diene and δ -amorphene. *Phytochem.* **1997**, *44*, 1291-1296.
39. Beechan, C. M.; Djerassi, C.; Eggert, H. Terpenoids-LXXIV: The sesquiterpenes from the soft coral *sinularia mayi*. *Tetrahedron* **1978**, *34*, 2503-2508.
40. Wise, M. L.; Urbansky, M.; Helms, G. L.; Coates, R. M.; Croteau, R. Syn Stereochemistry of Cyclic Ether Formation in 1,8-Cineole Biosynthesis Catalyzed by Recombinant Synthase from *Salvia officinalis*. *J. Am. Chem. Soc.* **2002**, *124*, 8546-8547.
41. Jongedijk, E.; Cankar, K.; Buchhaupt, M.; Schrader, J.; Bouwmeester, H.; Beekwilder, J. Biotechnological production of limonene in microorganisms. *Appl. Microbiol. Biotechnol.* **2016**, *100*, 2927-2938.
42. Phillips, M. A.; Wildung, M. R.; Williams, D. C.; Hyatt, D. C.; Croteau, R. cDNA isolation, functional expression, and characterization of (+)- α -pinene synthase and (–)- α -pinene synthase from loblolly pine (*Pinus taeda*): Stereocontrol in pinene biosynthesis. *Arch. Biochem. Biophys.* **2003**, *411*, 267-276.
43. Prosser, I.; Altug, I. G.; Phillips, A. L.; König, W. A.; Bouwmeester, H. J.; Beale, M. H. Enantiospecific (+)- and (–)-germacrene D synthases, cloned from goldenrod, reveal a functionally active variant of the universal isoprenoid-biosynthesis aspartate-rich motif. *Arch. Biochem. Biophys.* **2004**, *432*, 136-144.
44. Sherman, D. H.; Tsukamoto, S.; Williams, R. M. Comment on “Asymmetric syntheses of sceptrin and massadine and evidence for biosynthetic enantiodivergence”. *Science* **2015**, *349*, 149-149.
45. van Santen, J. A.; Poynton, E. F.; Iskakova, D.; McMann, E.; Alsup, Tyler A.; Clark, T. N.; Fergusson, C. H.; Fewer, D. P.; Hughes, A. H., . . . Linington, R. G. The Natural Products Atlas 2.0: a database of microbially-derived natural products. *Nucleic Acids Res.* **2022**, *50*, D1317-D1323.
46. Zeng, T.; Liu, Z.; Zhuang, J.; Jiang, Y.; He, W.; Diao, H.; Lv, N.; Jian, Y.; Liang, D., . . . Wu, R. TeroKit: A Database-Driven Web Server for Terpenome Research. *J. Chem. Inf. Model.* **2020**, *60*, 2082-2090.

47. Tantillo, D. J. Biosynthesis via carbocations: theoretical studies on terpene formation. *Nat. Prod. Rep.* **2011**, *28*, 1035-53.
48. Tantillo, D. J. Importance of Inherent Substrate Reactivity in Enzyme-Promoted Carbocation Cyclization/Rearrangements. *Angew. Chem. Int. Ed.* **2017**, *56*, 10040-10045.
49. Zhang, Q.; Tiefenbacher, K. Terpene cyclization catalysed inside a self-assembled cavity. *Nat. Chem.* **2015**, *7*, 197-202.
50. Pahima, E.; Zhang, Q.; Tiefenbacher, K.; Major, D. T. Discovering Monoterpene Catalysis Inside Nanocapsules with Multiscale Modeling and Experiments. *J. Am. Chem. Soc.* **2019**, *141*, 6234-6246.
51. Hong, Y. J.; Tantillo, D. J. Branching Out from the Bisaboly Cation. Unifying Mechanistic Pathways to Barbatene, Bazzanene, Chamigrene, Chamipinene, Cumacrene, Cuprenene, Dunniene, Isobazzanene, Iso- γ -bisabolene, Isochamigrene, Laurene, Microbiotene, Sesquithujene, Sesquisabinene, Thujopsene, Trichodiene, and Widdradiene Sesquiterpenes. *J. Am. Chem. Soc.* **2014**, *136*, 2450-2463.
52. Xu, B.; Tantillo, D. J.; Rudolf, J. D. Mechanistic Insights into the Formation of the 6,10-Bicyclic Eunicellane Skeleton by the Bacterial Diterpene Synthase Bnd4. *Angew. Chem. Int. Ed.* **2021**, *60*, 23159-23163.
53. Bergeler, M.; Simm, G. N.; Proppe, J.; Reiher, M. Heuristics-Guided Exploration of Reaction Mechanisms. *J. Chem. Theory Comp.* **2015**, *11*, 5712-5722.
54. Chow, J.-Y.; Tian, B.-X.; Ramamoorthy, G.; Hillerich, B. S.; Seidel, R. D.; Almo, S. C.; Jacobson, M. P.; Poulter, C. D. Computational-guided discovery and characterization of a sesquiterpene synthase from *Streptomyces clavuligerus*. *Proc. Natl. Acad. Sci. U.S.A.* **2015**, *112*, 5661-5666.
55. Steiner, M.; Reiher, M. A human-machine interface for automatic exploration of chemical reaction networks. *Nat. Commun* **2024**, *15*, 3680.
56. Zeng, T.; Hess, B. A., Jr.; Zhang, F.; Wu, R. Bio-inspired chemical space exploration of terpenoids. *Brief Bioinform* **2022**, *23*.
57. Klucznik, T.; Syntrivanis, L.-D.; Baś, S.; Mikulak-Klucznik, B.; Moskal, M.; Szymkuć, S.; Mlynarski, J.; Gadina, L.; Beker, W., . . . Grzybowski, B. A. Computational prediction of complex cationic rearrangement outcomes. *Nature* **2024**, *625*, 508-515.
58. Dickschat, J. S. Modern Aspects of Isotopic Labellings in Terpene Biosynthesis. *Eur. J. Org. Chem.* **2017**, *2017*, 4872-4882.
59. Martin, J. M. L.; Santra, G. Empirical Double-Hybrid Density Functional Theory: A 'Third Way' in Between WFT and DFT. *Isr. J. Chem.* **2020**, *60*, 787-804.
60. Zev, S.; Gupta, P. K.; Pahima, E.; Major, D. T. A Benchmark Study of Quantum Mechanics and Quantum Mechanics-Molecular Mechanics Methods for Carbocation Chemistry. *J. Chem. Theory Comput.* **2022**, *18*, 167-178.
61. Zhu, X.; Thompson, K. C.; Martínez, T. J. Geodesic interpolation for reaction pathways. *J. Chem. Phys.* **2019**, *150*, 164103.
62. Ásgeirsson, V.; Birgisson, B. O.; Bjornsson, R.; Becker, U.; Neese, F.; Riplinger, C.; Jónsson, H. Nudged Elastic Band Method for Molecular Reactions Using Energy-Weighted Springs Combined with Eigenvector Following. *J. Chem. Theory Comput.* **2021**, *17*, 4929-4945.
63. Durairaj, J.; Girolamo, A. D.; Bouwmeester, H. J.; Ridder, D. d.; Beekwilder, J.; Dijk, A. D. v. An analysis of characterized plant sesquiterpene synthases. *Phytochem.* **2019**, *158*, 157-165.
64. Durairaj, J.; Melillo, E.; Bouwmeester, H. J.; Beekwilder, J.; Ridder, D. d.; Dijk, A. D. J. v. Integrating structure-based machine learning and co-evolution to investigate specificity in plant sesquiterpene synthases. *PLoS Comput. Biol.* **2021**, *17*, e1008197.
65. Ansbacher, T.; Freud, Y.; Major, D. T. Slow-Starter Enzymes: Role of Active-Site Architecture in the Catalytic Control of the Biosynthesis of Taxadiene by Taxadiene Synthase. *Biochemistry* **2018**, *57*, 3773-3779.
66. Dixit, M.; Weitman, M.; Gao, J.; Major, D. T. Chemical Control in the Battle against Fidelity in Promiscuous Natural Product Biosynthesis: The Case of Trichodiene Synthase. *ACS Catal.* **2017**, *7*, 812-818.
67. Pan, X.; Yang, J.; Van, R.; Epifanovsky, E.; Ho, J.; Huang, J.; Pu, J.; Mei, Y.; Nam, K., . . . Shao, Y. Machine-Learning-Assisted Free Energy Simulation of Solution-Phase and Enzyme Reactions. *J. Chem. Theory Comp.* **2021**, *17*, 5745-5758.

68. Kubař, T.; Elstner, M.; Cui, Q. Hybrid Quantum Mechanical/Molecular Mechanical Methods For Studying Energy Transduction in Biomolecular Machines. *Annu. Rev. Biophys.* **2023**, *52*, 525-551.
69. Yang, Y.; Yu, H.; York, D.; Cui, Q.; Elstner, M. Extension of the Self-Consistent-Charge Density-Functional Tight-Binding Method: Third-Order Expansion of the Density Functional Theory Total Energy and Introduction of a Modified Effective Coulomb Interaction. *J. Phys. Chem. A* **2007**, *111*, 10861-10873.
70. Hwang, W.; Austin, S. L.; Blondel, A.; Boittier, E. D.; Boresch, S.; Buck, M.; Buckner, J.; Caflisch, A.; Chang, H.-T., . . . Karplus, M. CHARMM at 45: Enhancements in Accessibility, Functionality, and Speed. *J. Phys. Chem. B* **2024**, *128*, 9976-10042.
71. Satorras, V. G.; Hoogeboom, E.; Welling, M., E(n) Equivariant Graph Neural Networks. In *Proceedings of the 38th International Conference on Machine Learning*, Virtual Event, 2021; Vol. 139.
72. Batzner, S.; Musaelian, A.; Sun, L.; Geiger, M.; Mailoa, J. P.; Kornbluth, M.; Molinari, N.; Smidt, T. E.; Kozinsky, B. E(3)-equivariant graph neural networks for data-efficient and accurate interatomic potentials. *Nat. Commun* **2022**, *13*, 2453.
73. Kiani, T.; Caciularu, A.; Zev, S.; Major, D. T.; Goldberger, J. In *Utilizing Perturbation of Atoms' Positions for Equivariant Pre-Training in 3D Molecular Analysis*, 2023 IEEE 33rd International Workshop on Machine Learning for Signal Processing (MLSP), IEEE: 2023; pp 1-6.
74. One Thousand Plant Transcriptomes Initiative. One thousand plant transcriptomes and the phylogenomics of green plants. *Nature* **2019**, *574*, 679-685.
75. Priya, P.; Yadav, A.; Chand, J.; Yadav, G. Terzyme: a tool for identification and analysis of the plant terpenome. *Plant Methods* **2018**, *14*, 4.
76. Ryu, J. Y.; Kim, H. U.; Lee, S. Y. Deep learning enables high-quality and high-throughput prediction of enzyme commission numbers. *Proc. Natl. Acad. Sci. U.S.A.* **2019**, *116*, 13996-14001.
77. Wang, H.; Wang, Q.; Liu, Y.; Liao, X.; Chu, H.; Chang, H.; Cao, Y.; Li, Z.; Zhang, T., . . . Jiang, H. PCPD: Plant cytochrome P450 database and web-based tools for structural construction and ligand docking. *Synth. Syst. Biotechnol.* **2021**, *6*, 102-109.
78. Fansher, D. J.; Besna, J. N.; Fendri, A.; Pelletier, J. N. Choose Your Own Adventure: A Comprehensive Database of Reactions Catalyzed by Cytochrome P450 BM3 Variants. *ACS Catal.* **2024**, *14*, 5560-5592.
79. Abramson, J.; Adler, J.; Dunger, J.; Evans, R.; Green, T.; Pritzel, A.; Ronneberger, O.; Willmore, L.; Ballard, A. J., . . . Jumper, J. M. Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature* **2024**, *630*, 493-500.
80. Eswar, N.; John, B.; Mirkovic, N.; Fiser, A.; Ilyin, V. A.; Pieper, U.; Stuart, A. C.; Marti-Renom, M. A.; Madhusudhan, M. S., . . . Yerkovich, B. Tools for comparative protein structure modeling and analysis. *Nucleic Acids Res.* **2003**, *31*, 3375-3380.
81. Freud, Y.; Ansbacher, T.; Major, D. T. Catalytic Control in the Facile Proton Transfer in Taxadiene Synthase. *ACS Catal.* **2017**, *7*, 7653-7657.
82. Himpich, S.; Ringel, M.; Schwartz, R.; Dimos, N.; Driller, R.; Helmer, C. P. O.; Kumar Gupta, P.; Haack, M.; Thomas Major, D., . . . Loll, B. How Can the Diterpene Synthase CotB2V80L Alter the Product Profile? *ChemCatChem* **2024**, e202400711.
83. Whittington, D. A.; Wise, M. L.; Urbansky, M.; Coates, R. M.; Croteau, R. B.; Christianson, D. W. Bornyl diphosphate synthase: structure and strategy for carbocation manipulation by a terpenoid cyclase. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99*, 15375-80.
84. Chen, M.; Al-lami, N.; Janvier, M.; D'Antonio, E. L.; Faraldos, J. A.; Cane, D. E.; Allemann, R. K.; Christianson, D. W. Mechanistic insights from the binding of substrate and carbocation intermediate analogues to aristolochene synthase. *Biochemistry* **2013**, *52*, 5441-53.
85. Schlichting, I.; Berendzen, J.; Chu, K.; Stock, A. M.; Maves, S. A.; Benson, D. E.; Sweet, R. M.; Ringe, D.; Petsko, G. A., . . . Sligar, S. G. The Catalytic Pathway of Cytochrome P450cam at Atomic Resolution. *Science* **2000**, *287*, 1615-1622.
86. Poulos, T. L.; Finzel, B. C.; Howard, A. J. High-resolution crystal structure of cytochrome P450cam. *J. Mol. Biol.* **1987**, *195*, 687-700.
87. Bařler, J.; Paternoga, H.; Holdermann, I.; Thoms, M.; Granneman, S.; Barrio-Garcia, C.; Nyarko, A.; Lee, W.; Stier, G., . . . Hurt, E. A network of assembly factors is involved in remodeling rRNA elements during preribosome maturation. *J. Cell Biol.* **2014**, *207*, 481-498.

88. Gu, M.; Wang, M.; Guo, J.; Shi, C.; Deng, J.; Huang, L.; Huang, L.; Chang, Z. Crystal structure of CYP76AH1 in 4-PI-bound state from *Salvia miltiorrhiza*. *Biochem. Biophys. Res. Commun.* **2019**, *511*, 813-819.
89. Yehorova, D.; Di Geronimo, B.; Robinson, M.; Kasson, P. M.; Kamerlin, S. C. L. Using residue interaction networks to understand protein function and evolution and to engineer new proteins. *Curr. Opin. Struct. Biol.* **2024**, *89*, 102922.
90. Xu, Y.; Verma, D.; Sheridan, R. P.; Liaw, A.; Ma, J.; Marshall, N. M.; McIntosh, J.; Sherer, E. C.; Svetnik, V.; . . . Johnston, J. M. Deep dive into machine learning models for protein engineering. *J. Chem. Inf. Model.* **2020**, *60*, 2773-2790.
91. Ribeiro, A. J. M.; Riziotis, I. G.; Tyzack, J. D.; Borkakoti, N.; Thornton, J. M. EzMechanism: an automated tool to propose catalytic mechanisms of enzyme reactions. *Nat. Meth.* **2023**, *20*, 1516-1522.
92. Tian, B.-X.; Wallrapp, F. H.; Holiday, G. L.; Chow, J.-Y.; Babbitt, P. C.; Poulter, C. D.; Jacobson, M. P. Predicting the Functions and Specificity of Triterpenoid Synthases: A Mechanism-Based Multi-intermediate Docking Approach. *PLoS Comput. Biol.* **2014**, *10*, e1003874.
93. O'Brien, T. E.; Bertolani, S. J.; Tantillo, D. J.; Siegel, J. B. Mechanistically informed predictions of binding modes for carbocation intermediates of a sesquiterpene synthase reaction. *Chem. Sci.* **2016**, *7*, 4009-4015.
94. O'Brien, T. E.; Bertolani, S. J.; Zhang, Y.; Siegel, J. B.; Tantillo, D. J. Predicting Productive Binding Modes for Substrates and Carbocation Intermediates in Terpene Synthases-Bornyl Diphosphate Synthase as a Representative Case. *ACS Catal.* **2018**, *8*, 3322-3330.
95. Kästner, J. Umbrella sampling. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2011**, *1*, 932-942.
96. Mandelli, D.; Hirshberg, B.; Parrinello, M. Metadynamics of Paths. *Phys. Rev. Lett.* **2020**, *125*, 026001.
97. Balasubramani, S. G.; Schwartz, S. D. Transition Path Sampling Based Calculations of Free Energies for Enzymatic Reactions: The Case of Human Methionine Adenosyl Transferase and Plasmodium vivax Adenosine Deaminase. *J. Phys. Chem. B* **2022**, *126*, 5413-5420.
98. Huang, J.; Rauscher, S.; Nawrocki, G.; Ran, T.; Feig, M.; Groot, B. L. d.; Grubmüller, H.; Alexander D. MacKerell, J. CHARMM36m: an improved force field for folded and intrinsically disordered proteins. *Nat. Meth.* **2017**, *14*, 71-73.
99. Vanommeslaeghe, K.; Hatcher, E.; Acharya, C.; Kundu, S.; Zhong, S.; Shim, J.; Darian, E.; Guvench, O.; Lopes, P., . . . Mackerell, A. D., Jr. CHARMM general force field: A force field for drug-like molecules compatible with the CHARMM all-atom additive biological force fields. *J. Comput. Chem.* **2010**, *31*, 671-690.
100. Zev, S.; Raz, K.; Schwartz, R.; Tarabeh, R.; Gupta, P. K.; Major, D. T. Benchmarking the Ability of Common Docking Programs to Correctly Reproduce and Score Binding Modes in SARS-CoV-2 Protease Mpro. *J. Chem. Inf. Model.* **2021**, *61*, 2957-2966.
101. Jo, S.; Kim, T.; Iyer, V. G.; Im, W. CHARMM-GUI: A Web-based Graphical User Interface for CHARMM. *J. Comput. Chem.* **2008**, *29*, 1859-1865.
102. Das, S.; Nam, K.; Major, D. T. Rapid Convergence of Energy and Free Energy Profiles with QM Size in QM/MM Simulations of Proton Transfer in DNA. *J. Chem. Theory Comp.* **2018**, *14*, 1695-1705.
103. Major, D. T.; Gupta, P. K.; Gao, J. Origin of Catalysis by Nitroalkane Oxidase. *J. Phys. Chem. B* **2023**, *127*, 151-162.