# High-Resolution Mapping of Cytochrome P450 Functional Landscapes: A Comparative Computational Framework for Synthetic Biology and Protein Engineering

The systematic classification of the Cytochrome P450 (CYP) superfamily has historically relied on primary sequence identity, employing a heuristic threshold where enzymes sharing more than 40% sequence similarity are grouped into the same family, and those with more than 55% similarity are categorized within the same subfamily.[1] While this nomenclature provides a robust framework for evolutionary tracking, it frequently fails to predict the substrate specificity, regioselectivity, and metabolic function of these heme-thiolate monooxygenases. This discrepancy arises from the inherent plasticity of the CYP fold, particularly the flexible loops—such as the B-C and F-G segments—that define the access channels and the catalytic cavity.[1] To bridge the gap between sequence-based classification and functional reality, a novel methodology utilizing binding site vectors (BSVs) has been developed, providing a high-resolution, multidimensional descriptor of the active site's geometric and electrostatic landscape.[1] This report details the codebase, analytical tools, and comparative strategies utilized in this framework, with specific emphasis on its application to both human and plant enzymatic systems and its potential implementation in novel biosynthetic contexts such as the metabolic engineering of *Tacca* species.

## The Binding Site Vector Framework and Software Ecosystem

The investigation into CYP functional landscapes is centered on a specific codebase and a suite of integrated computational tools designed to handle the complexities of protein dynamics and pocket topography. The core methodology presented in the research is not merely a theoretical construct but is supported by an accessible and well-documented software infrastructure.

### The Software Codebase and Integration

The research is underpinned by two primary software repositories that facilitate the generation and comparison of binding site descriptors. The first is the GROMOS++ suite of programs, and the second is the Chemical Data Processing Toolkit (CDPKit).[1]

The primary tool for the initial generation of the binding site vectors is the program pocket, which is part of the GROMOS++ package.[1] GROMOS++ is a comprehensive collection of C++ programs specialized for the pre-processing and post-analysis of molecular dynamics (MD) trajectories. It is designed to work seamlessly with the GROMOS simulation engine but is also compatible with GROMACS trajectories following appropriate coordinate conversion.[8] The pocket program implements the icosahedral projection logic described in the study, allowing researchers to automate the tracing of vectors from a central anchor point (typically the heme iron) to the van der Waals surface of the protein residues.[1]

Complementary to GROMOS++ is the CDPKit library, which provides the Python-based analytical layer required for high-level data processing and comparative modeling. The researchers have specifically integrated their methodology into CDPKit through two key scripts:

1. gen_kuvek_bp_descr.py: This script handles the automated generation of the binding pocket descriptors from structural data, incorporating both the geometric length of the vectors and the associated partial charges of the intersected atoms.[1]
2. cmp_kuvek_bp_descrs.py: This tool is utilized for the large-scale pairwise comparison of these descriptors, calculating the normalized root-mean-square differences (RMSD) between different protein conformations or different enzyme isoforms.[1]

CDPKit is an open-source cheminformatics toolkit implemented in C++ with a comprehensive Python-interfacing layer (CDPL).[6] Its inclusion in the workflow allows for the integration of binding site similarity analysis with other cheminformatics tasks, such as ligand-receptor interaction pharmacophore generation, molecular fragmentation, and machine learning-based site-of-metabolism predictions.[6]

## Computational Tools and Simulation Infrastructure

The research utilized a diverse array of computational tools to manage the structural data, perform the atomistic simulations, and analyze the resulting conformational ensembles. The following table provides a structured overview of the tools and their specific roles within the study.

| Tool Category | Specific Software/Force Field | Functional Role in CYP Analysis |
|---|---|---|
| **MD Engines** | GROMOS, GROMACS (2020) | Execution of 500 ns production simulations across 23 isoforms.[1] |

| | | |
|---|---|---|
| **Force Field** | GROMOS 54a8 | Parameterization of amino acids and the heme-cysteine complex.[1] |
| **Water Model** | SPC (Simple Point Charge) | Solvation of protein systems for dynamic ensemble generation.[1] |
| **Structural Correction** | PDBFixer | Reconstruction of missing loops and residues in PDB crystal structures.[1] |
| **Coordinate Handling** | OpenBabel | Assignment of van der Waals radii for vector intersection logic.[1] |
| **Superposition** | PyMOL (align/cealign) | Structural alignment of diverse CYPs to a common coordinate system.[1] |
| **Clustering** | scikit-learn (Affinity Propagation) | Identification of representative conformational exemplars from MD data.[1] |
| **Trajectory Analysis** | MDAnalysis | Processing of massive simulation trajectories for cluster assignments.[1] |
| **Phylogenetics** | Clustal Omega | Multiple sequence alignment for phylogenetic tree construction.[1] |

The integration of these tools allows for a "Dynamic Characterization" pipeline. In this workflow, static structures from the Protein Data Bank (PDB) or predicted models from AlphaFold2 are first curated and then subjected to MD simulations. The resulting trajectories are sampled to extract thousands of conformations (15,625 per isoform in this study), which are then encoded as binding site vectors.[1] This approach acknowledges that the "active site" of a CYP is not a fixed cavity but a fluctuating environment governed by the principle of conformational selection.[1]

## Comparative Methodologies for Non-Plant and Plant

# Cytochromes P450

A fundamental challenge in the study was the significant disparity in the structural data available for human (non-plant) versus plant CYPs. This necessitated distinct methodological approaches for structural sourcing, alignment, and the handling of the essential heme cofactor.

## Structural Sourcing and Provenance

For the non-plant dataset, the research focused on human CYPs due to the wealth of experimental data. A total of 285 human CYP structures were retrieved from the PDB.[1] These structures benefit from high-resolution experimental validation, with heme groups and, in many cases, ligands already present in the binding pocket. The selection criteria for these structures were stringent, requiring the presence of a heme iron within 4 Å of a cysteine sulfur and a backbone RMSD of $\leq 7$ Å relative to a reference CYP3A4 structure.[1]

In contrast, the plant CYP dataset relied almost exclusively on computational models. The researchers retrieved 343 AlphaFold2-predicted structures from UniProt, as the number of plant CYPs with resolved crystal structures is remarkably low.[1] While AlphaFold2 provides high-confidence models (high pLDDT scores) for the conserved CYP fold, these models lack cofactors like heme and ligands.[1] This structural "emptiness" required a specific alignment strategy to define the binding site vectors.

## Alignment and the Heme Anchor Strategy

Because the binding site vectors radiate from the heme iron, the absence of heme in plant models presented a geometric problem. The researchers resolved this by aligning all structures—human and plant—to a single reference structure: CYP3A4 (PDB ID: 4I3Q).[1]

- **For Human CYPs:** The heme iron was already present. The protein was translated and rotated so that the heme iron sat at the coordinate origin ($0, 0, 0$) and the heme plane lay in the xy-plane.[1]
- **For Plant CYPs:** Following the backbone-based alignment to the reference CYP3A4, the heme iron was assumed to be located at the origin.[1] This assumption is justified by the highly conserved orientation of the heme relative to the I-helix and the cysteine-containing heme-binding loop across the superfamily.

## MD Simulation and Ensemble Averaging

The study further differentiated the analysis by performing MD simulations on a subset of enzymes (8 human and 15 plant) to evaluate if dynamic data provided better functional insights than static structures.

| Feature | Human (Non-Plant) Approach | Plant Approach |
|---|---|---|
| **Primary Structural Data** | Experimental X-ray structures from PDB.[1] | AlphaFold2-predicted structures from UniProt.[1] |
| **Total Structures (Static)** | 285 structures across 11 families.[1] | 343 structures across 45 families.[1] |
| **Cofactor Management** | Heme coordinates directly used from crystal data.[1] | Heme anchor point assumed via alignment to reference.[1] |
| **Validation Metric** | Experimental resolution and R-factors. | AlphaFold2 pLDDT confidence scores.[1] |
| **Functional Context** | Well-defined substrate preferences (xenobiotic metabolism).[1] | Largely unknown substrate specificity (specialized metabolism).[1] |

The results of these comparative approaches revealed that for human CYPs, binding site vectors could accurately group enzymes involved in specific metabolic tasks, such as the steroidogenic enzymes CYP17A1 and CYP19A1, which clustered together despite low sequence similarity.[1] For plant CYPs, the analysis identified family-level variations, such as the internal diversity within the CYP81A family, suggesting that despite a conserved backbone, the binding sites have evolved for distinct functions in specialized metabolism.[1]

# Mathematical Basis of the Vector Descriptors

The binding site vectors encode two primary properties: shape ($l$) and electrostatics ($q$). The similarity between two binding sites, $i$ and $j$, is quantified using a normalized RMSD calculation. This normalization is critical because length and charge have different units and scales.

The combined RMSD ($RMSD_{ij}^{sc}$) is calculated as follows:

$$RMSD_{ij}^{sc} = \sqrt{\frac{1}{N}\left[\sum_{k=1}^{N}\left(\frac{l_{k,i}-l_{k,j}}{\sigma_l}\right)^2 + \sum_{k=1}^{N}\left(\frac{q_{k,i}-q_{k,j}}{\sigma_q}\right)^2\right]}$$

In this formula, $N$ represents the 260 vectors of the active-site hemisphere.[1] The variables $\sigma_l$ and $\sigma_q$ are the standard deviations of all vector lengths and charges across the dataset, respectively. For human CYPs, the researchers found $\sigma_l = 3.742$ Å and $\sigma_q = 0.198\ e$, while for plant CYPs, $\sigma_l = 3.833$ Å and $\sigma_q = 0.213\ e$.[1]

The framework also allows for feature weighting using a parameter $d$:

$$w_c = (1+d)w_s$$
$$w_s = \frac{2}{2+d}$$
$$w_c = \frac{2(1+d)}{2+d}$$

By adjusting $d$, a researcher can prioritize geometric fit ($d < 0$) or electrostatic complementarity ($d > 0$).[1] This is particularly relevant in synthetic biology, where a designer may want to engineer a binding site that specifically accommodates the bulky hydrophobic rings of a phytosterol (prioritizing shape) or one that stabilizes a polar intermediate (prioritizing charge).

# Application to Tacca CYP450 Research: Substrate Affinity and MD Analysis

For the evaluation of *Tacca* CYP450 enzymes against phytosterols and taccalonolide intermediates, the binding site vector methodology offers a distinct advantage over traditional docking. The user's dataset of sequences and structures can be integrated into this framework to evaluate substrate compatibility and functional overlap.

## Can the BSV Approach Evaluate Affinity?

It is vital to understand that the binding site vector framework does not directly calculate a $K_d$ or a $\Delta G$ of binding in the traditional sense of a scoring function. Instead, it provides a "Binary Docking Criterion".[1] This criterion assesses whether a given ligand pose is geometrically and electrostatically compatible with the pre-existing conformations of the

enzyme's binding pocket.

The core logic for this evaluation is as follows:

1. **Conformational Sampling:** Perform 500 ns MD simulations for the *Tacca* CYPs to generate a conformational ensemble. This is necessary because P450s are flexible, and a single static structure (especially a predicted one) may not show an "open" or "fit" pocket.[1]
2. **Ligand Vector Generation:** The ligand (phytosterol or intermediate) is also described by vectors originating from the same anchor point (heme iron). For each direction, the vector extends to the point on the ligand surface farthest from the origin.[1]
3. **The Fitting Evaluation:** For each conformation in the MD ensemble, the ligand vectors are compared to the binding site vectors. A "fit" is achieved if the ligand vector is shorter than the binding site vector in every direction (indicating no steric clash) and if the charges are complementary.[1]
4. **Affinity Proxy:** The percentage of the conformational ensemble that can accommodate the ligand serves as a proxy for the enzyme's affinity or selectivity. If an enzyme samples "fitting" conformations 10% of the time, it is more likely to bind the substrate than an enzyme that only samples such states 0.1% of the time.[1]

## Ligand Preparation and Input Formats

The user specifically inquired about the necessity of SMILES. While SMILES is a standard format for chemical notation, the binding site vector calculation requires 3D coordinates to determine intersection points.

1. **From SMILES to 3D:** SMILES strings must be converted into 3D structures. CDPKit provides routines for 3D coordinate generation and conformer ensembles.[6]
2. **Docking Poses:** A ligand must be placed in a putative binding mode above the heme. Traditional docking tools (like AutoDock or GLIDE) or machine learning-based docking (like DiffDock) can be used to generate these initial poses.[18]
3. **Vector Processing:** Once a pose is established, the gen_kuvek_bp_descr.py script can be used to generate the ligand vectors.[1]

| Ligand Type | Recommended Preparation Steps | Key Descriptors |
|---|---|---|
| **Phytosterols** | Generate 3D conformers from SMILES; Dock to Tacca CYP; Orient sterol ring above heme iron.[14] | Steric bulk of the ring system (Shape-driven fit).[1] |

| | | |
|---|---|---|
| **Taccalonolide Intermediates** | Curate functional group protonation states; Establish distance to heme iron (typically 2.2 Å for inhibitory/reactive atoms).[1] | Electrostatic complementarity of oxygenated groups (Charge-driven fit).[1] |

# Dynamic Insights and Conformational Diversity

One of the most profound insights from the research is that "average" binding site similarity (Case Study 2) is often explained by the distribution of conformations across distinct clusters (Case Study 3).

The researchers used the Affinity Propagation algorithm to cluster 15,625 frames for each of the 23 simulated isoforms into seven representative clusters.[1]

- **Cluster 1:** Heavily populated by human CYPs (3A4, 2C9, 2C19), representing a pocket capable of accommodating large, promiscuous ligands.[1]
- **Cluster 3:** Characterized by conformations of steroid-metabolizing enzymes (CYP17A1, CYP19A1). Interestingly, the plant enzyme CYP79E1 populated this cluster, suggesting its potential involvement in similar steroidal or tyrosine-based metabolism.[1]
- **Cluster 5:** A taxon-specific cluster for plant CYPs, illustrating the distinct evolutionary path of the plant specialized metabolome.[1]

For a metabolic engineer working on *Tacca*, this clustering approach is invaluable. By performing a similar clustering on *Tacca* CYP MD simulations, one can identify which "clusters" or conformational states are unique to specific stages of taccalonolide biosynthesis. If a *Tacca* CYP shares a cluster with human CYP17A1 (steroid metabolism), it provides strong evidence for its role in modifying the phytosterol core.

# Technical Considerations and Limitations

While the binding site vector methodology is powerful, it has specific technical prerequisites and inherent limitations that must be addressed in any protein engineering workflow.

## Structural Alignment Sensitivity

The comparison of binding sites relies entirely on accurate backbone alignment. Because vectors are compared by index (e.g., vector 1 of protein A is compared to vector 1 of protein B), any misalignment of the backbone will cause vectors to point at different regions of the pocket, leading to false divergence.[1] The use of PyMOL's cealign or GROMOS++ stable secondary structure element alignment is recommended to ensure consistency.[1]

## Geometrically Occluded Regions

The vector-based representation terminates at the first atomic surface encountered. Consequently, "subpockets" that lie behind bulky residues or secondary structure elements are not captured by a single set of vectors.[1] For complex *Tacca* CYP pockets with multiple access channels, it may be necessary to anchor multiple vector sets at different points (e.g., at the center of the pocket and at the entrance of a known channel) to map the entire volume.[1]

### Conformational Selection vs. Induced Fit

The methodology effectively captures the conformational selection model—sampling pre-existing states. However, it does not explicitly model "induced fit," where the protein undergoes a major structural rearrangement only upon ligand binding.[1] For CYPs, this is generally less of a concern as they are known to sample their binding-competent states in the apo form, but for highly rigid proteins or those with gated access, the BSV method might underestimate binding potential.[1]

## Synthesis and Recommendations for Tacca CYP450 Analysis

The binding site vector framework provides a sophisticated alternative to sequence identity for mapping the functional landscapes of CYPs. For a researcher possessing *Tacca* CYP450 structural data and sequences, the implementation of this approach is highly feasible and scientifically robust.

### Actionable Steps for Implementation

1. **Code Acquisition:** Download and install GROMOS++ and CDPKit. Ensure the Python CDPL bindings are correctly configured.[6]
2. **Ensemble Generation:** Use GROMACS or GROMOS to perform at least 500 ns MD simulations for each *Tacca* CYP. Use the GROMOS 54a8 force field and SPC water to match the validated study parameters.[1]
3. **Vector Mapping:**
    - Align MD frames to a reference CYP heme.
    - Run the pocket program (GROMOS++) or gen_kuvek_bp_descr.py (CDPKit) to generate BSVs for each frame.[1]
4. **Substrate Compatibility:**
    - Convert taccalonolide intermediate SMILES to 3D SDF format.
    - Identify a reasonable binding pose using docking software.
    - Generate ligand vectors and apply the binary docking criterion to the Tacca MD ensemble.[1]
5. **Functional Grouping:** Perform hierarchical clustering based on the combined BSV RMSD to determine which *Tacca* CYPs share similar functional landscapes, regardless of their

family nomenclature.[1]

By following this methodology, the researcher can move beyond the "sequence-similarity-equals-functional-similarity" dogma and identify the true catalysts for taccalonolide production based on the three-dimensional geometric and electrostatic realities of the enzyme-substrate complex. This high-resolution mapping represents the current frontier in the integration of computational biology and metabolic engineering for the exploitation of complex plant specialized metabolism.

**Works cited**


1. binding-site-vectors-enable-mapping-of-cytochrome-p450-functional-landscapes.pdf
2. Binding Site Vectors Enable Mapping of Cytochrome P450 Functional Landscapes, accessed February 3, 2026, https://www.researchgate.net/publication/400180053_Binding_Site_Vectors_Enable_Mapping_of_Cytochrome_P450_Functional_Landscapes
3. A Computational Pipeline Observes the Flexibility and Dynamics of Plant Cytochrome P450 Binding Sites - PubMed Central, accessed February 3, 2026, https://pmc.ncbi.nlm.nih.gov/articles/PMC11545509/
4. Binding Site Vectors Enable Mapping of Cytochrome P450 Functional Landscapes, accessed February 3, 2026, https://pubmed.ncbi.nlm.nih.gov/41587116/
5. Binding Site Vectors Enable Mapping of Cytochrome P450 Functional Landscapes, accessed February 3, 2026, https://pubs.acs.org/doi/10.1021/acs.jcim.5c02705
6. molinfo-vienna/CDPKit: The Chemical Data Processing Toolkit - GitHub, accessed February 3, 2026, https://github.com/molinfo-vienna/CDPKit
7. GROMOS++Software for the Analysis of Biomolecular Simulation Trajectories, accessed February 3, 2026, https://research.vu.nl/en/publications/gromossoftware-for-the-analysis-of-biomolecular-simulation-trajec/
8. Biomolecular Simulation - The GROMOS Software, accessed February 3, 2026, https://www.gromos.net/
9. A Suite of Advanced Tutorials for the GROMOS Biomolecular Simulation Software [Article v1.0], accessed February 3, 2026, https://livecomsjournal.org/index.php/livecoms/article/download/v2i1e18552/967/2757
10. The GROMOS Software for (Bio)Molecular Simulation, accessed February 3, 2026, https://ethz.ch/content/dam/ethz/special-interest/chab/imps/csms-dam/doc/CSCBP-res/CSBMS_gro_man_v7_HS22.pdf
11. Computer tools and databases - VLS3D.COM, accessed February 3, 2026, https://www.vls3d.com/computers-tools.html
12. Introduction — CDPKit 1.2.3 documentation, accessed February 3, 2026, https://cdpkit.org/introduction.html

13. CDPL Python Cookbook — CDPKit 1.1.1 documentation, accessed February 3, 2026, https://cdpkit.org/v1.1.1/cdpl_python_cookbook/index.html
14. How to generate a 3D molecular structure from a SMILES string - CCDC, accessed February 3, 2026, https://www.ccdc.cam.ac.uk/discover/blog/smiles-to-3d-chemical-structure-csd/
15. The GROMOS Interface - ChemShell, accessed February 3, 2026, https://chemshell.org/static_files/tcl-chemshell/manual/gromos.html
16. (PDF) A Suite of Advanced Tutorials for the GROMOS Biomolecular Simulation Software [Article v1.0] - ResearchGate, accessed February 3, 2026, https://www.researchgate.net/publication/347919152_A_Suite_of_Advanced_Tutorials_for_the_GROMOS_Biomolecular_Simulation_Software_Article_v10
17. Visualizing Molecular Structure From SMILES Using RDKit | by Nazish Javeed - Medium, accessed February 3, 2026, https://medium.com/@nazishjaveed164/visualizing-molecular-structure-from-smiles-using-rdkit-3ff070f53763
18. PocketGen: Generating Full-Atom Ligand-Binding Protein Pockets | bioRxiv, accessed February 3, 2026, https://www.biorxiv.org/content/10.1101/2024.02.25.581968v1.full-text
19. DockFormer: Affinity Prediction and Flexible Docking with Pair Transformer | bioRxiv, accessed February 3, 2026, https://www.biorxiv.org/content/10.1101/2024.11.25.625135v2.full-text
20. A Protein-Ligand Interaction-focused 3D Molecular Generative Framework for Generalizable Structure-based Drug Design - ChemRxiv, accessed February 3, 2026, https://chemrxiv.org/doi/pdf/10.26434/chemrxiv-2023-jsjwx
21. Binding Affinity via Docking: Fact and Fiction - PMC, accessed February 3, 2026, https://pmc.ncbi.nlm.nih.gov/articles/PMC6222344/