SMARTpy: A Python Package for the Generation of Cavity-Specific Steric Molecular Descriptors and Applications to Diverse Systems

Beck R. Miller, Ryan C. Cammarota, Matthew S. Sigman*

Department of Chemistry, University of Utah, 315 South 1400 East, Salt Lake City, UT 84112, USA

KEYWORDS. Molecular descriptors, cavity, pocket, open-source programming.

ABSTRACT: Steric molecular descriptors designed for machine learning (ML) applications are critical for connecting structure-function relationships to mechanistic insight. However, many of these descriptors are not suitable for application to complex systems, such as catalyst reactive site pockets. In this context, we recently disclosed a new set of 3D steric molecular descriptors that were originally designed for dirhodium(II) tetra-carboxylate catalysts. Herein, we expand the Spatial Molding for Rigid Targets (SMART) descriptor toolkit by releasing SMARTpy; an automated, open-source Python API package for computational workflow integration of SMART descriptors. The impact of the structure of the molecular probe for generation of SMART descriptors was analyzed. Resultant SMART descriptors and pocket features were found to be highly dependent upon probe selection, and do not scale linearly. Flexible probes with smaller substituents can explore narrow pocket regions resulting in a higher resolution pocket imprint. Macrocyclic probes with larger substituents are more applicable to larger cavities with smooth boundaries, such as dirhodium paddlewheel complexes. In these cases, SMARTpy provides comparable descriptors to the original calculation method using UCSF Chimera. Finally, we analyzed a series of case studies demonstrating how SMART descriptors can impact other areas of catalysis, such as organocatalysis, biocatalysis, and protein pocket analysis.

INTRODUCTION

Structure-function relationships are leveraged to provide mechanistic insight into the connections between catalyst structural features and observed experimental outcome. A diverse array of steric molecular descriptors has historically captured structural features of 3D-representations for application to statistical modeling and machine learning (ML) prediction of reaction performance. Traditional steric descriptors, including Sterimol[1,2] (*L*, B_1, B_5) and buried volume[3,4] (V_{Bur}), are successfully applied to diverse areas of catalysis and provide unique insight into structure-function relationships from resultant ML models. However, limitations of many steric molecular descriptors prevent their application to certain complex systems.

We recently developed a set of steric molecular descriptors tailored for dirhodium paddlewheel catalysts.[5–7] These are privileged catalyst scaffolds with large, conical reactive pockets, that provide a confined environment conducive to selective transformations.[8–10] As a result of these complex 3D-conformations, steric features of these catalysts cannot be adequately parametrized using traditional molecular descriptors.[5] For instance, Sterimol descriptors are highly dependent upon the selection of the *L*-axis. This can be difficult to apply to systems with multiple bridging ligands and distinct axial binding sites (Figure 1a). Similarly, V_{Bur} assumes a spherical binding environment around the metal center of interest, and the radius of search space is often too small to encompass the distal ligand environment (Figure 1b). As a result, these steric descriptors that were designed for small molecule catalysts were found to be insufficient to describe the complex cavity environments in dirhodium catalysts.

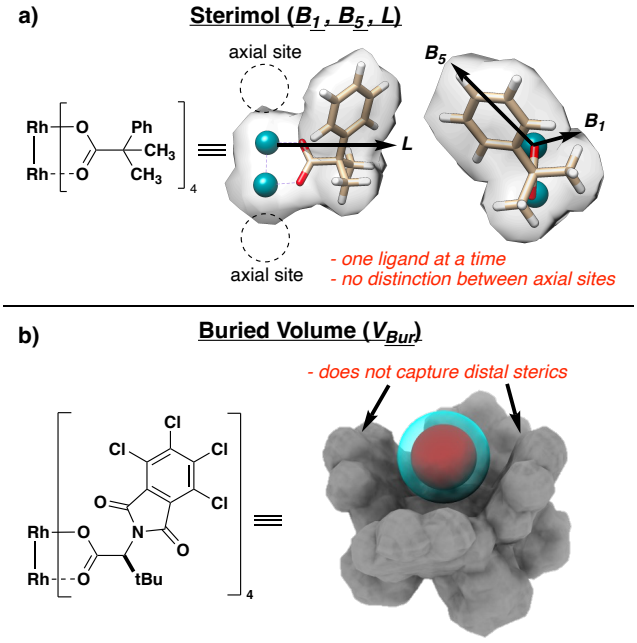


Figure 1: a) Challenges of dirhodium catalysts for Sterimol descriptors. b) Challenges of dirhodium catalysts for V_{Bur} descriptors.

Free and open-source toolkits for assessing the size and shape of catalytically active pockets are well established[11–14] and utilized in fields such as protein docking.[15–18] These

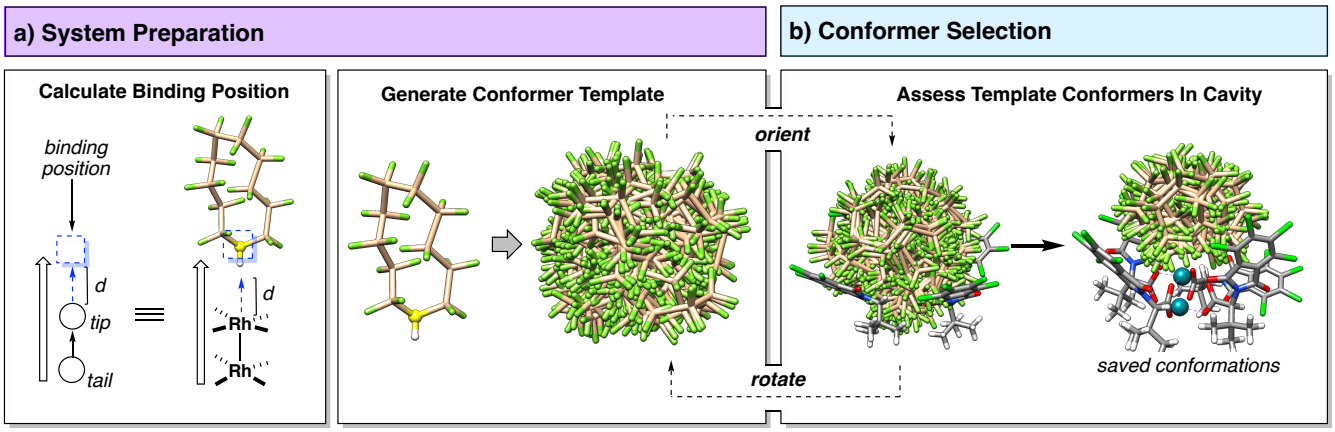
toolkits also have limitations that prevent flexible application to a diverse set of structures. First, many of these programs do not allow for pocket analysis of structures containing subunits beyond the scope of amino acids or DNA-bases, and thus cannot be applied to many small molecule transition metal catalysts. Second, these programs are designed to analyze pockets encompassed within or between larger molecules and can struggle to provide a reasonable cut off for pockets with a wide entry. Finally, most methods interpretate pocket accessibility on the basis of solvent access. [19,20,21]This method typically relies on generating a "space filling"[22] model of points with assigned Van der Waals radii to parametrize an active cavity through its interactions with solvent models.

Other approaches for pocket description have been explored, including generating representations based on ligand docking, electron density maps, grid-based approaches[23], and machine learning algorithms[24–26]. These established methods can still overestimate the size and accessibility of specific regions within a pocket from the perspective of a bound molecule. Approaches based on experimental assessment of a series of docked molecules can inherently limit the domain of applicability of the pocket information to structurally similar molecules.[27,28]Thus, a general method to generate quantitative pocket representations remain of interest.

Spatial Molding for Approachable Rigid Targets (SMART) descriptors quantify structural features at the reactive pockets of catalysts, such as cavity volume (V_{CAVITY}), entry surface area (*ESA*), and contact surface area (*CSA*) with the surrounding ligands. These descriptors are obtained through conformational sampling of reactive site space using a generalized molecular probe. SMART descriptors were initially applied to quantify the origins of regioselectivity in dirhodium C–H functionalization of donor/acceptor carbenes[5] and diastereoselectivity in dirhodium C–H insertion of donor/donor carbenes[7]. Although we envisioned broader applicability to diverse areas of catalysis, the original implementation of SMART was challenging for widespread adoption, including a significant reliance on user input and the necessity for commercial software. These two factors have prevented the rapid analysis of larger data sets and limited the accessibility of the tool to a broader community of potential users.

Herein, we release SMARTpy; a Python suite uniting open-source computational packages in a fully automated workflow for the generation of SMART descriptors. In addition to description of the construction of the SMART cavities we evaluated the impact of probe design on resultant descriptors. Finally, we demonstrate the applicability of SMART descriptors through a series of case studies. This code is freely available and open-sourced on GitHub (https://github.com/SigmanGroup/SMART-molecular-descriptors.git). A detailed description of the API is supplied in the Supplementary Information, and all structures analyzed are available in the Git repository.

Scheme 1. SMART template conformational search protocol.



WORKFLOW

Original Workflow for SMART Descriptor Calculation. The workflow for generating SMART descriptors has been partially disclosed by Davies and Sigman.[5] In this workflow, molecular probes were added to catalysts, checked for atomic overlap with the structure, then conformer searched, all requiring manual user input for every step. This implementation was time consuming and limited the possibility for high throughput catalyst parametrization. The most significant limitation of the original workflow is that probe conformer ensembles were generated using the OPLS3e forcefield[29] and a torsional Monte Carlo (MC) algorithm implemented in MacroModel, a commercial software distributed by Schrödinger. Molecular descriptors were then calculated using the free program UCSF Chimera[30]. SMARTpy employs exclusively free and open-source Python modules to generate conformer ensembles. Additionally, the package employs multiple methods for computing an array of steric descriptors.

Molecular Probe Conformational Generation in SMARTpy. The initial method for conformer searching implemented in the SMART package was a simple torsional search algorithm that rejected moves based on Van der Waals overlaps with the structure. This method performed well for acyclic probes with freely rotatable bonds, but conformational searching for macrocycles was not possible using this method. Macrocyclic structures are a known limitation of torsional algorithms as rotating one bond along a macrocycle causes multiple other bonds to

simultaneously rotate on the structure in different directions to maintain atomic geometry. This makes the search space difficult to explore by simple torsional methods.

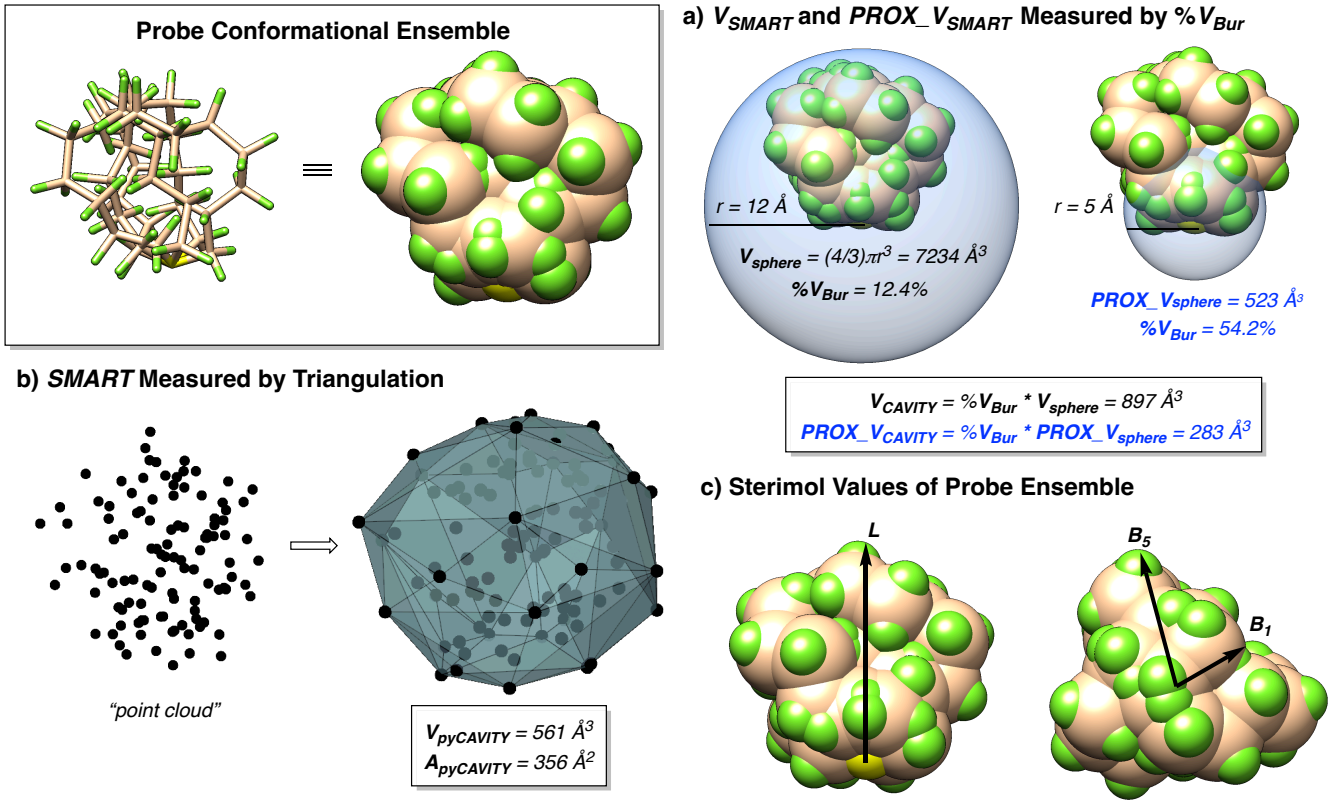


Figure 2. a) V_{CAVITY} and $proxV_{CAVITY}$ measured using V_{Bur} computed at different radii (12 Å and 5 Å). b) V_{CAVITY} and A_{CAVITY} measured by triangulation of the probe ensemble point cloud. c) Sterimol values (L, B_1, B_5) measured for the cavity ensemble.

MacroModel conformational searching employs a version of the ConfGen algorithm disclosed by Watts et. al.[31] to expand applicability and speed up conformer searching. ConfGen employs a template-based method where substructures of the molecule of interest are matched to precomputed templates of conformer ensembles. Inspired by the format of the ConfGen algorithm, a similar approach was employed in SMARTpy.

Using the RDKit function EmbedMultipleMolecules command, a conformer ensemble template is first generated for the free probe using the MMFF forcefield. This represents the accessibility of space to the probe unhindered by a catalyst structure (Scheme 1). This template is then fit into the pocket of interest aligned to a defined binding axis vector, and conformers are saved or rejected based on Van der Waals overlap with the structure. The orientation of the probe template is rotated about the binding axis stochastically, and the fitting and assessing process is repeated for a user-defined number of steps. The saved conformers from each fitting iteration are compiled into a single ensemble and returned as an object or optionally saved to an SDF file for later analysis.

Molecular Descriptor Computational Methods Available in SMARTpy. In UCSF Chimera, the command molmap was used to enclose the probe conformers in a molecular surface from which V_{CAVITY} and A_{CAVITY} were computed (Figure SX). The *molmap* function in UCSF Chimera is a density-based computation that computes a surface around select atoms in a manner proportional to the atomic numbers. Open-source Python packages were implemented instead for either speed or expanded functionality to compute SMART descriptors from probe ensembles.

Volume descriptors, such as V_{CAVITY}, can be computed through two different methods. In the first method, algebraic triangulation and the alpha method[19] to compute a surface encompassing all atoms of the probe ensemble using PyVista[32] (Figure 3b). Proximal ($proxV_{CAVITY}$) and distal ($distV_{CAVITY}$) volume can be computed by defining a radius for spherical intersection with the ensemble and computing the space taken up by separate portions of the cavity (Figure 3a). This first method was implemented for speed of descriptor calculation, as assessment of the probe ensembles proved to be the fastest (Table S1).

In the second method, V_{Bur} is first calculated for the total probe ensemble using Morfeus. To accomplish this, the ensemble is enclosed within a large sphere and the percentage of sphere volume occupied by the conformers is computed (Figure 3a). This method is significantly slower than the first (Table S1), but is implemented for[19] the opportunity to compute an extended array of SMART descriptors. The cavity space can be further subdivided into quadrants ($V_{QUADRANT}$) and octants (V_{OCTANT}). Sterimol descriptors can also be computed for the conformational

ensemble, as an interpretable method to parametrize the shape of the cavity by the maximum (B_5) and minimum (B_1) widths perpendicular to the structure binding axis (L) (Figure 3c).

METHODS

Structures for Analysis. Computed dirhodium(II) catalyst structures from a study by Shaw and Sigman[7] were used to assess the impact of probe features on SMART descriptors. A subset of conformers was selected with symmetrical, asymmetrical, and chiral ligands with the intent to maximize representative ligand feature diversity (Figure SX). All molecular probes (Table S2) employed for analysis have tetrahedral Si core atoms functionalized with either H or F. The tether atom that binds to the structure is S with a dummy H atom that is removed after initial docking. The choice of Si was initially practical for ease of pocket manipulation in UCSF Chimera with the legacy method, but many molecular units can now be used as a molecular probe core using SMARTpy.

Computation of Case Study Structures. Each case study is adapted from a literature data set or series of literature data sets. Protein and enzyme structures were obtained from the RCSB Protein Data Bank (PDB). A subset of 1,1′-bi-2-napthol (BINOL) and 1,1′-spirobiindane-7,7′-diol (SPINOL) catalysts were selected from a published computational study on BINOL catalysts to represent a diverse set of substituent steric environments.[33] Initial structures of all chiral phosphoric acid (CPA) catalysts were optimized by xTB-GFN2 using the ALPB solvation method in dichloromethane. All 3D images are generated in UCSF Chimera.

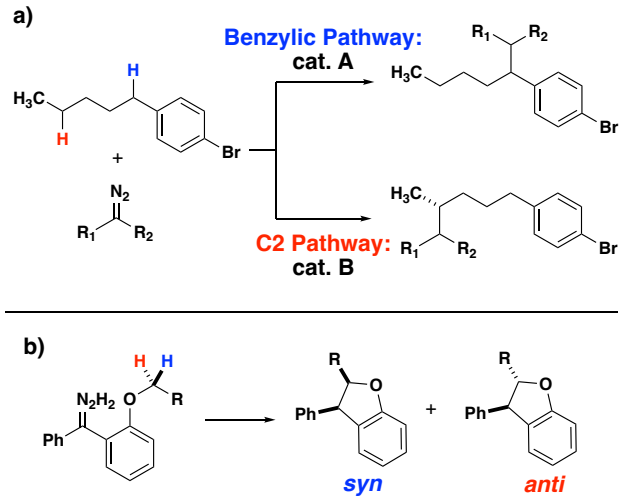


Figure 3. Dirhodium(II) catalyzed reactions for SMART descriptor application. a) First disclosure of SMART descriptors in C-H functionalization of 1-bromo-4-pethylbenzene. b) Subsequent application and expansion of SMART descriptors in diastereoselective C-H insertion. Subsequent application of SMART descriptors.

GENERAL UTILITY GUIDE

The Case for General Cavity Descriptors. The first application of SMART descriptors aided in mechanistic understanding and modeling for dirhodium(II) catalyzed site-selective C-H functionalization of 1-bromo-4-pethylbenzene via donor/acceptor carbenes (Figure 4a).[5] This initial study explicitly quantified that more confined and rigid catalysts allowed for functionalization at the less hindered C2 site. The authors noted direct comparisons showing that traditional Sterimol and V_{bur} steric descriptors were unable to capture peripheral steric hindrance, the flexibility of catalyst shape, and the resulting variable accessibility of the bound carbene to the approaching substrate C–H bonds.

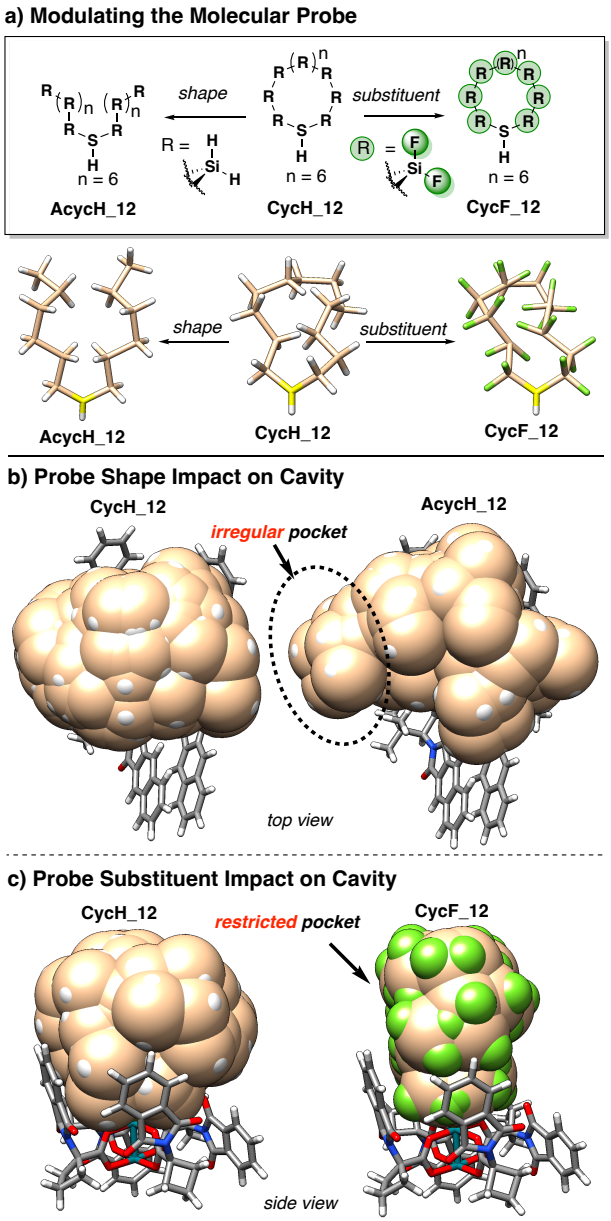


Figure 4. a) Opportunities for modulating the generalized molecular probe; the core shape (left) and substituents (right). b) The probe shape greatly impacts the shape of the cavity. Flexible probes (**AcycH_12**) can explore more hindered regions of a cavity than rigid probes (**CycH_12**). c) Probe substituents also impact the size of the cavity. Small substituents (**CycH_12**) can parametrize more space than larger substituents (**CycF_12**).

SMART descriptors were subsequently used to model diastereoselectivity in the C–H insertion of donor/donor carbenes for the cyclization of benzodihydrofurans (Figure 4b).[7] However, due to the steric demands of the intramolecular cyclization transition state, this system required a different molecular probe and a set of proximal and distal SMART descriptors. In this general utility guide, we present a mechanistic analysis of the different SMART methods utilized in these two applications to contextualize the practical considerations analyzed.

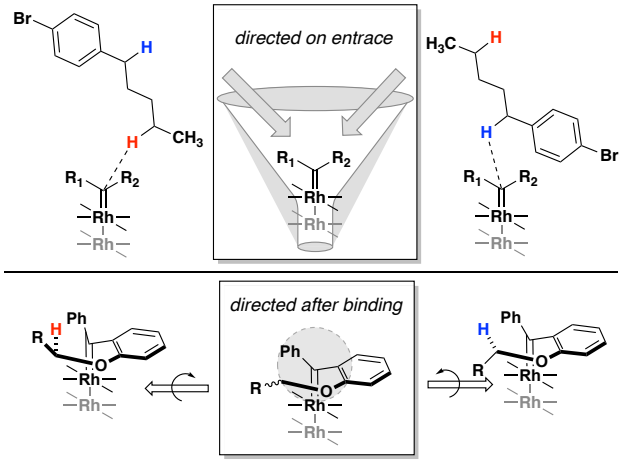


Figure 5. Ligands can direct intermolecular regioselectivity during entrance of a substrate into the pocket (top). On the other hand, ligands can direct intramolecular diastereoselectivity within the pocket after binding (bottom).

Parametrizing Dirhodium(II) Cavity Subspace. In the initial SMART application, the full cavity space was parametrized. This proved to be advantageous for an intermolecular C–H insertion as the second substrate enters the catalyst cavity and is directed towards the rhodium carbene (Figure 5, top). In the intramolecular cyclization, the site for C–H insertion is already within the pocket upon carbene formation, thus the space proximal to the rhodium is likely to be most influential to selectivity (Figure 5, bottom).

This analysis prompted the division of space within the SMART cavity into proximal vs distal with respect to the rhodium. Excluding the large, distal portion of the pocket allows for focused parametrization of the proposed active space of the cavity for the diastereoselectivity determining step. To accomplish this, a sphere was centered 2.0 Å from the rhodium (along the Rh-Rh vector) to simulate the position of a bound donor/donor carbene. The proximal cavity space was then separately parametrized from the full space.

It is generally recommended that the position of the probe be determined using information about the structure via computational or experimental methods. If a mechanistically guided "docking point" is not available, then consistency of the positioning and distance between the structure binding point and the molecular probe should be conserved across a data set.

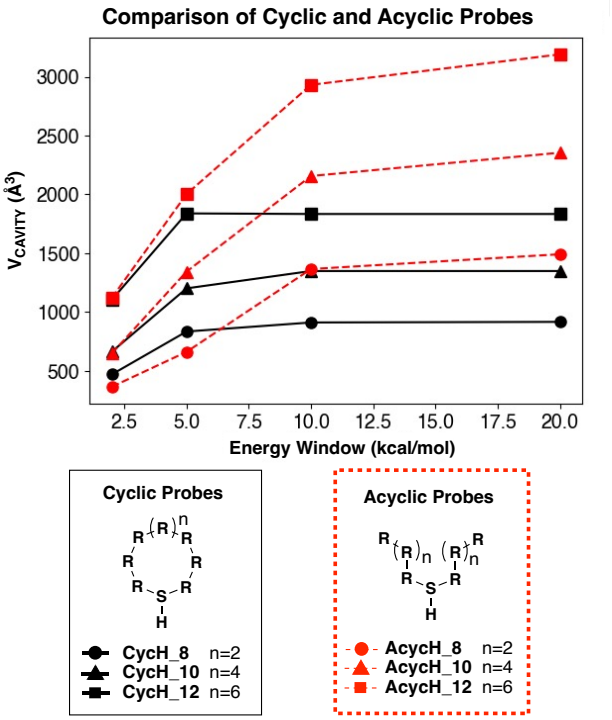


Figure 6. Comparison of V_{CAVITY} for molecular probes with an increasing energy window. Conformational sampling was performed using MacroModel.

DISCUSSION

Analyzing Molecular Probe Design. Users may wish to carefully tailor the probe structure to a specific system of interest, thus the careful design of a molecular probe is essential. The general SMART molecular probe is a feature with two main modes of modularity: shape and substituent radius (Figure 4a). Probe shape can significantly influence the determination of accessible pocket space. Acyclic probes allow for exploration of smaller areas with more hindrance, such as between dirhodium ligands, resulting in a more irregular pocket than macrocyclic probes (Figure 4b). Though both studies using SMART utilize macrocyclic probes acyclic probes are noteworthy variants that may be preferred in certain applications where high flexibility is essential, such as shape-dependent analysis.

Cavities generated using acyclic probes generally result in larger values for SMART descriptors due to their increased flexibility and therefore larger search space compared to macrocyclic probes. This is shown to impact V_{CAVITY} when varying the conformational search energy window (Figure 6).

Macrocyclic probes (**CycH_8**, **CycH_10**, **CycH_12**) reach maximum V_{CAVITY} quickly, and higher energy conformers are unable to continue to parametrize additional space by further window increases beyond 5.0 kcal/mol. These probes are more constrained in shape, generally resulting in more regular, spherical pockets. Acyclic probes (**AcycH_8**, **AcycH_10**, **AcycH_12**) explore more space (larger V_{CAVITY}) with higher conformer energy windows. The flexibility and narrow side arms of acyclic probes can access smaller cavities within a pocket of interest, such as

gaps and channels between ligands, parametrizing unique cavity space compared to macrocyclic probes.

Substituents bound to probes can also determine how small of space is accessible to the probe, and consequently the amount of detail in the resultant pocket information. Probes with H and F substituents from literature probes were compared as test cases. Smaller substituents (H) allow for exploration of space closer to the surrounding structure, resulting in a larger pocket on average. V_{CAVITY} computed by probes **CycH_12** and **CycF_12** show poor correlation at low V_{CAVITY}, indicating that they are disparately parametrizing highly confined cavities (Figure 7). The smaller **CycH_12** substituents increase flexibility of the probe, allowing it to explore tighter spaces more completely.

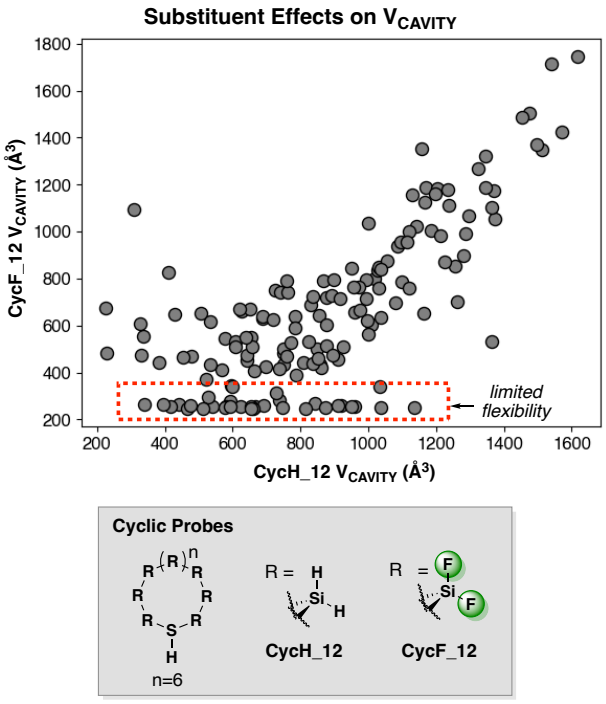


Figure 7. Comparison of H and F probe substituents. Descriptors do not correlate as well at low values of V_{CAVITY}. The lowest volumes are more limited using **CycF_12** instead of **CycH_12** due to less flexibility.

Comparison of SMARTpy Descriptors to Legacy. SMART descriptors computed trhough SMARTpy were shown to perform comparably to the original UCSF Chimera implementation. V_{CAVITY} and $proxV_{CAVITY}$ are well correlated between the two methods (Figure 11). Two outliers are observed that are not well correlated due to the number of probe conformers comprising the cavity ensemble. SMART descriptors calculated on sparse ensembles are more variable and dependent upon the conformation of the probe. The reduction of cavity featurization to a single conformer also eliminated the generality of the pocket information, as this conformer is more representative of where the molecular probe can go as opposed to a substrate.

SMARTpy computed V_{CAVITY} and $proxV_{CAVITY}$ are found to correlate well to Chimera-computed descriptors (Figure 8a). A few interesting outliers are observed in these correlations (Figure 8a, dashed line). These structures were visually assessed and found to have highly hindered pockets, resulting in only a single probe conformer fit. Such small probe ensembles are hypothesized to give disparate V_{CAVITY} due to the significant dependence upon the exact probe conformation, which are fit into the pocket using a stochastic algorithm.

V_{AREA} and *ESA* are also shown to correlate well to UCSF Chimera descriptors (Figure 8b). This correlation does not hold for smaller areas (Figure 8b , gray region), attributed again to the high variability of SMART descriptors for sparse probe ensembles. We again attribute this to the area of the conformer ensemble being highly variable for sparse ensembles. V_{AREA} is thus found to be less stable than V_{CAVITY}, suggesting that area descriptors should only used for dense ensembles.

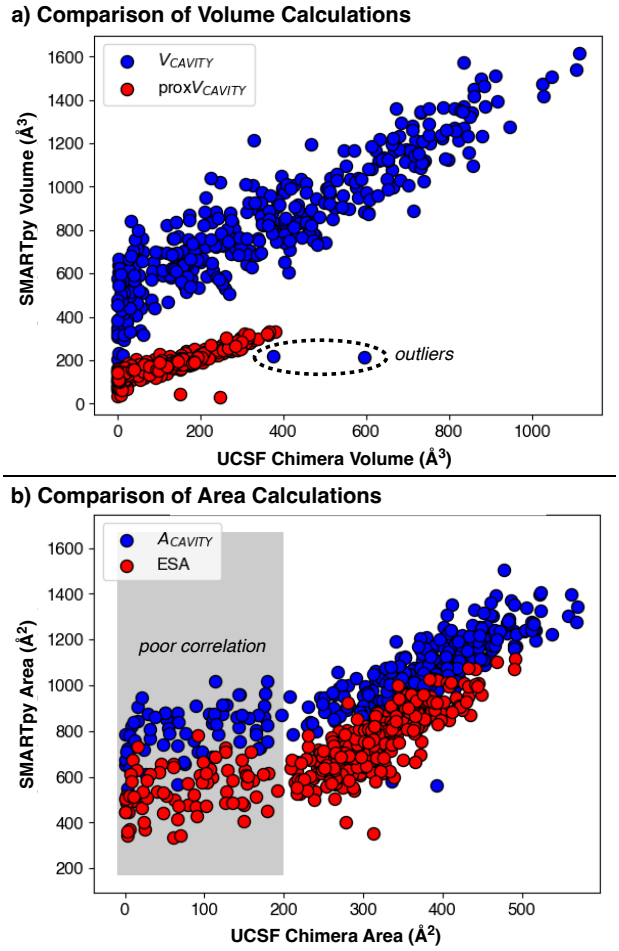


Figure 8. a) V_{CAVITY} (blue) and $proxV_{CAVITY}$ (red) computed using SMARTpy correlate well to the UCSF Chimera volume descriptors. The two outliers observed are thought to be an artifact of sparse conformational ensembles where the final V_{CAVITY} is more dependent on individual conformations than with larger ensembles. b) A_{CAVITY} (blue) and *ESA* (red) computed using SMARTpy correlate well to the UCSF Chimera area descriptors

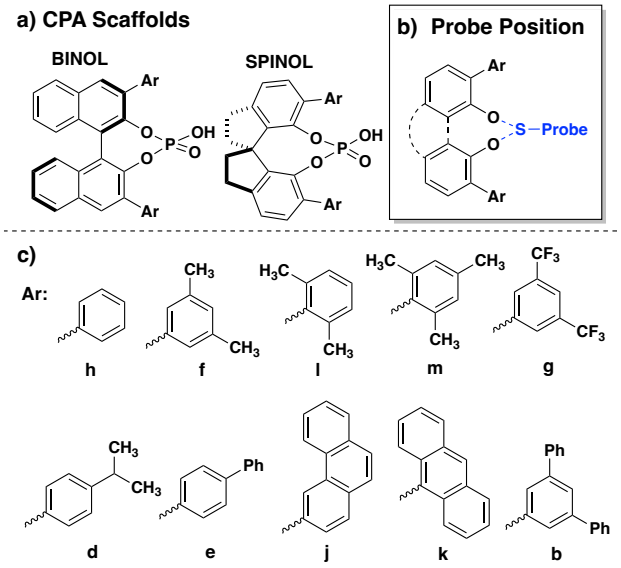


Figure 10. a) Structure of SPINOL and BINOL backbones. b) Probe positioning for CPA catalysts. The phosphoric acid was replaced by the molecular probe. c) Scope of substituents analyzed for both BINOL and SPINOL.

APPLICATIONS

SMART molecular descriptors are envisioned with broad applicability to the study and design of catalysts with irregular shapes. In this section, we demonstrate the utility of SMART for describing chiral phosphoric acids, enantioselective metalloenzyme catalysis, and protein side pockets by analyzing mechanistic implications of computed descriptors.

Chiral Phosphoric Acid Scaffolds. Chiral phosphoric acid (CPA) catalysts mediate a vast array of enantioselective transformations.[34] The axially chiral scaffold asymmetrically hinders the binding site around the phosphoric acid moiety, encouraging selectivity. Diverse CPA backbones and scaffolds have been designed to sterically modulate the phosphoric acid site. Some of the most employed scaffolds include BINOL and SPINOL backbones (Figure 10a). Variants of these scaffolds were considered to assess the ability of SMART to parametrize the steric hindrance of the reactive sites of CPAs (Figure 10c).

One design feature commonly leveraged is the confinement and rigidity of the binding pocket.[35] Similar to the dirhodium(II) catalysts, a more hindered CPA binding site is often connected to higher enantioselectivity. The dependence of CPA performance on 3,3' substitution was assessed by Goodman, showing that the positioning of steric bulk around the phosphoric acid controls reactivity by directing substrate orientation.[36,37] From this model of reactivity it was hypothesized that SMART descriptors could aid in the comparison and selection of sterically hindered CPA structures.

Due to the proposed proximal influence of the steric environment around the phosphoric acid moiety on selectivity, the probe was docked taking the place of the P atom in the BINOL and SPINOL backbones (Figure 10b). The original **CycH_12** molecular probe was implemented in the SMART workflow for these structures. Upon visual inspection of the docked scaffolds the probe was determined to be too long and would likely parametrize redundant space far from the binding site (Figure SX). While this could be resolved during the descriptor computation step by only considering *proxV*$_{CAVITY}$, we employed a shorter probe (**CycH_10**) to increase the speed of conformer generation.

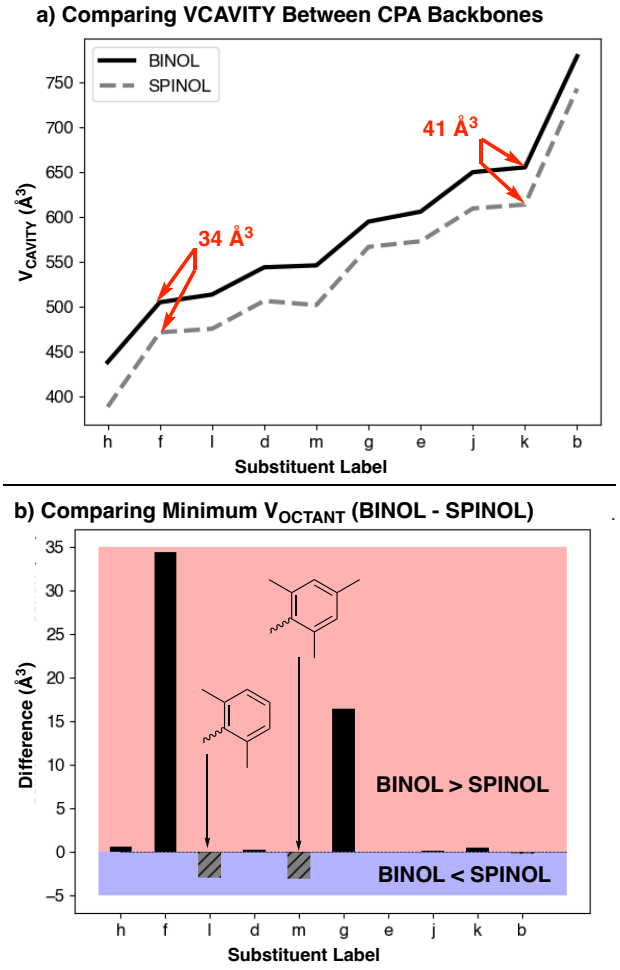


Figure 11. Difference between the minimum octant V$_{OCTANT}$ between BINOL and SPINOL backbones. Bars in the red region represent catalysts where the minimum BIONOL octant is larger than the SPINOL. Bars in the blue region represent catalysts where the minimum SPINOL octant is larger than the BINOL.

SMART Descriptor Analysis for Phosphoric Acid Catalysts. SPINOL catalysts were initially designed to provide more constrained and rigid reactive cavities than their BINOL analogs. Analysis of the V$_{CAVITY}$ for various substituted SPINOL and BINOL catalysts shows a linear correlation between backbones (Figure 11a) supporting linear scaling of substituent bulk between backbone scaffolds. SPINOL catalysts generally have a smaller V$_{CAVITY}$ than BINOL analogs (Figure Xb), supporting the initial design impetus for SPINOL scaffolds. A more complex trend between BINOL and SPINOL became apparent through analysis of *V*$_{OCTANT}$.

It is hypothesized that quadrant bulk plays an important role in substrate orientation. To remove bias in cavity

volume arising from the backbone scaffolds, octant analysis was performed for each catalyst, where only the positive-Z octants were considered for analysis (Figure SXa). The two backbone aryl C atoms were assigned as the XZ plane, and the probe tether atom defined the Z-axis (Figure SXb). The minimum V_{OCTANT} was found to vary in magnitude depending on both the substituent and the backbone. The minimum V_{OCTANT} for the BINOL backbone is larger than the SPINOL for substituents with a strict 3,5-substitution pattern (Figure 11b, red). Substituents with a 1,6- pattern generate a larger minimum octant for SPINOL backbones than BINOL (Figure 11b, blue).

Protein Binding Pockets. The binding of small molecules to protein receptors is fundamental to many biological processes. Assessment of the binding environment in protein active sites is crucial to the design of small molecule ligands and pharmaceuticals. Important features to assess in docking studies include the size and shape of the active cavity as shape matching influences binding. To illustrate the utility of SMART for quantifying protein binding pockets, the structure of the G-coupled protein receptor (GPR) was selected for analysis.

GPRs are responsible for a variety of biological functions,[38,39] and design of small molecule antagonists for GPRs is of interest in the field of computational drug design.[39,40] The structure and dynamics of the side binding pocket of the GPR101-Gs complex (PDB: 8W8R) have been shown to influence binding in computational antagonist design.[38] The structure of the GPR101–Gs protein was obtained from the PDB (PDB: 8W8R) and truncated to the side binding pocket of GPR101. Water molecules and ions were removed from the structure to allow space for the molecular probe. Multiple conformations of these proteins were not considered to reduce computational cost, but in principle this workflow could be applied to analyze pockets changes across conformational ensembles.

The significance of residues around the binding site can be difficult to discern, as dynamic, noncovalent interactions between the substrate and protein are influential to docking. Three residues were selected along the binding pocket to capture the local environments at different depths, represented by noncovalent attachment to different types of residues (Figure 12). The centroid of P30 was used as the binding reference to assess environment around the N-terminus. The C2 of the W441 residue was selected to parametrize the transmembrane domain between the N-terminus and the deeper region of the pocket. Finally, T111 was selected to probe the deeper region of the larger binding cavity. The default probe **CycH_12** was unable to dock in the side pocket without overlapping with protein residues. Due to the narrow shape of the binding pocket, a linear probe (**LinH_6**) was utilized for protein descriptor calculation.

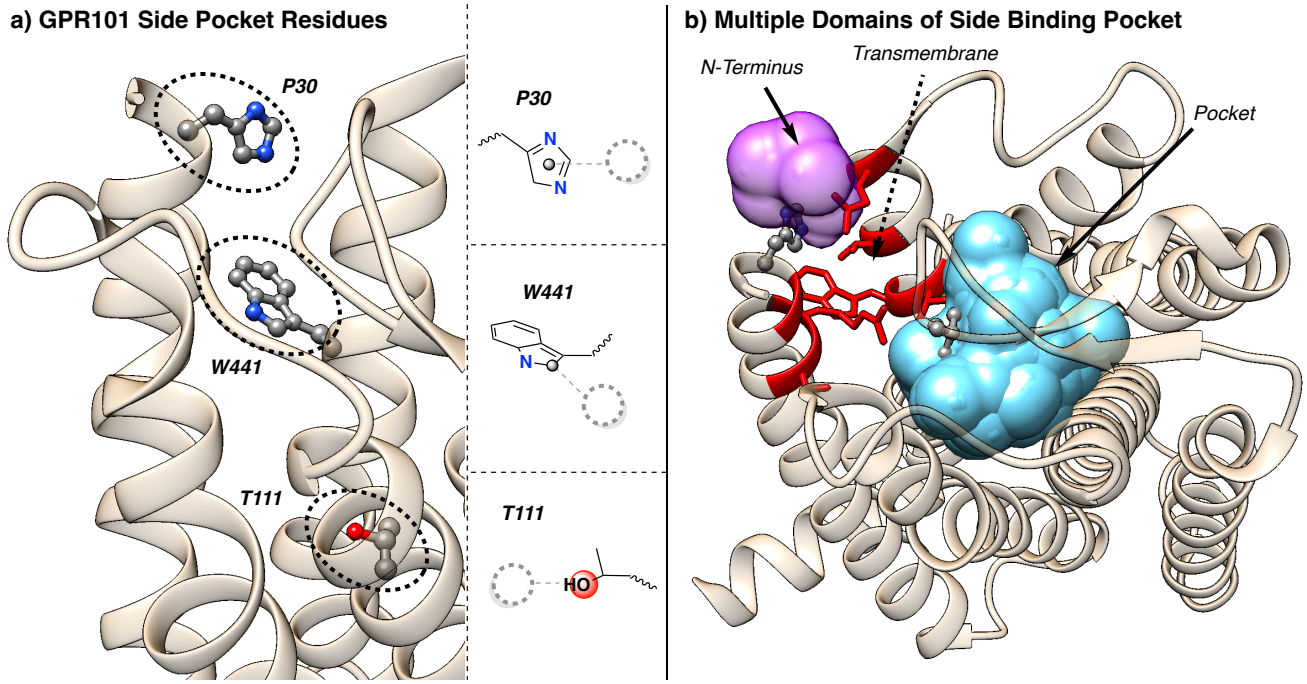


Figure 12. a) Select GPR101 (PDB: 8W8R) residues along the transmembrane domain and binding pocket. Depiction of probe positioning for each residue (center). b) SMART cavities for GPR101 (PDB: 8W8R) side biding pocket at different residues. The N-terminus is very hindered resulting in a small cavity (purple). The deepest region of the pocket (blue) is larger and likely has more flexibility in binding molecule features. The transmembrane between the two pockets is too hindered for the probe to enter. Either conformational dynamics or favorable electrostatic interactions are hypothesized to dictate binding in this domain.

The N-terminus is shown to be significantly hindered in GPR101 (Figure X, purple). Additionally, the environment around residue P30 is too hindered to fit a general molecular probe (Figure X, red). Based on this analysis, entrance of a small molecule into the side pocket is likely dictated by either protein flexibility to open the transmembrane domain, or by favorable electrostatic interactions with neighboring residues. The deepest part of the pocket is shown to be large and irregular in shape, which may promote the binding of diverse antagonists.

Selectivity in Fe-Porphyrin Enzymes. Enzymes with engineered reactive sites can induce highly selective transformations.[39–42] One well established transformation is intermolecular carbene insertion bio-catalyzed by Fe-porphyrin residues.[43–45] The orientation and approach of the substrate to the Fe-carbene intermediate influences the observed selectivity, thus the residues around the porphyrin site are often specifically targeted for mutations.

In 2022, Arnold disclosed a site selective C-H functionalization using engineered enzyme catalysts derived from the P411-PFA variant.[46] Three variants were assessed to provide insight into the structural relationship between active site residues and observed reactivity. We reasoned that SMART descriptors could provide additional insight into the porphyrin site proximal to Fe, representative of the approach of N-phenyl-morpholine to a Fe-carbene intermediate.

Enzyme structures were obtained from the PDB (IDs: 5UCW, 8DSG) and truncated to a single chain (A) for SMART analysis. Due to the structural significance of the bridging water (**w0**) in P411-PFA (8DSG), this molecule was retained in the truncated structure.[46] Remaining water, ion, and non-covalently bound residues were removed from each enzyme to allow for assessment of the empty cavity. The linear probe, **LinH_6**, was docked at the axial position of the Fe site to represent a bound carbene intermediate.

The major structural difference between P411-PFA and the P-4 variant used previously for selective amination is the perturbation of the helix directly over the binding site. In P4110PFA, a residue mutation induces a flip in orientation resulting in a site of increased steric hindrance. This artifact hinders the distal portion of the binding cavity, shown by a decrease in $distV_{CAVITY}$ from 260Å^3 to 237Å^3 (Figure 15a,b). This distal hindrance around the porphyrin may influence observed selectivity by restricting the approach of the N-phenyl-morpholine substrate.

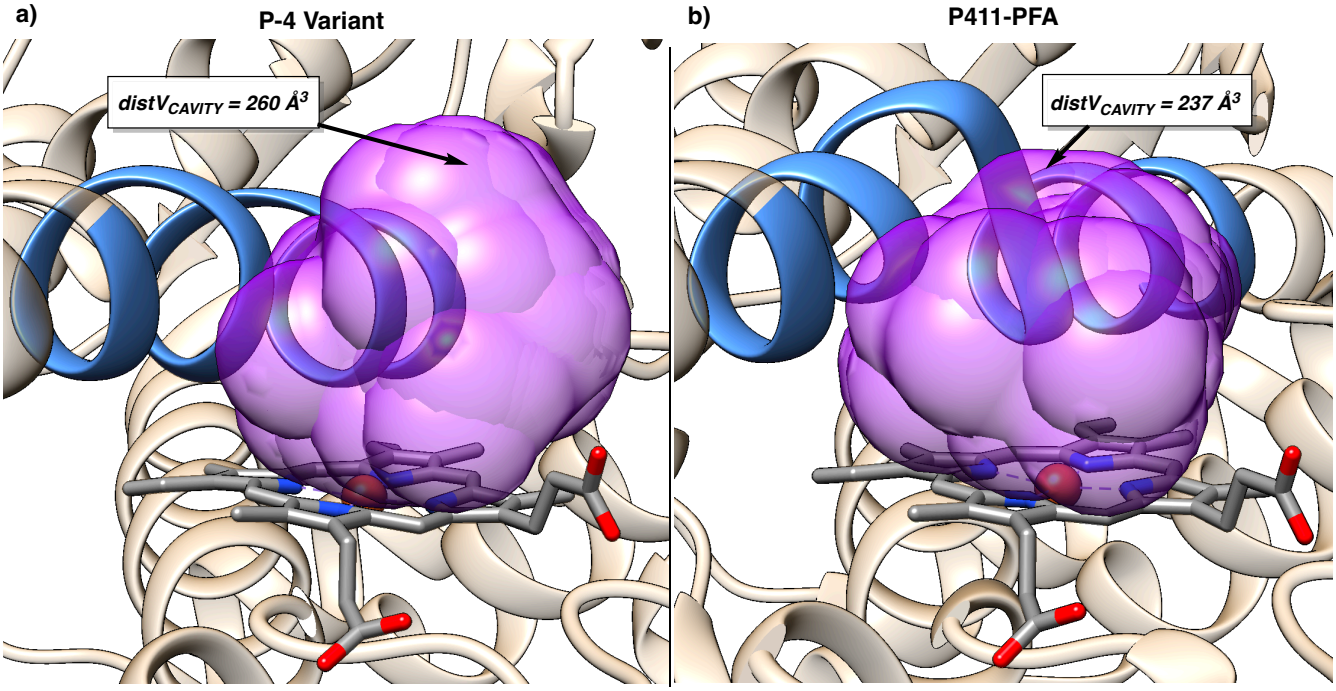


Figure 13. a) SMART cavity (purple) for P-4 variant (PDB: 5UCW). b) SMART cavity (purple) for P411-PFA enzyme (PDB: 8DSG). The bulge in the helix proximal to the porphyrin site exerts more hindrance on the cavity than in 5UCW.

CONCLUSION

Reactive cavities are difficult to sterically parametrize in mechanistically meaningful ways using traditional molecular descriptors. A free, open-source Python package, SMARTpy, is introduced to compute SMART molecular descriptors that have been disclosed in application to dirhodium(II) selectivity. SMART descriptors provide information about the steric environment within a reactive cavity from the perspective of a bound or docked substrate. Though designed for dirhodium catalysts, we envision a broad scope of applicability to diverse systems.

SMARTpy performs a template-based conformational search that generates an ensemble representative of the topology of the cavity. The choice of molecular probe is shown to influence the information obtained from SMART parameters. Acyclic probes are shown to generate highly irregular cavities, parametrizing the space between ligands. Macrocyclic probes generate regular, more spherical pockets due to rotational barriers. The flexibility of macrocyclic probes can be increased by the selection of small substituents bound to the core, such as H. This allows the probe to explore space closer to the ligands, resulting in a "high definition" representation of the cavity. Depending on the flexibility of the substrates coming together within the pocket in the transformation of interest, smaller or larger probes may be more suitable for generating SMART descriptors.

SMART descriptors were found to capture salient trends across BINOL and SPINOL CPAs. Lower V_{CAVITY} in

SPINOL catalysts supports the prevalence of higher selectivity compared to BINOL catalysts. SMARTpy was also demonstrated with a GPR101-Gs side binding pocket. The hindrance of the N-terminus and transmembrane domain are emphasized, suggesting that favorable non-covalent interactions are likely responsible for the initial procedure of small molecule binding. Finally, two enzymes used for different selective transformations are shown to differ in the distal region of the porphyrin cavity. The more hindered distal cavity observed in P411-PFA is hypothesized to constrain the approach of substrates to the Fe-carbene, directing selectivity for C-H functionalization.

In summary, SMART provides a convenient tool for the precise quantification of steric environments for complex, irregularly shaped 3D cavities, which are critical for controlling reactivity in disparate chemical and biochemical systems. Although non-covalent interactions are not currently supported for the generation of molecular probe ensembles, this is an area of current development.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website.

All Python codes, case study structures, SMART conformer ensembles, and computed descriptors used in case studies can be found on GitHub (https://github.com/SigmanGroup/SMART-molecular-descriptors.git)

Supporting Information (PDF)

AUTHOR INFORMATION

Corresponding Author

*Matthew S. Sigman – Department of Chemistry, University of Utah, Salt Lake City, Utah 84112, United States; orcid.org/0000-0002-5746-8830; Email: matt.sigman@utah.edu

Present Addresses

†If an author's address is different than the one given in the affiliation line, this information may be included here.

Author Contributions

All authors have given approval to the final version of the manuscript.

Funding Sources

Any funds used to support the research of the manuscript should be placed here (per journal style).

Notes

Any additional relevant notes should be placed here.

ACKNOWLEDGMENT

We are grateful for the insight from our beta-testers, especially Dr. James Howard. Support for this project was provided by the National Science Foundation (CHE- 2154502). The support and resources from the Center for High-Performance Computing at the University of Utah are gratefully acknowledged.

ABBREVIATIONS

SMART, Spatial Molding for Approachable Rigid Targets; CPA, chiral phosphoric acid;

REFERENCES

(1) Verloop, A. *The Sterimol Approach: Further Development of the Method and New Applications*; International Union of Pure and Applied Chemistry, 1983. https://doi.org/10.1016/b978-0-08-029222-9.50051-2.

(2) Brethomé, A. V.; Fletcher, S. P.; Paton, R. S. Conformational Effects on Physical-Organic Descriptors: The Case of Sterimol Steric Parameters. *ACS Catal* **2019**, *9* (3), 2313–2323. https://doi.org/10.1021/acscatal.8b04043.

(3) Clavier, H.; Nolan, S. P. Percent Buried Volume for Phosphine and N-Heterocyclic Carbene Ligands: Steric Properties in Organometallic Chemistry. *Chemical Communications* **2010**, *46* (6), 841–861. https://doi.org/10.1039/b922984a.

(4) Wu, K.; Doyle, A. G. Parameterization of Phosphine Ligands Demonstrates Enhancement of Nickel Catalysis via Remote Steric Effects. *Nat Chem* **2017**, *9* (8), 779–784. https://doi.org/10.1038/NCHEM.2741.

(5) Cammarota, R. C.; Liu, W.; Bacsa, J.; Davies, H. M. L.; Sigman, M. S. Mechanistically Guided Workflow for Relating Complex Reactive Site Topologies to Catalyst Performance in C − H Functionalization Reactions. **2021**, *2*. https://doi.org/10.1021/jacs.1c12198.

(6) Boni, Y. T.; Cammarota, R. C.; Liao, K.; Sigman, M. S.; Davies, H. M. L. Leveraging Regio- and Stereoselective C(Sp3)-H Functionalization of Silyl Ethers to Train a Logistic Regression Classification Model for Predicting Site-Selectivity Bias. *J Am Chem Soc* **2022**, *144* (34), 15549–15561. https://doi.org/10.1021/jacs.2c04383.

(7) Souza, L. W.; Miller, B. R.; Cammarota, R. C.; Lo, A.; Lopez, I.; Shiue, Y.-S.; Bergstrom, B. D.; Dishman, S. N.; Fettinger, J. C.; Sigman, M. S.; Shaw, J. T. Deconvoluting Nonlinear Catalyst–Substrate Effects in the Intramolecular Dirhodium-Catalyzed C–H Insertion of Donor/Donor Carbenes Using Data Science Tools. *ACS Catal* **2023**, 104–115. https://doi.org/10.1021/acscatal.3c04256.

(8) Qin, C.; Davies, H. M. L. Role of Sterically Demanding Chiral Dirhodium Catalysts in Site-Selective C-H Functionalization of Activated Primary C-H Bonds. *J Am Chem Soc* **2014**, *136* (27), 9792–9796. https://doi.org/10.1021/ja504797x.

(9) Hansen, J.; Davies, H. M. L. High Symmetry Dirhodium(II) Paddlewheel Complexes as Chiral Catalysts. *Coord Chem Rev* **2008**, *252* (5–7), 545–555. https://doi.org/10.1016/j.ccr.2007.08.019.

(10) Davies, H. M. L.; Morton, D. Guiding Principles for Site Selective and Stereoselective Intermolecular C–H Functionalization by Donor/Acceptor Rhodium Carbenes. *Chem Soc Rev* **2011**, *40* (4), 1857–1869. https://doi.org/10.1039/c0cs00217h.

(11) Le Guilloux, V.; Schmidtke, P.; Tuffery, P. Fpocket: An Open Source Platform for Ligand Pocket Detection. *BMC Bioinformatics* **2009**, *10*, 1–11. https://doi.org/10.1186/1471-2105-10-168.

(12) Kochnev, Y.; Durrant, J. D. FPocketWeb: Protein Pocket Hunting in a Web Browser. *J Cheminform*

2022, *14* (1), 1–7. https://doi.org/10.1186/s13321-022-00637-0.

(13) Tian, W.; Chen, C.; Lei, X.; Zhao, J.; Liang, J. CASTp 3.0: Computed Atlas of Surface Topography of Proteins. *Nucleic Acids Res* **2018**, *46* (W1), W363–W367. https://doi.org/10.1093/nar/gky473.

(14) Durrant, J. D.; Votapka, L.; Sørensen, J.; Amaro, R. E. POVME 2.0: An Enhanced Tool for Determining Pocket Shape and Volume Characteristics. *J Chem Theory Comput* **2014**, *10* (11), 5047–5056. https://doi.org/10.1021/ct500381c.

(15) Liang, J.; Edelsbrunner, H.; Woodward, C. Anatomy of Protein Pockets and Cavities: Measurement of Binding Site Geometry and Implications for Ligand Design. *Protein Science* **1998**, *7* (9), 1884–1897. https://doi.org/10.1002/pro.5560070905.

(16) Ebalunode, J. O.; Ouyang, Z.; Liang, J.; Zheng, W. Novel Approach to Structure-Based Pharmacophore Search Using Computational Geometry and Shape Matching Techniques. *J Chem Inf Model* **2008**, *48* (4), 889–901. https://doi.org/10.1021/ci700368p.

(17) Wirth, M.; Volkamer, A.; Zoete, V.; Rippmann, F.; Michielin, O.; Rarey, M.; Sauer, W. H. B. Protein Pocket and Ligand Shape Comparison and Its Application in Virtual Screening. *J Comput Aided Mol Des* **2013**, *27* (6), 511–524. https://doi.org/10.1007/s10822-013-9659-1.

(18) Kahraman, A.; Morris, R. J.; Laskowski, R. A.; Thornton, J. M. Shape Variation in Protein Binding Pockets and Their Ligands. *J Mol Biol* **2007**, *368* (1), 283–301. https://doi.org/10.1016/j.jmb.2007.01.086.

(19) Edelsbrunner, H.; Mücke, E. P. Three-Dimensional Alpha Shapes. *ACM Transactions on Graphics (TOG)* **1994**, *13* (1), 43–72. https://doi.org/10.1145/174462.156635.

(20) Tian, W.; Liang, J. On Quantification of Geometry and Topology of Protein Pockets and Channels for Assessing Mutation Effects. *2018 IEEE EMBS International Conference on Biomedical and Health Informatics, BHI 2018* **2018**, *2018-Janua* (March), 263–266. https://doi.org/10.1109/BHI.2018.8333419.

(21) Edelsbrunner, H.; Facello, M.; Liang, J. On the Definition and the Construction of Pockets in Macromolecules. *Discrete Appl Math (1979)* **1998**, *88* (1–3), 83–102. https://doi.org/10.1016/S0166-218X(98)00067-5.

(22) Richards, F. M. AREAS, VOLUMES, PACKING, AND PROTEIN STRUCTURE. *Ann. Rev. Biphys. Bioneg.* **1977**, *6*, 151–176.

(23) Zahrt, A. F.; Henle, J. J.; Rose, B. T.; Wang, Y.; Darrow, W. T.; Denmark, S. E. Prediction of Higher-Selectivity Catalysts by Computer-Driven Workflow and Machine Learning. *Science (1979)* **2019**, *363* (6424). https://doi.org/10.1126/science.aau5631.

(24) Meller, A.; Ward, M.; Borowsky, J.; Kshirsagar, M.; Lotthammer, J. M.; Oviedo, F.; Ferres, J. L.; Bowman, G. R. Predicting Locations of Cryptic Pockets from Single Protein Structures Using the PocketMiner Graph Neural Network. *Nat Commun* **2023**, *14* (1), 1–15. https://doi.org/10.1038/s41467-023-36699-3.

(25) Shen, L.; Fang, J.; Liu, L.; Yang, F.; Jenkins, J. L.; Kutchukian, P. S.; Wang, H. Pocket Crafter: A 3D Generative Modeling Based Workflow for the Rapid Generation of Hit Molecules in Drug Discovery. *J Cheminform* **2024**, *16* (1), 1–17. https://doi.org/10.1186/s13321-024-00829-w.

(26) Feng, W.; Wang, L.; Lin, Z.; Zhu, Y.; Wang, H.; Dong, J.; Bai, R.; Wang, H.; Zhou, J.; Peng, W.; Huang, B.; Zhou, W. Generation of 3D Molecules in Pockets via a Language Model. *Nat Mach Intell* **2024**, *6* (1), 62–73. https://doi.org/10.1038/s42256-023-00775-6.

(27) Kudo, G.; Hirao, T.; Yoshino, R.; Shigeta, Y.; Hirokawa, T. Pocket to Concavity: A Tool for the Refinement of Protein-Ligand Binding Site Shape from Alpha Spheres. *Bioinformatics* **2023**, *39* (4), 1–3. https://doi.org/10.1093/bioinformatics/btad212.

(28) Wang, L.; Bai, R.; Shi, X.; Zhang, W.; Cui, Y.; Wang, X.; Wang, C.; Chang, H.; Zhang, Y.; Zhou, J.; Peng, W.; Zhou, W.; Huang, B. A Pocket-Based 3D Molecule Generative Model Fueled by Experimental Electron Density. *Sci Rep* **2022**, *12* (1), 1–14. https://doi.org/10.1038/s41598-022-19363-6.

(29) Roos, K.; Wu, C.; Damm, W.; Reboul, M.; Stevenson, J. M.; Lu, C.; Dahlgren, M. K.; Mondal, S.; Chen, W.; Wang, L.; Abel, R.; Friesner, R. A.; Harder, E. D. OPLS3e: Extending Force Field Coverage for Drug-Like Small Molecules. *J Chem Theory Comput* **2019**, *15* (3), 1863–1874. https://doi.org/10.1021/acs.jctc.8b01026.

(30) Pettersen, E. F.; Goddard, T. D.; Huang, C. C.; Couch, G. S.; Greenblatt, D. M.; Meng, E. C.; Ferrin, T. E. UCSF Chimera - A Visualization System for Exploratory Research and Analysis. *J Comput Chem* **2004**, *25* (13), 1605–1612. https://doi.org/10.1002/jcc.20084.

(31) Watts, K. S.; Dalal, P.; Murphy, R. B.; Sherman, W.; Friesner, R. A.; Shelley, J. C. ConfGen: A Conformational Search Method for Efficient Generation of Bioactive Conformers. *J Chem Inf Model* **2010**, *50* (4), 534–546. https://doi.org/10.1021/ci100015j.

(32) Sullivan, C.; Kaszynski, A. PyVista: 3D Plotting and Mesh Analysis through a Streamlined Interface for the Visualization Toolkit (VTK). *J Open Source Softw* **2019**, *4* (37), 1450. https://doi.org/10.21105/joss.01450.

(33) Li, J.; Grosslight, S.; Miller, S. J.; Sigman, M. S.; Toste, F. D. Site-Selective Acylation of Natural Products with Binol-Derived Phosphoric Acids. *ACS Catal* **2019**, *9* (11), 9794–9799. https://doi.org/10.1021/acscatal.9b03535.

(34) Parmar, D.; Sugiono, E.; Raja, S.; Rueping, M. Complete Field Guide to Asymmetric BINOL-Phosphate Derived Brønsted Acid and Metal Catalysis: History and Classification by Mode of Activation; Brønsted Acidity, Hydrogen Bonding, Ion Pairing, and Metal Phosphates. *Chem Rev* **2014**, *114* (18), 9047–9153. https://doi.org/10.1021/cr5001496.

(35) Schreyer, L.; Properzi, R.; List, B. IDPi Catalysis. *Angewandte Chemie - International Edition* **2019**, *58* (37), 12761–12777. https://doi.org/10.1002/anie.201900932.

(36) Reid, J. P.; Goodman, J. M. Goldilocks Catalysts: Computational Insights into the Role of the 3,3′ Substituents on the Selectivity of BINOL-Derived Phosphoric Acid Catalysts. *J Am Chem Soc* **2016**, *138* (25), 7910–7917. https://doi.org/10.1021/jacs.6b02825.

(37) Reid, J. P.; Simón, L.; Goodman, J. M. A Practical Guide for Predicting the Stereochemistry of Bifunctional Phosphoric Acid Catalyzed Reactions of Imines. *Acc Chem Res* **2016**, *49* (5), 1029–1041. https://doi.org/10.1021/acs.accounts.6b00052.

(38) Yang, Z.; Wang, J. Y.; Yang, F.; Zhu, K. K.; Wang, G. P.; Guan, Y.; Ning, S. L.; Lu, Y.; Li, Y.; Zhang, C.; Zheng, Y.; Zhou, S. H.; Wang, X. W.; Wang, M. W.; Xiao, P.; Yi, F.; Zhang, C.; Zhang, P. J.; Xu, F.; Liu, B. H.; Zhang, H.; Yu, X.;

Gao, N.; Sun, J. P. Structure of GPR101–Gs Enables Identification of Ligands with Rejuvenating Potential. *Nat Chem Biol* **2023**, *20* (April). https://doi.org/10.1038/s41589-023-01456-6.

(39) Athavale, S. V.; Gao, S.; Das, A.; Mallojjala, S. C.; Alfonzo, E.; Long, Y.; Hirschi, J. S.; Arnold, F. H. Enzymatic Nitrogen Insertion into Unactivated C-H Bonds. *J Am Chem Soc* **2022**, *144* (41), 19097–19105. https://doi.org/10.1021/jacs.2c08285.

(40) Chen, K.; Huang, X.; Jennifer Kan, S. B.; Zhang, R. K.; Arnold, F. H. Enzymatic Construction of Highly Strained Carbocycles. *Science (1979)* **2018**, *360* (6384), 71–75. https://doi.org/10.1126/science.aar4239.

(41) Prier, C. K.; Zhang, R. K.; Buller, A. R.; Brinkmann-Chen, S.; Arnold, F. H. Enantioselective, Intermolecular Benzylic C-H Amination Catalysed by an Engineered Iron-Haem Enzyme. *Nat Chem* **2017**, *9* (7), 629–634. https://doi.org/10.1038/nchem.2783.

(42) Liu, Z.; Calvó-Tusell, C.; Zhou, A. Z.; Chen, K.; Garcia-Borràs, M.; Arnold, F. H. Dual-Function Enzyme Catalysis for Enantioselective Carbon–Nitrogen Bond Formation. *Nat Chem* **2021**, *13* (12), 1166–1172. https://doi.org/10.1038/s41557-021-00794-z.

(43) Rogge, T.; Zhou, Q.; Porter, N. J.; Arnold, F. H.; Houk, K. N. Iron Heme Enzyme-Catalyzed Cyclopropanations with Diazirines as Carbene Precursors: Computational Explorations of Diazirine Activation and Cyclopropanation Mechanism. *J Am Chem Soc* **2024**, *146* (5), 2959–2966. https://doi.org/10.1021/jacs.3c06030.

(44) Yang, Y.; Arnold, F. H. Navigating the Unnatural Reaction Space: Directed Evolution of Heme Proteins for Selective Carbene and Nitrene Transfer. *Acc Chem Res* **2021**, *54* (5), 1209–1225. https://doi.org/10.1021/acs.accounts.0c00591.

(45) Chen, K.; Arnold, F. H. Engineering Cytochrome P450s for Enantioselective Cyclopropenation of Internal Alkynes. *J Am Chem Soc* **2020**, *142* (15), 6891–6895. https://doi.org/10.1021/jacs.0c01313.

(46) Zhang, J.; Maggiolo, A. O.; Alfonzo, E.; Mao, R.; Porter, N. J.; Abney, N. M.; Arnold, F. H. Chemodivergent C(Sp 3)–H and C(Sp 2)–H Cyanomethylation Using Engineered Carbene Transferases. *Nat Catal* **2023**, *6* (2), 152–160. https://doi.org/10.1038/s41929-022-00908-x.