Title
Computational Chemistry Studies Relevant to Medicinal Chemistry

Author
Gingrich, Phillip W

Publication Date
2023

Peer reviewed|Thesis/dissertation

Computational Chemistry Studies Relevant to Medicinal Chemistry

By

Phillip Gingrich

DISSERTATION

Submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

Chemistry

in the

OFFICE OF GRADUATE STUDIES

of the

UNIVERSITY OF CALIFORNIA

DAVIS

Approved:

Dean J. Tantillo, Ph.D. (Co-Chair)

Justin B. Siegel, Ph.D. (Co-Chair)

David B. Goodin, Ph.D.

Committee in Charge

2022

Acknowledgements

The last few years have been a rewarding experience in my journey toward becoming a better scientist. While I have gained new skills and insights that I will carry with me, I am most thankful for those who supported, challenged, and allowed me to pursue my doctoral degree.

First, I appreciate my management team at my full-time employer for believing that advancing my education was a worthwhile endeavor for the organization. Through the conduct of my research and coursework, I feel that I have bolstered my background in biology and data science, and that has equipped me to work in a more interdisciplinary fashion for our team. I know I have already been able to leverage my newfound knowledge in my position, and I look forward to further contributing to collaborative projects in the future.

Second, I want to thank my advisors Dean and Justin for their willingness to take me on as a non-traditional graduate student and provide me the benefit of their experiences. In all our interactions, I felt like I was engaged professionally and congenially, and that empowered me to try new things and follow the science wherever it took me. The diversity between their research interests, as well as those across the entire Chemistry department, exposed me to areas of chemistry with which I was unfamiliar. That diversity afforded me opportunities think critically and to collegiately question the work presented by others considering my own background and experience. I'm especially appreciative of the opportunity to expand my knowledge in collaborative projects, notably with David Olson and his students.

Lastly and most importantly, I'm thankful for the support of my wife and family. Between graduate school, my full-time employment, and my service in the Army National Guard, I have felt stretched thin across multiple fronts. My children have been understanding

and patient, and Alicia's love and support sustained me when the piles of work seemed rather tall. As I rebalance my time and priorities with graduate school ending, I hope to repay my family for all their contributions to my professional development.

Abstract

This dissertation summarizes original work relevant to product predictions for Cytochrome P450 (CYP450) catalyzed transformations using a combination of computationally affordable methods, specifically modern force field and semi-empirical methods and protein-ligand docking. Additionally, it highlights multiple applications of Density Functional Theory (DFT) in collaboration with our synthetic chemistry colleagues to explore and explain photoisomerization, redox chemistry, and reaction mechanisms.

Firstly, Cytochrome P450s (CYP450) are metabolically and synthetically important enzymes, catalyzing an array of oxidative transformations across all kingdoms of life. The prediction of oxidative products resulting from CYP450 catalyzed transformations is historically challenging and often relies only on enzyme-substrate fit and binding affinity estimates while neglecting measures of reactivity. Herein we present computationally affordable methodology for estimating epoxidation and hydroxylation barriers. When predicted hydroxylation barriers are paired with traditional protein-ligand docking, we improve on previously published prediction success rates and open the door to enzyme design in CYP450s for the purpose of achieving novel biosynthetic outcomes.

In Chapter 1, epoxidation barriers were predicted using a multiple linear regression model with the fractional occupation number weighted density (FOD) and orbital weighted Fukui index (f_w^+) as descriptors localized to the vinylic carbon atom involved in the initial C–O bond formation event during epoxidation. Relative to previously computed epoxidation barriers using density functional theory in a panel of 36 compounds, mean absolute errors of 0.66 and 0.70 kcal/mol were achieved in the training and test sets, respectively, with coefficients of

determination of ca. 0.80 were. This was done at the GFN2-xTB//GFN-FF level of theory. By performing electronic structure calculations on force field generated geometries, this approach is highly scalable.

In Chapter 2, a single linear regression model was built to predict hydrogen atom transfer (HAT) barriers following the formation of Compound I, relevant to CYP450-mediated hydroxylations. The C–H bond dissociation energy involving a "frozen radical" – that is the removal of a hydrogen atom from an sp^3 hybridized carbon in the substrate followed by single point energy calculations as doublets for the resulting unoptimized substrate radical and hydrogen atom – was found to correlate well with hydrogen atom transfer barriers previously computed with density functional theory. At the GFN2-xTB//GFN-FF level of theory for a panel of 24 sp^3 hybridized carbon atoms across 21 substrates, mean absolute errors of ca. 1 kcal/mol were achieved in both training and test sets. By again leveraging force field and semi-empirical methods, this approach will scale to thousands of structures on even a modest computing resource.

In Chapter 3, hydroxylation product predictions are made by combining enzyme-substrate docking and HAT barrier regression modelling. Hydroxylated product formation certainly relies significantly on the fit and binding affinity of a substrate with a CYP450 enzyme and not on the HAT barrier with Compound I alone. To this end, HAT barriers predicted using regression modeling were combined with O_{heme}–$H_{substrate}$ constrained docking and pose clustering to make product predictions on a set of 25 substrates for the CYP101A1 camphor 5-monooxygenase enzyme. Using RxDock as an example utility used in high throughput virtual screening (HTVS), the prediction success rate for any hydroxylation product was 84% in the top

two predictions when HAT barriers were included, compared to only 80% without the inclusion of HAT barriers. Combining HAT barriers and docking scores from Rosetta, any hydroxylation product was successfully predicted in the top two predictions in 92% of the 25 substrates studied. More importantly, the primary hydroxylation product prediction success rate was 84% in the top two predictions. Collectively, these findings meet or exceed the performance of previously published results in a non-parametric fashion. More importantly, the performance using Rosetta indicates our combination of docking and HAT barriers holds tremendous promise in the application of enzyme design.

In the second half of this work, theoretical calculations were employed to rationalize experimental outcomes. Such retrospective analysis tends to be employed when experimental observations fail to meet our preconceived chemical intuition. By coupling wet experiment with quantum chemical theory, we can gain insight into underlying electronic structures and, in doing so, better understand and even predict spectroscopic or thermochemical properties in our systems of interest. To this end and in collaboration with the laboratory of David Olsen, we explored three series of experimental findings using Density Functional Theory in the sort of post-hoc fashion described above. These three efforts focused on the spectroscopic properties (and limitations) of azobenzene photoswitches,[1] oxidation potentials relevant to Baeyer-Mills reactions,[2] and the samarium-mediated rearrangement of vinyl aziridines to afford more complex heterocycles. In all three cases, computational efforts followed behind the experiment and aimed to generate models that explained the Olsen's labs findings, as well as affording methodology that could be used to further expand their work ahead of experimental efforts.

In Chapter 4, the photoisomerization of acylhydrazone-functionalized azobenzene derivatives is explored. With seemingly two photoswitchable motifs present, our collaborators only observed *E* to *Z* photoisomerization across the azobenzene substructure. Using Time-dependent Density Functional Theory (TD-DFT), the π to π* transition at approximately 380 nm is predicted to have a strongly localized electron density difference over the azo motif between the ground and excited state, with no discernable difference predicted over the acyl hydrazone functionality. Additionally, substituent effects were studied for a handful of electron withdrawing and donating cases explored synthetically, all showing no π to π* transition near 300 nm, inconsistent with other acylhydrazone photoswitches. This study rationalized the findings of our collaborators, and while retrospective analysis is useful, this study further highlights the opportunity to leverage computational techniques prospectively to guide synthetic efforts.

In Chapter 5, retrospective analysis of synthetic findings was again conducted. In this application, the Bayer-Mills reaction is a traditional route to azobenzenes by way of a condensation reaction. However, azoxybenzene side products are also formed. Here, we attempted to correlate the formation of azoxybenzene with one electron oxidation potentials computed with DFT. In this work, electron-rich aniline derivatives with low oxidation potentials were found to produce undesirable levels of the azoxybenzene product, and we demonstrate that the computed oxidation potential from DFT with implicit solvation is a useful descriptor in predicting the outcome of the Baeyer-Mills reaction for given reactants.

Lastly in Chapter 6, access to vinyl aziridines is explored mechanistically using traditional stationary point searching with DFT. Our collaborators discovered that vinyl aziridines could

undergo ring expansion in the presence of samarium (II) iodide. While simple Lewis-acid promoted expansions are known, we explored a radical mechanism consistent with samarium (II) iodide mediated single electron transfer reactions observed in reductions and cross-couplings of ketones. From our analyses at the PBE0-D3BJ/def2-TZVP (ECP = Sm, I; SMD = toluene)//PBE0/def2-SVP (ECP = Sm, I) level of theory, a radical mechanism on the septet spin surface is achievable thermally at room temperature, with an overall free energy barrier of 25.1 kcal/mol and a strong thermodynamic driving force to favor the product-catalyst complex by 22.1 kcal/mol, both relative to the reactant-catalyst complex. These findings corroborate those of our synthetic colleagues and suggest that the transformation occurs according to a single electron transfer mechanism. This affords a mechanistically differentiated route to substituted 3-pyrrolines.

In all, this work showcases multiple applications of computational chemistry that are relevant to protein engineering and medicinal chemistry, with an aim toward increased prospective use in the design of experiments.

Chapter 1: Prediction of Epoxidation Barriers

Introduction

Of the multiple oxidative transformations that cytochrome P450s catalyze,[3] hydroxylation is best known and most widely studied. Often, hydroxylation is critical for the overall metabolism of xenobiotics within humans (and other organisms), where the increased hydrophilicity of the hydroxylated product is necessary for excreting downstream metabolites through the urine, or hydroxylated products are further functionalized and marked for removal. In addition to hydroxylation, other reactivities, such as N- or S-oxidation, dealkylation, and dehalogenation, are known to be catalyzed by P450s.[4] Among these, epoxidation of alkenes (and arenes) is particularly common.[5]

Safety profiling is important for epoxides as they are not typically innocuous. Given the intrinsic strain in the 3-membered ring, coupled with the strongly electronegative oxygen atom, the carbon atoms in an epoxide are considerably electrophilic, potentially making epoxide containing metabolites strong alkylating agents toward biologically important compounds such as DNA.[6] For example, aflatoxin B1 exo-8,9-epoxide covalently binds to guanine residues at the N7 position, making aflatoxin B1 (via this epoxide) a known hepatocarcinogen.[7] To this end, the prediction of epoxidation products from P450-catalyzed transformations is especially important for drug design and metabolism predictions, as such predictions could help elucidate the origins of, and even anticipate, off-target effects.

To predict downstream metabolites, docking studies are often included in high throughput screening campaigns to evaluate binding modes and to predict associated binding affinities.[8] However, docking alone neglects the importance of the electronic structure of a

substrate or residence time[9] in determining its susceptibility toward epoxidation by Compound 0 or 1 in the P450 catalytic cycle (Figure 1-1).[4, 10, 11]

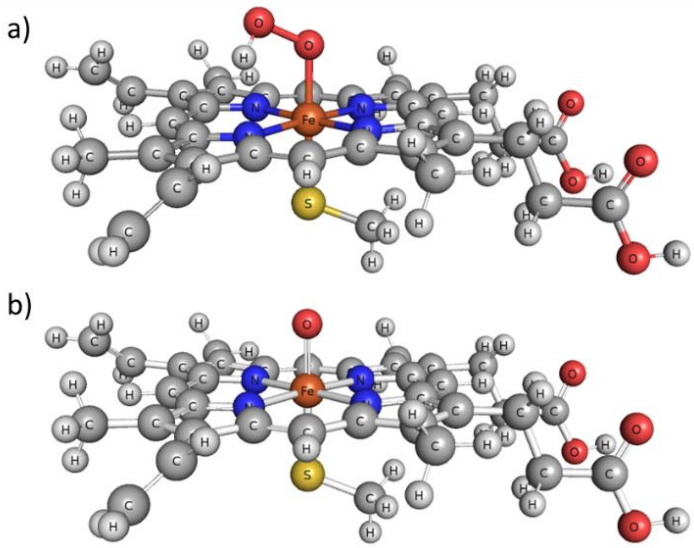

Figure 1-1 Structures of a) Compound 0 and b) Compound 1, the potential oxidants in the catalytic cycle of a cytochrome P450 enzymes. Each is shown ligated to a methyl thiolate axially and with both propionate substituents in their protonated form.

The inclusion of reactivity information, particularly in the form of epoxidation barriers, alongside binding affinities from docking trials, could prove useful in assessing a docking pose as productive or unproductive. Traditionally, such barriers would be computed using density functional theory (DFT) through stationary point analyses using a truncated Compound 1 model.[4, 12] However, within high throughput virtual screening, such calculations are prohibitively expensive, creating a need for more affordable methods. Cheminformatics and regression approaches fit nicely in this space.

Promising work by Zhang and Liu demonstrated a correlation between DFT-computed epoxidation barriers and ionization potentials for a panel of 36 alkenes with varying electron-withdrawing and -donating groups present (Figure 1-2).[13] In that study, computed adiabatic

ionization potentials (IP) in continuum solvent were used to build two linear models based on substrate polarity as determined by computed dipole moments. Because of the level of theory chosen and the geometry optimizations required for calculating non-vertical ionization potentials, Zhang and Liu's exact approach is too expensive for routine use, though it is much faster than transition state searching methods. Moreover, a unified model that does not depend on a compound's computed dipole moment would be preferable from a simplicity standpoint, if for no other reason. Additionally, setting an exact threshold for the molecular dipole moment to assess polarity is open to subjective assessment, and the calculation of the molecular dipole moment will vary with the selected level of theory. In their work, models for polar and non-polar compounds account for more than 95% of the variability in the epoxidation barrier by the IP alone.[13] When polar and non-polar compounds were combined from the entire data set into a single model, however, the coefficient of determination was only 0.768 and a mean absolute error (MAE) of 0.96 kcal/mol was observed. Further, the removed electron in an IP calculation originates from a molecular orbital that may not correspond to a π-type bonding orbital localized to the alkene. For example, compounds containing aliphatic amines or thioethers would likely ionize by way of an electron being removed from a non-bonding (lone pair) orbital localized to such heteroatoms. An ionization involving a non-bonding electron from a heteroatom electronically isolated from the alkene of interest would not be a useful descriptor as it would fail to capture the electronic character of the alkene undergoing epoxidation. A more localized approach is required.

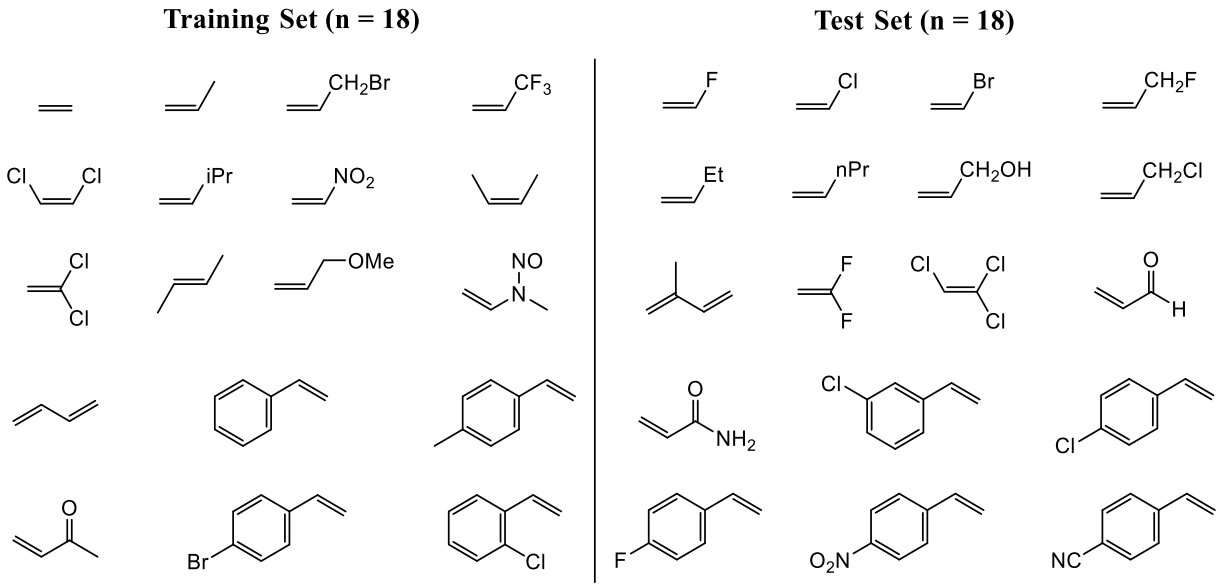

Figure 1-2 Training and test set compounds for which epoxidation barriers involving a truncated Compound 1 model were previously computed using DFT.[13]

To address these challenges, we have developed a computationally affordable method to accurately estimate epoxidation barriers combining two local descriptors, the fractional occupation number weighted density (FOD)[14] and the orbital weighted Fukui index (f_{w}^{+}),[15] through a multiple linear regression (MLR) model. As true values, we reuse, in accordance with FAIR data principles,[16] the computed zero-point energy corrected potential energy barriers on the quartet surface from Zhang and Liu's work (which were provided in the supporting information while not utilized in their presented models) for those compounds in Figure 1-2.[13] Our work assumes Compound 1 to be the responsible oxidant, though we recognize the preceding hydroperoxo intermediate as a competent oxidant.[17] This assumption is made as we employ single task regression models that we anticipate using alongside predictive models built for hydroxylation barriers following the hydrogen atom transfer mechanism between a substrate and Compound 1. Either a separate predictive model would need built for the

hydroperoxo mechanism or a multi-task learning approach would be required. By computing the required descriptors with Grimme's GFN family of methods, we provide a validated and rapid approach for systematically estimating P450-mediated epoxidation barriers for inclusion in high throughput screening protocols. The work presented in this chapter has been previously published,[18] and the associated text and content is used with permission.

Computational Methods

All compounds were first prepared in Avogadro 1.2.0[19] and initially optimized using the MMF94[20] force field.

Grimme's crest[21] (version 2.11.1) and xtb[22] (version 6.4.1) programs were used for all semi-empirical calculations. Conformer sampling was first done in crest using GFN2-xTB[23] and resulting conformers were sorted according to their gas phase free energies using the "--prop hess" flag with the required thermochemical calculations performed at standard temperature and pressure. The lowest free energy conformer for each compound was then optimized using GFN1-xTB[24] or GFN-FF[25] to generate the equilibrium structures at those respective levels of theory for further use. All equilibrium structures were found utilizing the "vtight" convergence criteria and the absence of imaginary vibrational frequencies was confirmed following vibrational analyses.

The lowest energy conformers were then used to compute the FOD on the sp^2-hybridized carbon involved in the initial C–O bond formation event during epoxidation in xtb using the "--fod" flag. Additionally, molden[26] input files were generated using the "--molden" flag at the GFN1-xTB or GFN2-xTB level of theory for *N*, *N+1*, and *N-1* electron states for Conceptual Density Functional Theory[27] (CDFT) calculations. The molden input files were then

read with Multiwfn.[28] Hirshfeld[29] and Mulliken[30] atomic charges were determined for the *N*-electron state. Condensed traditional[31] and orbital-weighted[15] Fukui Indices were determined.

Ordinary least squares linear regressions were performed in python, utilizing the scikit-learn,[32] pandas,[33] and statsmodels[34] packages. To create training and test sets, a random 50/50 split was made to place 18 compounds in each set. Min-max scaling was used to scale the predictor variables between 0 and 1 according to the training set. To select features for multiple linear regression (MLR) modeling, a Lasso regression using k-fold cross validation for hyperparameter tuning was performed over the entire dataset. An ordinary least squares MLR model was then fit on the training set and evaluated on the test set. The variance inflation factor for each descriptor was computed in the case of MLR models to check for co-linearity between the descriptors.[35] In the final regression analyses, residuals were verified to be normally distributed according to a Shapiro-Wilk normality test.[36]Stationary point analyses for the initial C–O bond formation event for ethylene, vinyl chloride, and nitroethylene were performed in Gaussian 16[37] on the quartet surface at the B3LYP[38, 39]/LACVP**[40, 41] level of theory in the gas phase. Default integration grids and geometry convergence criteria were utilized. Reaction complexes and intermediates were confirmed as adjoining minima through intrinsic reaction coordinate calculations. Equilibrium geometries for the reaction complex and first intermediate were confirmed as minima by the absence of imaginary frequencies, and transition state structures were verified to have a singular imaginary frequency corresponding to the C–O bond formation vibration. Hirshfeld[29] charges were computed within Gaussian and summed over the substrate fragment.

Similarly, zero-point corrected potential energy barriers on the quartet surface for initial C–O bond formation events were computed using DFT. These were calculated at the B3LYP/Wachters+f[42] (Fe)/TZVP[43]//B3LYP/LACVP** level of theory, with barrier values taken relative to the separated substrate and Compound 1 model. Subsequently, our MLR models using GFN2-xTB and GFN2-xTB//GFN-FF derived descriptors were validated using these DFT-computed barriers.

All semi-empirical calculations were performed on a workstation equipped with an Intel Core i7-4790 and 16 GB of RAM, highlighting the affordability of the methods herein. DFT calculations were performed on a 48 core Intel Xeon Gold 6126 processor with 128 GB of RAM.

Results and Discussion

To build our data set for model generation, we examined the difference between barriers on the doublet and quartet surfaces for our substrate panel as presented by Zhang and Liu.[13] Figure 1-3 provides for a visual inspection of the barrier correlation between spin surfaces. For a more rigorous comparison, we performed a paired t-test between the zero-point corrected potential energy barriers on the doublet and quartet surfaces for all 36 substrates.

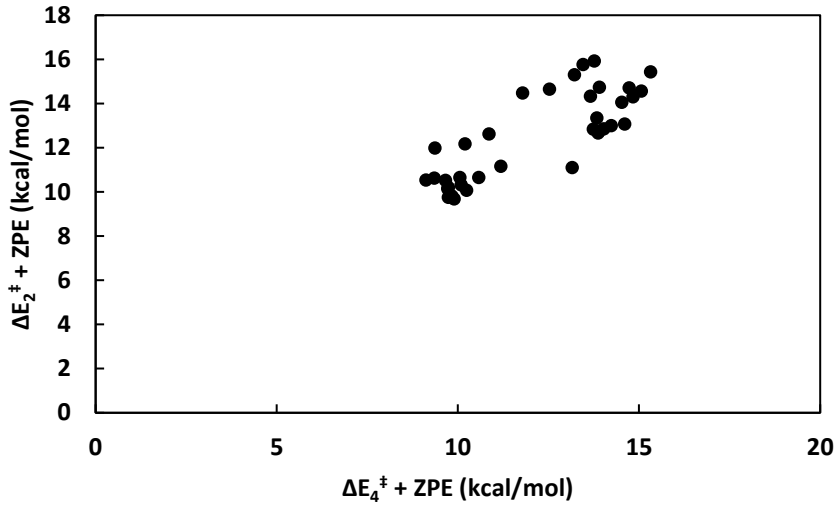

Figure 1-3 Correlation between zero-point corrected potential energy barriers on the doublet and quartet surfaces. Barriers were computed at B3LYP/Wachters+f (Fe)/TZVP// B3LYP/LACVP** for both spin states. Data reused from Zhang and Liu.[13]

While no statistically significant difference could be found (p=0.095), the quartet surface gave an average barrier 0.36 kcal/mol lower than that on the doublet surface. For that reason, we used the quartet surface zero-point corrected potential energy barriers presented by Zhang and Liu as our "true" values.

As previously mentioned, the π-type molecular orbital across an alkene of interest may not always be the HOMO associated with the calculation of the IP. To add localization information, we examined atom-centered descriptors that could be incorporated into either a single or multivariate regression model for barrier prediction.

As it is widely held that alkene epoxidation occurs by a stepwise radical mechanism, a measure of radical character could provide a useful descriptor.[3] One such descriptor is the fractional occupation number weighted density (FOD). As described by Bauer and co-workers, FOD is useful to identify statically correlated and chemically reactive (what the authors called

"hot") electrons.[14] For our panel of substrates, we computed the FOD at the sp^2 carbon involved initially in C–O bond formation using Grimme's GFN family of methods. Figure 2-4 shows the univariate correlation between DFT-computed epoxidation barriers and the FOD on the alkene carbon involved in C–O bond formation at the GFN2-xTB level of theory.

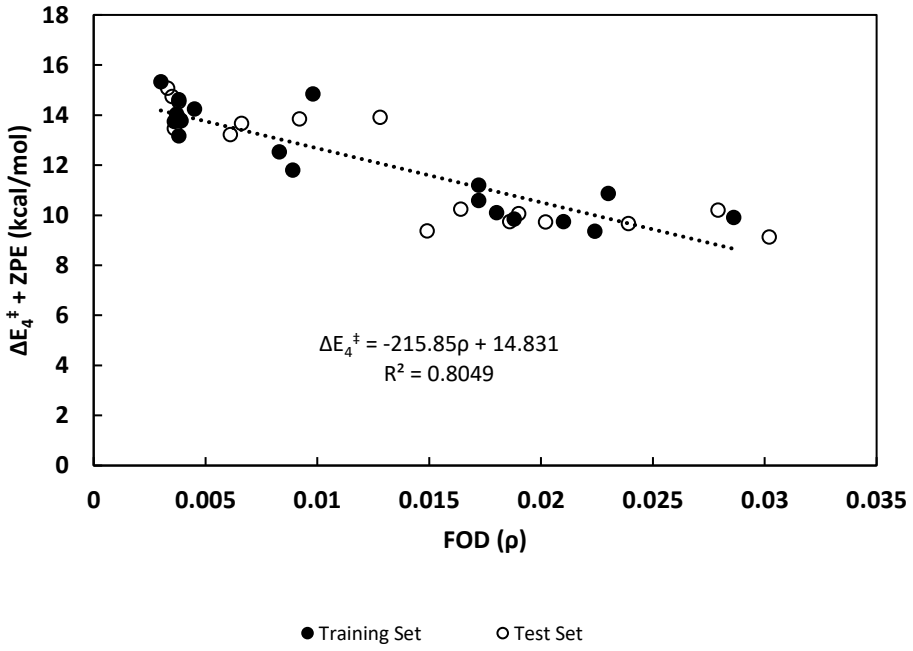

Figure 1-4 Correlation between GFN2-xTB fractional occupation number weighted densities on the vinyl carbon involved in C–O bond formation.

An increase in the FOD corresponds to an increase in local radical character and to a decrease in the barrier to epoxidation. Provided a radical mechanism for epoxidation, the observed trend matches our chemical intuition. When applied in a single linear regression model, FOD as a descriptor afforded MAEs of 0.85 and 0.71 kcal/mol in the training and test sets, respectively, at the GFN2-xTB level of theory. This result alone affords a singular model (without regard for substrate polarity) that recapitulates DFT computed epoxidation barriers.

Additionally, it is worth noting that the computational cost per structure by this approach is measured in milliseconds, making the approach highly affordable.

Still, to explore the possibility of further reducing the computational cost and/or improving our predictive power, we repeated the above analyses using GFN1-xTB, as well as utilizing GFN-FF generated geometries and then calculating the FODs with GFN1-xTB or GFN2-xTB. These results are summarized in Table 1-1. In each model, good correlations between FODs and the DFT computed epoxidation barriers are observed with MAEs well below 1 kcal/mol. Given the similarity between the metrics in Table 1-1 and the possibility that different training sets may result in improved performance, we would not conclude that one approach is definitively preferred over the other. The results originating for GFN-FF geometries show that highly comparable results are achievable at a significantly reduced computational cost, owing to the low cost of utilizing a force field for geometry generation.

Table 1-1 Coefficients of determination and mean absolute errors for linear regression models between FOD values and DFT computed epoxidation barriers.

	Training Set		Test Set	
Method	**R^2**	**MAE (kcal/mol)**	**R^2**	**MAE (kcal/mol)**
GFN2-xTB	0.80	0.85	0.79	0.71
GFN1-xTB	0.82	0.71	0.83	0.59
GFN1-xTB//GFN-FF	0.80	0.79	0.82	0.64
GFN2-xTB//GFN-FF	0.76	0.88	0.78	0.81

Traditional Condensed Fukui Indices

Seeking further improvement, additional descriptors were examined for use in a multivariate regression. We surmised that Conceptual Density Functional Theory[27] might be

useful and that, specifically, condensed Fukui indices would be physically relevant.[44] Summarily, Fukui indices aim to quantify the local change in electron density as electron density is added to or removed from a system. The condensed indices assign the changes in electron density to atoms in the molecule as the number of electrons in the molecule is incremented by ± 1. In this way, the indices serve as descriptors of susceptibility of the atom to be attacked by nucleophilic, electrophilic, or radical species. These reactivities correspond to the *f(-)*, *f(+)*, and *f(0)* indices, respectively. In the context of cytochrome P450 mediated epoxidation, we consider the possible mechanisms in Figure 1-5.

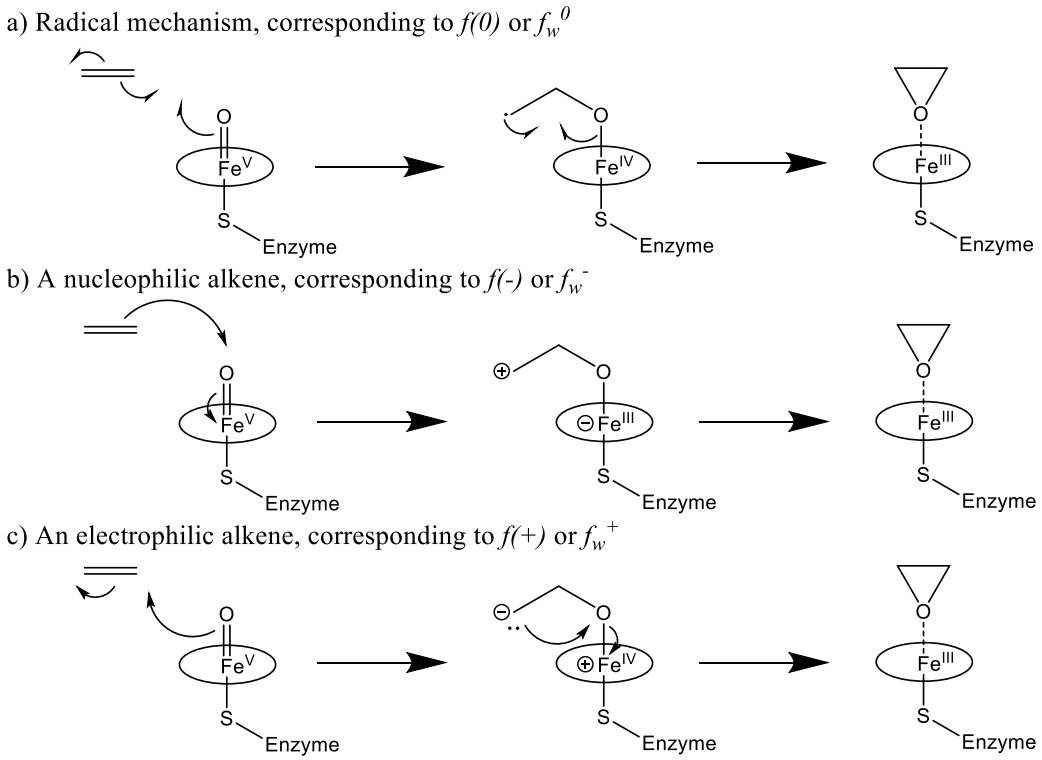

Figure 1-5 Possible mechanisms for the epoxidation of ethylene with the alkene treated as a a) radical, b) nucleophile, or c) electrophile. The protoporphyrin portion of Compound 1 has been abbreviated by the ring about the iron for simplicity.

While it is widely held (and we believe) that the epoxidation mechanism occurs according to a radical mechanism (Figure 1-5a), the alkene substrate could also be treated as a nucleophile (Figure 1-5b) or as an electrophile (Figure 1-5c). With these reactivity paradigms in mind, we are equipped to rationalize relationships between Fukui indices and epoxidation barriers.

Assuming a radical mechanism, we expected the *f(0)* index to correlate with epoxidation barriers. However, the *f(0)* index for the sp^2 carbon atom involved in initial C–O bond formation yielded a MAE of 1.19 kcal/mol compared to the computed epoxidation barriers in the test set. Even worse performance was realized with the *f(-)* index in the test set (MAE = 1.61 kcal/mol). The *f(+)* index, however, correlated reasonably with epoxidation barriers (Figure 1-6). In general, the predictive power of traditional condensed Fukui indices by a linear model went as *f(+)* > *f(0)* > *f(-)*.

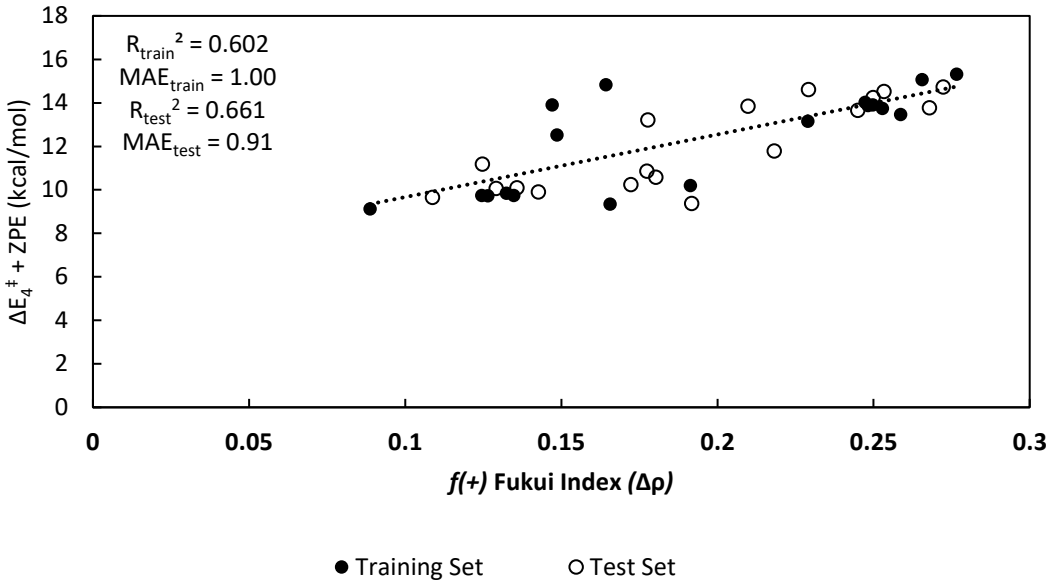

Figure 1-6 Correlation between the condensed *f(+)* Fukui index and computed epoxidation barriers. The remaining Fukui indices as computed at GFN2-xTB yielded MAEs of >> 1 kcal/mol.

While the comparatively poor performance of *f(0)* for predicting epoxidation barriers is surprising, the findings regarding the remaining indices perhaps match our expectations. Treating the alkene as a nucleophile, as in Figure 1-5b, would generate a carbocation intermediate. For many of our substrates, this would be a secondary carbocation and be a generally unfavorable intermediate. An intermediate with cationic character may explain the lack of barrier correlation to *f(-)* in our data set containing principally electron withdrawing substituents. While our substrate panel is lacking strongly π-donating conjugated substituents, these also are not generally found among nature's CYP450 substrates, perhaps due to competing dealkylation mechanisms.[4] Alternatively, the alkene may be considered as an electrophile (Figure 1-5c). This viewpoint diverges from the dogma of a radical mechanism, with the intermediate following C–O bond formation being a carbanion. An intermediate with

anionic character will be reasonably stabilized by the substituents found in our dataset. Indeed, this trend is observed between the *f(+)* index and epoxidation barriers (Figure 1-6). Substrates such as acrolein and nitroethylene, among others, have epoxidation barriers ca. 5 kcal/mol less than that for ethylene. Additionally, it is documented that the axial thiolate ligand coordinated to the heme iron is a particularly strong donor.[45] The removal of this electron "push" has been studied with neutral serine P450 mutants that exhibit altered reactivity (e.g., the carbene transferase).[46] Considering this electron donating interaction alongside observed substituent effects, assigning electrophilic character to the alkene is reasonable in our assessment and rationalizes the observed correlation between epoxidation barriers and the *f(+)* index.

To further examine this point, we investigated the charge evolution during the initial C–O bond formation event using traditional stationary point analysis with B3LYP/LACVP**. Using ethylene, vinyl chloride, and nitroethylene, summed Hirshfeld charges in the substrate fragment were examined in the reaction complex, transition state, and intermediate structures.

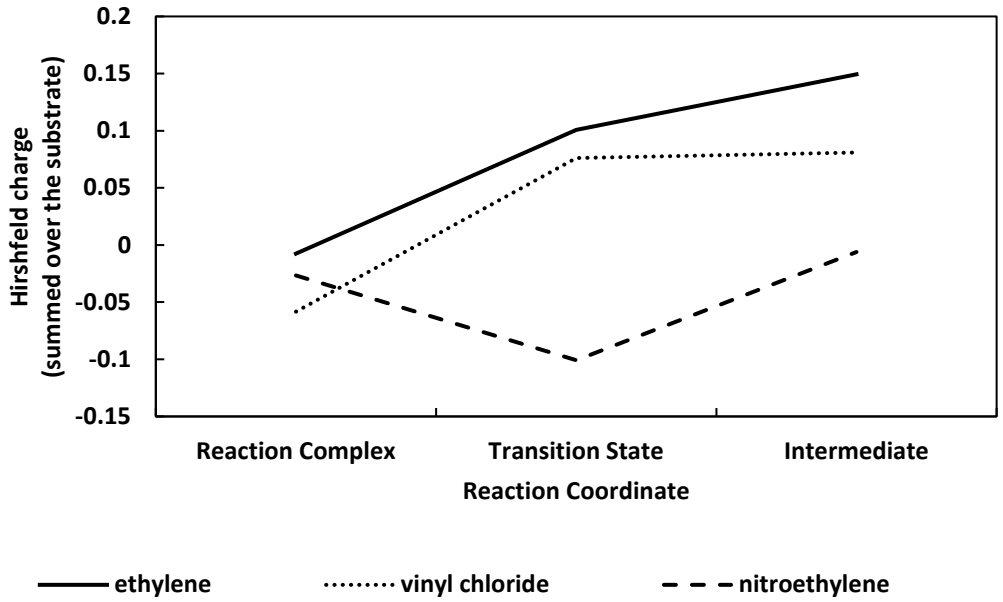

Figure 1-7 Evolution of the summed Hirshfeld charges on the substrate fragment across the reaction coordinate for the initial C–O bond formation event in ethylene, vinyl chloride, and nitroethylene.

As seen in Figure 1-7, there is discernable charge separation between the substrate fragment and the heme (given a total neutral charge for the modeled system) in the transition state for all three compounds. Additionally, the decrease in charge in the substrate fragment relative to the ethylene system in the transition state qualitatively follows the strength of the electron withdrawing substituents (perhaps as indicated by Hammett σ_p values[47]), with nitroethylene yielding the most negatively charged substrate fragment. While these data do not support, nor is it our aim to argue, that carbanions are intermediates in these reactions, these findings suggest the radical mechanism in Figure 1-5a involves significant charge separation, at least for those substrates described here. We believe the barrier correlation with the *f(+)* index indicates the ability of electron (and spin) density to be delocalized away from the carbon

involved in the C–O bond formation, rather than pointing toward the formation of a localized anion in the intermediate preceding epoxide ring closure.

Orbital-Weighted Fukui Indices

While reasonable, the predictive power of the traditional *f(+)* index only barely results in a MAE of less than 1 kcal/mol in the test set (Figure 1-6). It is known that traditional Fukui indices may be misleading in symmetric systems or those with nearly or fully degenerate frontier molecular orbitals, and orbital weighted Fukui indices are not susceptible to such issues.[15]

As multiple substrates in our data set belong to higher order point groups and may have (quasi-)degenerate frontier molecular orbitals, we explored orbital weighted Fukui indices using the same combinations of geometries and electronic structure calculations as in Table 1-1. In doing so, the same trend for predictive performance ($f_w^+ > f_w^0 > f_w^-$) was observed between the three indices, and the results for the f_w^+ index are summarized in Table 1-2. Performance was slightly biased toward the training set. While reasonable structural diversity is present in both training and test sets, structural space is not comprehensively sampled. Nonetheless, f_w^+ taken alone would lack broad applicability based on these findings. One noted benefit of orbital weighted Fukui indices is that their calculation does not require additional single point calculations for the N-1 and N+1 electron states, making orbital weighted Fukui indices more affordable computationally.

Table 1-2 Coefficients of determination and mean absolute errors for linear regression models between f_w^+ indices and DFT computed epoxidation barriers.

	Training Set		Test Set	
Method	**R^2**	**MAE (kcal/mol)**	**R^2**	**MAE (kcal/mol)**
GFN2-xTB	0.80	0.76	0.64	0.83
GFN1-xTB	0.72	0.82	0.53	0.91
GFN1-xTB//GFN-FF	0.74	0.78	0.54	0.90
GFN2-xTB//GFN-FF	0.81	0.73	0.64	0.83

Multiple Regression Analysis

Lastly, we considered the application of a multiple regression model to further reduce MAEs for the test set by combining all Fukui indices, atomic charges, and FOD values. Through a Lasso regression for feature selection,[48] we found both the FOD and f_w^+ descriptors at the GFN2-xTB level of theory to be retained as important descriptors with non-zero coefficients amongst all sampled descriptors. After feature selection and through an ordinary least squares MLR built using the training set, both descriptors were found to be statistically significant (p_{FOD}=0.024 and p_{fw+}=0.019) when GFN2-xTB was employed for the required calculations. Similar statistical significance was obtained using GFN2-xTB//GFN-FF. However, when GFN1-xTB replaced GFN2-xTB for MLR model evaluation, the f_w^+ index became insignificant (p = 0.702). Figure 1-8 depicts the correlation between Zhang and Liu's DFT computed epoxidation barriers and those predicted by our MLR approach using descriptors generated using GFN2-xTB//GFN-FF. Table 1-3 summarizes the performance metrics at all levels of theory. Again, both GFN2-xTB and GF2-xTB//GFNFF performed comparably. Using force field generated geometries has an obvious advantage with respect to computing time and for that reason might be the preferred

approach in high throughput screening. Upon evaluation, the variance inflation factor for each descriptor was found to be ~1.1 at all levels of theory, suggesting the absence of co-linearity between descriptors. This is expected following the Lasso regression for feature selection.

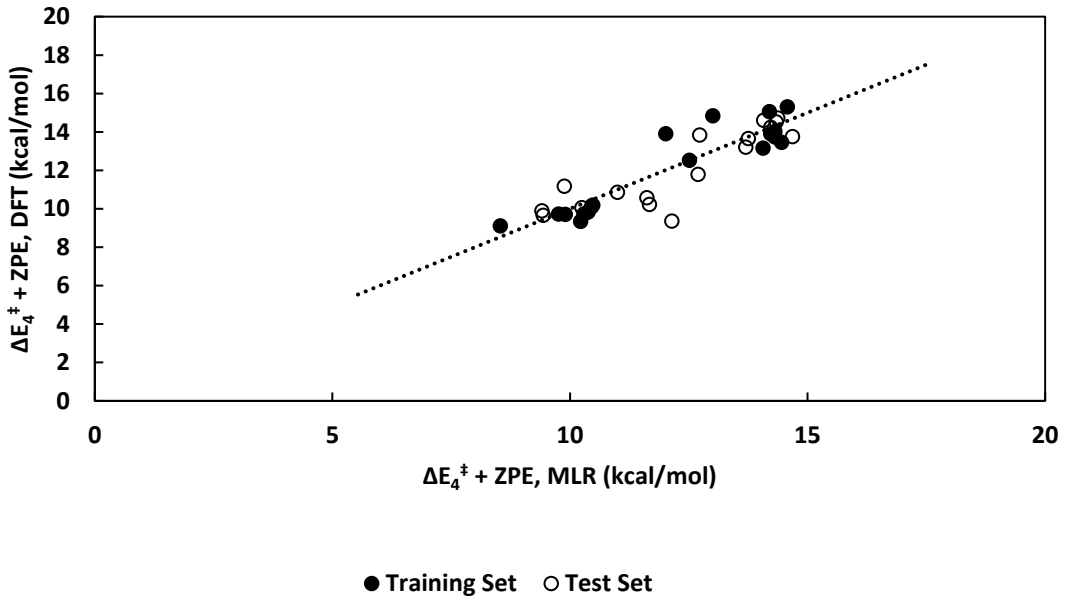

Figure 1-8 Correlation between P450-mediated epoxidation barriers previously computed with DFT[13] and those estimated by our MLR model in this work using substrate-centric descriptors at the GFN2-xTB//GFN-FF level of theory.

Several key advantages are realized by our regression approach. First and most obviously, the amount of computing time to calculate the required substrate-centric descriptors using Grimme's family of GFN methods is orders of magnitude faster than traditional stationary point analysis for the C–O bond formation event in a stepwise epoxidation mechanism. Those calculations would typically take hours (at least) with any reasonable level of DFT on the same computing resource using the typical truncated Compound 1 model. Because the required semi-empirical geometry optimizations take milliseconds with only modest computing

hardware, it is possible to apply these calculations to thousands of structures, such as would be generated from docking simulations. Consequently, we believe quantitative reactivity information can be coupled with binding affinity estimations from docking simulations. Lastly, our model is constructed without regard for substrate polarity as assessed by a compound's overall dipole moment, providing for a simplified application. Polarity and shape are certainly important factors for substrate fit within the context of an enzyme active site, and such properties would be addressed when combining reactivity with accessibility (such as through docking).[49]

Table 1-3 Adjusted coefficients of determination and mean absolute errors for MLR models using FOD values and f_w^+ indices to predict epoxidation barriers.

	Training Set		Test Set	
Method	**R^2**	**MAE (kcal/mol)**	**R^2**	**MAE kcal/mol)**
GFN2-xTB	0.83	0.68	0.76	0.67
GFN2-xTB//GFN-FF	0.84	0.66	0.76	0.70
GFN1-xTB[a]	0.79	0.70	0.81	0.59
GFN1-xTB//GFN-FF[a]	0.81	0.70	0.78	0.66

[a] The f_w^+ index was statistically insignificant using GFN1-xTB or GFN1-xTB//GFN-FF, but the results above are presented for completeness.

One shortcoming in this data set is the absence of tetrasubstituted alkenes. Our review of the literature failed to discover examples of tetrasubstituted alkenes that are epoxidized by P450s, and some literature suggests that tetrasubstituted alkenes are too sterically crowded to undergo epoxidation in a P450.[50] Even peroxo ligated iron porphyrin catalysts, that may not have the same steric limitations as an enzyme active site, are unable to oxidize tetramethylethylene to the corresponding epoxide.[51] As with any predictive model, the appropriateness of the model for a system of interest must be carefully evaluated.

Further Validation

To that end, we further tested our approach by examining the compounds in Figure 1-9 using GFN2-xTB and GFN2-xTB//GFN-FF. These eight compounds were selected as they are either known substrates of CYP101A1[52,53,54] (Figure 1-9a) that undergo epoxidation as well as electron-rich and/or sterically crowded alkenes (Figure 1-9b). Zero-point corrected potential energy barriers were computed at the B3LYP/Wachters+f (Fe)/TZVP//B3LYP/LACVP** level of theory on the quartet surface, with the relative zero taken as the separated substrate and Compound 1 model. To estimate the barrier for these eight compounds, a multiple linear regression model was trained on all 36 compounds in Figure 1-2.

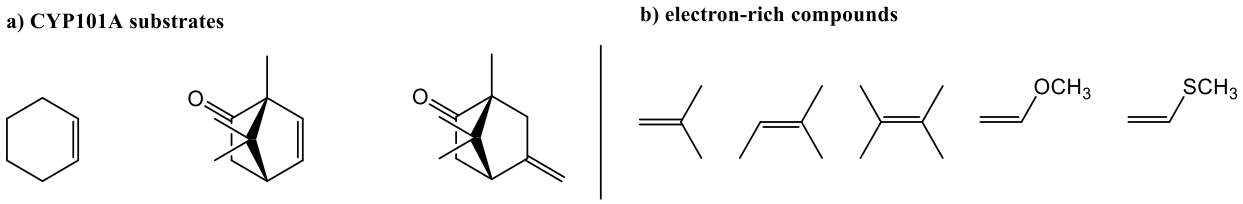

Figure 1-9 CYP101A1 epoxidation substrates and electron-rich alkenes used for further method validation. Electron-rich compounds were used to probe the limitations of the approach herein.

Table 1-4 DFT-computed and MLR-predicted epoxidation barriers (in kcal/mol) for validation compounds in Figure 1-9.

Substrate	DFT[a]	MLR Prediction[b] (GFN2-xTB)	MLR Prediction[b] (GFN2-xTB//GFN-FF)
cyclohexene	13.5	13.5	13.5
dehydrocamphor	11.9	11.1	11.2
5-methylenecamphor	11.7	11.6	11.6
2-methylpropene	12.6	13.5	13.6
trimethylethylene	11.6	13.4	13.3
tetramethylethylene	11.8	13.2	13.1
vinyl methyl ether	9.8	13.4	13.4
vinyl methyl thioether	8.5	11.3	11.3
MAE	n/a	1.42	1.40

[a] Zero-point corrected potential energy barriers were computed using B3LYP/Wachters+f (Fe)/TZVP//B3LYP/LACVP** on the quartet surface.
[b] A multiple linear regression model was fit on all 36 records in Figure 2 using the FOD and f_w^+ index as descriptors.

As seen in Table 1-4, performance consistent with the hold out validation above was observed in the case of the three known CYP101A1 substrates (MAE = 0.27 kcal/mol using GFN2-XTB//GFN-FF), while the more electron-rich compounds performed quite poorly (MAE = 2.06 kcal/mol using GFN2-xTB//GFN-FF). While the models reasonably predicted the barrier for 2-methylpropene epoxidation, trimethyl- and tetramethylethylene were poorly predicted. We surmise that steric hindrance about the alkene could explain this observation. The ethers included in the validation set are strongly electron donating and given such electron-rich alkenes are not represented in Figure 1-2, the inaccurate prediction of their epoxidation barriers is not surprising. These noted limitations further highlight the need to examine any model's suitability for systems of interest prior to use.

Conclusions

By coupling semi-empirical quantum chemical methods with linear regression modeling, it is possible to reliably estimate epoxidation barriers for alkene substrates in cytochrome P450 catalysis. Compared to the use of IPs, we employ descriptors that are localized and describe the radical nature (FOD) and electron deficiency (f_w^+) at the alkene carbon involved in C–O bond formation. With MAEs well below 1 kcal/mol and computational time requirements measured in milliseconds for each input structure, we believe this method is extensible for high throughput screening protocols and would fit nicely alongside protein-ligand docking where conformer ensembles are inexpensively generated using GFN-FF prior to docking. Docked poses could then be evaluated using GFN2-xTB to assess reactivity. In doing so, substrate fit data could be complimented by reactivity information, deepening data sets in efforts to make more reliable product predictions for P450-mediated catalysis.

Chapter 3: Hydrogen Atom Transfer Barrier Prediction

Introduction

Found across all kingdoms of life, cytochrome P450s play critical roles in metabolizing both exogenously (e.g., drugs) and endogenously (e.g., hormones) derived compounds. This is achieved through a number of possible oxidative transformations, including hydroxylation, epoxidation, sulfoxidation, aldehyde oxidation, and others.[3] Among them, hydroxylation at saturated sp^3 hybridized carbon atoms is perhaps the most widely studied. Mechanistically, the consensus is that hydroxylation occurs by way of an initial hydrogen atom transfer (HAT) event from a $H–C_{sp3}$ in the substrate to Compound I, where Compound I is believed to be the most capable and ultimate oxidant.[10, 55] We proceed forward using this mechanistic paradigm, though we acknowledge that oxidation by Compound 0 also has been shown to be energetically reasonable for hydroxylation by way of HAT.[56]

The C–H bond cleavage event during HAT is predicted by theory to be the most energetically demanding event after formation of Compound I, with radical rebound occurring with a smaller barrier or without a barrier at all.[57, 58] This model has been experimentally supported by kinetic isotope effect experiments.[57, 59] Because of the energetic demand, the HAT step is of particular interest for computational modeling. Traditionally, potential energy surface (PES) stationary point analyses using Density Functional Theory (DFT) have been used to compute HAT barriers,[60] with B3LYP being the most commonly employed functional.[38, 61-64]

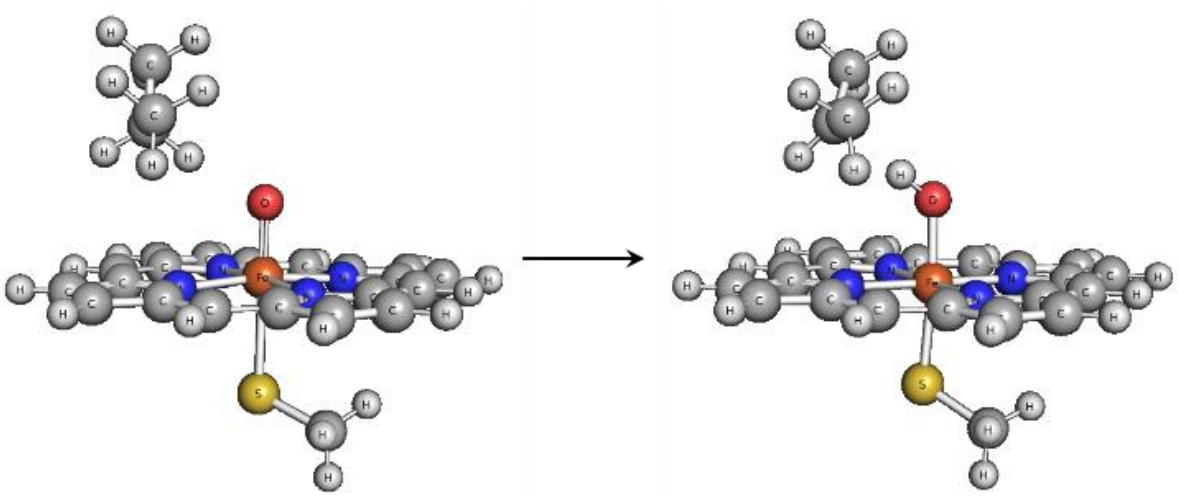

Figure 2-1 The hydrogen atom transfer event between propane at the C2 position and the prototypically truncated model for Compound I. The computational expense for modeling this event originates, in part, from the required geometry optimizations of the structures above, as well as the transition state structure connecting them. The barrier for this event is the quantity of interest in this work.

However, this approach is often not straightforward. First, the sheer size of the system (enzyme + substrate) makes such calculations prohibitively expensive for routine screening. Even the prototypical[65] Compound I model (Figure 2-1) involving a truncated porphyrin ligated by a methylthiolate consists of 43 atoms and hundreds of electrons, while the inclusion of a substrate of biologic relevance, such as a steroid[66] or terpene,[67] can result in a system of 100 atoms or more. Furthermore, substrate conformational flexibility must be considered; even largely rigid structures such as steroids may have multiple energetically accessible conformations that must be considered. Stationary point analyses for systems of this size with numerous heavy atoms can take days or even weeks to complete depending on multiple factors, including the intricacies of the system (e.g., subtleties of the PES), the quality of the

initial geometry guess (dependent on the modeler), the level of theory (the usual trade-off between accuracy and efficiency rears its head here), and the available computing hardware, to name a few. Consequently, computationally affordable statistical and informatics-based models can be tremendously useful, provided they are sufficiently accurate.

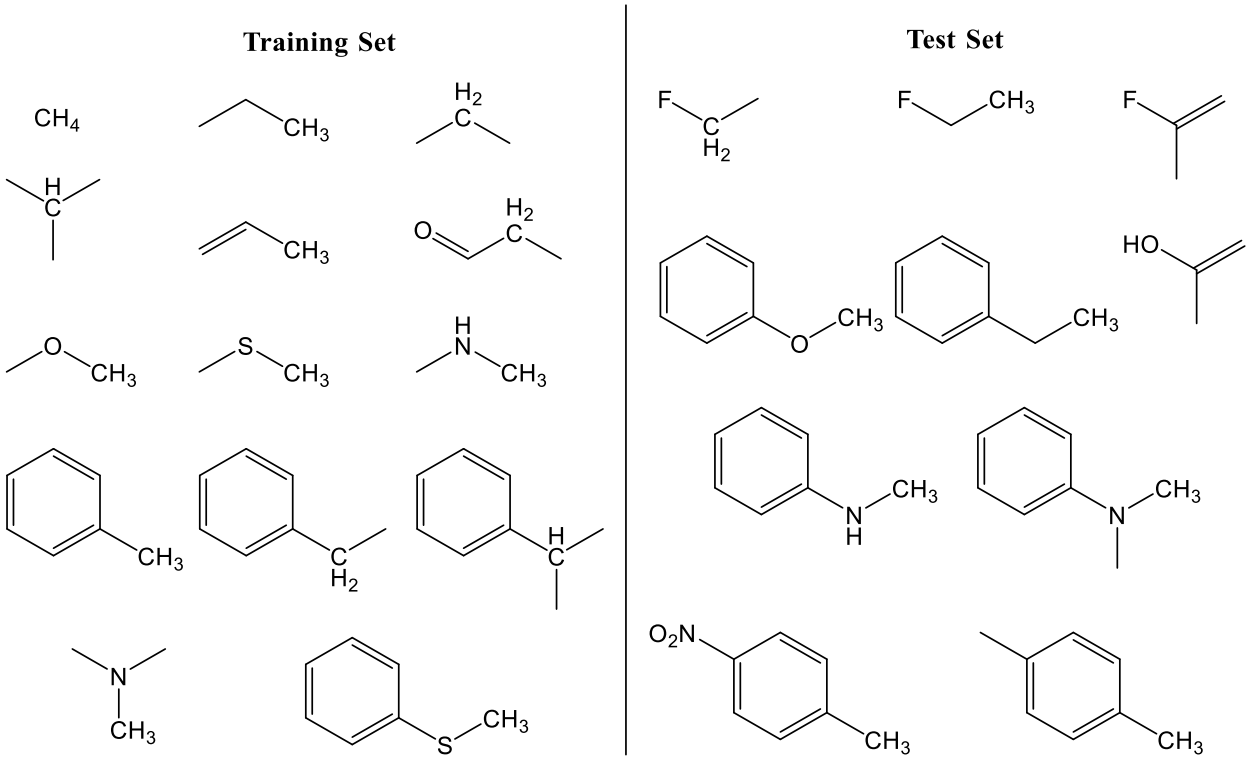

Figure 2-2 A chemically diverse panel of 21 compounds provide a total of 24 unique C-H moieties for model development. The panel was broken into training and test sets as indicated. The hydrogen to be abstracted is shown explicitly at the associated sp^3 hybridized carbon.

Previously, Olsen and co-workers demonstrated that easily computed descriptors could be used to predict HAT barriers that were calculated with DFT for a panel of 24 different C–H bonds across 21 compounds (Figure 2-2).[65] Among them, bond dissociation energies from a "frozen radical" (BDE_{fr}) – that is, the removal of a hydrogen atom from a fully optimized substrate structure followed by a single point energy calculation for the resulting unrelaxed

radical structures (see Figure 2-3) – linearly correlated with the computed HAT barrier. Specifically, BDE_{fr} values obtained with B3LYP/6-31G(d) yielded R^2 values of ~0.9 and a mean absolute error (MAE) of ~1 kcal/mol versus "true" zero-point corrected potential energy HAT barriers computed at B3LYP/6-311++G(2d,2p)//B3LYP/6-31G(d) (Fe = SVP). The required BDE_{fr} calculations took ~20 minutes per molecule, making such calculations still too costly for any high throughput computational screening workflow. Olsen attempted to reduce the cost of the BDE_{fr} calculations by using the AM1 semi-empirical method,[68] but with comparatively poor results (R^2 = 0.68 and MAE ≈ 1.6 kcal/mol).

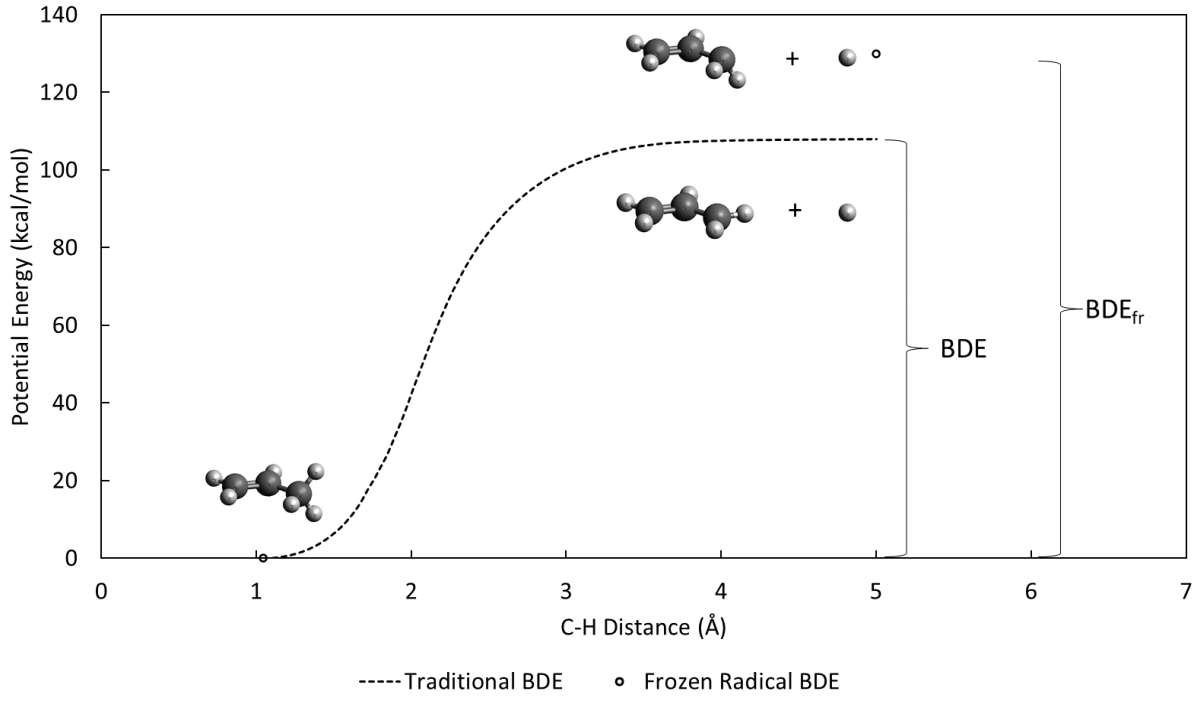

Figure 2-3 To calculate the BDE_{fr} value for propene, an allylic hydrogen atom is removed and an unoptimized ("frozen") substrate radical is formed. The resulting potential energy difference is computed between these species as doublets and the equilibrium geometry of the substrate as a singlet.

Since 2006, computing hardware, computational chemistry software, and available computational methods have advanced, particularly for semi-empirical methods in recent years.[22, 69] In this work, we apply recently developed methods in pursuit of more computationally affordable and accurate approaches for estimating HAT barriers in P450 enzymes. In building on Olsen and co-workers' seminal work, we deliver simple, systematic, and validated models for applications in high throughput screening workflows. The work presented in this chapter has been previously published,[70] and the associated text and content is used with permission.

Computational Methods

All calculations were performed on a personal workstation equipped with an Intel Core i7-4790 with 16 GB of RAM and a solid state storage drive. All structures were prepared in Avogadro 1.2.0[19] and initially optimized using the MMF94[20] forcefield.

Grimme's xtb program (version 6.4.1)[22] was used for all GFN semi-empirical (or force field, in the case of GFN-FF), and Gaussian 16[37] was used for all remaining calculations. Default integration grids and convergence criteria were used in all cases. Each optimized substrate geometry was verified as a minimum without any imaginary frequencies. Open Babel version 2.3.2[71] and bash scripting was used for converting structures and generating radicals. Each substrate structure first was fully optimized using a given method. The hydrogen to be abstracted was then removed. The resulting substrate and hydrogen radicals were then submitted to single point energy calculations (i.e., without further optimization) as doublets to calculate the BDE_{fr}.

Further descriptors were obtained using the GFN2-xTB.[23] The C–H bond local mode force constant was calculated in the local vibrational mode from +/- 0.02 Å about the equilibrium bond length for each substrate, sampling 10 points. A quadratic regression was fit in python using numpy to recover the force constant from the second order term. The standard bond dissociation energy (BDE) was calculated by optimizing each frozen radical structure as a doublet. Mulliken charges[30] and atomic polarizabilities[72] for the abstracted hydrogen and the sp^3 carbon were taken from optimized substrate structures, along with Wiberg bond orders.[73] The solvent accessible surface area for the abstracted hydrogen was computed using GFN2-xTB(ALPB=benzene).[74]

HAT barriers were computed in Gaussian 16 A.03.[37] for several substrates of CYP101A1, specifically (+)-camphor, norcamphor, and (+)-α-pinene, at experimentally observed sites of hydroxylation. This was done at the B3LYP/6-311++G(2d,2p)//B3LYP/6-31G(d) (Fe = SVP) level of theory following the approach taken by Olsen and coworkers.[41, 62, 65, 75] Summarily, the transition state was found on the quartet spin surface, and was confirmed as a transition state structure by the presence of a singular imaginary frequency corresponding to the hydrogen atom transfer motion. The barrier was taken as the zero-point corrected potential energy difference between the transition state structure and the separated reactants, with the Compound I model optimized as a quartet and the substrate as a singlet. These geometries were confirmed as true minima by the absence of any imaginary frequencies. Zero-point corrections were taken from the B3LYP/6-31G(d) (Fe=SVP) level of theory with the potential energy difference taken from the higher-level single point calculation. These barriers were then

predicted by the BDE_{fr} value as described above according to a univariate linear regression to further validate the use of this descriptor.

Ordinary least squares linear regressions were performed in python, utilizing the scikit-learn,[32] pandas,[33] and statsmodels[34] packages. The HAT barriers computed with DFT by Olsen[65] were used as the response variables in all statistical analyses, and for a more direct comparison to that work, the same training and test set split was used. Additionally, a repeated 3-fold cross-validation was performed, with the average of 10 repeats reported in Table 2-1. To select features for multiple linear regression (MLR) modeling, a Lasso regression using repeated k-fold cross validation for hyperparameter tuning was performed. An ordinary least squares MLR model was then fit on the training set and evaluated on the test set, with further reduction of descriptors performed sequentially by manual inspection on the basis on statistical significance or co-linearity. The variance inflation factor for each descriptor was computed in the case of MLR models to examine co-linearity between the descriptors.[35] In the final regression analyses, residuals were verified to be normally distributed according to a Shapiro-Wilk normality test.[36] Lastly, the MLR model was cross-validated using 10 repeats of a 3-fold cross validation and the coefficient of determination and the mean absolute error were measured as the average over those 10 repeats.

All substrate structures optimized at the GFN2-xTB level of theory have been provided in mol2 format. Likewise, a mol2 file is available of the DFT-generated transition state structures, substrates, and the Compound I model. Machine readable data files of computed descriptors and barriers are provided and python scripts for both feature selection and analysis are provided, as well.[70]

Results and Discussion

As an initial check, we began with the AM1 method to ensure we returned comparable results to those reported by Olsen, since substrate geometries were not provided in that publication.[65] Our returned coefficient of determination and mean absolute errors (Table 2-1) were similar to Olsen's (*vide supra*).

Employing the same hold out validation strategy depicted in Figure 2-2, we then turned our attention to sampling other semi-empirical methods, specifically PM3,[76] PDDG/PM3,[77] PM6,[78] PM6-D3,[79, 80] and PM7.[81] Unfortunately, none of these yielded sufficiently accurate correlations to HAT barriers (Table 2-1). As the target quantity for these parameterized methods is the heat of formation, this finding is perhaps not surprising.

Table 2-1 Coefficients of determination and mean absolute errors for linear regression models between BDE_{fr} values and DFT computed HAT barriers by method.

	Training Set (n=14)		Test Set (n=10)		Repeated 3-Fold Cross Validation[a]	
Method	**R^2**	**MAE[b]**	**R^2**	**MAE[b]**	**R_{CV}^2**	**MAE_{CV}[b]**
GFN-FF	0.77	1.39	0.93	0.91	0.71	1.24
GFN2-xTB	0.87	1.01	0.88	0.97	0.80	1.06
GFN2-xTB//GFN-FF	0.84	1.06	0.89	1.10	0.63	1.18
GFN1-xTB	0.85	1.05	0.91	0.92	0.73	1.08
GFN1-xTB//GFN-FF	0.88	1.05	0.95	0.72	0.85	0.96
GFN0-xTB	0.43	1.94	0.66	2.05	0.12	2.28
B3LYP/STO-3G	0.78	1.25	0.36	2.11	0.35	1.89
B3LYP-D3/STO-3G	0.84	1.03	0.19	2.53	0.11	2.02
HF/STO-3G	0.33	2.46	0.30	2.51	-0.17	2.80
PM3	0.48	2.11	0.46	2.22	0.26	2.30
PDDG/PM3	0.47	2.27	0.41	2.39	0.01	2.52
PM6	0.38	2.37	0.51	2.32	-0.05	2.53
PM6-D3	0.36	2.43	0.47	2.43	-0.55	2.68
PM7	0.48	2.21	0.42	2.38	0.17	2.42
AM1	0.72	1.58	0.63	1.61	0.51	1.78

[a] The average metrics from 10 repeats are reported
[b] Mean absolute error as measured in kcal/mol

Ab initio and DFT methods that would generally be considered computationally affordable were then examined. Olsen's best correlation to computed HAT barriers was obtained with BDE_{fr} values calculated with B3LYP/6-311++G(2d,2p)//B3LYP/6-31G(d). For this reason, B3LYP/STO-3G was used to retain the same functional while reducing the cost of the calculation through a substantially smaller basis set. A respectable correlation (R^2 = 0.78) was returned for the training set but not for the test set (R^2 = 0.36), making this model unsuitable.

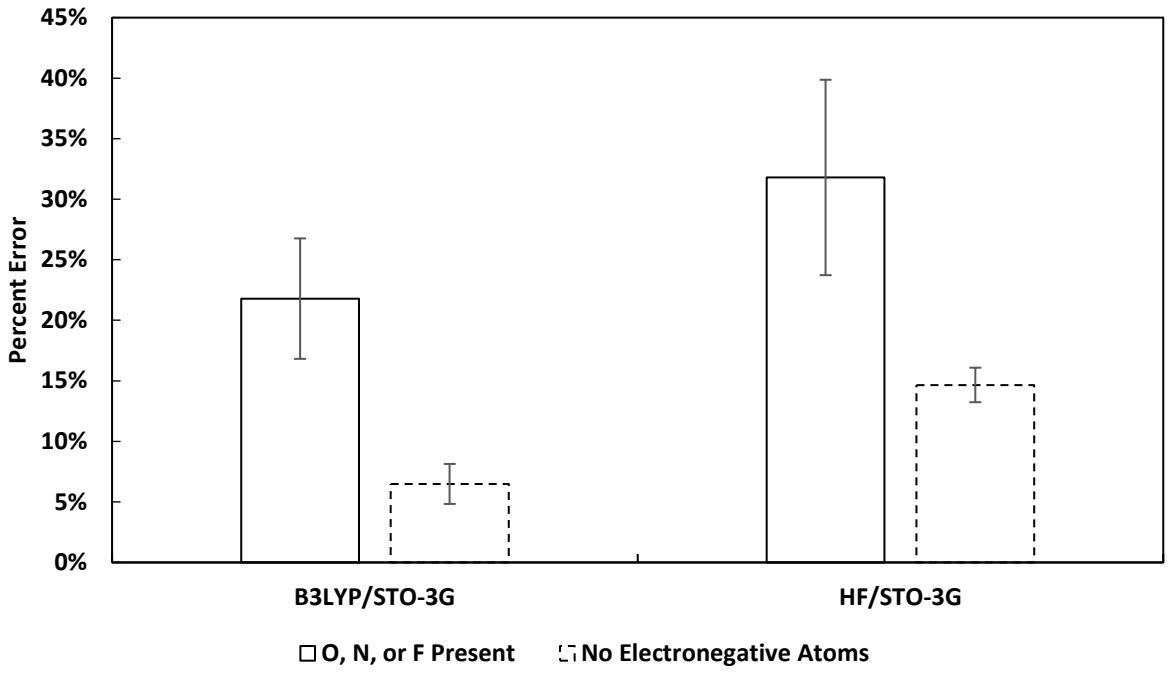

Figure 2-4 Comparison of the average percent error in the predicted HAT barriers by BDE_{fr} values in compounds with and without electronegative atoms in the entire compound panel. Error bars are reported as the standard error of the mean.

HF/STO-3G was also considered but failed to produce reasonable results. We attributed the (expected) poor performance of these methods primarily to the limited basis sets that lack polarization and diffuse functions, which are generally needed for electronegative atoms and radicals.[82] Figure 2-4 shows the mean absolute percent difference across the substrate panel between the predicted barriers by each method and Olsen's computed barriers. In both cases, the presence of electronegative atoms (taken as N, O, and F) yielded higher absolute average errors.

Grimme's extended tight binding methods that focus on returning geometries, frequencies, and non-covalent interactions (GFN) were then explored.[23, 83] This included GFN0-xTB, GFN1-xTB, and GFN2-xTB. Additionally, the partially polarizable force field (GFN-FF)[25] was

employed both as a standalone method and in conjunction with a semi-empirical method where substrate geometries (and by extension, the resultant frozen radical geometries) were generated with GFN-FF and the BDE_{fr} was computed with GFN1-xTB or GFN2-xTB. GFN1-xTB and GFN2-xTB yielded good correlations ($R^2 > 0.80$) and small mean absolute errors (MAE ≈ 1.0 kcal/mol) (Table 2-1). The linear regression for GFN2-xTB is plotted in Figure 2-5 for both the training and test sets. Both methods gave correlations and mean absolute errors as good as those reported by Olsen using B3LYP[61-63, 84] in conjunction with larger Pople basis sets.[41] GFN0-xTB failed to yield sufficient correlation between BDE_{fr} values and computed HAT barriers in the substrate panel. GFN-FF performed less consistently between the training and test sets, suggesting it should be employed alone with caution, if at all.

To better estimate the performance of these univariate models on new data points, 10 repeats of 3-fold cross validatoin were performed. The cross validated performance metrics, as measured by the average of 10 random repeats, are presented in Table 2-1. All values were found to be consistent with the hold out strategy employed by Olsen and coworkers. All BDE_{fr} values from all methods are available in the supporting information.

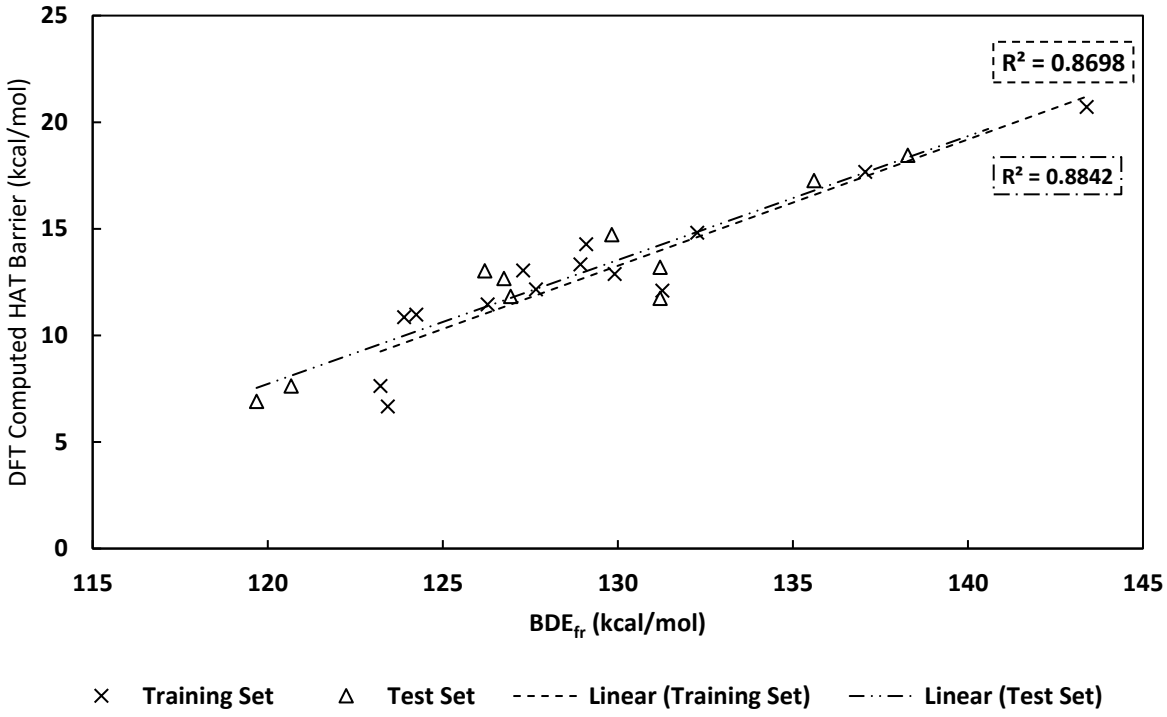

Figure 2-5 Linear regression in the training and test sets between computed DFT Barriers[65] and GFN2-xTB generated BDE_{fr} values.

As our aim was to identify an approach that can systematically and efficiently predict HAT barriers in thousands of substrate structures, computing times were compared between approaches with an MAE of ca. 1 kcal/mol. Figure 2-6 shows that the presented methods perform at the millisecond timescale per input structure (within our computing environment) and that respectable reductions in computational cost can be achieved by employing force field generated structures without significantly increasing the MAE in each model.

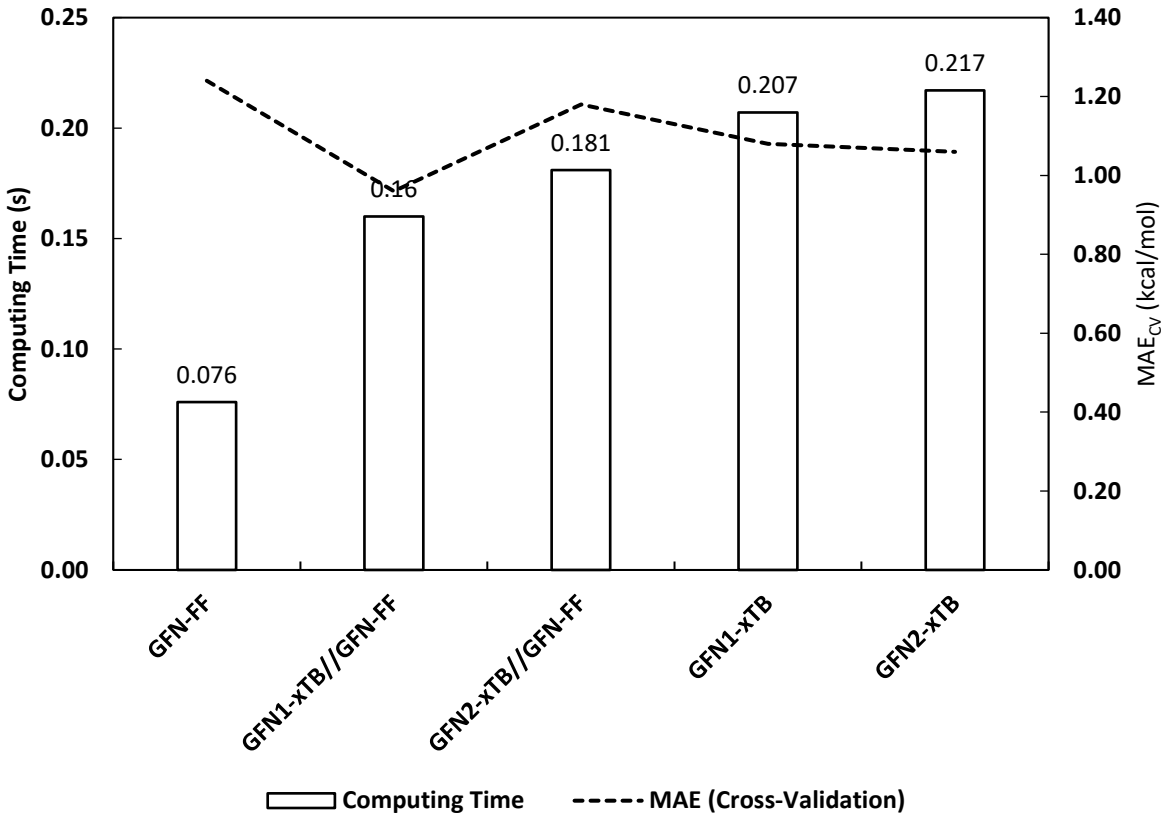

Figure 2-6 Comparison of the required computing time for the determination of the BDE_{fr} for p-nitrotoluene and the mean absolute error in cross-validation by method.

While the time savings enjoyed by force field generated geometries are not surprising, the low MAE values achieved with GFN1-xTB//GFN-FF and GFN2-xTB//GFN-FF were a welcome finding. Compared to GFN-FF alone, more consistent performance between the training and test sets were obtained when GFN1-xTB or GFN2-xTB was used for the required single point calculations in computing BDE_{fr} values from GFN-FF geometries. As measured by the total time to perform all necessary calculations for *p*-nitrotoluene, 10-20% time savings were achieved with both GFN1-xTB//GFN-FF and GFN2-xTB//GFN-FF when compared to each tight binding method alone. Performing such barrier estimations with GFN1-xTB//GFN-FF or GFN2-xTB/GFN-FF is particularly attractive where a compound of interest is conformationally flexible and a

conformational ensemble must be generated. In such cases, Grimme's crest program could naturally be incorporated into a screening workflow, employing the force field for efficiency.[85]

It is worth noting that the barriers computed by Olsen were from gas phase calculations without empirical dispersion included.[86, 87] Later reports showed that including dispersion effects has significant bearing on hydroxylation (as well as epoxidation) barriers.[88] As dispersion effects are included in all GFN methods, we surmised that dispersion corrections may contribute to the performance improvement over other semi-empirical methods sampled here. However, our B3LYP-D3/STO-3G and PM6D3 studies showed degraded performance compared to the respective methods without dispersion corrections. These findings, along with Olsen's correlation with a large Pople basis set, would suggest that the basis set and the target quantities of semi-empirical methods are more important than dispersion effects for this regression approach. This is exemplified by the errors observed in B3LYP/STO-3G, B3LYP-D3/STO-3G, GFN1-xTB, and GFN2-xTB. Both B3LYP approaches yielded absolute errors of ca. 9 kcal/mol for the predicted HAT barrier for fluoroethane at C1, while both tight binding methods produced more reasonable errors of ~1.5 kcal/mol. We attributed this to the inclusion polarization functions in the basis sets used by both GFN methods.

Other Descriptors

Given the speed of the calculations using Grimme's methods, more computationally demanding quantities, specifically standard BDE values and local mode C–H force constants, were examined for correlation to computed barriers using GFN2-xTB. Figure 2-7 shows that neither local mode force constants nor traditional BDEs correlated well with HAT barriers. The

finding of poor correlation with BDE values is consistent with Olsen's findings at the level of theories explored in that work.[65]

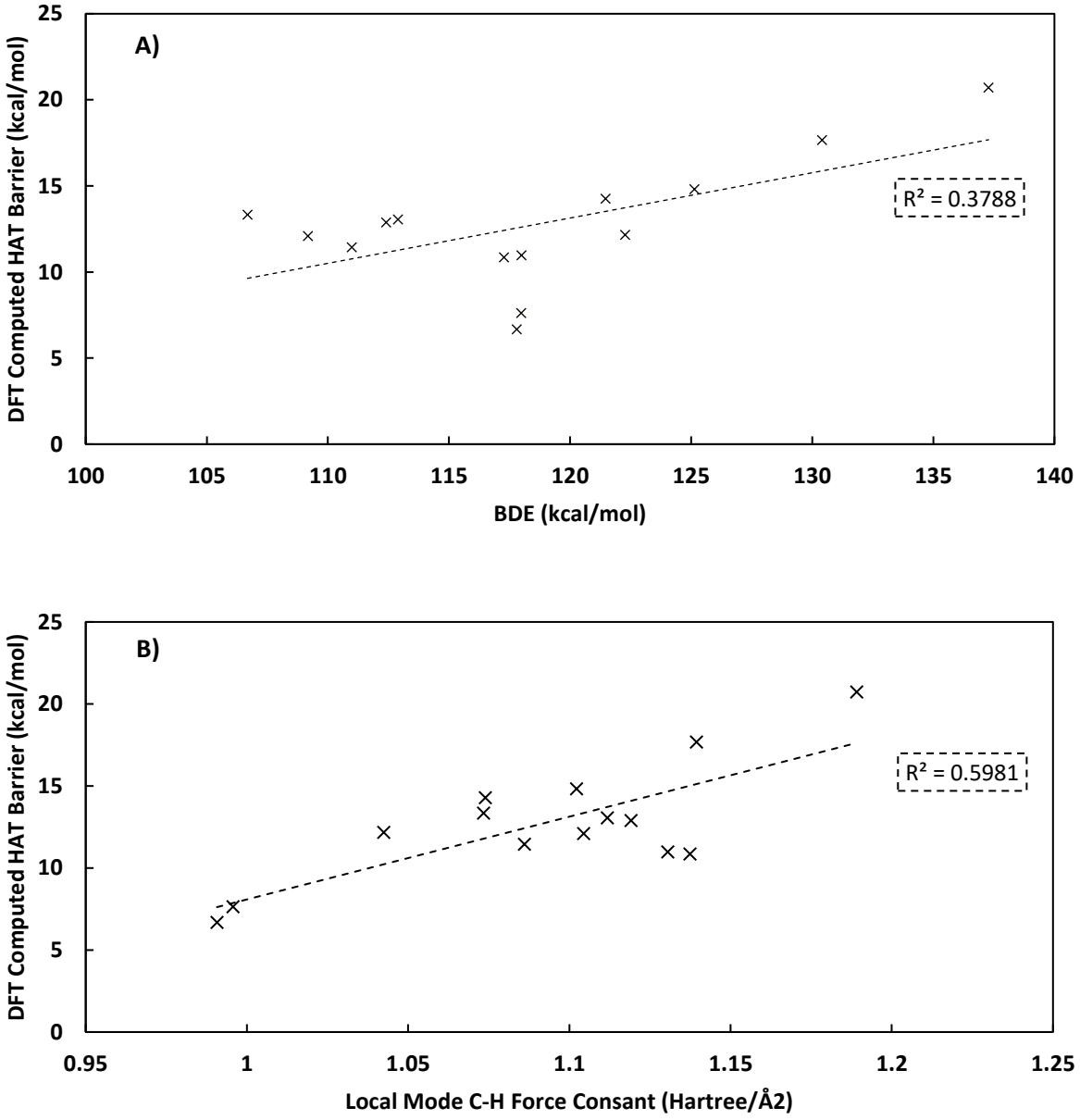

Figure 2-7 Linear regressions between DFT computed HAT barriers[65] in the training set and GFN2-xTB computed A) C-H BDE values and B) local mode C-H force constants revealed poor correlations as compared to BDE_{fr} values.

Multiple Linear Regression (MLR) Modeling

Lastly, we employed multiple linear regression using descriptors generated with GFN2-xTB. The descriptors screened for inclusion in a MLR model were the BDE, BDE_{fr}, C–H local mode force constant, Wiberg C–H bond order, the Mulliken charges and atomic polarizabilities on the hydrogen and carbon atoms involved, and the solvent-accessible surface area (SASA) on the hydrogen obtained from a GFN2-xTB (ALPB=benzene) single point calculation on the gas phase optimized substrate structure.[74] Employing a Lasso regression and k-fold repeated cross validation over the entire data set for feature selection, BDE, BDE_{fr}, C–H force constant, C–H bond order, and the Mulliken charges and atomic polarizabilities on the carbon atoms were retained as possible features. With a training set of only 14 records, only three descriptors can reasonably be employed. From ordinary least squares MLR models, C–H bond order, BDE, and the hydrogen atom charges and polarizabilities were removed sequentially, constructing a new MLR model from the training set after each descriptor was removed. C–H bond order was removed as it is reasonably co-linear with the BDE values (R^2=0.55). The Mulliken charge on the carbon atom was then dropped due to statistical insignficance (p=0.821), as was the BDE value (p=0.662). Ultimately, BDE_{fr}, the local mode force consant, and the carbon atom polarizability were retained on both the basis of their physical relevance and/or their statistical significance within the model. BDE_{fr} was retained due to its demonstrated performance in a univariate model both in our work and by Olsen.[65]

The local mode force constant was selected given at least the qualitative correlation between with the HAT barrier (as shown in Figure 2-7) and our belief that smaller force constants should correspond to more easily abstracted hydrogens. Lastly, we surmised a larger

carbon atom polarizability may correlate with a relatively more stable radical following the HAT event and perhaps a lower HAT barrier. BDE_{fr} and the local mode force constant were both found to be statistically significant at the 95% confidence interval in our model ($p = 0.00$ and 0.01, respectively), while the carbon atom polarizability was less significant ($p = 0.083$). As shown in Figure 2-8, the coefficient of determination between the MLR predicted barriers and Olsen's DFT computed barriers in both the training and test sets increased to greater than 0.9, and the MAE values were 0.62 kcal/mol and 0.83 kcal/mol in the training and test sets, respectively. 10 repeats of 3-fold cross-validation were performed for the MLR model, and the MAE_{CV} value was found to be 0.82 kcal/mol, consistent with the hold out strategy. The complete data set used for the MLR model exploration is included in the supporting information.

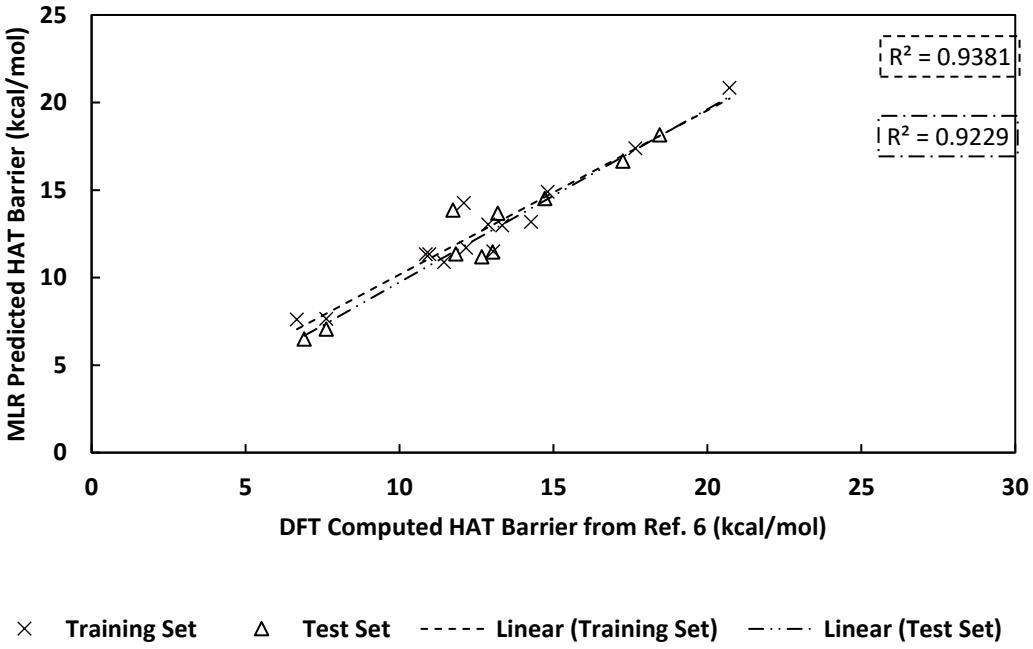

Figure 2-8 Correlation of MLR-predicted HAT Barriers from GFN2-xTB generated descriptors and DFT computed barriers. Descriptors include the BDE_{fr}, the C–H local mode force constant, and the atomic polarizability of the carbon atom.

While better performance was achieved by a multiple linear regression model, the improvement is modest as measured by the mean absolute error values. The MAE_{CV} value of 0.82 kcal/mol for the MLR model decreased from a value of 1.06 kcal/mol for the univarate model based on BDE_{fr} values, and in our opinion, such a limited performance improvement does not justify the added computational time required to scan the C–H bond length. While the additional sophistication may be useful in establishing correlative models to larger, more diverse data sets, the simplicity of the univariate approach is probably sufficient for systematically assessing reactivity, and we believe it will be easily scripted alongside docking in high throughpout virtual screening.

Further Validation: CYP101A1 Substrates

As our aim is to apply these computationally efficient calculations and associated regression models to docked substrate poses, we evaluated the univariate linear regression model using BDE_{fr} values to substrates of the CYP101A1 (P450cam) enzyme. Specifically, we examined (+)-camphor, norcamphor, and (+)-α-pinene at experimentally observed hydroxylation sites, providing for six unique HAT barriers. The experimentally observed outcomes are summarized in Figure 2-9. For (+)-camphor, hydroxylation occurs exclusively at the 5-exo position.[89] Hydroxylation of norcamphor is less selective, occuring at the 3-exo, 5-exo, and 6-exo positions.[90] Finally, (+)-α-pinene is hydroxylated at both allylic positions, yielding (+)-*cis*-verbenol and (+)-myrtenol.[91] Table 2-2 provides both the HAT barriers as computed by DFT and the predicted HAT barriers at the GFN2-xTB, GFN2-xTB//GFN-FF, GFN1-xTB, and GFN1-xTB//GFN-FF levels of theory utilizing a univariate linear regression trained on all 24 records from Figure 2-2. The findings in Table 2-2 show that the MAE values from these six HAT events are consistent with the cross validated MAE values in Table 2-1. This suggests the approach is generalizable to experimentally interesting substrates.

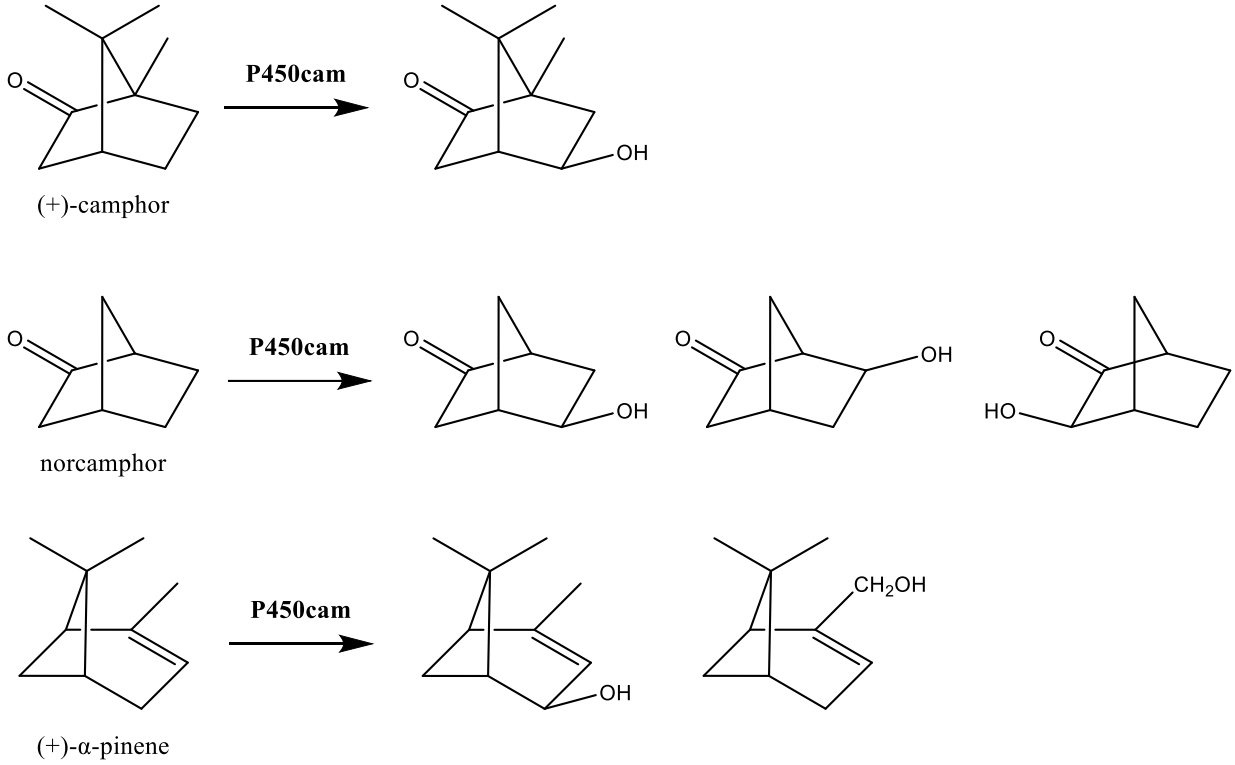

Figure 2-9 Three substrates of CYP101A were used for further validation of a univariate linear model using BDE_{fr} values generated using semi-empirical methods. The experimentally observed hydroxylations in the native enzyme are provided.

Table 2-2 Predicted HAT barriers for CYP101A1 substrates using BDE_{fr} values computed by semi-empirical methods and the linear models trained on the compounds from Figure 2-2.

		BDE_{fr} Method			
Substrate (position)	DFT Barrier[a]	GFN2-xTB	GFN2-xTB //GFN-FF	GFN1-xTB	GFN1-xTB //GFN-FF
(+)-Camphor (5-exo)	14.76	14.06	14.87	13.75	14.78
Norcamphor (3-exo)	14.07	11.17	12.46	11.10	12.47
Norcamphor (5-exo)	13.96	14.57	15.47	14.28	15.38
Norcamphor (6-exo)	12.73	12.37	13.89	12.26	13.99
(+)-α-pinene (4-exo)	10.85	9.79	10.18	8.69	9.45
(+)-α-pinene (2-methyl)	11.07	12.31	12.72	11.98	12.59
MAE	**-**	**1.15**	**1.12**	**1.31**	**1.20**

[a] Reference HAT barriers were computed at the B3LYP/6-311++G(2d,2p)// B3LYP/6-31G(d) (Fe = SVP) level theory. Each barrier is expressed as the potential energy difference between the transition state structure and the separated reactants at the B3LYP/6-311++G(2d,2p) level of theory, with zero-point corrections applied from the lower level of theory.

In the cases of norcamphor, (+)-α-pinene, and other compounds where multiple hydroxylation products are observed, other factors, notably residence time in a given conformation, can impact product ratios. Harris and co-workers computationally observed that norcamphor, owing to its smaller molecular volume in comparison to the native d-camphor substrate, is less rotationally encumbered in the active site, resulting in populated configurations that expose the C6 and C3 positions to the oxo ligand in Compound 1.[92] While HAT barriers alone would predict C6 hydroxylation to dominate experimental outcomes for norcamphor (where C5 and C6 hydroxylation are observed experimentally in roughly equal proportions), Harris' work demonstrated the importance of residence time configurations

leading to hydroxylation at the C5 position being most populated. To summarize, HAT barriers taken alone to predict P450-mediated hydroxylation neglects residence time and protein-ligand interaction. To accentuate this point in P450cam, a worthwhile future study, which is currently absent from the literature, is a MD-based analysis of 5,5-difluorocamphor in P450cam where hydroxylation is observered exclusively at the C9 position experimentally.[93]

Conclusions

Herein we have expanded Olsen's regression modeling approach for estimating hydrogen atom transfer barriers in the hydroxylation of substrates by cytochrome P450s using modern semi-empirical methods. Whereas other semi-empirical methods returned residuals upon cross-validation that were too large to be trustworthy, GFN1-xTB or GFN2-xTB performed well, with both single and multiple linear regression models returning mean absolute errors on the order of 1 kcal/mol or less. Furthermore, the utility of this method was further examined using HAT barriers computed with DFT for substrates of CYP101A1. In addition to adequately predicting HAT barriers, the employed methods are 2-3 orders of magnitude faster than the other "inexpensive" methods benchmarked in our work. As each GFN method performs the required calculations on the millisecond timescale with only modest computing hardware, this approach is extensible to high throughput screening workflows for examining hundreds to thousands of structures of interest with minimal computing resources.

Chapter 3: Hydroxylated Product Predictions in CYP101A1

Introduction

C–H bond activation remains one of the holy grails of organic chemistry.[94] Transformations of unactivated aliphatic C–H bonds are difficult, attributed in part to bond dissociation energies of approximately 100 kcal/mol. Despite bond dissociation being energetically demanding, CYP450s catalyze such transformations at room temperature with excellent regio- and stereoselectivity. For this reason, CYP450s are attractive biosynthetic tools and targets for enzyme design to achieve otherwise inaccessible transformations. Indeed, CYP450s have been engineered by directed evolution to modulate existing activity, as well as to produce novel ones, such as carbene and nitrene transferase activities.[95]

In humans, CYP450s in the liver are responsible for ~75% of xenobiotic metabolism, with the 3A4 isoform responsible for about half of such activity.[96] Given the centrality of CYP450s in ADMET processes, understanding CYP450-mediated metabolism for drug candidate optimization is critical. For example, the replacement of aliphatic hydrogens at reactive sites by fluorine is commonly used to alter P450-mediated outcomes to decrease the rate of metabolism.[97] Optimization efforts may also focus on improving the safety profiles of discovery leads. Off-target toxicological effects caused by oxidation products is known; the P450-mediated conversion of acetaminophen to N-acetyl-p-benzoquinone imine under acute overdose conditions is the classical example, leading to drug induced liver injury.[98] During drug discovery, an understanding of P450-mediated metabolism may facilitate inactivation efforts.[99]

Between motivations in both synthesis and drug design, reliable predictions of CYP450 oxidation products are necessary to accelerate discovery. Historically, several approaches have

been used. As described in Chapters 2, Density Functional Theory (DFT) calculations of hydrogen atom transfer (HAT) barriers using a truncated Compound I model have been performed, along with computing less expensive descriptors that can be used to predict HAT barriers.[65] Subsequent informatics-based approaches, most notably SmartCYP,[100] using such barrier estimates and making structural comparisons to compounds of interest have found utility while significantly reducing computational cost. The downfall of barrier calculations employing just the substrate and a reduced Compound 1 model is the absence of the enzyme's influence, specifically steric and electrostatic effects within the enzyme active site. For example, the barrier for the 5-exo hydroxylation of d-camphor using this truncated model was previously computed at over 21 kcal/mol by Kamachi and Yoshizawa.[101] A more complete model of the $P450_{cam}$ enzyme treated with a combined quantum mechanics/molecular mechanics (QM/MM) by Lonsdale and coworkers found the same barrier to be ~18 kcal/mol, while the inclusion of a Grimme's D2 dispersion correction[102] reduced the barrier even further to ~14 kcal/mol.[103] Jerome and coworkers found the same barrier to be 10 kcal/mol employing QM/MM with empirical dispersion and transition metal optimized localized orbital corrections, while the experimental barrier is known to be at most 10 kcal/mol.[104]

Beyond barrier modulation, the enzyme is most assuredly responsible for the regio- and stereoselectivity observed during camphor hydroxylation, as an example. In addition to computing the HAT barrier for the 5-exo hydroxylation of camphor using the typical reduced Compound I model, Kamachi and Yoshizawa predicted that the intermediate substrate radical following the HAT event at the C5 position in camphor is notably less stable than the radical generated at the C3 and C6 positions.[101] Based on their relative radical stabilities, we would

assume HAT events at the C3 and C6 positions would have lower HAT barriers than that predicted at the 5-exo position, yet only the 5-exo hydroxylated product is observed. B3LYP/6-311++G(2d,2p)// B3LYP/6-31G(d) (Fe = SVP). At the same level of theory in our own work exploring HAT barriers for norcamphor,[70] the 6-exo and 5-exo hydroxylated products would be predicted in a ratio of roughly 90:10; however, norcamphor is known to be hydroxylated to these two products in almost equal parts as the major products.[105] Collectively, all the above reports indicate the need to account for the enzyme.

To affordably account for the contributions of the enzyme, protein-ligand docking is performed to evaluate possible substrate binding modes.[9, 106] As a more costly option, molecular dynamics (MD) simulations using molecular mechanics (MM) force fields have been used to investigate substrate-enzyme interactions.[10, 90, 107] In the case of cytochrome P450s, MD simulations have also helped to explain conformational changes that occur during the catalytic cycle.[108] In the cases of docking and MM-based MD, reactivity information is absent. The computational cost for QM/MM/MD also is prohibitively high for routine analyses. Additionally, this approach requires carefully, and most often manually, assigning enzyme, substrate, and solvent atoms between the QM and MM layers. There exists a gap in reliable methodology that 1) considers reactivity, 2) accounts for the protein environment, and 3) is computationally affordable. This gap is further widened by the lack of CYP450 design studies in the literature using Rosetta. We aim here to close this gap while eyeing enzyme design as a future effort.

There is precedent in the literature for combining reactivity and binding mode predictions from docking. Specifically, Tyzack and co-workers combined bond order calculations at the B3LYP/6-31G(d,p) level of theory with tethered docking outcomes from GOLD through a

manually parameterized scoring approach.[109] While straightforward to implement, the parameterization is purely empirical, with parameters varying by enzyme isoform. In addition, the tethering approach converted the abstracted hydrogen atom to an oxygen atom bound to both the heme iron atom and a carbon atom on the substrate. Such tethering is more akin to docking an unreleased product in complex with the enzyme. Despite our perspective on their exact approach, the improvement in their predictions upon inclusion of bond order calculations was substantial and clearly demonstrates the value of combining electronic structure calculations with docking.

To this end, we present a generalizable workflow for the prediction of P450 hydroxylation products that combines docking and pose clustering with substrate-centric post-docking calculations using highly scalable semi-empirical tight binding and force field calculations. CYP101A1 (camphor 5-monooxygenase) served as our development system given the abundance of readily available experimental data for the native sequences. Additionally, given its bacterial origin and the number of active mutants known in the literature, it is a good candidate for future enzyme engineering efforts. Experimental hydroxylation outcomes for substrates of CYP101A1, including product ratios where applicable, were gathered from the literature for the 25 substrates in Figure 3-1. We demonstrate improved hydroxylation prediction performance over docking alone by combining reactivity information with docking in a way that will naturally extend to future enzyme design efforts.

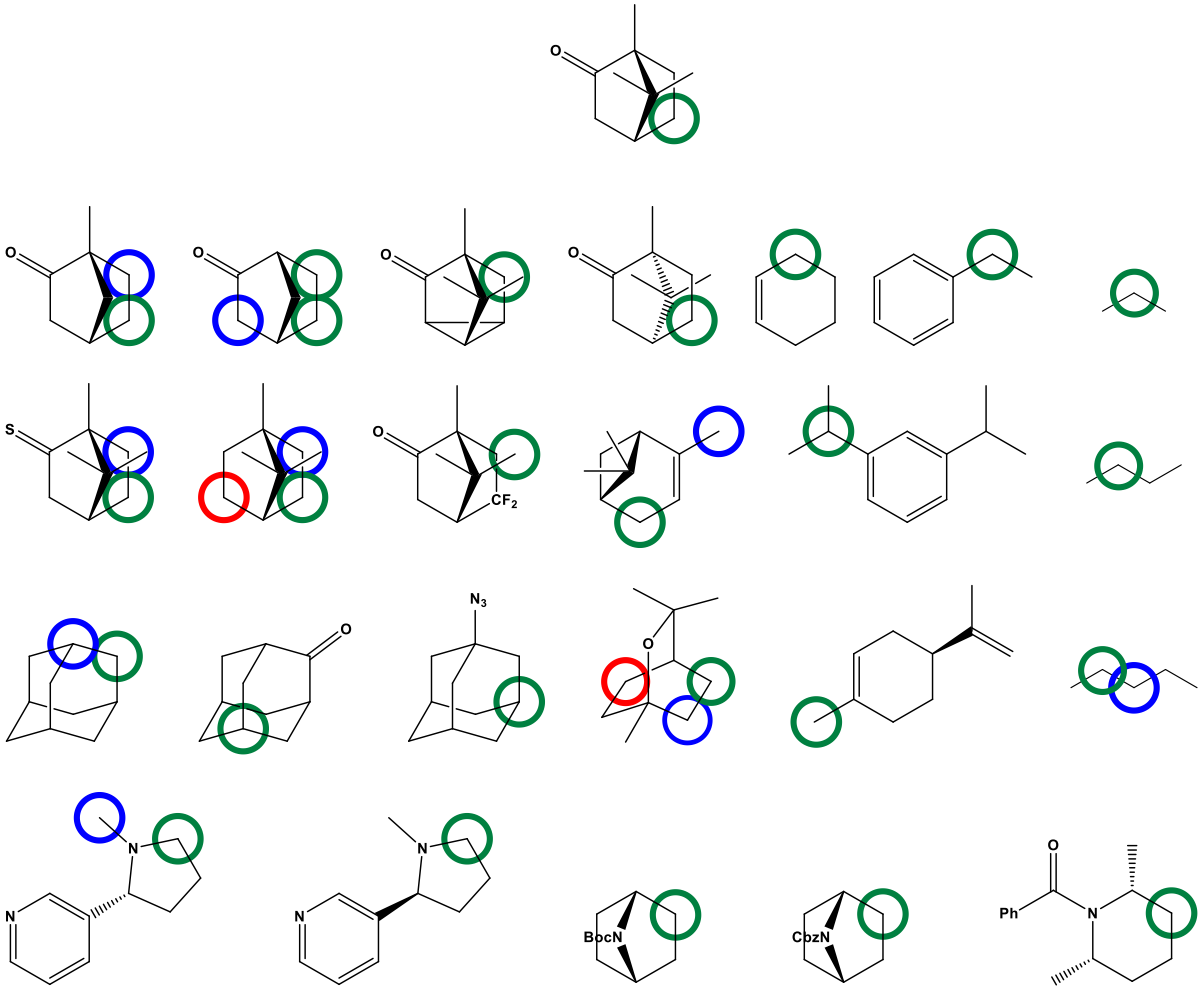

Figure 3-1 Substrates utilized for method development. Positions corresponding to the major (green) and minor (blue then red) hydroxylation sites are circled.

Computational Methods

All software used herein is open source and free for use, at least within academia. Avogadro[19] was used for initial substrate geometry generation along with the bundled MMF-94 force field.[20] In addition to Avogadro, PyMol[110] was used for visualization and enzyme structure protonation. Grimme's xtb[22] and crest[111] programs were used for substrate conformer generation at the GFN-FF level of theory. Docking was performed using RosettaScripts[112] within Rosetta 3.13 and RxDock 0.1.0.[113] Pose clustering was performed using Zhang's DockRMSD

utility.[114] All scripts were written in python or bash, and where possible, GNU parallel was utilized to parallelize docking operations.[115] Within python, pandas[116] was utilized for data filtering. Openbabel 3.1.1 was used throughout to convert between required file types.[71]

Conformer-Rotamer Ensemble Generation for Substrates

Substrates were initially built and minimized in Avogadro as indicated above. Using xtb[22] and crest[111] with GFN2-xTB[23] as the level of theory and the default search algorithm within crest, a conformer-rotamer ensemble was generated, retaining conformers and rotamers within 12.0 kcal/mol of the best identified conformation. All structures were then reoptimized within xtb using GFN-FF[25] using the "vtight" gradient convergence criterion. This was done as post-docking calculations for thousands of structures are more scalable with the force field than with GFN2-xTB. The resulting structures were sorted according to their potential energy and all those within a 3.0 kcal/mol window relative to the best structure were retained for docking. In this way, only conformations that should be reasonably populated will be docked into the receptor. This same conformer/rotamer ensemble was used for docking in both RxDock and Rosetta.

Receptor Selection

The 1DZ8[117] accession into the Protein Data Bank was used as the protein receptor. While it is possible, or even probable, that the 1DZ8 structure is the reduced hydroperoxyl form of the enzyme as suggested by Nagano and Poulos, the authors also acknowledge their own 2A1M "… wild-type dioxygen complex structure is very much the same as reported previously…" by Sligar and co-workers.[118] With this in mind, we proceeded with the 1DZ8 structure as representing an oxygen bound, later-stage catalytic intermediate of the enzyme.

In this work, we worked mechanistically backwards from the peroxo-ligated intermediate to account for the space occupied by molecular oxygen upon binding to reduce the size of the search space during stochastic docking. Our aim was to reduce the occurrence of poses that would not accommodate the binding of molecular oxygen and that, as a result, would not be hydroxylated according to our mechanistic paradigm. While a quality crystal structure the oxo-ligated complex is not available, future work could explore the utility of starting from a later catalytic intermediate once, specifically Compound I, should a quality structure become available. This would also require explicit modeling of catalytically important water molecules. [119]

Receptor Preparation for RxDock

From the 1DZ8 structure, all hydrogens were added using PyMol. Additionally, D297 and H355 were protonated to complete the hydrogen bonding interactions to the propionate substituents on the heme. Each molecular oxygen atom was then assigned a formal charge of -1, and this state will be referred to as the ferric-peroxo state hereafter. To generate a structure more consistent with the ferric/resting state of the heme prior to substrate binding, the dioxygen ligand was removed from the structure without any additional modifications. While we acknowledge a five-coordinate heme could be in the ferrous state following the first electron transfer to the enzyme and before the binding of molecular oxygen, this state is referred to as the ferric state from this point forward.

Cavities for each state of the receptor were prepared using the reference ligand method within RxDock. A radius of 6 Å around each atom of camphor in its native binding orientation in

the 1DZ8 structure was used, and a single cavity was prepared with a small sphere radius of 0.5 Å with a minimum cavity volume of 100 $Å^3$.

Receptor Preparation and Parameter File Generation in Rosetta

From the 1DZ8 crystal structure in the Protein Data Bank, the dioxygen-ligated heme was extracted along with C357. The C357 residue was then truncated to a methylthiolate in PyMol. By default, Rosetta deprotonates aspartic acid residues to yield carboxylated and histidine residues to give their neutral form. Therefore, both proprionate substituents on the heme were protonated to account for Rosetta's default handling of ionizable residues while also getting the overall local charge and proton accounting correct provided D297 and H355 will be deprotonated within Rosetta during docking. The remaining hydrogens were also added in PyMol. Within xtb, a constrained minimization was performed as a dianion with GFN-FF to consistently place hydrogens while fixing all heavy atoms. From this constrained geometry, a conformer/rotamer ensemble was not constructed because the local protein environment around the heme affords salt bridges to the propionate side chains and steric limitations on the vinyl substituents on the heme that restrict the over configuration of the heme. After removal of the methylthiolate ligand, the remaining structure was parameterized for Rosetta in standard fashion using the molfile_to_params.py script, using the "--recharge" flag to set the total charge to -1 (with the other negative charge being formally assigned to the thiolate residue).[120] In this way, the charges on the heme should be consistent with the ferric-peroxo state of the enzyme following the second electron transfer event. In the resulting parameter file, the peroxo ligand oxygen atoms were converted to virtual atoms as the peroxo ligand was parameterized

separately. This heme structure was assigned a chain code of “H” and a three letter residue code of “HM3,” differentiating it from Rosetta’s native heme definition.

The peroxo ligand from the 1DZ8 structure was then also parameterized in standard fashion.[120] The oxygen and iron atoms were extracted to a mol2 file and parameterized, with the iron atom converted to a virtual atom and assigned a partial charge of zero. The partial charges from the heme preparation above for the oxygen atoms were then manually assigned to the peroxo ligand here and the chain was assigned a three letter residue code of O2M and a chain identifier of “O.”

Finally for the enzyme, we generated a “noncanonical” amino acid (NCAA) representation of C357 as a thiolate anion as Rosetta will not deprotonate the thiol by default. Rather than the standard Rosetta protocol for NCAA parameterization, we utilized the 3-chloroalanine NCAA already available in the Rosetta database. The acetylated N-methyl amidated cysteine thiolate was constructed from C357 in the 1DZ8 structure, as is typically done for NCAA generation in Rosetta, and was optimized at the GFN2-xTB level of theory as a singlet anion. Starting from the 3-chloroalanine parameter file, bond lengths and angles were manually copied from the SEQM-optimized structure to the parameter file. Mulliken charges from the optimized geometry were also assigned as the partial atomic charges in the parameter file. The rotamer library from 3-chloroalanine was not used, and a rotamer library was not generated since the cysteine side chain is not expected to occupy alternative conformations. This residue was assigned the three-letter code of “CYA” to differentiate it from other cysteine residues.

Finally, each substrate was parameterized in standard fashion by converting the conformer-rotamer ensemble into a mol2 file and parameterizing the substrate in standard fashion. [120]

All four full-atom parameter files were provided to Rosetta by the "-extra_res_fa" flag and without modification of the default database.

Constraint Generation and Implementation

Estimates of local geometric parameters of a substrate complexed with the ferric-peroxo intermediate were made for use as pharmacophoric constraints during docking. Using a truncated heme model in the ferric-peroxo dianionic state and model substrates, potential energy surface scans using GFN2-xTB were performed along the $O_{proximal}$–$H_{substrate}$ distance on the doublet surface. Ethane, propane, isobutane, and propene were used as models for primary, secondary, tertiary, and allylic hydrogens, respectively. The scan was performed from 2.0 Å to 2.4 Å in 40 steps. The minimum identified along the scan coordinate was used to measure the $O_{proximal}$–$H_{substrate}$ distance, Fe_{heme}–$O_{proximal}$– $H_{substrate}$ angle, and the $O_{proximal}$–$H_{substrate}$– $C_{substrate}$ angle.

Docking with RxDock

A pharmacophoric constraint was used to ensure a nonpolar hydrogen atom docked within 2.22 Å of the proximal oxygen atom's coordinates. RxDock only differentiates between polar and nonpolar hydrogen atoms and cannot limit the pharmacophoric constraint to only aliphatic hydrogen atoms. Additionally, RxDock only affords the ability to set an upper bound on the constraint distance, so the maximum value for the $O_{proximal}$–$H_{substrate}$ distance from constraint generation was used. This constraint was applied in both the ferric-peroxo and ferric

states with a penalty weight of 100.0 arbitrary units per $Å^2$ (although RxDock alleges energy units of kJ / mol). A quadratic penalty is the only option in RxDock.

Each substrate was docked for a total of 1000 trials per conformer in the generated ensemble. With rotamers included in the conformer/rotamer ensemble, the total number of trials was divided over all rotamers in the ensemble and parallelized using GNU parallel.[115] Stochastic docking was first performed in the ferric-peroxo state using the default docking algorithm and desolvated scoring function in RxDock. Then, each pose was minimized in the ferric state using the pose minimization algorithm in RxDock with the desolvated scoring function. During docking, substrate dihedral angles were sampled ± 30° from the conformer provided for docking.

Docking with RosettaScripts

Overall, docking in Rosetta was performed in a similar manner to RxDock, though many more options exist within Rosetta and those differences are annotated below.

Match style pharmacophoric constraints[121] were used with the range of all three geometric parameters described above. Specifically, the $O_{proximal}$–$H_{substrate}$ distance was constrained between 2.08 and 2.22 Å, the Fe – O – H angle was constrained between 122.2° and 128.6°, and the O – H – C angle was constrained between 167.7° and 179.5°. During development, it was noted that the salt bridge between R112 and the heme propionate would frequently be lost during sidechain rotamer sampling. This interaction is highly conserved in the PDB, so a distance constraint of 2.0 ± 0.2 Å between a proton on the arginine residue and a propionate oxygen atom was set. Similarly, the salt bridge between H355 and the heme propionate were also routinely lost during side chain sampling, so again a distance constraint of

2.0 ± 0.2 Å was applied. For all constraints, a conservative penalty weight of 50 Rosetta Energy Units (REU) / Å was applied.

Side chains were first saved using the SaveAndRetreiveSidechains mover. All side chains except prolines and glycines were then converted to alanine. The Transform mover was used to stochastically dock the substrate in the polyalanine active site using a box size of 5 Å, a step size of 0.1 Å, and a rotation angle of 360°. This was done with Rosetta temperature of 5.0 for at least 10 cycles or until a negative score was returned. The original side chains were then restored. After applying the match style constraints above, the substrate was minimized. Three rounds of side chain sampling of residues with 12 Å of the substrate, including minimization of side chains, backbone atoms, and the substrate. Through this, the heme and peroxo ligand were treated rigidly to conserve the orientation of salt bridges and the peroxo ligand. The peroxo ligand was then removed and the system within 12 Å of the substrate, including the heme, were minimized one last time. The InterfaceScoreCalculator was used to compute the interface energy between the substrate and the enzyme. The number of docking trials per abstractable hydrogen bonded to an sp^3 hybridized carbon was 50 times the number of conformers in the ensemble for no more than 250 trials.

Post-Docking Calculations

An important limitation within RxDock is that explicitly tethering each hydrogen atom in the substrate to the proximal oxygen atom is not possible. Instead, a pharmacophoric distance constraint between the proximal oxygen atom and any hydrogen atom is possible, allowing potentially multiple hydrogen atoms to satisfy the constraint. To predict the HAT barrier for each pose, local geometric parameters between each hydrogen atom in the substrate and the

proximal oxygen atom were computed. First, only hydrogens with a Fe_{heme}–$O_{proximal}$–$H_{substrate}$ angle of less than 150° but more than 100° were considered. Next, only hydrogens with a $O_{proximal}$–$H_{substrate}$ distance at most 2.2 Å was considered. If a pharmacophoric constraint penalty was assess, the closest hydrogen was considered, while still minding the previously mentioned Fe_{heme}–$O_{proximal}$–$H_{substrate}$ angle constraint. From these hydrogens in a geometrically reasonable orientation relative to the proximal oxygen atom, the one resulting in the lowest HAT barrier was taken as the abstracted hydrogen for that pose. In addition to recording this most reactive hydrogen and its associated HAT barrier, the closest hydrogen was also recorded for product prediction without using these post-docking descriptors.

Within Rosetta, explicit tethering of each hydrogen atom bonded to an sp^3-hydridized carbon atom was performed. Therefore, the abstracted hydrogen atom in each pose was assigned according to which hydrogen atom was explicitly tethered.

For both RxDock- and Rosetta-generated poses, the gas phase HAT barrier in kcal/mol for the pose was estimated using our regression model (Equation 3-1) at the GFN2-xTB//GFN-FF level of theory as previously described.[70]

$$\Delta E^{\ddagger}_{4,HAT} = \left(BDE_{\ C-H,FR} * 0.5674 + 0.0956\right) * \ 627.5096 \qquad \text{Eqn. 3-1}$$

In addition to estimating the HAT barrier, the single point potential energy of each pose was computed in the gas phase using GFN-FF to calculate the conformation's relative potential energy as compared to the best conformer in the generated ensemble used for docking. Next, each pose was minimized using GFN-FF to estimate the distortion relative to the nearest adjoining minimum in the gas phase ($\Delta E_{distortion}$). We assume, but cannot guarantee, that the adjoining minimum is a conformer in the initially generated ensemble. While not used in

analysis, this quantity was recorded. Post-docking calculations could be accelerated by forgoing these geometry optimizations. Prior to pose clustering, poses with a relative potential energy greater than 4.0 kcal/mol were filtered out as physically improbable. 4.0 kcal/mol was chosen to afford 1.0 kcal/mol of flexibility about substrate dihedral angles over the 3.0 kcal/mol energy window set in conformer/rotamer ensemble generation. In our view, this was particularly important for the analysis of Rosetta results as the ref2015 scoring function in Rosetta negates intramolecular repulsion for ligands.

Pose Clustering

Clusters of substrate poses were generated using heavy atom pairwise root mean square deviation (RMSD) clustering using DockRMSD.[114] All poses for each substrate were first sorted according to their interface score. In RxDock, the intermolecular score term was utilized, and for Rosetta, the interface score used was the "interface_delta_X" term from the InterfaceScoreCalculator mover.[120] These will be generally and interchangeably be referred to as "interface scores" hereafter. The best (lowest) scoring pose served as the seed structure for the first cluster. The RMSDs between the seed and all other poses were determined, and those poses within 1.0 Å were assigned to the first cluster. Tukey's fence method[122] was used to remove high interface score outliers in each cluster, with an upper fence set to 3.0 times the interquartile range over the third quartile of interface score in the cluster. Removed outliers were recycled back into the remaining pose list for ultimate clustering. A second cluster was formed as above using the best scoring pose of the remaining poses as the seed structure, and clustering was repeated in this way until all poses were assigned to a cluster. Enzyme conformational differences and the total complex score from Rosetta were not considered.

Prediction Assignment

For RxDock, assignment of the abstracted hydrogen was first performed without the consideration of HAT barriers calculated for docked poses. Generated clusters for each substrate were rank ordered according to each cluster's lowest interface score, and the hydrogen atom closest to the proximal oxygen atom in the peroxo ligand was assigned as the abstracted hydrogen. Where the closest hydrogen was bound to a sp^2 hybridized carbon atom, that cluster was manually skipped over in rank ordering. This manual intervention was required due to the limitations of RxDock's pharmacophoric constraint features. Such predictions in the absence of reactivity information were not performed using Rosetta.

For both Rosetta and RxDock results, prediction assignments were then made using the estimated HAT barrier of each pose during clustering. Once clustered as above, the abstracted hydrogen atom was assigned for the cluster according to the lowest estimated HAT barrier within a cluster. In this way, the assignment in the cluster was made.

Lastly, clusters, and thus the abstracted hydrogen atom predictions, were rank ordered according to the best interface score in each cluster. For completeness in record keeping by both approaches, the RMSD between the top-ranked cluster and the remaining clusters were computed as previously described, using the seed structure within each cluster. Additionally, the number of poses in each cluster was recorded.

Results and Discussion

Pharmacophoric Constraint Generation

Relaxed surface scans of the O-H coordinate for four model substrates were performed at the GFN2-xTB[23] level of theory. The O-H coordinate was chosen as the HAT event with

Compound 1 proceeds along this motion. We assess a minimum along this coordinate in this reduced model is representative of the Van der Waals complex between the substrate and the enzyme following the second electron transfer step and relevant as the substrate should be positioned to be product once Compound 1 is formed. The structure along each scan with the lowest potential energy was used to extract the geometric parameters indicated in Figure 3-2.

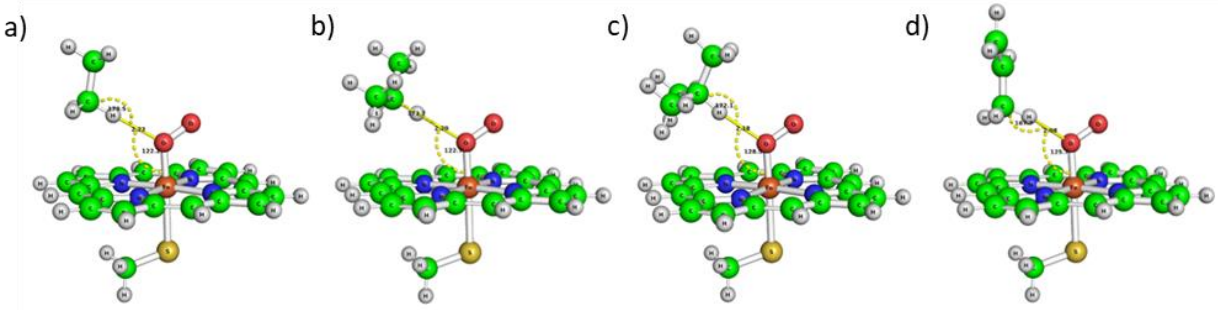

Figure 3-2 Minima structures along relaxed surface scans of the O-H coordinate for a) ethane, b) propane, c) isobutane, and d) propene in complex with a truncated heme model in the ferric-peroxo state. The $O_{proximal}$–$H_{substrate}$ distance, $O_{proximal}$–$H_{substrate}$–$C_{substrate}$ angle, and Fe_{heme}–$O_{proximal}$–$H_{substrate}$ angle are labeled for each. Carbon, nitrogen, oxygen, sulfur, hydrogen, and iron atoms are green, blue, red, yellow, white, and bronze, respectively.

The generation of reasonable geometric parameters to use as pharmacophoric constraints during docking is necessary to limit the size of configurational space that is searched during any docking algorithm. While many QM studies focus on the HAT event between Compound 1 and the substrate, we are unaware of any studies that model the substrate in complex at the ferric-peroxo state in the catalytic cycle. To gauge the reasonableness of our constraints, we compared the $O_{proximal}$–$C_{substrate}$ distance of the propane complex above (Figure 3-2b) with the $O_{proximal}$–$C_{substrate}$ distance in the 1DZ8 and 2A1M[118] crystal structures. These comparisons were made as all three represent hydrogen abstraction at a secondary aliphatic

carbon. The $O_{proximal}$–$C_{substrate}$ distance for propane of 3.29 Å was approximately 0.2 Å longer than the 3.11 Å and 3.09 Å distances measured in the 1DZ8 and 2A1M structures, respectively. Additionally, the Fe_{heme}–$O_{proximal}$–$C_{substrate}$ angle in the propane complex was found to be 123.1°, which was bound by the angles of 122.8° and 126.1° in the 1DZ8 and 2A1M structures, respectively. Given both structures are assessed as the ferrous dioxygen complex following the first electron transfer step and not as the ferric-peroxo state, we assessed our geometric parameters as reasonable and continued.

Similar modeling of reduced systems could have likewise been performed using the ferric-hydroperoxo (the Compound 0) state. However, as we proceeded forward with docking, the location of the "catalytic" water[119] implicated in protonating the ferric-peroxo state would need to be included. Incorporating explicit water molecules is an area of active research, both in docking algorithm development as in the recently released AutoDock Vina 1.2.0[123] and in applications of machine learning to improve scoring functions and their handling of relevant entropic effects.[124] However, the incorporation of explicit water within Rosetta for enzyme design is nontrivial, with the introduction of bias on part of the researcher being our chief concern. For this reason, our effort to consider late-stage catalytic intermediates only went as far as the ferric-peroxo state.

Rigid Receptor Docking in RxDock

While our aim is to build methodology that can be used for enzyme design within Rosetta, we first employed a more traditional docking utility that could find use in virtual screening and binding mode prediction. While AutoDock Vina is arguably (perhaps irrefutably) the most common with over 20,000 citations of the initial publication,[125] the clustering

algorithm is run automatically, and we wished to recover all generated poses, not just the clustered results. Secondly, neither nonpolar hydrogen atoms nor bond orders, which could be used to replace hydrogens in post-processing, are retained in the output pdbqt file format used by AutoDock Vina, and ligand hydrogens are required for estimating HAT barriers. Lastly, pharmacophoric constraints during docking is not trivially performed within AutoDock Vina. While lesser known, RxDock[113] overcomes each of these limitations.

Traditionally, high throughput virtual screenings might only consider and dock into the ferric state of P450cam with all crystallographic waters removed. As our groups[126] and others[127] have previously described for terpene synthases, the incorporation of mechanistic information by way of evaluating multiple catalytic intermediates yields improved predictions. The central premise of these studies is that docked poses of intermediates along a catalytic reaction coordinate must be consistent with the adjoining catalytic states for the pose in the initial state to be a productive binding mode. We extended this logic to this work in that poses in complex with the varied iron-oxygen intermediates of the heme must be configurationally consistent. We assume that drastic geometric differences in substrate orientation between catalytic intermediates are improbable and that a correct, productive binding mode in the ferric state is likely to accommodate the binding of molecular oxygen.

To this end, we hypothesized that performing stochastic docking into the ferric-peroxo state followed by minimization of these poses after the removal of the peroxo ligand would afford poses in the ferric state of the enzyme that would accommodate molecular oxygen. While multiple electron transfer events occur following substrate binding,[10] it is unlikely that any docking program and associated scoring function could reliably capture the nuanced

electronic differences between the ferric-peroxo and ferrous-dioxygen states and between the ferrous and ferric states of the enzyme. Therefore, we docked only into the ferric-peroxo state and worked backwards, minimizing in the ferric state.

While RxDock allows for controlled sampling of dihedral angles to prevent gross deviations from input conformations, we additionally evaluated the need to incorporate post-docking calculations of relative substrate potential energies affected product predictions. We visually examined the distribution of all intramolecular terms from RxDock and relative potential energies from post-docking calculations for the twelve substrates with rotatable dihedrals other than methyl rotations through a pairplot. 94% of the poses in this subset of substrates had distortion energies less than 4.0 kcal/mol, with the bicycloheptane and piperidine derivatives having the most poses with high relative potential energies as compared to the best discovered conformer as seen in Figure 3-3. The lack of correlation between force field-derived relative potential energies in post-docking calculations and the ligand intramolecular scoring term from RxDock agrees with Huang's assessment of the challenges in considering ligand flexibility during virtual screening.[128] These collective findings point to this as an area for improvement, at least in RxDock and those packages sampled by Huang (AutoDock Vina,[125] DOCK,[129] and MDock[130]). Because of this disagreement and the extensive benchmarking performed in the development of GFN-FF,[25] we sought to apply a post-docking filter based on the GFN-FF computed relative potential energy of a pose, and we only considered the intermolecular score term from RxDock for our predictions.

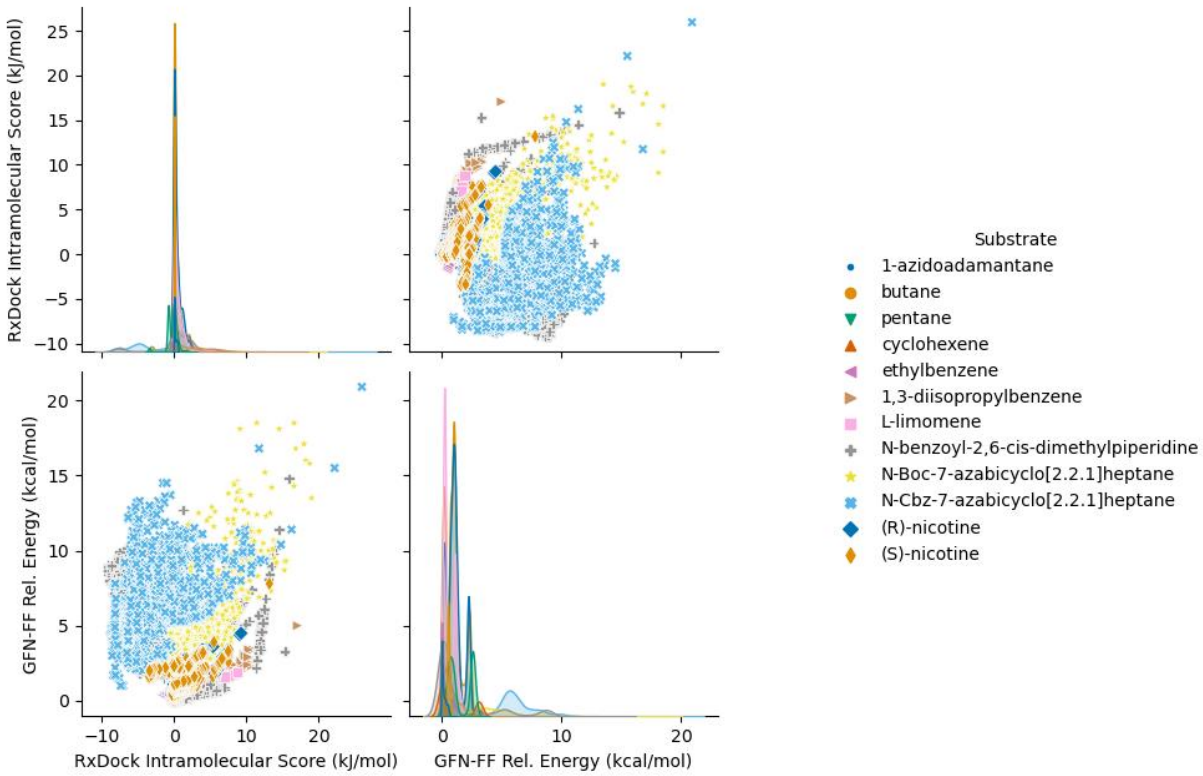

Figure 3-3 Pairplot of RxDock intramolecular scores and GFN-FF relative potential energies for flexible substrates in the panel. Flexible substrates were taken as those with rotatable bonds other than methyl rotations.

From our inspection, we established a post-docking filtering threshold of 4.0 kcal/mol using the relative potential energy of a pose prior to clustering. While a threshold selection of 4.0 kcal/mol afforded conformers within 3.0 kcal/mol range of the lowest energy conformer to modestly distort further, we additionally assessed 4.0 kcal/mol as a reasonable threshold based on n-butane as a conformationally flexible model and known substrate of our system. Where previous studies have placed the *syn* conformation of butane ca. 5.5 kcal/mol higher in potential energy relative to the *anti* conformation,[131] a threshold of 4.0 kcal/mol would allow for exploration of most of butane's configurational space.

With a filter in place to remove unreasonable distorted compound poses, we gauged the benefit of incorporating estimated HAT barriers into our analysis by comparing the results from RxDock both with and without considering the HAT barrier during clustering as described in our methods. As shown in Figure 3-4 below, a substantive improvement in our prediction success rate for the primary hydroxylation product was observed when HAT barriers were used during pose clustering to assign the abstracted hydrogen. These benefits were most notable in the top one and two ranked predictions, were prediction success rates increased by 20% and 12%, respectively. This finding suggests that between clustered poses, the correct hydrogen to be abstracted is the one with the lower barrier to abstraction rather than the hydrogen closest to the proximal oxygen atom.

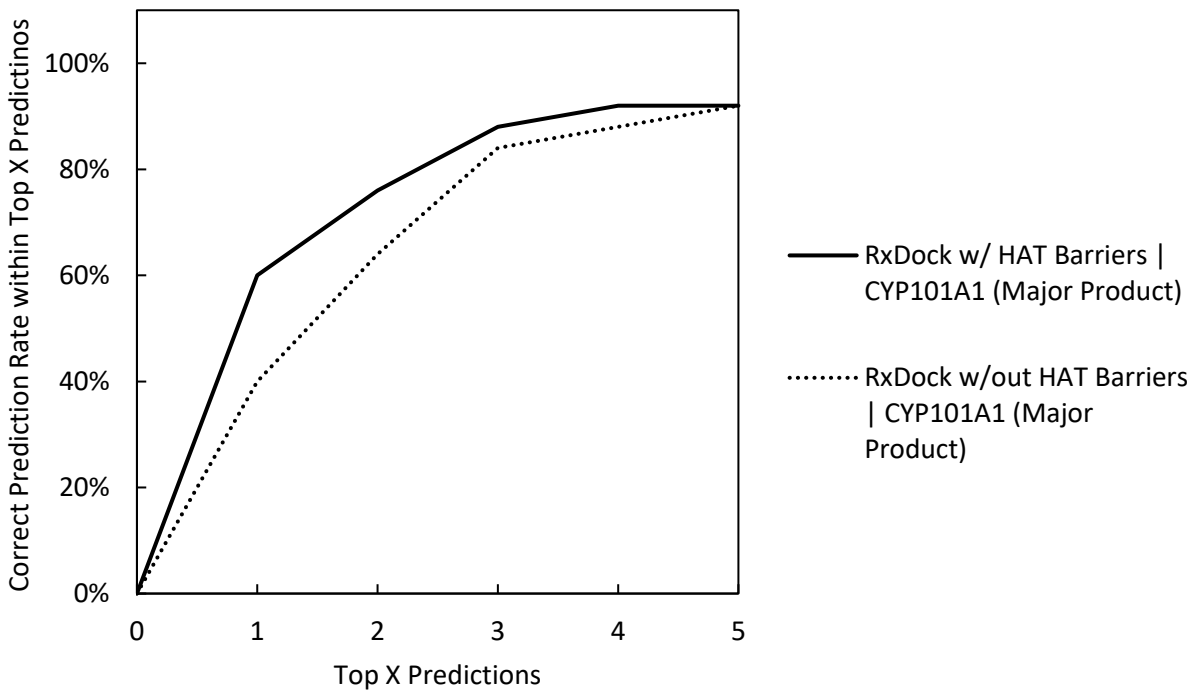

Figure 3-4 RxDock prediction success rates for primary hydroxylation products across the top 5 ranked predictions, comparing the inclusion and exclusion of HAT barriers to assign the abstracted hydrogen.

To gauge the comparative performance of our approach, we examined our success rates alongside those of Tyzack.[109] For a more direct comparison, we computed our success rate in predicting any experimentally observed product in the first five predictions across our substrate panel to compare to the success rate for the top two and three predictions reported by Tyzack. Our approach achieves performance on par with or modestly improved over the inclusion of bond order alone as employed in their work.[109]

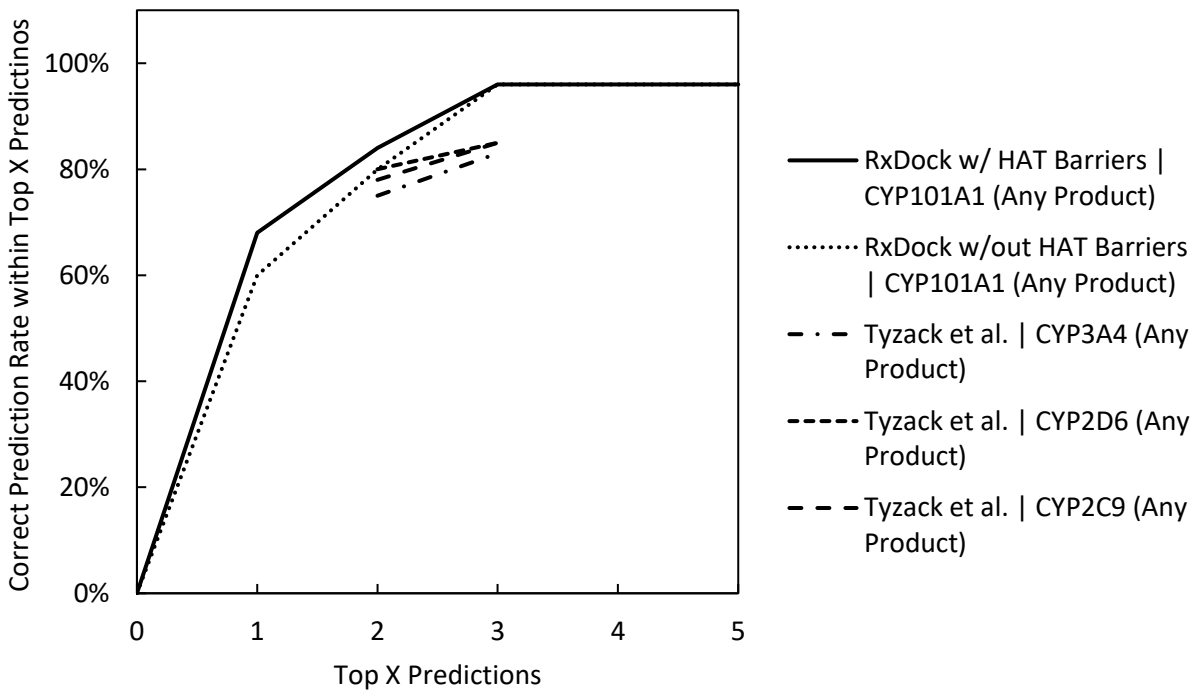

Figure 3-5 Prediction success rate comparison between work presented here and that previously reported in the literature.[109]

We further considered the improvement of each approach over the use of docking scores alone in the top 3 predictions. The incorporation of estimated HAT barriers yielded comparable improvements to those achieved with reactivity measured employed by Tyzack. In general, the agreement between our improvements promotes the importance of considering reactivity in addition to estimated binding affinities and substrate fit considered through docking alone. While comparable in performance, we propose our approach to be more generalizable given we do not employ scoring parameters or tethering distances that are varied according to the enzyme isoform in question. Additionally, our approach is more computationally affordable, relying on semi-empirical and force field methods as opposed to density functional theory is moderately sized Pople basis sets.[41]

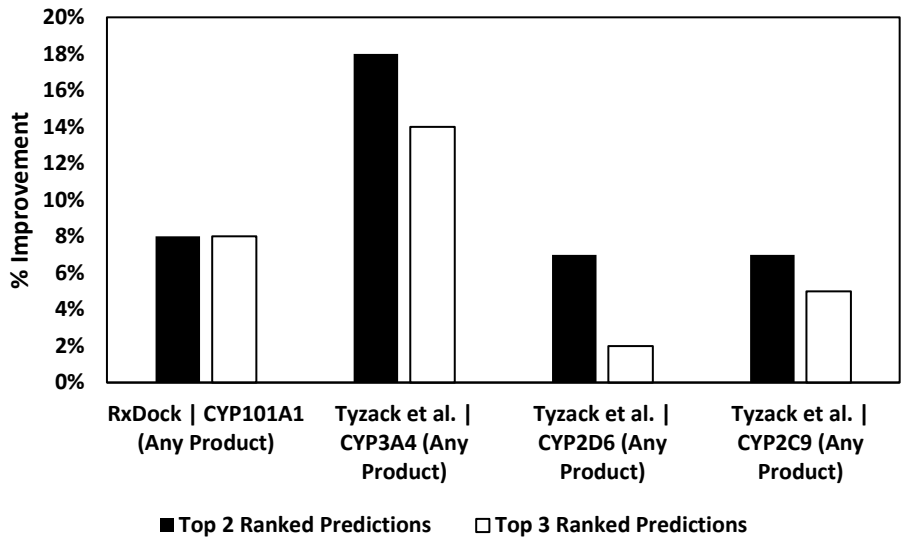

Figure 3-6 Comparison of improvements to prediction success rates through the inclusion of reactivity measures over docking alone. Reactivity measures included estimated HAT barriers employed in this work alongside RxDock outputs and bond orders or SET scores in the work reported by Tyzack and co-workers.[109]

While perhaps critical of the exact approach taken by Tyzack and co-workers to re-parameterize pose scoring using C–H bond orders for HAT reactivity, their work is undoubtedly inspirational and respectful of the need to consider electronic structure and reactivity alongside substrate fit and binding affinity estimates from docking. While their study examines a larger substrate panel in multiple human isoforms of cytochrome P450, ours focuses on the prototypical, bacterially originated CYP101A1 as an enzyme system that is known to be amenable to mutation for protein engineering.[132] With our comparison to previously published work complete using a docking package more suited for virtual screening, we turned our attention to implementing this work within Rosetta with an ultimate goal of developing methodology for enzyme design.

Product Predictions in Rosetta

While pharmacophoric constraints were limited to the O_{heme}–$H_{substrate}$ distance in RxDock, all distance and angular parameters between the heme, the proximal oxygen atom (or its coordinates) and a substrate hydrogen atom are accessible in Rosetta through match style constraints. Therefore, we employed all the geometric parameters recovered from SQM calculations above in Figure 3-2as a penalty-free range from the minimum to the maximum value for each parameter. Two additional parameters are notably absent. Both the $X_{substrate}$–$C_{substrate}$–$H_{substrate}$–O_{heme} and $H_{substrate}$–O_{heme}–Fe_{heme}–N_{heme} dihedrals were not restricted in our approach. We believe limitations about these degrees of freedom should be controlled by the scoring function within Rosetta. Specifically, the rigid body treatment of the ferric-peroxo heme should limit the accessible angular space within the active site with respect to the $H_{substrate}$–O_{heme}–Fe_{heme}–N_{heme} dihedral. Likewise, the $X_{substrate}$–$C_{substrate}$–$H_{substrate}$–O_{heme} dihedral, and the conformation of substrate in general, should controlled by the supplied conformer/rotamer ensemble and interactions in the enzyme active site. In general, we believe as few constraints as are a mechanistically relevant or physically reasonable should be included, with all remaining degrees of freedom controlled by the scoring function to reduce bias introduced by the modeler. However, as mentioned previously with regards to substrate flexibility, we expect the Rosetta framework and chosen scoring function to struggle with substrate intramolecular repulsion. As pointed out by Smith and Meiler,[133] ref2015[134] and three other scoring functions within Rosetta are not the top performers in the Comparative Assessment of Score Functions 2016 (CASF-2016)[135] benchmark with respect to recovering the native binding mode in top ranked predictions. While GALigandDock is the latest iteration in substrate pose prediction and

shows improvement native binding mode prediction,[136] ref2015 was employed here as the default scoring function used for enzyme design efforts and our aims to apply this work in that area.

Again, we examined the distribution of relative potential energies for flexible substrates. 23% of all flexible substrate poses are above ca. 4 kcal/mol, with 10% of poses above ca. 16 kcal/mol (Figure 3-7). These extremes are significant increases over our observations with RxDock. Where discrete control of ligand flexibility in RxDock is possible, the same degree of granular control is not available within the enzyme design framework in Rosetta. The generation of physically unreasonable poses, based on extremely high potential energies relative to the best discovered conformer for a substrate, was again most notable for the piperidine and bicycloheptane derivatives. While improvements have been made within Rosetta for docking specifically,[136] such distorted structures highlight the need for further development in this area across the community, particularly for applications in enzyme design.

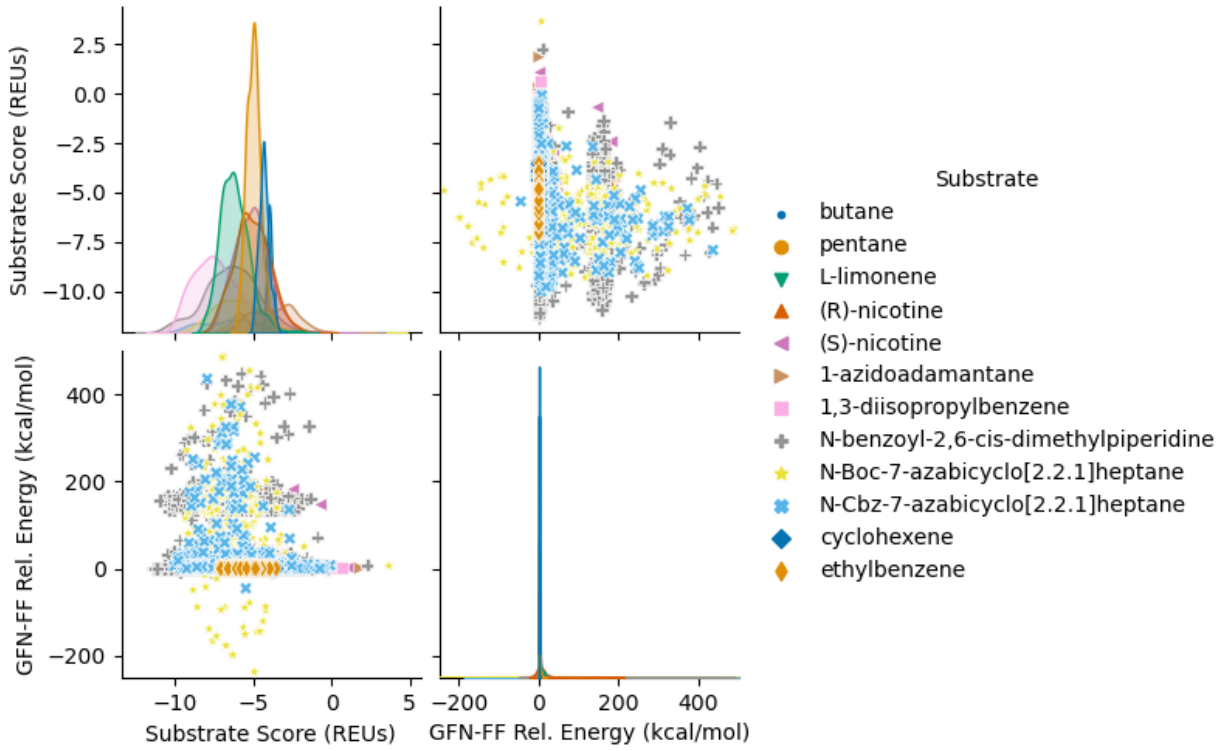

Figure 3-7 Pairplot of relative potential energies and substrate distortion energies as calculated with GFN-FF in post-docking for Rosetta-produced poses in flexible substrates. Flexible substrates were again taken as those with rotatable dihedrals other than methyl rotations.

For consistency with our treatment of docked poses from RxDock, we proceeded with a threshold of 4.0 kcal/mol for relative potential energy as computed post-docking using GFN-FF. We applied our complete filtering and clustering approach as before. Figure 3-8 presents our prediction success rate for primary hydroxylation products in our substrate panel. In our panel of 25 substrates, we achieved a correct prediction rate of 92% for any observed product in the top 2 ranked predictions. While perhaps qualitatively improved over the findings of Tyzack and co-workers,[109] we acknowledge the aim of their inspirational work was focused on enabled pharmacokinetic predictions rather than developing methodology to ultimately extend to in silico enzyme engineering. Additionally, our substrate panel is more modest in size as our

implemented protocol in Rosetta will not scale readily to hundreds or thousands of compounds given the degree of allowed protein flexibility.

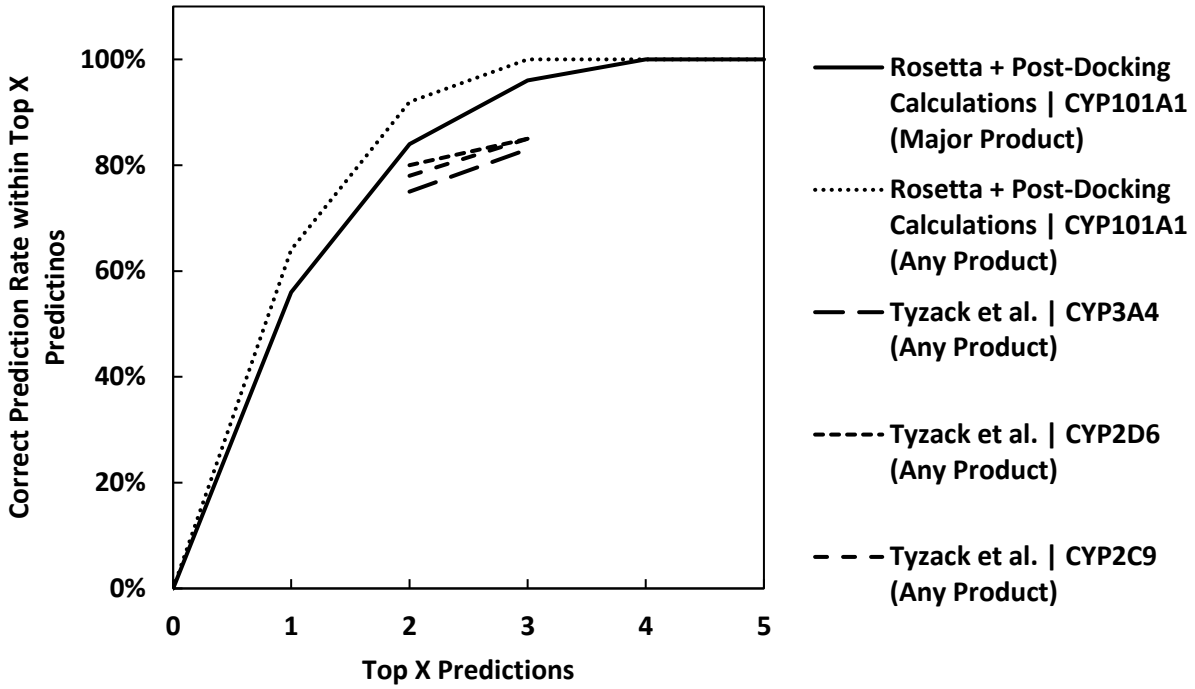

Figure 3-8 Hydroxylation product prediction performance using Rosetta in conjunction with post-docking ligand-centric calculations and clustering. Previously published work is plotted for comparison.

Binding Mode Prediction

The correct prediction success rates presented above are indeed encouraging. While enthusiastic about correctly predicting any observed hydroxylation product with a high success rate in the top two and three ranked poses, the correct prediction of the major product is more relevant to enzyme engineering efforts, and we assess an 84% success rate as promising. However, equally important to the prediction success rate is that predicted poses for major product formation are consistent with experimentally observed orientations where data is

available. Unfortunately, many CYP101A1 crystallographic structures are in complex with camphor or related analogs. Still, making what limited comparisons are available is useful to assess, at least qualitatively, the performance of our approach with respect to binding mode prediction. Figure 3-9 presents our top generated pose consistent with major product prediction for camphor, camphane, (S)-nicotine, and thiocamphor as compared to experimental findings.

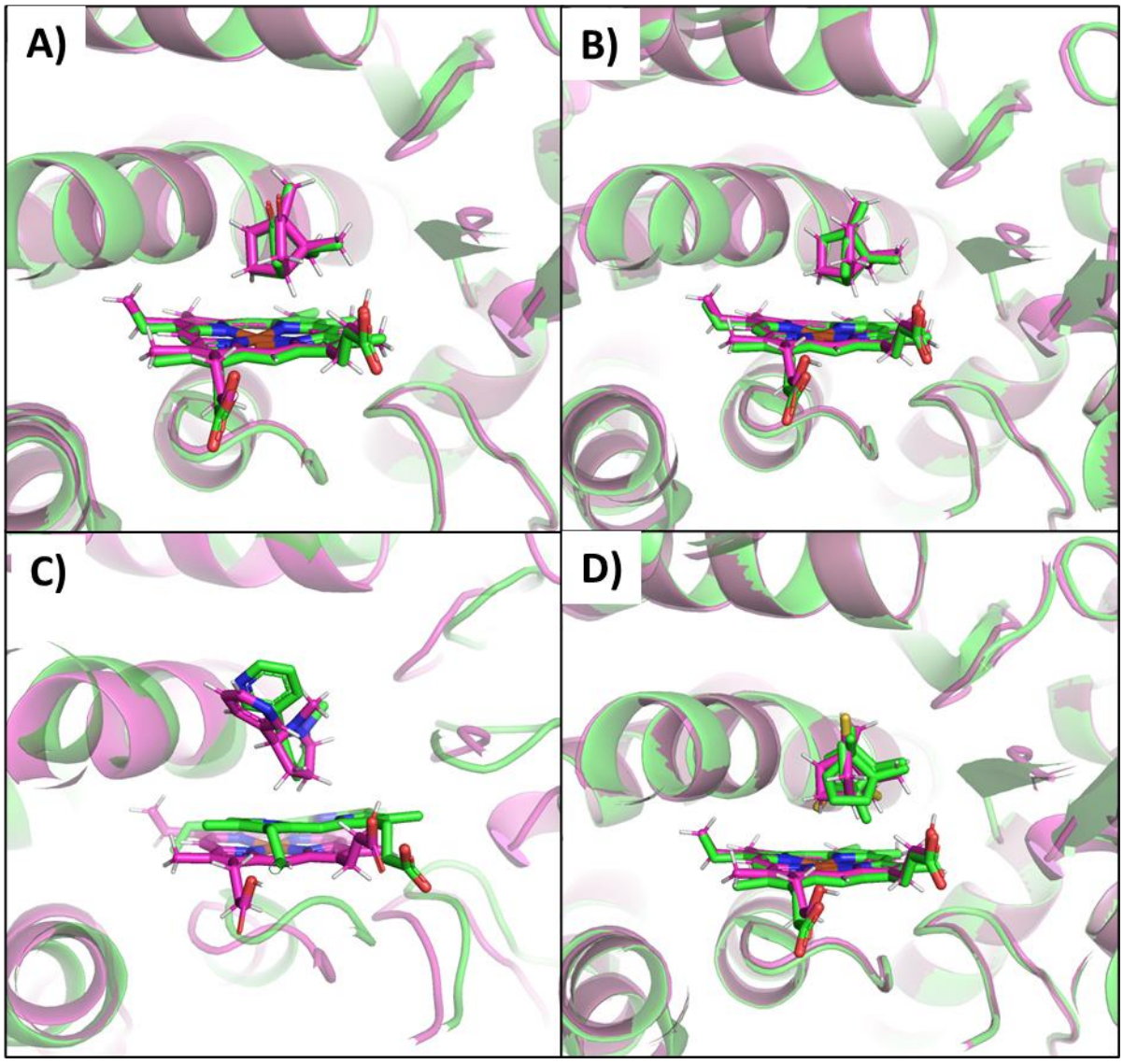

Figure 3-9 Binding pose prediction (magenta) comparisons for select compounds where crystallographic data (green) is available. A) Camphor in complex with CYP101A1 (2A1M).[118] B) Camphane in complex with CYP101A1 (6CPP).[137] C) (S)-nicotine in complex with CYP2A6 (4EJJ).[138] D) Thiocamphor in complex with CYP101A1 (8CPP).[137]

Perhaps unsurprisingly, camphor and its closely related analogs are predicted to bind in modes consistent with known structures. As our ultimate aim is to leverage enzymes such as CYP101A1 to transform non-native substrates, we considered the binding mode of (S)-nicotine

in CYP2A6. [138] In our work here, we recovered a predicted binding mode that is qualitatively consistent with that in the CYP2A6 system but rotates the pyridinyl substructure to satisfy a hydrogen bond with the Y97 residue in CYP101A1. While a quantitative comparison between the two structures would be preferable, differences in sequence identity around the active site, notably the absence of a hydrogen bond donor in CYP2A6 that maps to Y97 in CYP101A1, understandably results in a differed binding mode between the two complexes. Lastly, experimental structures may not capture productive binding modes that lead to major hydroxylation product formation. Such is the case for thiocamphor in CYP101A1 in the 8CPP structure[137] wherein thiocamphor occupies two separate binding modes, neither being consistent with formation of the 5-exo hydroxylated major product. Given the limited number of experimental comparisons that can be made to assess binding mode prediction performance in our panel of substrates, careful inspection of selected binding modes should be made with a critical eye. Whether for the purpose of product prediction in virtual screening as immediately shown in this work or in projected enzyme design efforts, the employment of complimentary methods for quantitative assessment of the filtered results may be needed to further validate our performance. Pose clustering followed by molecular dynamics simulations have been benchmarked in the literature and suggested for the reduction of false positive hits from docking.[139] While too expensive for routine use early in a pipeline with many drug candidates or enzyme mutants to screen, our presented work may benefit from MD as a final step in a comprehensive pipeline to confidently predict product formation or to evaluate proposed mutants to focus wet lab efforts on those leads with the highest probability of success.

Conclusions

The inclusion of a measure of HAT reactivity is critical in predicting hydroxylation outcomes from CYP450 enzymes. This is consistent between Tyzack and Glenn's work in human CYP450 isoforms[109] and our studies with both RxDock and Rosetta presented herein. We have demonstrated that systematic estimates of HAT barriers are useful in the prediction of CYP450-mediated hydroxylation products for the CYP101A1 enzyme. Furthermore, our approach does not rely on parameterized rescoring and should generalize to other docking packages and the analysis of their outputs. In performing substrate-centric gas phase calculations with modern force field and semi-empirical methods, the computations required are affordable and scalable alongside docking and pose clustering, being orders of magnitude less expensive than even modest density functional theory on the same computing resource.

Additionally, the importance of assessing and limiting substrate flexibility presented above is consistent with the generalized assessment from Huang.[128] By systematically assessing substrate distortion, even if only in post-processing, we present a systematic approach to evaluating the reasonableness of docked ligand poses. Relative potential energies in our panels of more than 20 kcal/mol were observed, and while such gross deviations from the lowest energy conformer in the generated ensemble may at times be obvious, our implementation of post-hoc filtering using GFN-FF computed potential energies provides an affordable, quantifiable, and objective measure by which to remove unreasonable poses.

In total, prediction success rates using both RxDock and Rosetta along with the additional descriptors, filtering, and clustering we have described rival or exceed those previously reported.[109] As it specifically relates to enzyme design, our methodology is the first

reported utilizing Rosetta, with the explicit aim of engineering CYP450 enzymes. Given a primary product prediction success rate of over 90% in the top three predictions, we believe our approach within RosettaScripts will naturally lend itself to enzyme design as both leverage the same movers. Such protein engineering efforts are of immediate interest to our groups and our collaborators, and proof of concept studies to this end are forthcoming.

Chapter 4: Predicting Photoisomerization Under Biologically Relevant Conditions

Introduction

Molecular photoswitches, molecules that photoconvert from one stable isomeric state to another (meta)stable state, are of significant interest in microscopy applications. In particular, those that undergo large conformational changes, that demonstrate fatigue resistance in an aqueous environment, and that possess a high quantum yield are attractive tools in to probe, and even perturb, biological systems.[140] Among such biologically relevant photoswitches, azo compounds, particularly azobenzene, are perhaps the oldest known[141] and most widely studied.[142] In recent years, however, heteroaryl variants of azobenzene have gained increased interest to further tune the electronic properties of the photoswitch, extending beyond just substituent modifications as is the case of azobenzene.[143] Nevertheless, azobenzenes remain relevant as important tools to probe biological systems, and further synthetic techniques are warranted to expand their utility. In addition, acylhydrazones have been developed as modular and tunable photoswitches.[144] Acylhydrazones are generally hydrolytically stable and tunable, making them an attractive motif to couple with azobenzenes to create molecular photoswitches with multiple photo-induced isomerizations.

Zhu and co-workers expanded the viable synthetic routes to functionalized azobenzenes following a redox isomerization strategy that incorporate both the azobenzene and acylhydrazone functionalities.[1] In doing so, the desire was to develop a selective multimodal photoswitch to target pharmacological targets according to the attached pharmacophore. Isomerization selectively was assumed based on the presence of two otherwise photoisomerizable functional groups, the azobenzene and the acylhydrazone motifs, being

present. Figure 4-1 shows their generalized product, where they sought to install pharmacophores (generalized below as *R)* proximally to the acylhydrazone.

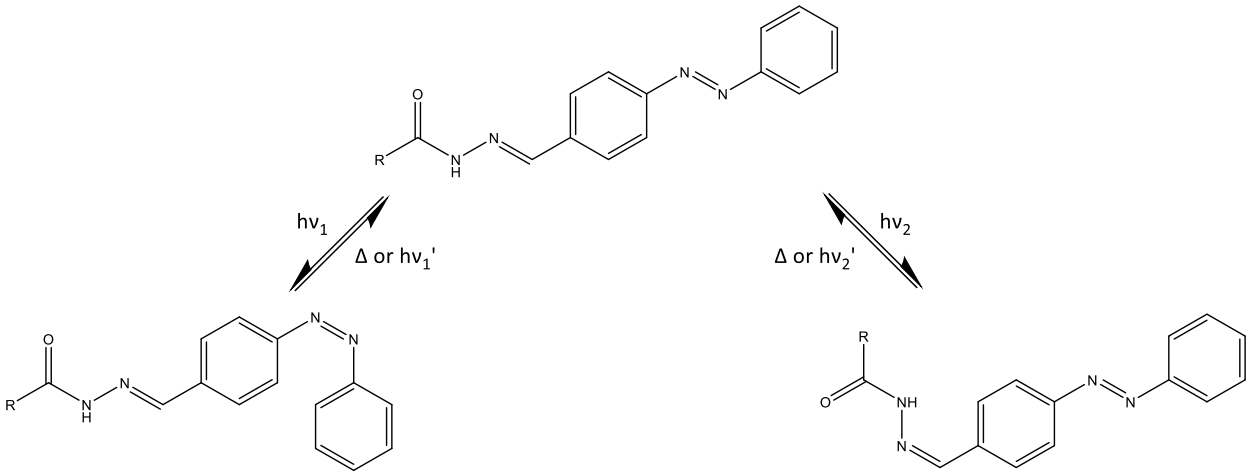

Figure 4-1 Proposed photoswitch containing two photoswitchable motifs.

However, Zhu and co-workers were unable to isomerize the acylhydrazone moiety but showed excellent efficiency in isomerizing the azobenzene motif. To rationalize this observation, we employed time-dependent density functional theory (TD-DFT) to rationalize the inaccessible isomerization of the acylhydrazone functionality. The work presented in this chapter has been previously published,[1] and the associated text and content is used with permission.

Computational Methods

DFT calculations were performed using Gaussian 16 Revision A.03,[37] and surfaces and predicted absorption spectra were rendered using default settings GaussView 6. Figure 4-2 provides the three structures for which calculations were performed as representative structures where electron-donating or electron-withdrawing groups were included by Zhu and co-workers proximal to the azobenzene motif to tune the photoswitch, and a phenyl group was included proximal to the acylhydrazone to represent a pharmacophore substituent. For **1**, **2**,

and **3** below (Figure 4-2), *Z* and *E* isomers of both the azo and acylhydrazone groups were modeled.

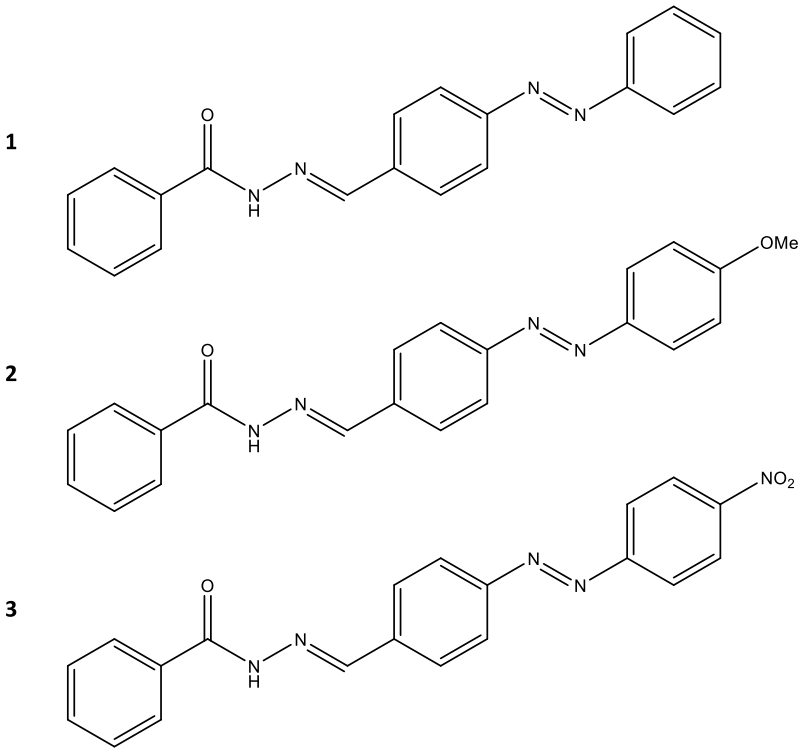

Figure 4-2 Photoswitches of interest used to computationally explore the effects of electron donating and withdrawing groups.

Ground state structures were first optimized at B3LYP[38, 39, 61]/3-21G PCM(DMSO). cam-B3LYP[145]-GD3BJ[146]/6-311+G(d,p) PCM(DMSO) was then utilized for further geometry refinement and for TD-DFT calculations of the excited state. The coulomb-attenuating method was chosen due to the degree of conjugation in our systems. Grimme's D3 empirical dispersion[79] with Becke and Johnson damping[147] was included as dispersion effects are likely important in Z-isomers of the switchable moieties given the steric crowding that prevents planarity of the system. PCM[148] solvation corrections were included in both the ground and excited states[149] using the standard linear response solvation method within Gaussian, and DMSO was selected as the solvent to

mirror experimental conditions. Absorption spectra were predicted for the first 10 singlet excitations for each isomer. Overall, this level of theory accurately reproduced experimentally observed spectra for (E)-1,2-diphenyldiazene and (E)-N'-benzylidenebenzhydrazide (not presented) as a quality check. Excited state densities were calculated for the π to π* transition for E-azo isomers and for the n to π* transition for Z-azo isomers as the transitions most consistent with the excitation wavelength (365 nm) used experimentally. Differential electron densities between the excited and ground states were then visualized as a surface in GaussView to better understand the localization of electron density changes between states and the origin of the excited electrons. In this way, the observation of only isomerizing the azo group may be rationalized.

Results and Discussion

Absorption spectra for **1**, **2**, and **3** are presented in Figure 4-3 for both the E-azo and Z-azo isomers. Zhu and co-workers achieved an enriched 9:1 *Z*-azo:*E*-azo photostationary state upon excitation at 365 nm. Shorter wavelength excitations were not explored in their work due to the biological incompatibility of shorter UV wavelengths and the desired biological application of these compounds. As shown in Figure 4-3, TD-DFT calculations performed here predicted an excitation maximum at approximately 365 nm, coinciding with the Zhu's findings. As expected, the inclusion of the nitro group modestly red-shifted the π to π* transition to a predicted wavelength of 381 nm, with the methoxy group not predicted to significantly shift this transition. Our calculations show that the π to π* transitions for the *E*-azo isomers of **1**, **2**, and **3** all have relatively poor spectral overlap with the n to π* transitions at ca. 453 nm for the

corresponding Z-azo isomers. This contributes to the ability to access photostationary states of **1**, **2**, and **3** and selectively isomerize these compounds across the azo motif.

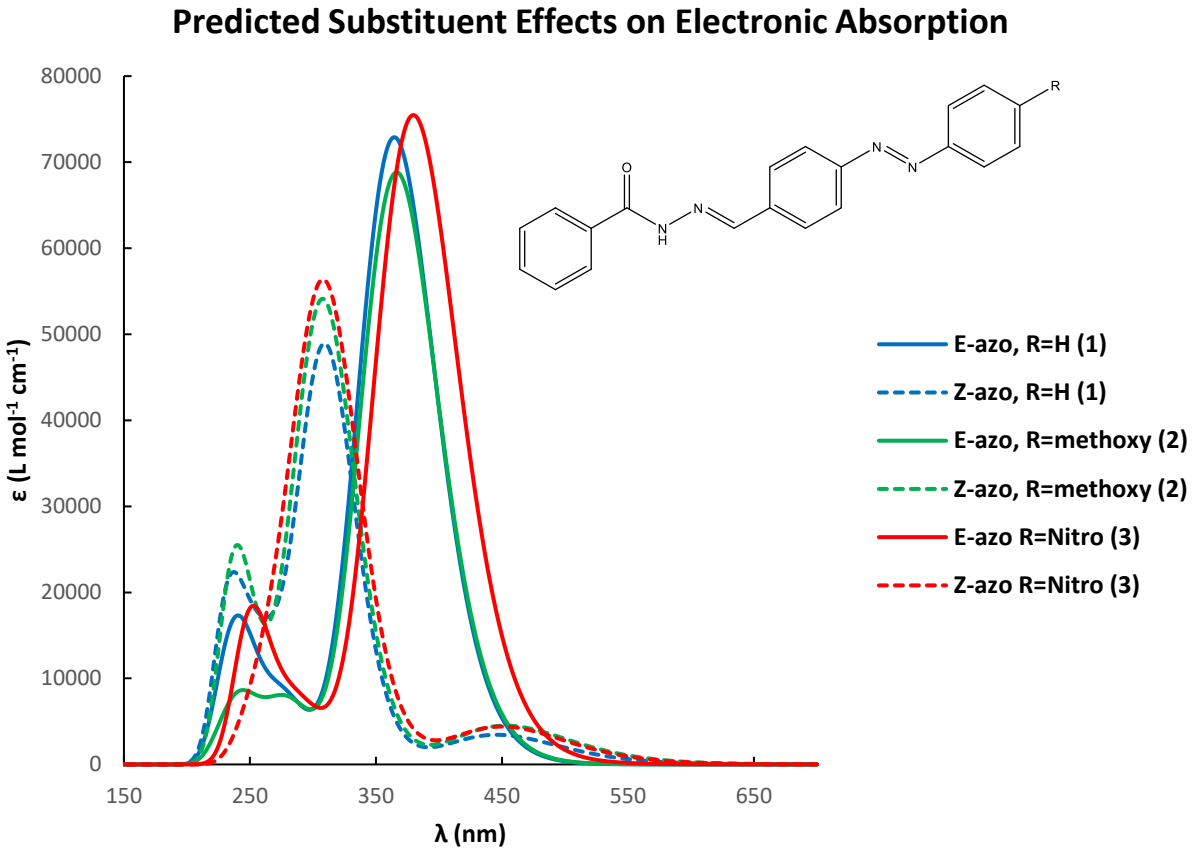

Figure 4-3 Calculated electron absorption spectra from the first 10 predicted electronic transitions for the *Z*-azo and *E*-azo isomers of the compounds presented in Figure 4-2.

Still, our collaborators were surprised that no isomerization was observed across the acylhydrazone motif at the excitation wavelength of 365 nm. To explain this, electron density difference surfaces were computed for the *E*-azo and *Z*-azo isomers of **2** between the ground and first excited state for both the photostable *Z*-azo isomer and the *E*-azo isomer. **2** was selected for modelling as an electron donating group promotes the Baeyer-Mills condensation[150] reaction in the synthesis of the azobenzene building dock used. For the *E*-azo

isomer, the π to π* transition at 365 nm was modeled. As shown in Figure 4-4a, the difference in electron density between the ground and excited state is greatest, and increased (red), over the azo motif. Upon visual inspection, a node is observed between the azo nitrogen atoms, consistent with the excitation to a π* type orbital. The occupation of such an antibonding orbital would correspond to a decrease in the bond order over the azo, facilitating the isomerization to the photostable *Z*-azo isomer. The absence of a significant difference in electron density over the acylhydrazone in Figure 4-4a is consistent with the lack of experimentally observed isomerization at that motif. For the *Z*-azo isomer, the n to π* transition at 450 nm was modeled (Figure 4-4b). The difference in electron density between the ground and excited states is well-localized to the azo motif. Zhu and co-workers were unable to experimentally test the *Z*-to-*E* isomerization photochemically.

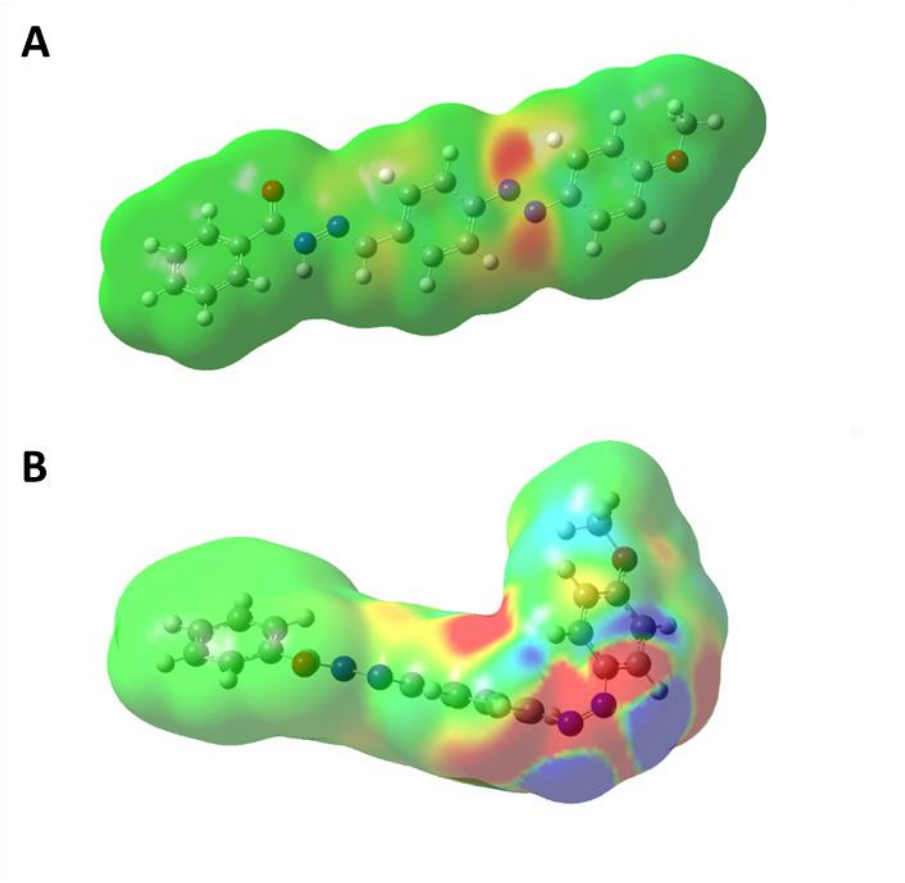

Figure 4-4 Differential electron density surfaces for the methoxy substituted photoswitch in Figure 4-2. The difference is taken as the density in the excited state minus the density in the ground state. Decreased electron density in the ground state is indicated by blue coloring, while increased density in the ground state is indicated by red coloring.

Conclusions

While our computational findings rationalize the experimental observations of our collaborators, the significance of this work is that it affords a framework with which to computationally screen conceived photoswitches. These TD-DFT simulations could precede synthetic efforts and prioritize experimentally testing for only those leads that are predicted to photoisomerize as desired under the limitations of biologically compatible conditions. While the

level of theory used herein was chosen based upon replicating experimental findings for related compounds, an exhaustive benchmarking of various functionals and basis sets could be performed on these and other photoswitchable systems with the aim of reducing the computational cost. In reducing the computational cost, this approach could become more accessible to the synthetic chemistry community for affordable incorporation into design workflows.

Chapter 5: Predicting Baeyer-Mills Reaction Outcomes with Calculated Oxidation Potentials

Introduction

The Baeyer-Mills reaction is a common and important route to the diazene functionality as found in azobenzenes involving the condensation of an arylamine with a nitrosoarene.[150, 151] However, Baeyer-Mills reactions do not only result in the desired diazene moiety. Using nitrosobenzene as a starting material, the formation of azoxybenzene is observed as a side or major product.

Tombari and co-workers experimentally explored the dependence of azoxybenzene formation through substituent screening as depicted in Figure 5-1 below, where the substituted aniline was electronically modulated by varied electron withdrawing and donating groups.[2] They hypothesized that this modulation would result in systematically varied product of the undesired azoxybenzene product.

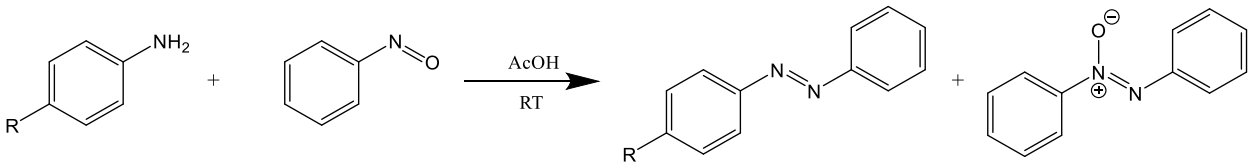

Figure 5-1 The Baeyer-Mills reaction, leading to the formation of substituted azobenzenes, as well as azoxybenze as an undesired side product.

As indicated in Figure 5-1, the azoxybenzene side product was not substituted in any of the experimental conditions tested. This was confirmed by deuterium labeling the aniline derivative.[2] Additionally, the reaction of N,N-dimethylaniline, which lacks exchangeable protons required for the condensation reaction with nitrosobenzene, led to formation of azoxybenzene in 52% yield.[2] Based on these findings and collaboratively with Tombari and co-workers, we hypothesized that the oxidation potential of the supplied aniline derivative was an important

factor in predicting the experimental outcome, specifically for the production of the undesired azoxybenzene side product. Because the reaction is conducted in neat acetic acid, the experimental determination of the oxidation potential using cyclic voltammetry is not feasible, owing to electrode degradation. To this end, we computed oxidation potentials for substituted aniline and indole derivatives, with the aim of correlating azoxybenzene product to the predicted oxidation potential in acetic acid. The work presented in this chapter has been previously published,[2] and the associated text and content is used with permission.

Computational Methods

Single electron oxidation potentials for phenol and aniline derivatives using implicit solvation models have been reported previously in the literature[152] using the B3LYP[38, 39, 61]//6-311++G(2d,2p)[41] level of theory and with solvation free energies computed for gas phase geometries using the Solvation Model based on Density (SMD).[153] Summarily, neutral reactant geometries were constructed in Avogadro[19] and optimized using the MMFF-94 force field.[20] For computational speed in trend analysis, B3LYP/6-31+G(d,p) was used for both neutral and radical cation geometry optimization with default integration grids and convergence criteria in Gaussian 16 A.03.[37] Frequency analyses were performed to verify the optimized geometries as minima by the absence of imaginary frequencies and to compute the gas phase free energy of each aniline derivative. Single point calculations were then performed for these geometries using B3LYP/6-31+G(d,p) with an implicit SMD solvation model of water or acetic acid to compute the solvation free energy in each medium. Using a traditional thermodynamic cycle to apply solvation corrections, the oxidation potential for each aniline or indole derivative of

interest was computed according to the Nernst equation[154] using a Standard Hydrogen Electrode (SHE) voltage of 4.28 V, as done in previous work.[155]

Results and Discussion

To benchmark our chosen level of theory against those values computed in the literature[152] in an aqueous continuum, we computed the oxidation potential for the aniline derivatives in Table 5-1. B3LYP/6-311+G(2d,2p) (SMD=water)//B3LYP/6-311++G(2d,2p) has previously been used in the calculation of oxidation potentials and afforded a strong correlation (R^2 = 0.835) between calculated[152] and experimentally determined[152, 156] oxidation potentials in a set of 25 aniline derivatives. As seen in Figure 5-2, excellent correlation to previously computationally oxidation potentials[152] in water are observed when the level of theory is reduced to B3LYP/6-31+G(d,p) (SMD=Water)//B3LYP/6-31+G(d,p). From these correlations, we concluded that this level of theory is appropriate for computing oxidation potentials in acetic acid, for which there are no experimentally determined oxidation potentials.

Table 5-1 Computed oxidation potentials in acetic acid for analines of interest to benchmark previously computed oxidation potentials against a less expensive level of theory.

Compounds	E_{ox} (V vs SHE) Calculated in this work[a]	E_{ox} (V vs SHE) Previously calculated[152,b]
aniline	0.88	1.02
2,6-dimethoxyaniline	0.39	0.47
4-ntitroanline	1.53	1.64
4-aminophenol	0.54	0.49
4-mtethoxyaniline	0.50	0.64
4-methylaniline	0.70	0.84
4-chloroaniline	0.95	1.05
4-cyanoylaniline	1.24	1.35
2,6-dinitro-4-methylaniline	1.82	1.91
2,4-dinitroaniline	2.05	2.20

[a] B3LYP/6-31+G(d,p) (SMD=Water)//B3LYP/6-31+G(d,p)
[b] B3LYP/6-311+G(2d,2p) (SMD=water)//B3LYP/6-311++G(2d,2p)

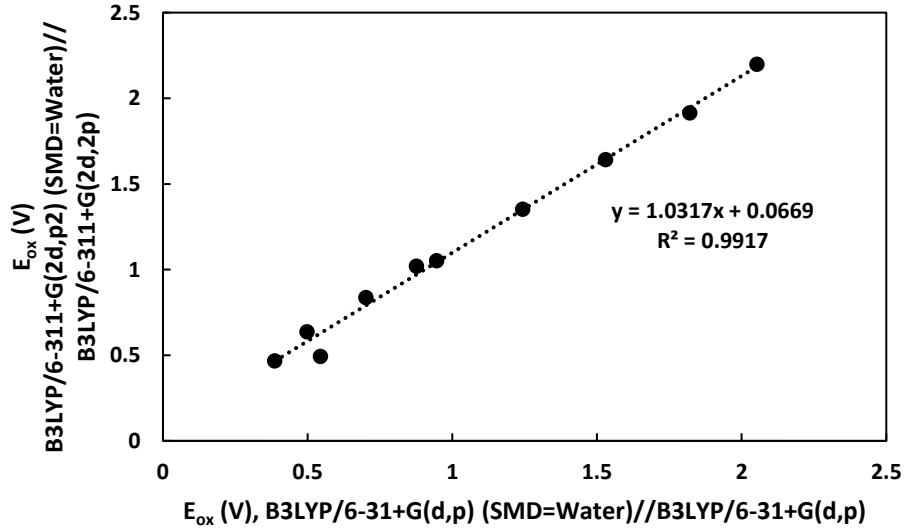

Figure 5-2 Correlation between single electron oxidation potentials computed at B3LYP/6-31+G(d,p) (SMD=Water)//B3LYP/6-31+G(d,p) in this work and B3LYP/6-311++G(2d,2p) (SMD=Water)//B3LYP/6-311+G(2d,2p) previously.[152]

Provided the strong correlation with previously computed values in an implicit water model, we computed the single electron oxidation potentials for the aniline derivatives in Table 5-2, implicitly modeling acetic acid as the solvent used experimentally. The computed oxidation potentials, Hammett σ_p^+ values,[47] and the experimental findings[2] of our collaborators are presented below.

Table 5-2 Computed oxidation potentials in acetic acid for analine derivatives screened in the Baeyer-Mills condensation with nitrosobenzene, along with associated Hammett σ_p^+ values and experimental outcomes.

Compounds	$E_{ox,\ calc}$ [a] (V vs SHE)	σ_p^+ [47]	Azobenzene % Yield[2]	Azoxybenzene % Yield[2]
aniline	1.2	0	≥95	≤5
4-iodoaniline	1.16	0.14	≥95	≤5
methyl 4-aminobenzoate	1.49	0.49	82	6
4-trifluoromethylaniline	1.67	0.61	≥95	≤5
4-cyanoaniline	1.63	0.66	64	7
4-nitiroaniline	1.9	0.79	19	≤5
4-methylaniline	0.97	−0.31	95	≤5
4-methoxyaniline	0.77	−0.78	≥95	5
4-aminophenol	0.83	−0.92	34	69
4-aminoaniline	0.38	−1.30	28	43
4-dimethylaminoaniline	0.18	−1.70	34	45
2-aminoaniline	0.73	N/A	8	91
2,6-dimethoxyaniline	0.77	N/A	46	35
t-butyl (4-aminophenyl)carbamate	0.8	N/A	87	6
2-methoxyaniline	0.93	N/A	82	13
t-butyl (2-aminophenyl)carbamate	1.04	N/A	85	12
2-ethylaniline	1.08	N/A	77	20
2-bromoaniline	1.46	N/A	49	8
2,6-difluoroaniline	1.55	N/A	12	≤5
2-nitroaniline	1.88	N/A	≤5	≤5

[a] B3LYP/6-31+G(d,p) (SMD=Water)//B3LYP/6-31+G(d,p)

The yield of both azobenzene and azoxybenzene were examined visually as a function of the oxidation potential (Figure 5-3). No quantitative relationship (linear or otherwise) between the predicted oxidation potential and either product's yield is apparent, but inferences may be made to inform further investigation and to qualitatively predict the formation of the desired product. First, the proposed mechanism of azoxybenzene formation first includes a reduction of nitrosobenzene, presumably by the single electron oxidation of the accompanying aniline derivative. The results in Figure 5-3 would suggest that aniline derivatives with oxidation potentials higher than ca. 0.9 V cannot be oxidized by nitrosobenzene, preventing the formation of azoxybenzene in any appreciable yield. More electron-rich aniline derivatives with more negative oxidation potentials, such as amino-substituted anilines, are required. Additionally, electron deficient anilines, such as the nitroanilines sampled here, produce poor yields of the desired azobenzene product. This is likely owed to the reduced nucleophilicity of the amino group required in the initial $N_{aniline}$–$N_{nitrosobenzene}$ formation event.

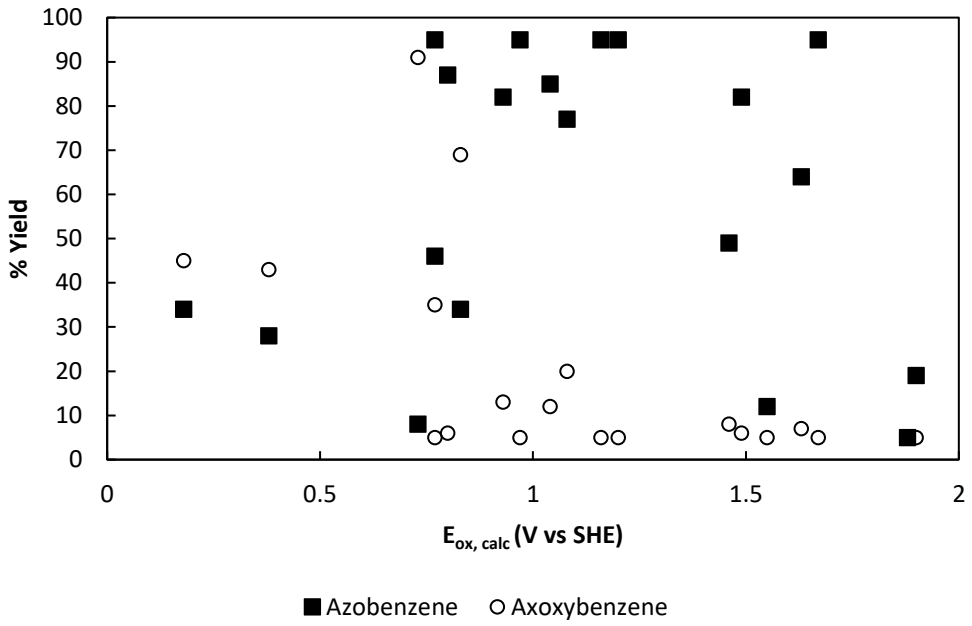

Figure 5-3 Azobenzene and azoxybenzene yields versus computed oxidation potential for compounds listed in Table 5-2.

Hammett σ_p^+ values were also compared against our computed oxidation potentials as a parameter more commonly used by synthetic chemists for rationalizing observed reactivity in conjugated systems owing to electron donating and withdrawing groups. Figure 5-4 shows that excellent correlation exists between the computed oxidation potentials and σ_p^+ values. This finding matches our intuition as strongly electron donating or withdrawing groups should reduce or increase the oxidation potential, respectively. This is also in agreement with previously published findings wherein single electron oxidation potentials in water were strongly correlated with σ_p^+ values.[157] These correlations are useful and compliment the utility of Hammett constants that is commonplace in substituent screening.

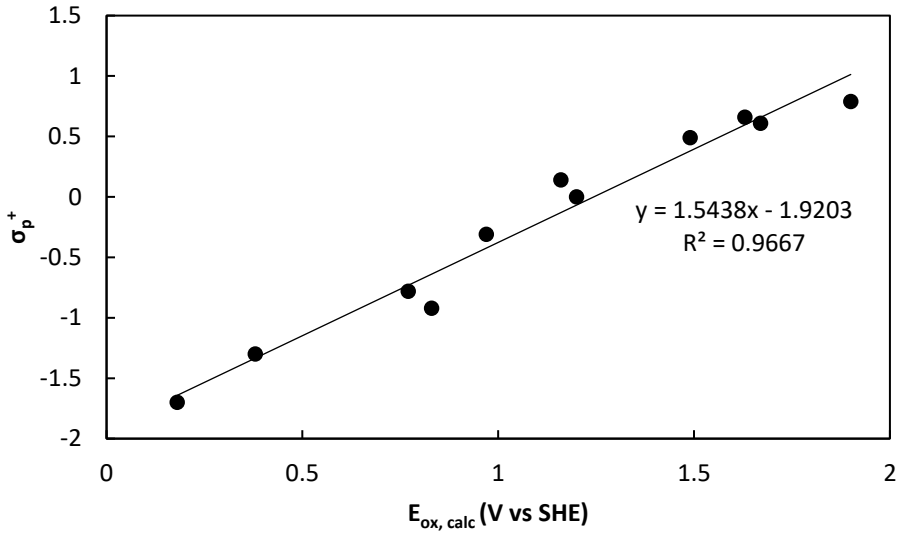

Figure 5-4 Correlation between computed oxidation potentials in acetic acid and σ_p^+ values presented in Table 5-2.

While Hammett constants are widely used and easily interpreted, they are not available for all substituents and not applicable to compounds with core structures that are more complex than a benzene ring. To that end, our collaborators also explored the N1-substituted 5-aminoindoles in Table 5-3.

Table 5-3 Computed oxidation potentials in acetic acid for aniline derivatives screened in the Baeyer-Mills condensation with nitrosobenzene, along with associated Hammett σ_p^+ values and experimental outcomes.

Compounds	$E_{ox,\ calc}$ (V) [a]	σ_p^+	Azoheteroarene % Yield[2]	Azoxybenzene % Yield[2]
1H-indol-5-amine	0.7	N/A	6	42
t-butyl 5-amino-1H-indole-1-carboxylate	0.89	N/A	58	18
1-acetyl-1H-indol-5-amine	0.99	N/A	72	17
1-tosyl-1H-indol-5-amine	1.01	N/A	63	10

[a] B3LYP/6-31+G(d,p) (SMD=Water)//B3LYP/6-31+G(d,p)

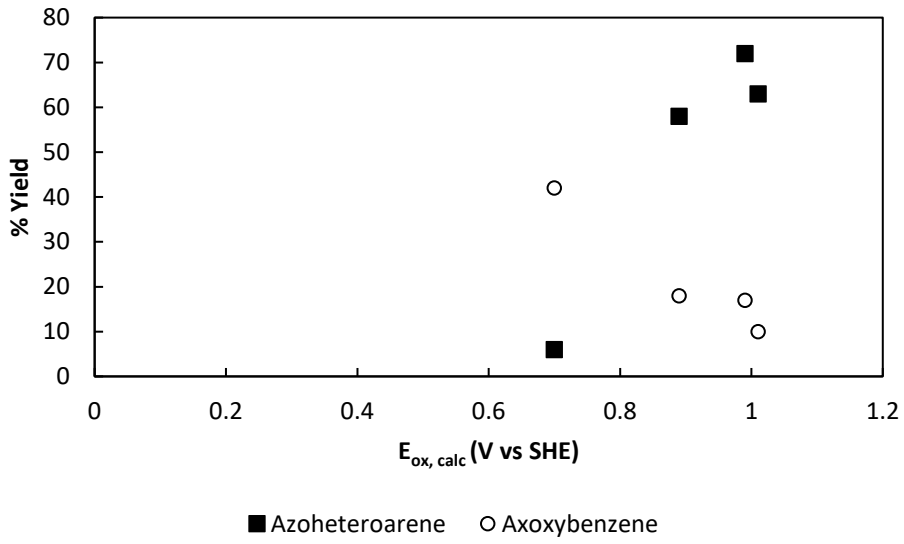

Figure 5-5 Azoheteroarene and azoxybenzene yields versus computed oxidation potential for compounds listed in Table 5-3.

As seen in Figure 5-5, reduced azoxybenzene formation was again observed for those indoles with oxidation potentials greater than ca. 0.9, with respectable azoheteroarene yields observed for those same compounds. While the dataset is limited with only four compounds, these data do suggest broader utility for utilizing computed oxidation potentials to predict Baeyer-Mills reaction outcomes where Hammett constants are not available.

Of note for this work, protonation states were not considered in this study as the aniline derivative must have a deprotonated primary amine to undergo the desired Baeyer-Mills condensation reaction. However, we acknowledge that both the nucleophilicity of the deprotonated amine involved in N–N bond formation in the Baeyer-Mills reaction and the acidity of its conjugate acid are likely to influence reaction outcomes. Additionally, the protonation state of the primary amine will certainly modulate the oxidation potential of the aniline derivative. This collection of factors is deserving of further investigation and will require

computational benchmarking and further mining of the literature for comparative experimental data.

Conclusions

Oxidation potentials computed using Density Functional Theory with an implicit solvation model afforded a useful metric to predict the formation of both azobenzene in the Baeyer-Mills condensation of aniline or indole derivatives with nitrosobenzene, as well as predict the formation of azoxybenzene as a side product. A slightly reduced basis set still resulted in excellent correlation compared to the levels of theory previously benchmarked against experimental values. This reduced the computational cost of the calculations while still allowing trend analysis alongside experimental data, making this a useful approach for the predominantly synthetic chemist where Hammett parameters are not available or applicable.

Chapter 6: Samarium Promoted Rearrangement of Vinyl Aziridines

Introduction

From amino acids to nucleic acids, heterocycles are prevalent in biology, with a believed 85% of biologically active chemical species containing a heterocycle.[158] For this reason, the synthesis of functionalized heterocycles is an area of active interest in medicinal chemistry for both therapeutic and diagnostic applications. Among nitrogen-containing heterocycles, the 5-member class of pyrrolines is of interest, both as a terminal synthetic motif and as they can be readily transformed into pyrrolidines or pyrroles.[159]

Synthetic routes to pyrrolines have been recently reviewed,[160] but of particular interest to our synthetic collaborators, David Olson and colleagues, was the ring expansion of a vinyl aziridine (Figure 6-1). Such ring expansions have been catalyzed by Lewis acids such as $Cu(hfacac)_2$, with the aziridine commonly being phthalimide-protected.[159] While Lewis acid catalysts have competently facilitated this transformation, increasingly specific catalysts are needed in multistep syntheses to conserve yields and improve stereochemical outcomes. Cheung and co-workers, while screening available Lewis acids in the Olson lab, screened samarium (II) iodide and found quantitative yields for a small panel of phthalimide-protected vinyl aziridines to afford expanded heterocycles. Samarium (II) iodide is known to catalyze cross-couplings through radical intermediates in carbonyl compounds,[161] and its use in the synthesis of nitrogen-containing heterocycles is known,[162] although that work did not include aziridine ring expansion and still promoted C–C coupling. Cheung's experimental findings prompted our groups to ask whether the transformation in Figure 6-1 occurred by traditional Lewis acid/base chemistry or by a radical mechanism. To understand this and to compliment

any further spectroscopic or radical clock studies, we performed stationary point analyses using DFT for the reaction in Figure 6-1 to propose a reasonable mechanism for the observed transformation under the given experimental conditions.

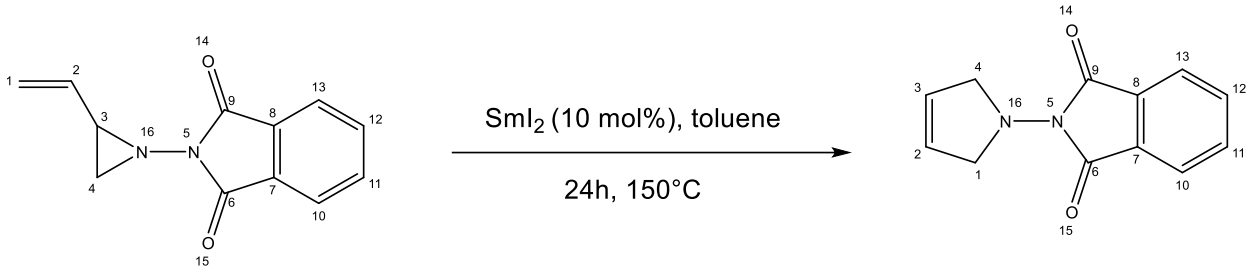

Figure 6-1 Samarium (II) iodide catalyzed ring expansion of a phthalimide-protected vinyl aziridine.

Computational Methods

The prototypical vinyl aziridine of interest in the reaction complex with SmI_2 in Figure 6-1 was first constructed in Avogadro[19] and optimized using MMFF-94.[20] Guesses for the subsequent intermediates as shown in Figure 6-2 were then also made, as well as the associated transition state structure for each event. The gas phase geometry for each minimum and saddle point was optimized in Gaussian A.03[37] at PBE0[163]/def2-SVP[164], using core potentials on both samarium and iodine, and vibrational analyses were performed. This was performed for both the quintet and septet spin states. Minima were confirmed as true minima by the absence of imaginary frequencies and transition state structures were confirmed as first order saddle points by the presence of a single imaginary frequency corresponding to the bond breaking or forming motion being modeled. Intrinsic reaction coordinate calculations were performed, confirming that the discovered transition state structures indeed connected the two adjoining minima. Solvated,[153] single point energies with empirical dispersion[86, 147]

corrections were then computed on both spin surfaces at the PBE0-D3BJ/def2-TZVP (ECP = Sm, I; SMD = toluene)//PBE0/def2-SVP (ECP = Sm, I) level of theory. Spin density surfaces on the septet surface at the lower level of theory were generated in GaussView, and Natural Bond Orbital[165] analysis was also performed.

Results and Discussion

To start, we proposed the mechanism below (Figure 6-2) to explain the SmI_2-catalyzed ring expansion of the protected vinyl aziridine of interest (Figure 6-1). This was based on known radical chemistry initiated by samarium (II) iodide in carbonyl-containing compounds.

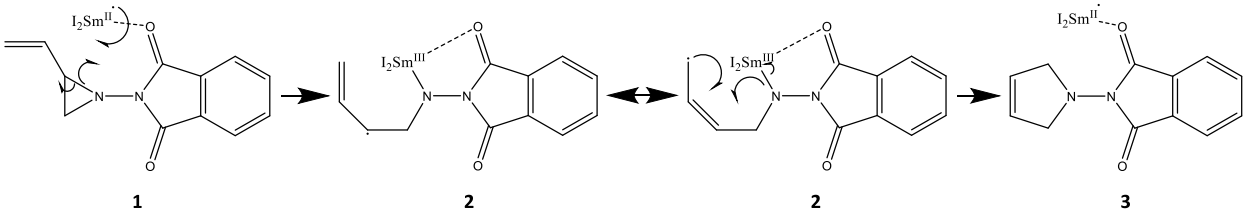

Figure 6-2 Proposed mechanism for the ring expansion of a phthalimide-protected vinyl aziridine as catalyzed by SmI_2.

From the stationary point analysis, the free energy diagram corresponding to the mechanism in Figure 6-2 was constructed (Figure 6-3). From this, reactivity was predicted to occur principally on the septet spin surface, with both spin states being comparable in energy for intermediate **2**. This is consistent with Hund's Rule of Maximum Multiplicity and given iodine is a weak field ligand. Afforded this observation, the free energy change along the reaction coordinate on the septet surface was replotted at the PBE0-D3(BJ)/def2-TZVP (ECP = Sm, I; SMD = toluene)//PBE0/def2-SVP (ECP = Sm, I) level of theory to account for solvation and dispersion effects. From these findings and given quantitative yields were achieved in 24 hours

at 150 °C, we concluded that the proposed mechanism was consistent with the experimental conditions given the computed free energy barriers for both events.

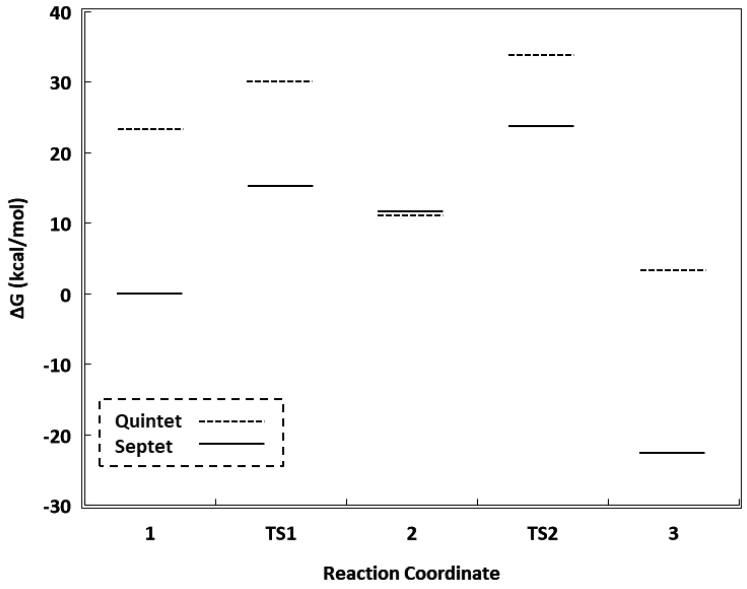

Figure 6-3 Gas phase free energy diagram for the mechanism shown in Figure 6-2 on both the quintet and septet surfaces at the PBE0/def2-SVP (ECP = Sm and I) level of theory.

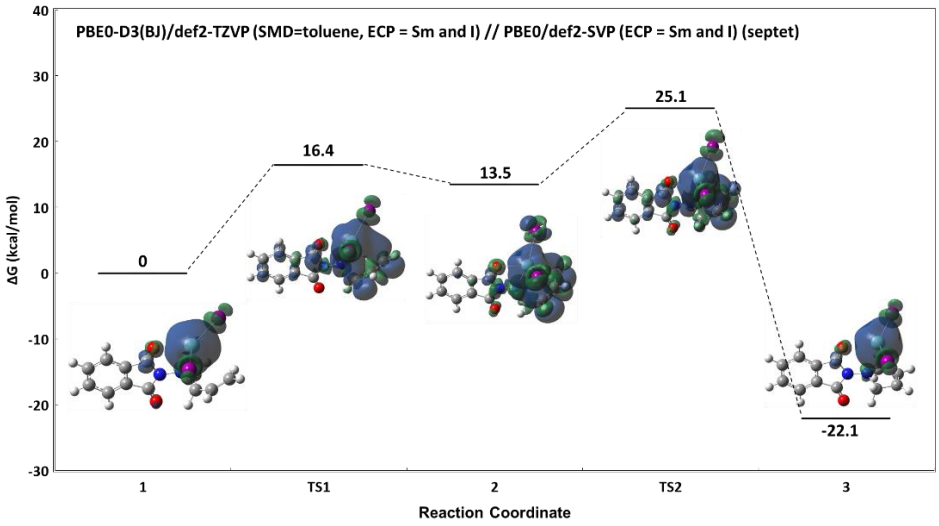

Figure 6-4 Solvated and dispersion corrected free energy diagram for the ring expansion mechanism proposed in Figure 6-2. Spin density surfaces for each structure show the evolution of spin and radical localization across the reaction coordinate.

Further consideration was then given to spin density localization visually as shown in Figure 6-4, as well as plotting the condensed spin densities on several key atoms across the reaction coordinate in Figure 6-5. Specifically, the samarium atom was followed given its catalytic role. Carbons 1, 3, and 9 were also considered as localized radical character would exists on those atoms provided the Lewis structures as drawn in the mechanism in Figure 6-2. Visually, the accumulation of spin density (and radical character) is predicted over the vinyl carbons as ring expansion progress through intermediate **2**, with spin density returning principally to the samarium atom and the proximal carbonyl carbon in product complex **3**. In the transition state structures, spin delocalization is predicted in the conjugated phthalimide protecting group, consistent with the belief that the formation of a radical in the aziridine fragment would be stabilized by the presence of the protecting group. Additionally, the condensed spin densities along the reaction coordinate are in quantitative agreement with our visual inspection. As seen in Figure 6-5, spin density decreases on samarium progressing from reactant complex **1** to intermediate **2**, but then rises back to approximately the reactant-complex density in the product complex **3**. Accordingly, spin density rises on C1 and C3 progressing to intermediate **2**, and then declines as ring expansion completes. While there is visible spin density on C9 in Figure 6-4, there is relatively little spin density predicted on C9 over the course of the reaction coordinate. We attribute this in part to the delocalization into the aromatic phthalimide substructure. Additionally, a radical on C9 would require a one electron transfer between the reactant fragment and samarium, which would yield an unfounded samarium (IV) subsequently upon Sm–C3 bond formation. Samarium (II) and samarium (III) are the only known charged oxidation states of samarium. Figure 6-5 is consistent with oxidation

states proposed in Figure 6-2. In reactant complex **1**, samarium (II) would be assigned 6 unpaired electrons, consistent with the predicted total spin density. In intermediate **2**, samarium (III) would have five unpaired electrons, with a delocalized radical shared over the allylic fragment. Our computational finding is consistent with a samarium (III). As ring expansion completes in product complex **3**, spin density should return principally to the samarium atom, which is what was predicted computationally.

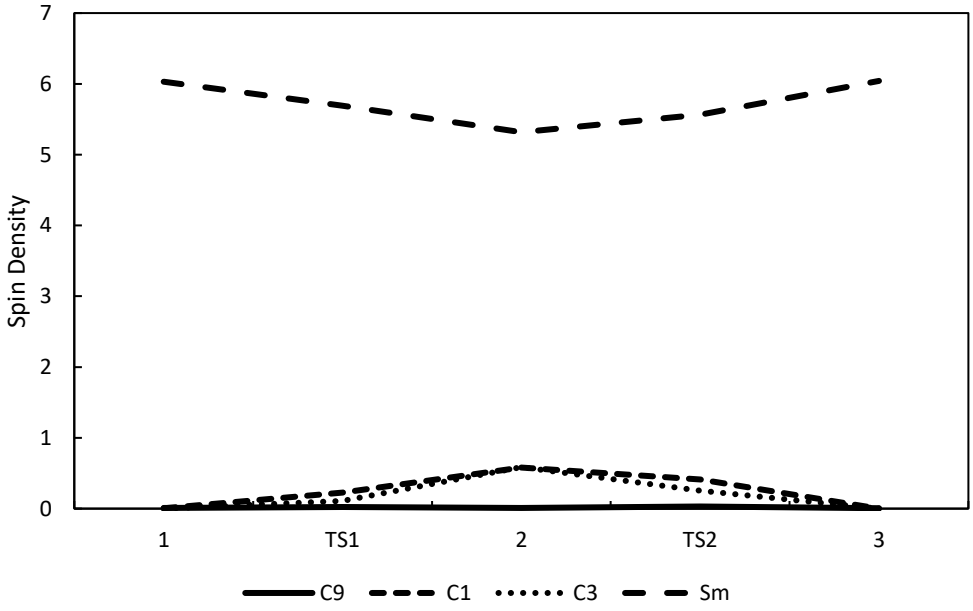

Figure 6-5 Condensed spin densities for select atoms in the ring expansion mechanism proposed in Figure 6-2.

Lastly, we examined the Wiberg Bond Indices originating from NBO[165] analysis on the septet surface at the PBE0/def2-SVP level of theory. As seen in Figure 6-6, at least a partial covalent bond with an Wiberg Bond Index of > 0.5 is predicted between Sm and the aziridinyl nitrogen. Also, the interactions between Sm, C1, C2, and C3, coupled with the planar orientation of the attached hydrogen atoms indicates an η-3 orientation of the allylic

substructure donating to the samarium (III) ion. Given the positive bond indices between the allylic fragment carbon atoms and samarium, we might predict that intermediate **2** is a quintet, owing to the spin pairing of the electron density from the allylic fragment and the density being contributed to the covalent interaction from the samarium atom. Based solely on computational predictions in the absence of spectroscopic data, further studies with all electron basis sets for samarium and iodine and/or coupled cluster single point calculations on DFT geometries could be conducted to confidently assign the multiplicity of intermediate **2**.

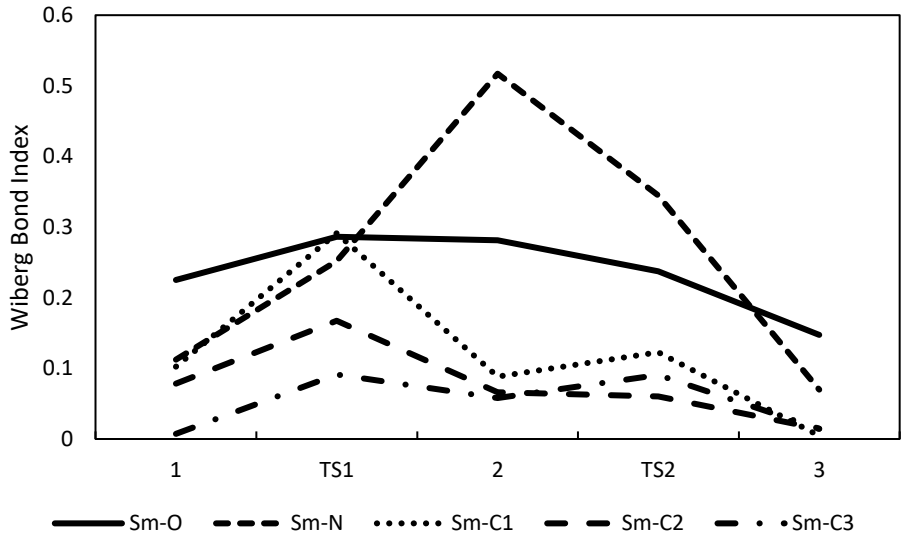

Figure 6-6 Wiberg Bond Indices for select bonds across the reaction coordinate, showing a significant and increasing Sm–N interaction leading into intermediate **2** and an overall decline in bonding character in the catalyst-product complex for all annotated interactions.

Conclusions

Summarily, we have proposed and computationally explored a mechanism for the samarium-mediated ring expansion of vinyl aziridines that is consistent with the currently available experimental findings of our collaborators. Based on local condensed spin density

differences at each evaluated stationary point, we predict the ring expansion of this, and other structurally similar, phthalimide-protect vinyl aziridines to occur according to a single electron transfer mechanism that is consistent with samarium (II) iodide's known reactivity in carbonyl containing compounds. The phthalimide protecting group allows for spin density delocalization across the reaction coordinate and, in conjunction with the aziridinyl nitrogen's covalent interactions with the samarium atom, orients the samarium such that ring expansion may occur via a radical pathway. Additional experimental work, specifically the design and execution of a radical clock experiment, should be performed to confirm a radical mechanism. Our computational findings along with the preliminary experimental work of our collaborators suggests this approach is a promising synthetic route to functionalized pyrrolidines and other heterocycles with, potentially, interesting biological applications.

Citations

(1) Zhu, J. S.; Larach, J. M.; Tombari, R. J.; Gingrich, P. W.; Bode, S. R.; Tuck, J. R.; Warren, H. T.; Son, J.-H.; Duim, W. C.; Fettinger, J. C. A Redox Isomerization Strategy for Accessing Modular Azobenzene Photoswitches with Near Quantitative Bidirectional Photoconversion. *Organic letters* **2019**, *21* (21), 8765-8770.
(2) Tombari, R. J.; Tuck, J. R.; Yardeny, N.; Gingrich, P. W.; Tantillo, D. J.; Olson, D. E. Calculated oxidation potentials predict reactivity in Baeyer–Mills reactions. *Organic & Biomolecular Chemistry* **2021**, *19* (35), 7575-7580.
(3) Meunier, B.; de Visser, S. P.; Shaik, S. Mechanism of Oxidation Reactions Catalyzed by Cytochrome P450 Enzymes. *Chemical Reviews* **2004**, *104* (9), 3947-3980. DOI: 10.1021/cr020443g.
(4) Shaik, S.; Cohen, S.; Wang, Y.; Chen, H.; Kumar, D.; Thiel, W. P450 Enzymes: Their Structure, Reactivity, and Selectivity—Modeled by QM/MM Calculations. *Chemical Reviews* **2010**, *110* (2), 949-1017. DOI: 10.1021/cr900121s.
(5) Vaz, A. D.; McGinnity, D. F.; Coon, M. J. Epoxidation of olefins by cytochrome P450: evidence from site-specific mutagenesis for hydroperoxo-iron as an electrophilic oxidant. *Proceedings of the National Academy of Sciences* **1998**, *95* (7), 3555-3560.
(6) Ehrenberg, L.; Hussain, S. Genetic toxicity of some important epoxides. *Mutation Research/Reviews in Genetic Toxicology* **1981**, *86* (1), 1-113.
(7) Massey, T. E.; Stewart, R. K.; Daniels, J. M.; Liu, L. Biochemical and molecular aspects of mammalian susceptibility to aflatoxin B1 carcinogenicity. *Proceedings of the Society for Experimental Biology and Medicine* **1995**, *208* (3), 213-227.
(8) Santos, R.; Hritz, J.; Oostenbrink, C. Role of Water in Molecular Docking Simulations of Cytochrome P450 2D6. *Journal of Chemical Information and Modeling* **2010**, *50* (1), 146-154. DOI: 10.1021/ci900293e.
(9) Chen, Y.-C. Beware of docking! *Trends in pharmacological sciences* **2015**, *36* (2), 78-95.
(10) Dubey, K. D.; Shaik, S. Cytochrome P450—The Wonderful Nanomachine Revealed through Dynamic Simulations of the Catalytic Cycle. *Accounts of Chemical Research* **2019**. DOI: 10.1021/acs.accounts.8b00467.
(11) Hirao, H.; Kumar, D.; Thiel, W.; Shaik, S. Two states and two more in the mechanisms of hydroxylation and epoxidation by cytochrome P450. *Journal of the American Chemical Society* **2005**, *127* (37), 13007-13018.
(12) Shaik, S.; Hirao, H.; Kumar, D. Reactivity of High-Valent Iron–Oxo Species in Enzymes and Synthetic Reagents: A Tale of Many States. *Accounts of Chemical Research* **2007**, *40* (7), 532-542. DOI: 10.1021/ar600042c. Shaik, S.; Hirao, H.; Kumar, D. Reactivity patterns of cytochrome P450 enzymes: multifunctionality of the active species, and the two states–two oxidants conundrum. *Natural Product Reports* **2007**, *24* (3), 533-552, 10.1039/B604192M. DOI: 10.1039/B604192M.
(13) Zhang, J.; Ji, L.; Liu, W. In silico prediction of cytochrome P450-mediated biotransformations of xenobiotics: a case study of epoxidation. *Chemical Research in Toxicology* **2015**, *28* (8), 1522-1531.
(14) Bauer, C. A.; Hansen, A.; Grimme, S. The Fractional Occupation Number Weighted Density as a Versatile Analysis Tool for Molecules with a Complicated Electronic Structure. *Chemistry – A European Journal* **2017**, *23* (25), 6150-6164. DOI: https://doi.org/10.1002/chem.201604682.
(15) Pino-Rios, R.; Yañez, O.; Inostroza, D.; Ruiz, L.; Cardenas, C.; Fuentealba, P.; Tiznado, W. Proposal of a simple and effective local reactivity descriptor through a topological analysis of an orbital-weighted fukui function. *Journal of Computational Chemistry* **2017**, *38* (8), 481-488. DOI: https://doi.org/10.1002/jcc.24699.

(16) Wilkinson, M. D.; Dumontier, M.; Aalbersberg, I. J.; Appleton, G.; Axton, M.; Baak, A.; Blomberg, N.; Boiten, J.-W.; da Silva Santos, L. B.; Bourne, P. E. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific data* **2016**, *3* (1), 1-9.
(17) Kells, P. M.; Ouellet, H.; Santos-Aberturas, J.; Aparicio, J. F.; Podust, L. M. Structure of cytochrome P450 PimD suggests epoxidation of the polyene macrolide pimaricin occurs via a hydroperoxoferric intermediate. *Chem Biol* **2010**, *17* (8), 841-851. DOI: 10.1016/j.chembiol.2010.05.026 PubMed.
(18) Gingrich, P. W.; Siegel, J. B.; Tantillo, D. J. Assessing Alkene Reactivity toward Cytochrome P450-Mediated Epoxidation through Localized Descriptors and Regression Modeling. *Journal of Chemical Information and Modeling* **2022**, *62* (8), 1979-1987. DOI: 10.1021/acs.jcim.1c01567.
(19) Hanwell, M. D.; Curtis, D. E.; Lonie, D. C.; Vandermeersch, T.; Zurek, E.; Hutchison, G. R. Avogadro: an advanced semantic chemical editor, visualization, and analysis platform. *Journal of cheminformatics* **2012**, *4* (1), 17.
(20) Halgren, T. A. Merck molecular force field. III. Molecular geometries and vibrational frequencies for MMFF94. *Journal of computational chemistry* **1996**, *17* (5-6), 553-586.
(21) Grimme, S. Exploration of Chemical Compound, Conformer, and Reaction Space with Meta-Dynamics Simulations Based on Tight-Binding Quantum Chemical Calculations. *Journal of Chemical Theory and Computation* **2019**, *15* (5), 2847-2862. DOI: 10.1021/acs.jctc.9b00143.
(22) Bannwarth, C.; Caldeweyher, E.; Ehlert, S.; Hansen, A.; Pracht, P.; Seibert, J.; Spicher, S.; Grimme, S. Extended tight-binding quantum chemistry methods. *WIREs Computational Molecular Science* **2021**, *11* (2), e1493. DOI: https://doi.org/10.1002/wcms.1493.
(23) Bannwarth, C.; Ehlert, S.; Grimme, S. GFN2-xTB—An Accurate and Broadly Parametrized Self-Consistent Tight-Binding Quantum Chemical Method with Multipole Electrostatics and Density-Dependent Dispersion Contributions. *Journal of Chemical Theory and Computation* **2019**, *15* (3), 1652-1671. DOI: 10.1021/acs.jctc.8b01176.
(24) Grimme, S.; Bannwarth, C.; Shushkov, P. A Robust and Accurate Tight-Binding Quantum Chemical Method for Structures, Vibrational Frequencies, and Noncovalent Interactions of Large Molecular Systems Parametrized for All spd-Block Elements (Z = 1–86). *Journal of Chemical Theory and Computation* **2017**, *13* (5), 1989-2009. DOI: 10.1021/acs.jctc.7b00118.
(25) Spicher, S.; Grimme, S. Robust Atomistic Modeling of Materials, Organometallic, and Biochemical Systems. *Angewandte Chemie International Edition* **2020**, *59* (36), 15665-15673. DOI: https://doi.org/10.1002/anie.202004239.
(26) Schaftenaar, G.; Noordik, J. H. Molden: a pre- and post-processing program for molecular and electronic structures*. *Journal of Computer-Aided Molecular Design* **2000**, *14* (2), 123-134. DOI: 10.1023/A:1008193805436.
(27) Geerlings, P.; Chamorro, E.; Chattaraj, P. K.; De Proft, F.; Gázquez, J. L.; Liu, S.; Morell, C.; Toro-Labbé, A.; Vela, A.; Ayers, P. Conceptual density functional theory: status, prospects, issues. *Theoretical Chemistry Accounts* **2020**, *139* (2), 36. DOI: 10.1007/s00214-020-2546-7.
(28) Lu, T.; Chen, F. Multiwfn: a multifunctional wavefunction analyzer. *Journal of computational chemistry* **2012**, *33* (5), 580-592.
(29) Hirshfeld, F. L. Bonded-atom fragments for describing molecular charge densities. *Theoretica chimica acta* **1977**, *44* (2), 129-138. DOI: 10.1007/BF00549096.
(30) Mulliken, R. S. Electronic population analysis on LCAO–MO molecular wave functions. I. *The Journal of Chemical Physics* **1955**, *23* (10), 1833-1840.
(31) Fukui, K.; Yonezawa, T.; Shingu, H. A molecular orbital theory of reactivity in aromatic hydrocarbons. *The Journal of Chemical Physics* **1952**, *20* (4), 722-725. Fukui, K.; Yonezawa, T.; Nagata, C.; Shingu, H. Molecular orbital theory of orientation in aromatic, heteroaromatic, and other conjugated molecules. *The Journal of Chemical Physics* **1954**, *22* (8), 1433-1442.

(32) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V. Scikit-learn: Machine learning in Python. *the Journal of machine Learning research* **2011**, *12*, 2825-2830.
(33) *pandas-dev/pandas: Pandas 1.3.4*; 2021. (accessed.
(34) Seabold, S.; Perktold, J. statsmodels: Econometric and statistical modeling with python. In 9th Python in Science Conference, 2010.
(35) Salmerón, R.; García, C.; García, J. Variance inflation factor and condition number in multiple linear regression. *Journal of Statistical Computation and Simulation* **2018**, *88* (12), 2365-2384.
(36) Shapiro, S. S.; Wilk, M. B. An analysis of variance test for normality (complete samples). *Biometrika* **1965**, *52* (3/4), 591-611.
(37) *Gaussian 16 Rev. A.03*; Wallingford, CT, 2016. (accessed.
(38) Becke, A. D. A new mixing of Hartree–Fock and local density-functional theories. *The Journal of chemical physics* **1993**, *98* (2), 1372-1377.
(39) Becke, A. D. Becke's three parameter hybrid method using the LYP correlation functional. *J. Chem. Phys* **1993**, *98* (492), 5648-5652.
(40) Hay, P. J.; Wadt, W. R. Ab initio effective core potentials for molecular calculations. Potentials for the transition metal atoms Sc to Hg. *The Journal of Chemical Physics* **1985**, *82* (1), 270-283. DOI: 10.1063/1.448799.
(41) Francl, M. M.; Pietro, W. J.; Hehre, W. J.; Binkley, J. S.; Gordon, M. S.; DeFrees, D. J.; Pople, J. A. Self-consistent molecular orbital methods. XXIII. A polarization-type basis set for second-row elements. *The Journal of Chemical Physics* **1982**, *77* (7), 3654-3665.
(42) Hay, P. J. Gaussian basis sets for molecular calculations. The representation of 3d orbitals in transition-metal atoms. *The Journal of Chemical Physics* **1977**, *66* (10), 4377-4384. DOI: 10.1063/1.433731.
(43) Schäfer, A.; Huber, C.; Ahlrichs, R. Fully optimized contracted Gaussian basis sets of triple zeta valence quality for atoms Li to Kr. *The Journal of Chemical Physics* **1994**, *100* (8), 5829-5835. DOI: 10.1063/1.467146.
(44) Yang, W.; Mortier, W. J. The use of global and local molecular parameters for the analysis of the gas-phase basicity of amines. *Journal of the American Chemical Society* **1986**, *108* (19), 5708-5711.
(45) Ogliaro, F.; de Visser, S. P.; Shaik, S. The 'push' effect of the thiolate ligand in cytochrome P450: a theoretical gauging. *Journal of Inorganic Biochemistry* **2002**, *91* (4), 554-567. DOI: https://doi.org/10.1016/S0162-0134(02)00437-3. Dawson, J. H.; Holm, R. H.; Trudell, J. R.; Barth, G.; Linder, R. E.; Bunnenberg, E.; Djerassi, C.; Tang, S. C. Magnetic circular dichroism studies. 43. Oxidized cytochrome P-450. Magnetic circular dichroism evidence for thiolate ligation in the substrate-bound form. Implications for the catalytic mechanism. *Journal of the American Chemical Society* **1976**, *98* (12), 3707-3709. DOI: 10.1021/ja00428a054.
(46) Coelho, P. S.; Wang, Z. J.; Ener, M. E.; Baril, S. A.; Kannan, A.; Arnold, F. H.; Brustad, E. M. A serine-substituted P450 catalyzes highly efficient carbene transfer to olefins in vivo. *Nature Chemical Biology* **2013**, *9* (8), 485-487. DOI: 10.1038/nchembio.1278.
(47) Hansch, C.; Leo, A.; Taft, R. W. A survey of Hammett substituent constants and resonance and field parameters. *Chemical Reviews* **1991**, *91* (2), 165-195. DOI: 10.1021/cr00002a004.
(48) Muthukrishnan, R.; Rohini, R. LASSO: A feature selection technique in predictive modeling for machine learning. In *2016 IEEE International Conference on Advances in Computer Applications (ICACA)*, 24-24 Oct. 2016, 2016; pp 18-20. DOI: 10.1109/ICACA.2016.7887916.
(49) Leth, R.; Ercig, B.; Olsen, L.; Jørgensen, F. S. Both Reactivity and Accessibility Are Important in Cytochrome P450 Metabolism: A Combined DFT and MD Study of Fenamic Acids in BM3 Mutants. *Journal of Chemical Information and Modeling* **2019**, *59* (2), 743-753. DOI: 10.1021/acs.jcim.8b00750.

(50) Gorycki, P. D.; Macdonald, T. L. The Oxidation of Tetrasubstituted Alkenes by Cytochrome P450. *Chemical Research in Toxicology* **1994**, *7* (6), 745-751. DOI: 10.1021/tx00042a006.
(51) Selke, M.; Sisemore, M. F.; Valentine, J. S. The Diverse Reactivity of Peroxy Ferric Porphyrin Complexes of Electron-Rich and Electron-Poor Porphyrins. *Journal of the American Chemical Society* **1996**, *118* (8), 2008-2012. DOI: 10.1021/ja953694y.
(52) Yoshioka, S.; Takahashi, S.; Ishimori, K.; Morishima, I. Roles of the axial push effect in cytochrome P450cam studied with the site-directed mutagenesis at the heme proximal site. *Journal of inorganic biochemistry* **2000**, *81* (3), 141-151.
(53) Gelb, M. H.; Malkonen, P.; Sligar, S. G. Cytochrome P450cam catalyzed epoxidation of dehydrocamphor. *Biochemical and Biophysical Research Communications* **1982**, *104* (3), 853-858.
(54) Jin, S.; Bryson, T. A.; Dawson, J. H. Hydroperoxoferric heme intermediate as a second electrophilic oxidant in cytochrome P450-catalyzed reactions. *JBIC Journal of Biological Inorganic Chemistry* **2004**, *9*, 644-653.
(55) Rittle, J.; Green, M. T. Cytochrome P450 Compound I: Capture, Characterization, and C-H Bond Activation Kinetics. *Science* **2010**, *330*, 933-937.
(56) Derat, E.; Kumar, D.; Hirao, H.; Shaik, S. Gauging the Relative Oxidative Powers of Compound I, Ferric-Hydroperoxide, and the Ferric-Hydrogen Peroxide Species of Cytochrome P450 Toward C−H Hydroxylation of a Radical Clock Substrate. *Journal of the American Chemical Society* **2006**, *128* (2), 473-484. DOI: 10.1021/ja056328f.
(57) Ortiz de Montellano, P. R. Hydrocarbon Hydroxylation by Cytochrome P450 Enzymes. *Chemical Reviews* **2010**, *110* (2), 932-948. DOI: 10.1021/cr9002193.
(58) Choe, Y. K.; Nagase, S. Effect of the axial cysteine ligand on the electronic structure and reactivity of high-valent iron (IV) oxo-porphyrins (Compound I): A theoretical study. *Journal of computational chemistry* **2005**, *26* (15), 1600-1611. de Visser, S. P.; Ogliaro, F.; Sharma, P. K.; Shaik, S. What factors affect the regioselectivity of oxidation by cytochrome P450? A DFT study of allylic hydroxylation and double bond epoxidation in a model reaction. *Journal of the American Chemical Society* **2002**, *124* (39), 11809-11826. de Visser, S. P.; Kumar, D.; Cohen, S.; Shacham, R.; Shaik, S. A predictive pattern of computed barriers for C− H hydroxylation by compound I of cytochrome P450. *Journal of the American Chemical Society* **2004**, *126* (27), 8362-8363. Kumar, D.; de Visser, S. P.; Sharma, P. K.; Cohen, S.; Shaik, S. Radical Clock Substrates, Their C− H Hydroxylation Mechanism by Cytochrome P450, and Other Reactivity Patterns: What Does Theory Reveal about the Clocks' Behavior? *Journal of the American Chemical Society* **2004**, *126* (6), 1907-1920.
(59) Guengerich, F. P.; Krauser, J. A.; Johnson, W. W. Rate-Limiting Steps in Oxidations Catalyzed by Rabbit Cytochrome P450 1A2. *Biochemistry* **2004**, *43* (33), 10775-10788. DOI: 10.1021/bi0491393.
(60) Shaik, S.; Kumar, D.; de Visser, S. P.; Altun, A.; Thiel, W. Theoretical perspective on the structure and mechanism of cytochrome P450 enzymes. *Chemical reviews* **2005**, *105* (6), 2279-2328.
(61) Becke, A. D. Density-functional thermochemistry. III. The role of exact exchange. *The Journal of Chemical Physics* **1993**, *98* (7), 5648-5652. DOI: 10.1063/1.464913.
(62) Lee, C.; Yang, W.; Parr, R. G. Development of the Colle-Salvetti correlation-energy formula into a functional of the electron density. *Physical review B* **1988**, *37* (2), 785.
(63) Stephens, P. J.; Devlin, F. J.; Chabalowski, C. F.; Frisch, M. J. Ab initio calculation of vibrational absorption and circular dichroism spectra using density functional force fields. *The Journal of physical chemistry* **1994**, *98* (45), 11623-11627.
(64) Tirado-Rives, J.; Jorgensen, W. L. Performance of B3LYP density functional methods for a large set of organic molecules. *Journal of chemical theory and computation* **2008**, *4* (2), 297-306.
(65) Olsen, L.; Rydberg, P.; Rod, T. H.; Ryde, U. Prediction of activation energies for hydrogen abstraction by cytochrome P450. *Journal of medicinal chemistry* **2006**, *49* (22), 6489-6499.

(66) Zhang, Y.-Y.; Yang, L. Interactions between human cytochrome P450 enzymes and steroids: physiological and pharmacological implications. *Expert Opinion on Drug Metabolism & Toxicology* **2009**, *5* (6), 621-629. DOI: 10.1517/17425250902967648.
(67) Janocha, S.; Schmitz, D.; Bernhardt, R. Terpene Hydroxylation with Microbial Cytochrome P450 Monooxygenases. In *Biotechnology of Isoprenoids*, Schrader, J., Bohlmann, J. Eds.; Springer International Publishing, 2015; pp 215-250.
(68) Dewar, M. J. S.; Zoebisch, E. G.; Healy, E. F.; Stewart, J. J. P. Development and use of quantum mechanical molecular models. 76. AM1: a new general purpose quantum mechanical molecular model. *Journal of the American Chemical Society* **1985**, *107* (13), 3902-3909. DOI: 10.1021/ja00299a024.
(69) Thiel, W. Semiempirical quantum–chemical methods. *Wiley Interdisciplinary Reviews: Computational Molecular Science* **2014**, *4* (2), 145-157.
(70) Gingrich, P. W.; Siegel, J. B.; Tantillo, D. J. Regression Modeling for the Prediction of Hydrogen Atom Transfer Barriers in Cytochrome P450 from Semi-empirically Derived Descriptors. *Chemistry-Methods* **2022**, e202100108.
(71) O'Boyle, N. M.; Banck, M.; James, C. A.; Morley, C.; Vandermeersch, T.; Hutchison, G. R. Open Babel: An open chemical toolbox. *Journal of cheminformatics* **2011**, *3* (1), 1-14.
(72) Applequist, J.; Carl, J. R.; Fung, K.-K. Atom dipole interaction model for molecular polarizability. Application to polyatomic molecules and determination of atom polarizabilities. *Journal of the American Chemical Society* **1972**, *94* (9), 2952-2960.
(73) Wiberg, K. B. Application of the pople-santry-segal CNDO method to the cyclopropylcarbinyl and cyclobutyl cation and to bicyclobutane. *Tetrahedron* **1968**, *24* (3), 1083-1096.
(74) Ehlert, S.; Stahn, M.; Spicher, S.; Grimme, S. A Robust and Efficient Implicit Solvation Model for Fast Semiempirical Methods. *ChemRxiv* **2021**, This content is a preprint and has not been peer-reviewed.
(75) Schäfer, A.; Horn, H.; Ahlrichs, R. Fully optimized contracted Gaussian basis sets for atoms Li to Kr. *The Journal of Chemical Physics* **1992**, *97* (4), 2571-2577. DOI: 10.1063/1.463096.
(76) Stewart, J. J. P. Optimization of parameters for semiempirical methods I. Method. *Journal of Computational Chemistry* **1989**, *10* (2), 209-220. DOI: https://doi.org/10.1002/jcc.540100208.
(77) Tubert-Brohman, I.; Guimarães, C. R. W.; Jorgensen, W. L. Extension of the PDDG/PM3 Semiempirical Molecular Orbital Method to Sulfur, Silicon, and Phosphorus. *Journal of Chemical Theory and Computation* **2005**, *1* (5), 817-823. DOI: 10.1021/ct0500287. Repasky, M. P.; Chandrasekhar, J.; Jorgensen, W. L. PDDG/PM3 and PDDG/MNDO: Improved semiempirical methods. *Journal of Computational Chemistry* **2002**, *23* (16), 1601-1622. DOI: https://doi.org/10.1002/jcc.10162.
(78) Stewart, J. J. P. Optimization of parameters for semiempirical methods V: Modification of NDDO approximations and application to 70 elements. *Journal of Molecular Modeling* **2007**, *13* (12), 1173-1213. DOI: 10.1007/s00894-007-0233-4.
(79) Grimme, S.; Antony, J.; Ehrlich, S.; Krieg, H. A consistent and accurate ab initio parametrization of density functional dispersion correction (DFT-D) for the 94 elements H-Pu. *The Journal of Chemical Physics* **2010**, *132* (15), 154104. DOI: 10.1063/1.3382344.
(80) Grimme, S. Supramolecular Binding Thermodynamics by Dispersion-Corrected Density Functional Theory. *Chemistry – A European Journal* **2012**, *18* (32), 9955-9964. DOI: https://doi.org/10.1002/chem.201200497.
(81) Stewart, J. J. Optimization of parameters for semiempirical methods VI: more modifications to the NDDO approximations and re-optimization of parameters. *Journal of molecular modeling* **2013**, *19* (1), 1-32.
(82) Zhang, R.-Q.; Fan, W.-J. Economical basis sets and their uses in ab initio calculations. *International Journal of Quantum Chemistry* **2015**, *115* (9), 570-577. DOI: https://doi.org/10.1002/qua.24830.
(83) Grimme, S.; Bannwarth, C.; Caldeweyher, E.; Pisarek, J.; Hansen, A. A general intermolecular force field based on tight-binding quantum chemical calculations. *The Journal of Chemical Physics* **2017**, *147*

(16), 161708. Grimme, S.; Bannwarth, C.; Shushkov, P. A robust and accurate tight-binding quantum chemical method for structures, vibrational frequencies, and noncovalent interactions of large molecular systems parametrized for all spd-block elements (Z= 1–86). *Journal of chemical theory and computation* **2017**, *13* (5), 1989-2009.
(84) Vosko, S. H.; Wilk, L.; Nusair, M. Accurate spin-dependent electron liquid correlation energies for local spin density calculations: a critical analysis. *Canadian Journal of physics* **1980**, *58* (8), 1200-1211.
(85) Pracht, P.; Bohle, F.; Grimme, S. Automated exploration of the low-energy chemical space with fast quantum chemical methods. *Physical Chemistry Chemical Physics* **2020**.
(86) Grimme, S. Density functional theory with London dispersion corrections. *Wiley Interdisciplinary Reviews: Computational Molecular Science* **2011**, *1* (2), 211-228.
(87) Caldeweyher, E.; Ehlert, S.; Hansen, A.; Neugebauer, H.; Spicher, S.; Bannwarth, C.; Grimme, S. A generally applicable atomic-charge dependent London dispersion correction. *The Journal of Chemical Physics* **2019**, *150* (15), 154122. DOI: 10.1063/1.5090222. Wagner, J. P.; Schreiner, P. R. London dispersion in molecular chemistry—reconsidering steric effects. *Angewandte Chemie International Edition* **2015**, *54* (42), 12274-12296. Liptrot, D. J.; Power, P. P. London dispersion forces in sterically crowded inorganic and organometallic molecules. *Nature Reviews Chemistry* **2017**, *1* (1), 1-12.
(88) Lonsdale, R.; Harvey, J. N.; Mulholland, A. J. Inclusion of Dispersion Effects Significantly Improves Accuracy of Calculated Reaction Barriers for Cytochrome P450 Catalyzed Reactions. *The Journal of Physical Chemistry Letters* **2010**, *1* (21), 3232-3237. DOI: 10.1021/jz101279n. Rydberg, P.; Lonsdale, R.; Harvey, J. N.; Mulholland, A. J.; Olsen, L. Trends in predicted chemoselectivity of cytochrome P450 oxidation: B3LYP barrier heights for epoxidation and hydroxylation reactions. *Journal of Molecular Graphics and Modelling* **2014**, *52*, 30-35. DOI: https://doi.org/10.1016/j.jmgm.2014.06.002.
(89) Collins, J. R.; Loew, G. H. Theoretical study of the product specificity in the hydroxylation of camphor, norcamphor, 5, 5-difluorocamphor, and pericyclocamphanone by cytochrome P-450cam. *Journal of Biological Chemistry* **1988**, *263* (7), 3164-3170.
(90) Loida, P. J.; Sligar, S. G.; Paulsen, M. D.; Arnold, G. E.; Ornstein, R. L. Stereoselective Hydroxylation of Norcamphor by Cytochrome P450cam EXPERIMENTAL VERIFICATION OF MOLECULAR DYNAMICS SIMULATIONS. *Journal of Biological Chemistry* **1995**, *270* (10), 5326-5330.
(91) Bell, S. G.; Chen, X.; Sowden, R. J.; Xu, F.; Williams, J. N.; Wong, L.-L.; Rao, Z. Molecular Recognition in (+)-α-Pinene Oxidation by Cytochrome P450cam. *Journal of the American Chemical Society* **2003**, *125* (3), 705-714. DOI: 10.1021/ja028460a.
(92) Harris, D.; Loew, G. Prediction of regiospecific hydroxylation of camphor analogs by cytochrome P450cam. *Journal of the American Chemical Society* **1995**, *117* (10), 2738-2746.
(93) Collins, J. R.; Loew, G. H. Theoretical study of the product specificity in the hydroxylation of camphor, norcamphor, 5,5-difluorocamphor, and pericyclocamphanone by cytochrome P-450cam. *Journal of Biological Chemistry* **1988**, *263* (7), 3164-3170. DOI: https://doi.org/10.1016/S0021-9258(18)69049-0.
(94) Arndtsen, B. A.; Bergman, R. G.; Mobley, T. A.; Peterson, T. H. Selective intermolecular carbon-hydrogen bond activation by synthetic metal complexes in homogeneous solution. *Accounts of chemical research* **1995**, *28* (3), 154-162. Huang, Z.; Dong, G. Site-Selectivity Control in Organic Reactions: A Quest To Differentiate Reactivity among the Same Kind of Functional Groups. *Accounts of Chemical Research* **2017**, *50* (3), 465-471. DOI: 10.1021/acs.accounts.6b00476.
(95) Liu, Z.; Arnold, F. H. New-to-nature chemistry from old protein machinery: carbene and nitrene transferases. *Current Opinion in Biotechnology* **2021**, *69*, 43-51. DOI: https://doi.org/10.1016/j.copbio.2020.12.005.
(96) Zanger, U. M.; Schwab, M. Cytochrome P450 enzymes in drug metabolism: Regulation of gene expression, enzyme activities, and impact of genetic variation. *Pharmacology & Therapeutics* **2013**, *138* (1), 103-141. DOI: https://doi.org/10.1016/j.pharmthera.2012.12.007.

(97) Park, B. K.; Kitteringham, N. R.; O'Neill, P. M. Metabolism of fluorine-containing drugs. *Annual review of pharmacology and toxicology* **2001**, *41* (1), 443-470. Purser, S.; Moore, P. R.; Swallow, S.; Gouverneur, V. Fluorine in medicinal chemistry. *Chemical Society Reviews* **2008**, *37* (2), 320-330.
(98) Patten, C. J.; Thomas, P. E.; Guy, R. L.; Lee, M.; Gonzalez, F. J.; Guengerich, F. P.; Yang, C. S. Cytochrome P450 enzymes involved in acetaminophen activation by rat and human liver microsomes and their kinetics. *Chemical Research in Toxicology* **1993**, *6* (4), 511-518. DOI: 10.1021/tx00034a019. Slikker Jr, W.; Andersen, M. E.; Bogdanffy, M. S.; Bus, J. S.; Cohen, S. D.; Conolly, R. B.; David, R. M.; Doerrer, N. G.; Dorman, D. C.; Gaylor, D. W. Dose-dependent transitions in mechanisms of toxicity: case studies. *Toxicology and applied pharmacology* **2004**, *201* (3), 226-294.
(99) Driscoll, J. P.; Sadlowski, C. M.; Shah, N. R.; Feula, A. Metabolism and Bioactivation: It's Time to Expect the Unexpected. *Journal of Medicinal Chemistry* **2020**.
(100) Olsen, L.; Montefiori, M.; Tran, K. P.; Jørgensen, F. S. SMARTCyp 3.0: Enhanced cytochrome P450 site-of-metabolism prediction server. *Bioinformatics* **2019**.
(101) Kamachi, T.; Yoshizawa, K. A theoretical study on the mechanism of camphor hydroxylation by compound I of cytochrome P450. *Journal of the American Chemical Society* **2003**, *125* (15), 4652-4661.
(102) Grimme, S. Semiempirical GGA-type density functional constructed with a long-range dispersion correction. *Journal of computational chemistry* **2006**, *27* (15), 1787-1799.
(103) Lonsdale, R.; Harvey, J. N.; Mulholland, A. J. Effects of Dispersion in Density Functional Based Quantum Mechanical/Molecular Mechanical Calculations on Cytochrome P450 Catalyzed Reactions. *Journal of Chemical Theory and Computation* **2012**, *8* (11), 4637-4645. DOI: 10.1021/ct300329h.
(104) Jerome, S. V.; Hughes, T. F.; Friesner, R. A. Successful application of the DBLOC method to the hydroxylation of camphor by cytochrome p450. *Protein Science* **2016**, *25* (1), 277-285. DOI: https://doi.org/10.1002/pro.2819.
(105) Loida, P. J.; Sligar, S. G.; Paulsen, M. D.; Arnold, G. E.; Ornstein, R. L. Stereoselective Hydroxylation of Norcamphor by Cytochrome P450cam: EXPERIMENTAL VERIFICATION OF MOLECULAR DYNAMICS SIMULATIONS (∗). *Journal of Biological Chemistry* **1995**, *270* (10), 5326-5330. DOI: https://doi.org/10.1074/jbc.270.10.5326.
(106) Olsen, L.; Oostenbrink, C.; Jørgensen, F. S. Prediction of cytochrome P450 mediated metabolism. *Advanced Drug Delivery Reviews* **2015**, *86*, 61-71. DOI: https://doi.org/10.1016/j.addr.2015.04.020.
(107) Filipovic, D.; Paulsen, M.; Loida, P.; Sligar, S.; Ornstein, R. Ethylbenzene hydroxylation by cytochrome P450cam. *Biochemical and biophysical research communications* **1992**, *189* (1), 488-495.
(108) Chuo, S.-W.; Wang, L.-P.; Britt, R. D.; Goodin, D. B. An Intermediate Conformational State of Cytochrome P450cam-CN in Complex with Putidaredoxin. *Biochemistry* **2019**, *58* (18), 2353-2361. DOI: 10.1021/acs.biochem.9b00192.
(109) Tyzack, J. D.; Williamson, M. J.; Torella, R.; Glen, R. C. Prediction of Cytochrome P450 Xenobiotic Metabolism: Tethered Docking and Reactivity Derived from Ligand Molecular Orbital Analysis. *Journal of Chemical Information and Modeling* **2013**, *53* (6), 1294-1305. DOI: 10.1021/ci400058s.
(110) Schrodinger, LLC. The PyMOL Molecular Graphics System, Version 1.8. 2015.
(111) Pracht, P.; Bohle, F.; Grimme, S. Automated exploration of the low-energy chemical space with fast quantum chemical methods. *Physical Chemistry Chemical Physics* **2020**, *22* (14), 7169-7192, 10.1039/C9CP06869D. DOI: 10.1039/C9CP06869D.
(112) Fleishman, S. J.; Leaver-Fay, A.; Corn, J. E.; Strauch, E.-M.; Khare, S. D.; Koga, N.; Ashworth, J.; Murphy, P.; Richter, F.; Lemmon, G. RosettaScripts: a scripting language interface to the Rosetta macromolecular modeling suite. *PloS one* **2011**, *6* (6).
(113) Ruiz-Carmona, S.; Alvarez-Garcia, D.; Foloppe, N.; Garmendia-Doval, A. B.; Juhos, S.; Schmidtke, P.; Barril, X.; Hubbard, R. E.; Morley, S. D. rDock: a fast, versatile and open source program for docking ligands to proteins and nucleic acids. *PLoS computational biology* **2014**, *10* (4), e1003571.

(114) Bell, E. W.; Zhang, Y. DockRMSD: an open-source tool for atom mapping and RMSD calculation of symmetric molecules through graph isomorphism. *Journal of Cheminformatics* **2019**, *11* (1), 1-9.
(115) Tange, O. *GNU parallel 2018*; Lulu. com, 2018.
(116) *pandas-dev/pandas: Pandas 1.4.2*; Zenodo: 2022. https://doi.org/10.5281/zenodo.6408044

http://dx.doi.org/10.5281/zenodo.6408044 (accessed.
(117) 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* (5458), 1615-1622.
(118) Nagano, S.; Poulos, T. L. Crystallographic Study on the Dioxygen Complex of Wild-type and Mutant Cytochrome P450cam: IMPLICATIONS FOR THE DIOXYGEN ACTIVATION MECHANISM*♦. *Journal of Biological Chemistry* **2005**, *280* (36), 31659-31663.
(119) Lee, Y.-T.; Glazer, E. C.; Wilson, R. F.; Stout, C. D.; Goodin, D. B. Three Clusters of Conformational States in P450cam Reveal a Multistep Pathway for Closing of the Substrate Access Channel. *Biochemistry* **2011**, *50* (5), 693-703. DOI: 10.1021/bi101726d.
(120) Lemmon, G.; Meiler, J. Rosetta Ligand docking with flexible XML protocols. In *Computational Drug Discovery and Design*, Springer, 2012; pp 143-155.
(121) Leaver-Fay, A.; Khare, S.; Bjelic, S.; Baker, D. De novo enzyme design using Rosetta3. *PLoS One* **2011**, *6* (5), e19230.
(122) Tukey, J. W. *Exploratory data analysis*; Reading, MA, 1977.
(123) Eberhardt, J.; Santos-Martins, D.; Tillack, A. F.; Forli, S. AutoDock Vina 1.2.0: New Docking Methods, Expanded Force Field, and Python Bindings. *Journal of Chemical Information and Modeling* **2021**, *61* (8), 3891-3898. DOI: 10.1021/acs.jcim.1c00203.
(124) Lu, J.; Hou, X.; Wang, C.; Zhang, Y. Incorporating Explicit Water Molecules and Ligand Conformation Stability in Machine-Learning Scoring Functions. *Journal of Chemical Information and Modeling* **2019**, *59* (11), 4540-4549. DOI: 10.1021/acs.jcim.9b00645.
(125) Trott, O.; Olson, A. J. AutoDock Vina: Improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *Journal of Computational Chemistry* **2010**, *31* (2), 455-461. DOI: https://doi.org/10.1002/jcc.21334.
(126) 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 Catalysis* **2018**, *8* (4), 3322-3330. DOI: 10.1021/acscatal.8b00342.
(127) 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. *Journal of Chemical Theory and Computation* **2019**, *15* (9), 5116-5134. DOI: 10.1021/acs.jctc.9b00366.
(128) Huang, S.-Y. Comprehensive assessment of flexible-ligand docking algorithms: current effectiveness and challenges. *Briefings in Bioinformatics* **2017**, *19* (5), 982-994. DOI: 10.1093/bib/bbx030 (acccessed 9/3/2022).
(129) Moustakas, D. T.; Lang, P. T.; Pegg, S.; Pettersen, E.; Kuntz, I. D.; Brooijmans, N.; Rizzo, R. C. Development and validation of a modular, extensible docking program: DOCK 5. *Journal of computer-aided molecular design* **2006**, *20* (10), 601-619.
(130) Huang, S.-Y.; Zou, X. Ensemble docking of multiple protein structures: Considering protein structural variations in molecular docking. *Proteins: Structure, Function, and Bioinformatics* **2007**, *66* (2), 399-421. DOI: https://doi.org/10.1002/prot.21214.
(131) Murcko, M. A.; Castejon, H.; Wiberg, K. B. Carbon−Carbon Rotational Barriers in Butane, 1-Butene, and 1,3-Butadiene. *The Journal of Physical Chemistry* **1996**, *100* (40), 16162-16168. DOI: 10.1021/jp9621742.

(132) Hernandez-Ortega, A.; Vinaixa, M.; Zebec, Z.; Takano, E.; Scrutton, N. S. A Toolbox for Diverse Oxyfunctionalisation of Monoterpenes. *Scientific Reports* **2018**, *8* (1), 14396. DOI: 10.1038/s41598-018-32816-1.
(133) Smith, S. T.; Meiler, J. Assessing multiple score functions in Rosetta for drug discovery. *PLOS ONE* **2020**, *15* (10), e0240450. DOI: 10.1371/journal.pone.0240450.
(134) Alford, R. F.; Leaver-Fay, A.; Jeliazkov, J. R.; O'Meara, M. J.; DiMaio, F. P.; Park, H.; Shapovalov, M. V.; Renfrew, P. D.; Mulligan, V. K.; Kappel, K.; et al. The Rosetta All-Atom Energy Function for Macromolecular Modeling and Design. *Journal of Chemical Theory and Computation* **2017**, *13* (6), 3031-3048. DOI: 10.1021/acs.jctc.7b00125.
(135) Su, M.; Yang, Q.; Du, Y.; Feng, G.; Liu, Z.; Li, Y.; Wang, R. Comparative Assessment of Scoring Functions: The CASF-2016 Update. *Journal of Chemical Information and Modeling* **2019**, *59* (2), 895-913. DOI: 10.1021/acs.jcim.8b00545.
(136) Park, H.; Zhou, G.; Baek, M.; Baker, D.; DiMaio, F. Force Field Optimization Guided by Small Molecule Crystal Lattice Data Enables Consistent Sub-Angstrom Protein–Ligand Docking. *Journal of Chemical Theory and Computation* **2021**, *17* (3), 2000-2010. DOI: 10.1021/acs.jctc.0c01184.
(137) Raag, R.; Poulos, T. L. Crystal structures of cytochrome P-450CAM complexed with camphane, thiocamphor, and adamantane: factors controlling P-450 substrate hydroxylation. *Biochemistry* **1991**, *30* (10), 2674-2684.
(138) DeVore, N. M.; Scott, E. E. Nicotine and 4-(methylnitrosamino)-1-(3-pyridyl)-1-butanone binding and access channel in human cytochrome P450 2A6 and 2A13 enzymes. *Journal of Biological Chemistry* **2012**, *287* (32), 26576-26585.
(139) Makeneni, S.; Thieker, D. F.; Woods, R. J. Applying Pose Clustering and MD Simulations To Eliminate False Positives in Molecular Docking. *Journal of Chemical Information and Modeling* **2018**, *58* (3), 605-614. DOI: 10.1021/acs.jcim.7b00588.
(140) Szymanski, W.; Beierle, J. M.; Kistemaker, H. A.; Velema, W. A.; Feringa, B. L. Reversible photocontrol of biological systems by the incorporation of molecular photoswitches. *Chemical reviews* **2013**, *113* (8), 6114-6178.
(141) Griefs, P. Vorläufige Notiz über die Einwirkung von salpetriger Säure auf Amidinitro-und Aminitrophenylsäure. *Justus Liebigs Annalen der Chemie* **1858**, *106* (1), 123-125.
(142) Bandara, H. D.; Burdette, S. C. Photoisomerization in different classes of azobenzene. *Chemical Society Reviews* **2012**, *41* (5), 1809-1825.
(143) Crespi, S.; Simeth, N. A.; König, B. Heteroaryl azo dyes as molecular photoswitches. *Nature Reviews Chemistry* **2019**, *3* (3), 133-146. Tuck, J. R.; Tombari, R. J.; Yardeny, N.; Olson, D. E. A Modular Approach to Arylazo-1, 2, 3-triazole Photoswitches. *Organic Letters* **2021**, *23* (11), 4305-4310.
(144) van Dijken, D. J.; Kovaříček, P.; Ihrig, S. P.; Hecht, S. Acylhydrazones as Widely Tunable Photoswitches. *Journal of the American Chemical Society* **2015**, *137* (47), 14982-14991. DOI: 10.1021/jacs.5b09519.
(145) Yanai, T.; Tew, D. P.; Handy, N. C. A new hybrid exchange–correlation functional using the Coulomb-attenuating method (CAM-B3LYP). *Chemical physics letters* **2004**, *393* (1-3), 51-57.
(146) Grimme, S.; Ehrlich, S.; Goerigk, L. Effect of the damping function in dispersion corrected density functional theory. *Journal of Computational Chemistry* **2011**, *32* (7), 1456-1465. DOI: https://doi.org/10.1002/jcc.21759.
(147) Becke, A. D.; Johnson, E. R. A density-functional model of the dispersion interaction. *The Journal of chemical physics* **2005**, *123* (15), 154101.
(148) Cancès, E.; Mennucci, B.; Tomasi, J. A new integral equation formalism for the polarizable continuum model: Theoretical background and applications to isotropic and anisotropic dielectrics. *The Journal of Chemical Physics* **1997**, *107* (8), 3032-3041. DOI: 10.1063/1.474659.

(149) Cossi, M.; Barone, V. Time-dependent density functional theory for molecules in liquid solutions. *The Journal of Chemical Physics* **2001**, *115* (10), 4708-4717. DOI: 10.1063/1.1394921.
(150) Baeyer, A. Nitrosobenzol und nitrosonaphtalin. *Berichte der deutschen chemischen Gesellschaft* **1874**, *7* (2), 1638-1640.
(151) Mills, C. XCIII.—Some new azo-compounds. *Journal of the Chemical Society, Transactions* **1895**, *67*, 925-933.
(152) Pavitt, A. S.; Bylaska, E. J.; Tratnyek, P. G. Oxidation potentials of phenols and anilines: correlation analysis of electrochemical and theoretical values. *Environmental Science: Processes & Impacts* **2017**, *19* (3), 339-349, 10.1039/C6EM00694A. DOI: 10.1039/C6EM00694A.
(153) Marenich, A. V.; Cramer, C. J.; Truhlar, D. G. Universal Solvation Model Based on Solute Electron Density and on a Continuum Model of the Solvent Defined by the Bulk Dielectric Constant and Atomic Surface Tensions. *The Journal of Physical Chemistry B* **2009**, *113* (18), 6378-6396. DOI: 10.1021/jp810292n.
(154) Stock, J. T.; Orna, M. V. *Electrochemistry, past and present*; ACS Publications, 1989.
(155) Tripkovic, V.; Björketun, M. E.; Skúlason, E.; Rossmeisl, J. Standard hydrogen electrode and potential of zero charge in density functional calculations. *Physical Review B* **2011**, *84* (11), 115452. DOI: 10.1103/PhysRevB.84.115452.
(156) Suatoni, J. C.; Snyder, R. E.; Clark, R. O. Voltammetric studies of phenol and aniline ring substitution. *Analytical Chemistry* **1961**, *33* (13), 1894-1897.
(157) Jovanovic, S. V.; Tosic, M.; Simic, M. G. Use of the Hammett correlation and. delta.+ for calculation of one-electron redox potentials of antioxidants. *The Journal of Physical Chemistry* **1991**, *95* (26), 10824-10827.
(158) Jampilek, J. Heterocycles in Medicinal Chemistry. *Molecules* **2019**, *24* (21), 3839. DOI: 10.3390/molecules24213839 PubMed.
(159) Brichacek, M.; Lee, D.; Njardarson, J. T. Lewis Acid Catalyzed [1,3]-Sigmatropic Rearrangement of Vinyl Aziridines. *Organic Letters* **2008**, *10* (21), 5023-5026. DOI: 10.1021/ol802123e.
(160) Medran, N. S.; La-Venia, A.; Testero, S. A. Metal-mediated synthesis of pyrrolines. *RSC Advances* **2019**, *9* (12), 6804-6844, 10.1039/C8RA10247C. DOI: 10.1039/C8RA10247C.
(161) Szostak, M.; Fazakerley, N. J.; Parmar, D.; Procter, D. J. Cross-coupling reactions using samarium (II) iodide. *Chemical reviews* **2014**, *114* (11), 5959-6039.
(162) Shi, S.; Szostak, M. Synthesis of Nitrogen Heterocycles Using Samarium(II) Iodide. *Molecules* **2017**, *22* (11), 2018.
(163) Adamo, C.; Barone, V. Toward reliable density functional methods without adjustable parameters: The PBE0 model. *The Journal of Chemical Physics* **1999**, *110* (13), 6158-6170. DOI: 10.1063/1.478522.
(164) Weigend, F.; Ahlrichs, R. Balanced basis sets of split valence, triple zeta valence and quadruple zeta valence quality for H to Rn: Design and assessment of accuracy. *Physical Chemistry Chemical Physics* **2005**, *7* (18), 3297-3305, 10.1039/B508541A. DOI: 10.1039/B508541A.
(165) Glendening, E.; Badenhoop, J.; Reed, A.; Carpenter, J.; Weinhold, F. NBO 3.1. *Theoretical Chemistry Institute, University of Wisconsin, Madison, WI* **1996**.