Assessing alkene reactivity toward Cytochrome P450 mediated epoxidation through localized descriptors and regression modeling

Phillip W. Gingrich, Justin B. Siegel, and Dean J. Tantillo\*

Department of Chemistry, University of California, Davis, One Shields Avenue, Davis, CA

Cytochrome P450, epoxidation, multiple linear regression, semi-empirical

ABSTRACT: The prediction of sites of epoxidation by cytochrome P450s during metabolism is particularly important in drug design, as epoxides are capable of alkylating biological macromolecules. Reliable methods are needed to quantitatively predict P450-mediated epoxidation barriers for inclusion in high throughput screening campaigns alongside protein-ligand docking. Utilizing the fractional occupation number weighted density (FOD) and orbital weighted Fukui index (*fw+*) as descriptors of local reactivity and a data set of 36 alkene epoxidation barriers computed with density functional theory (DFT), we developed and validated a multiple linear regression model for the reliable estimation of epoxidation barriers using only substrate structures as input. Using our recommended level of theory (GFN2-xTB//GFN-FF), mean absolute errors in the training and test sets were found to be 0.66 kcal/mol and 0.70 kcal/mol, respectively, with coefficients of determination of ca. 0.80. We demonstrate the utility of this approach on three known substrates of CYP101A1 and further show that is approach is inappropriate for particularly electron-rich alkenes. By employing a modern semi-empirical method on force field generated geometries, the required descriptors can be calculated on the millisecond timescale per structure, making the approach well-suited for incorporation into high throughput methodologies alongside docking.

Introduction

Of the multiple oxidative transformations that cytochrome P450s catalyze,1 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.2 Among these, epoxidation of alkenes (and arenes) is particularly common.3

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.4 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.5 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.6 However, docking alone neglects the importance of the electronic structure of a substrate or residence time7 in determining its susceptibility toward epoxidation by Compound 0 or 1 in the P450 catalytic cycle (Figure 1).2, 8, 9



Figure 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.2, 10, 11 However, within high throughput virtual screening, such calculations are prohibitively expensive, creating a need for more affordable methodology. Cheminformatics 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 2).12 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



Figure 2 Training and test set compounds for which epoxidation barriers involving a truncated Compound 1 model were previously computed using DFT.12

polarity is open to subjective assessment, and the calculation of the molecular dipole moment will obviously vary with the level of theory employed. In their work, models for polar and non-polar compounds account for more than 95% of the variability in the epoxidation barrier by the ionization potential alone.12 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 non-bonding (lone pair) orbitals 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.

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)13 and the orbital weighted Fukui index (*fw+*),14 through a multiple linear regression (MLR) model. As true values, we reuse, in accordance with FAIR data principles,15 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 2.12 Our work assumes Compound 1 to be the responsible oxidant, though we recognize the preceding hydroperoxo intermediate as a competent oxidant.16 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.

Computational Methods

All compounds were first prepared in Avogadro 1.2.017 and initially optimized using the MMF9418 force field.

Grimme’s CREST19 (version 2.11.1) and xtb20 (version 6.4.1) programs were used for all semi-empirical calculations. Conformer sampling was first done in CREST using GFN2-xTB21 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-xTB22 or GFN-FF23 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 sp2-hybridized carbon involved in the initial C–O bond formation event during epoxidation in xtb using the “--fod” flag. Additionally, molden 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 Theory24 (CDFT) calculations. The molden input files were then read with Multiwfn.25 Hirshfeld26 and Mulliken27 atomic charges were determined for the *N-*electron state. Condensed traditional28, 29 and orbital-weighted14 Fukui Indices were determined.

Ordinary least squares linear regressions were performed in python, utilizing the scikit-learn,30 pandas,31 and statsmodels32 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.33 In the final regression analyses, residuals were verified to be normally distributed according to a Shapiro-Wilk normality test.34Stationary point analyses for the initial C–O bond formation event for ethylene, vinyl chloride, and nitroethylene were performed in Gaussian 1635 on the quartet surface at the B3LYP36, 37/LACVP\*\*38, 39 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. Hirshfeld26 charges were computed within Gaussian and summed over the substrate fragment.

Similarly, zero-point corrected potential energy barriers on the quartet surface for the initial C–O bond formation event for the compounds in Figure 9 were computed using DFT. These were calculated at the B3LYP/Wachters+f40 (Fe)/TZVP41//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 personal 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.12 Figure 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 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.12

While no 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.1 One such descriptor is the FOD. As described by Bauer and co-workers, FOD is useful to identify statically correlated and chemically reactive (what the authors called “hot”) electrons.13 For our panel of substrates, we computed the FOD at the sp2 carbon involved initially in C–O bond formation using Grimme’s GFN family of methods. Figure 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 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. 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 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. Coefficients of determination and mean absolute errors for linear regression models between FOD values and DFT computed epoxidation barriers.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Training Set** | | **Test Set** | |
| **Method** | **R2** | **MAE (kcal/mol)** | **R2** | **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 Theory24 might be useful and that, specifically, condensed Fukui indices would be physically relevant.42 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 5.



Figure 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 5a), the alkene substrate could also be treated as a nucleophile (Figure 5b) or as an electrophile (Figure 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 sp2 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 6). In general, the predictive power of traditional condensed Fukui indices by a linear model went as *f(+)* > *f(0)* > *f(-)*.

Figure 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 would generate a carbocation intermediate. For many of our substrates, this would be a secondary carbocation and be generally unfavorable. 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 devoid of any strongly π-donating conjugated substituents, these also are not generally found among Nature’s P450 substrates, perhaps due to competing dealkylation mechanisms.2 Alternatively, the alkene may be considered as an electrophile (Figure 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 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.43, 44 The removal of this electron “push” has been studied with neutral serine P450 mutants that exhibit altered reactivity (e.g., the carbene transferase).45 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 7 Evolution of the summed Hirshfeld charge 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 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 σp values46), 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 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 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.14

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. In doing so, the same trend for predictive performance (*fw+* > *fw0 > fw-*)was observed between the three indices, and the results for the *fw+* index are summarized in Table 2. Performance was slightly biased toward the training set. While reasonable structural diversity is present in both training and test sets, there may be other training sets that prevent this. Nonetheless, taken alone, *fw+* 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 2. Coefficients of determination and mean absolute errors for linear regression models between *fw+* indices and DFT computed epoxidation barriers.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Training Set** | | **Test Set** | |
| **Method** | **R2** | **MAE (kcal/mol)** | **R2** | **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,47 we found both the FOD and *fw+* 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 (pFOD=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 *fw+* index became insignificant (p = 0.702). Figure 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 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 8 Correlation between P450-mediated epoxidation barriers previously computed with DFT12 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 semi-empirical 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 our required semi-empirical geometry optimizations take on the order of milliseconds with only modest computing hardware, the second advantage is that it is possible to apply this model to thousands of docking poses. Consequently, 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).48

Table 3. Adjusted coefficients of determination and mean absolute errors for MLR models using FOD values and *fw+* indices to predict epoxidation barriers.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Training Set** | | **Test Set** | |
| **Method** | **R2** | **MAE (kcal/mol)** | **R2** | **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-xTBa | 0.79 | 0.70 | 0.81 | 0.59 |
| GFN1-xTB//GFN-FFa | 0.81 | 0.70 | 0.78 | 0.66 |

a The *fw+* 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 that tetrasubstituted alkenes are not represented. While there may be examples of tetrasubstituted alkenes that are epoxidized by P450s, our review of the literature failed to uncover any, and some literature suggests that tetrasubstituted alkenes are too sterically crowded to undergo epoxidation in a P450.49 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.50 As with any predictive modeling approach, care should be given to assess the appropriateness of the model for systems of interest.

Further Validation

To that end, we further tested our approach by examining the compounds in Figure 9 using GFN2-xTB and GFN2-xTB//GFN-FF. These eight compounds were selected as they are either known substrates of CYP101A1 (Figure 9A) that undergo epoxidation as well as electron-rich and/or sterically crowded alkenes. 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 2.



Figure 9 Validation compounds. The CYP101A1 substrates are known to undergo epoxidation, while the remaining compounds were considered to probe the limitations of the approach herein.

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

|  |  |  |  |
| --- | --- | --- | --- |
| **Substrate** | **DFTa** | **MLR  Predictionb (GFN2-xTB)** | **MLR  Predictionb**  **(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 *fw+* index as descriptors.

As seen in Table 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 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 MLR 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 (*fw+*) 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. Such work is presently in development in our groups.

Data and Software Availability

All computed Fukui indices and FOD values are provided in the data.csv file in the supporting information, along with all “true” and MLR-predicted barriers. Grimme’s xtb (<https://github.com/grimme-lab/xtb>) and crest (<https://github.com/grimme-lab/crest>) are freely available on GitHub. Python 3.7.9 was used for all statistical modeling. The environment was managed in conda 4.10.1 and included pandas 1.1.3, scikit-learn 0.23.2, and statsmodels 0.12.0. Avogadro was used for molecular visualization and is freely available (<https://avogadro.cc/>).

ASSOCIATED CONTENT

Supporting Information.

Python scripts for feature selection and linear regressions

GFN2-xTB substrate geometries in mol2 format  
B3LYP/LACVP\*\* reaction coordinate geometries in mol2 the epoxidation of ethylene, vinyl chloride, nitroethylene, and the compounds in Figure 9

All computed FOD and CDFT descriptors at various levels of theory, along with “true” epoxidation barriers (data.csv file)

This material is available free of charge via the Internet at http://pubs.acs.org.

AUTHOR INFORMATION

Corresponding Author

\* Dean Tantillo, djtantillo@ucdavis.edu, ORCID: 0000-0002-2992-8844

Author Contributions

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

Funding Sources

The first author’s salary and tuition were funded by Bayer U.S. LLC.

ACKNOWLEDGMENT

We are grateful for the thoughtful critique of this work by multiple students and associates in our groups and department. While all calculations and statistical modeling were performed by the first author, our teams’ contributions, as always, were instrumental in the conduct of this research.

ABBREVIATIONS

DFT, Density Functional Theory; CYP450, Cytochrome P450; MMFF, Merck Molecular Force Field; GFN, Geometries, Frequencies, and Non-covalent Interactions; MLR, Multiple Linear Regression; SLR, Single Linear Regression

REFERENCES

1. Meunier, B.; de Visser, S. P.; Shaik, S., Mechanism of Oxidation Reactions Catalyzed by Cytochrome P450 Enzymes. *Chemical Reviews* **2004,** *104* (9), 3947-3980.

2. 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.

3. 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.

4. Ehrenberg, L.; Hussain, S., Genetic toxicity of some important epoxides. *Mutation Research/Reviews in Genetic Toxicology* **1981,** *86* (1), 1-113.

5. 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.

6. 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.

7. Chen, Y.-C., Beware of docking! *Trends in pharmacological sciences* **2015,** *36* (2), 78-95.

8. Dubey, K. D.; Shaik, S., Cytochrome P450—The Wonderful Nanomachine Revealed through Dynamic Simulations of the Catalytic Cycle. *Accounts of Chemical Research* **2019**.

9. 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.

10. 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.

11. 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.

12. 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.

13. 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.

14. 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.

15. 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.

16. 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.

17. 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.

18. 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.

19. 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.

20. 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.

21. 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.

22. 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.

23. Spicher, S.; Grimme, S., Robust Atomistic Modeling of Materials, Organometallic, and Biochemical Systems. *Angewandte Chemie International Edition* **2020,** *59* (36), 15665-15673.

24. 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.

25. Lu, T.; Chen, F., Multiwfn: a multifunctional wavefunction analyzer. *Journal of computational chemistry* **2012,** *33* (5), 580-592.

26. Hirshfeld, F. L., Bonded-atom fragments for describing molecular charge densities. *Theoretica chimica acta* **1977,** *44* (2), 129-138.

27. Mulliken, R. S., Electronic population analysis on LCAO–MO molecular wave functions. I. *The Journal of Chemical Physics* **1955,** *23* (10), 1833-1840.

28. 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.

29. 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.

30. 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.

31. Reback, J.; McKinney, W.; Den Van Bossche, J.; Augspurger, T.; Cloud, P.; Klein, A.; Roeschke, M.; Hawkins, S.; Tratner, J.; She, C. *pandas-dev/pandas: Pandas 1.3.4*, 2021.

32. Seabold, S.; Perktold, J., statsmodels: Econometric and statistical modeling with python. In *9th Python in Science Conference*, 2010.

33. 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.

34. Shapiro, S. S.; Wilk, M. B., An analysis of variance test for normality (complete samples). *Biometrika* **1965,** *52* (3/4), 591-611.

35. Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Petersson, G. A.; Nakatsuji, H.; Li, X.; Caricato, M.; Marenich, A. V.; Bloino, J.; Janesko, B. G.; Gomperts, R.; Mennucci, B.; Hratchian, H. P.; Ortiz, J. V.; Izmaylov, A. F.; Sonnenberg, J. L.; Williams; Ding, F.; Lipparini, F.; Egidi, F.; Goings, J.; Peng, B.; Petrone, A.; Henderson, T.; Ranasinghe, D.; Zakrzewski, V. G.; Gao, J.; Rega, N.; Zheng, G.; Liang, W.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Vreven, T.; Throssell, K.; Montgomery Jr., J. A.; Peralta, J. E.; Ogliaro, F.; Bearpark, M. J.; Heyd, J. J.; Brothers, E. N.; Kudin, K. N.; Staroverov, V. N.; Keith, T. A.; Kobayashi, R.; Normand, J.; Raghavachari, K.; Rendell, A. P.; Burant, J. C.; Iyengar, S. S.; Tomasi, J.; Cossi, M.; Millam, J. M.; Klene, M.; Adamo, C.; Cammi, R.; Ochterski, J. W.; Martin, R. L.; Morokuma, K.; Farkas, O.; Foresman, J. B.; Fox, D. J. *Gaussian 16 Rev. A.03*, Wallingford, CT, 2016.

36. Becke, A. D., A new mixing of Hartree–Fock and local density‐functional theories. *The Journal of Chemical Physics* **1993,** *98* (2), 1372-1377.

37. Becke, A. D., Becke’s three parameter hybrid method using the LYP correlation functional. *J. Chem. Phys* **1993,** *98* (492), 5648-5652.

38. 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.

39. 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.

40. 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.

41. 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.

42. 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.

43. 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.

44. 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.

45. 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.

46. 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.

47. Muthukrishnan, R.; Rohini, R. In *LASSO: A feature selection technique in predictive modeling for machine learning*, 2016 IEEE International Conference on Advances in Computer Applications (ICACA), 24-24 Oct. 2016; 2016; pp 18-20.

48. 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.

49. Gorycki, P. D.; Macdonald, T. L., The Oxidation of Tetrasubstituted Alkenes by Cytochrome P450. *Chemical Research in Toxicology* **1994,** *7* (6), 745-751.

50. 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.

Authors are required to submit a graphic entry for the Table of Contents (TOC) that, in conjunction with the manuscript title, should give the reader a representative idea of one of the following: A key structure, reaction, equation, concept, or theorem, etc., that is discussed in the manuscript. Consult the journal’s Instructions for Authors for TOC graphic specifications.

