Computational Study on Site-Selectivity of DMDO-Mediated C–H Hydroxylation

by

Yimin Chen

Bachelor of Science, Nanjing University, 2018

Submitted to the Graduate Faculty of the

Dietrich School of Arts and Sciences in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2023

UNIVERSITY OF PITTSBURGH

DIETRICH SCHOOL OF ARTS AND SCIENCES

This dissertation was presented

by

Yimin Chen

It was defended on

August 17, 2023

and approved by

Paul Floreancig, Professor, Chemistry, University of Pittsburgh

Rob Coalson, Professor, Chemistry, University of Pittsburgh

John Keith, Associate Professor, Chemical and Petroleum Engineering, University of Pittsburgh

Peng Liu, Professor, Chemistry, University of Pittsburgh

Computational Study on Site-Selectivity of DMDO-Mediated C–H Hydroxylation

Yimin Chen, PhD

University of Pittsburgh, 2023

C–H hydroxylation is an important type of transformation in organic synthesis. However, it is challenging to control the site-selectivity for structurally complex substrates with multiple C–H bonds of similar inherent reactivities. The site-selectivity of dimethyldioxirane (DMDO)-mediated C–H hydroxylation is investigated via computational chemistry tools and statistical methods to better understand factors affecting site-selectivity and to develop a predictive model for site-selectivity. In addition to the previously recognized electronic effects on site-selectivity, this study reveals the significance of steric effects and strain-release effects. A model capable of quantifying the electronic effects, steric effects, and strain-release effects on the selectivity of structurally complex compounds is developed. A reaction-specific descriptor based on solvent-accessible surface area (SASA) is developed to describe steric effects of C–H bonds. An activation function was found to be critical to improve the performance of the SASA descriptor. In addition, further application of the site-selectivity prediction model to a macrocyclic molecule is performed to examine the applicability of the model to structurally complex and conformationally flexible molecules.

Table of Contents

Preface ... xi

1.0 Introduction ... 1

1.1 DMDO-Mediated Late-Stage C–H Hydroxylation ... 1

1.2 Mechanistic Insights from Experimental and Computational Studies on DMDO-Mediated C–H Hydroxylation ... 8

1.3 Predictive Models of Reactivity and Selectivity of Chemical Reactions ... 15

2.0 Computational Methods ... 23

2.1 Computational Details of Generating Descriptors and Computing Activation Free Energies for Substrates in the Data Set ... 23

3.0 Developing a Predictive Model for DMDO-Mediated C–H Hydroxylation ... 25

3.1 Data Set Generation and Preliminary Analysis ... 25

3.2 Single-Descriptor Predictive Models ... 31

3.3 Multivariate Linear Regression Models ... 37

3.4 Using an Activation Function to Improve the Performance of Solvent Accessible Surface Area as a Reaction-Specific Steric Descriptor in Multivariate Regression ... 42

4.0 Further Applications of the Predictive Model to Structurally Complex Molecules ... 50

4.1 Challenges of Site-Selectivity Prediction for Conformationally Flexible Macrocyclic Molecules ... 50

4.2 Application of the Predictive Model to a Simplified Model Based on the Crystal Structure ... 54

4.3 Conclusions 59

Appendix A Code for performing quality threshold clustering 62

Appendix B Data Set.. 103

Bibliography 122

List of Tables

Table 3-1 List of substrate descriptors.. 28

Table 3-2 Use Akaike information criterion in the forward selection of descriptors........... 39

Table 4-1 Results of the crystal structure... 55

Appendix Table 1 Training and validation sets splitting results using the Kennard–Stone algorithm... 103

Appendix Table 2 Computed electronic descriptors 110

Appendix Table 3 Computed steric descriptors... 114

Appendix Table 4 Relaxed force constant, aSASA descriptor, and free energy of activation calculated by DFT 117

List of Figures

Figure 1-1 Different approaches of site-selective C–H hydroxylation 2

Figure 1-2 DMDO-mediated site-selective C–H hydroxylation 5

Figure 1-3 Experimental results of TFDO-mediated C–H hydroxylation 7

Figure 1-4 DMDO-mediated C–H hydroxylation is electrophilic. 9

Figure 1-5 DMDO-mediated C–H hydroxylation of (*R*)-2-phenylbutane 10

Figure 1-6 Mechanism of DMDO-mediated C–H hydroxylation 11

Figure 1-7 Bimodal correlations between Δ*H*‡ and BDE in DMDO-mediated C–H hydroxylation of relatively small substrates. 12

Figure 1-8 Steric effects in DMDO-mediated C–H hydroxylation on derivatives of bile acids 14

Figure 1-9 Strain-release effects promote the DMDO-mediated C–H hydroxylation of cyclohexane derivatives with axial substituents. 15

Figure 1-10 Recent applications of predictive models for C–H functionalization 18

Figure 1-11 Recent applications of predictive models for other reactions 19

Figure 1-12 Some commonly used steric descriptors in predictive models 21

Figure 2-1 Illustration of solvent accessible surface area (SASA) calculations. 24

Figure 3-1 Data set for developing a reactivity and selectivity model of DMDO-mediated C–H hydroxylation 26

Figure 3-2 Correlation matrix of descriptors from the data set 30

Figure 3-3 Chemical space of the data set 31

Figure 3-4 Univariate linear regression between electronic descriptor and free energy of activation 35

Figure 3-5 Univariate linear regression between steric descriptor and free energy of activation 36

Figure 3-6 Univariate linear regression between relaxed force constant and free energy of activation 37

Figure 3-7 Performance of two-, three-, and four-descriptor linear regression models with descriptors chosen from forward selection stepwise regression 40

Figure 3-8 Characteristics of the transition state in HAT step of DMDO-mediated C–H hydroxylation 42

Figure 3-9 Location of dividing line determined by piecewise linear regression between SASA and free energy of activation 44

Figure 3-10 Curve of the sigmoid function and location of the dividing line. 46

Figure 3-11 Activated solvent accessible surface area (aSASA) descriptor 47

Figure 3-12 Comparison of the performance of three-descriptor linear regression models with the standard SASA descriptor and with the aSASA descriptor using sigmoid activation function. 48

Figure 3-13 Performance of two models with different dividing lines 49

Figure 4-1 Bryostatin 1 and one of its analogues under study 51

Figure 4-2 DMDO-mediated C–H hydroxylation on analogues of bryostatin 1 52

Figure 4-3 Intramolecular hydrogen bonds in the crystal structure of bryostatin 1 53

Figure 4-4 A simplified model of the bryostatin analogue 1a used in computational study 54

Figure 4-5 Performance of the three-descriptor regression model for site-selectivity of DMDO-mediated C–H hydroxylation of a bryostatin analogue.. 57

Preface

I want to thank my advisor Peng Liu for kind support during my graduate years. Without his guidance, I would not be able to complete the graduate work. I want to thank my parents for appreciating my efforts. I want to thank group members who inspire me a lot and add color to my life during my graduate years. I want to thank my friends here at Pitt that have a lot of discussion with me. I want to thank my friend Yi Shi at Purdue University. I want to thank Ye-Cheng Wang and his advisor Mingji Dai for a fruitful collaboration between computation and experiments. I want to thank my undergraduate advisor Yong Liang and his colleague Zhongyue (John) Yang for leading me into the area of computational chemistry. I want to thank John Urbanic at Pittsburgh Supercomputing Center for offering workshops on high performance computing that helped me a lot. I want to thank Dr. Rob Coalson for lecturing statistical mechanics and being one of the committee members. I want to thank Dr. Paul Floreancig for guiding me in grading assignments and being one of the committee members. I want to thank Dr. John Keith for being one of the committee members. I want to thank the journal *Crux Mathematicorum*, which is delightful to read during my spare time.

1.0 Introduction

1.1 DMDO-Mediated Late-Stage C–H Hydroxylation

C–H hydroxylation is an important tool in late-stage diversification of drugs[1] and simulation of drug metabolism.[2] The hydroxylation reaction converts C–H bonds into C–OH bonds, which can alter the solubility and polarity of molecules,[3] make it easier to construct more complex core backbones, and create sites for further downstream functionalization, such as glycosylation. Organic oxidants,[4–6] transition metal catalysts,[7] electrochemical methods,[8,9] and enzymes[10] have been employed in this type of transformation. The control of site selectivity in C–H hydroxylation is being actively pursued by researchers because the chemical properties of different C–H bonds in a natural product can have little difference. Site-selective C–H hydroxylation can reduce protecting and deprotecting stages in synthetic sequences and reduce the usage of toxic or hazardous reagents. Different methods have been developed to perform selective C–H hydroxylation[5,7,9,10] while factors controlling site-selectivity of different C–H hydroxylation reactions are often unclear.

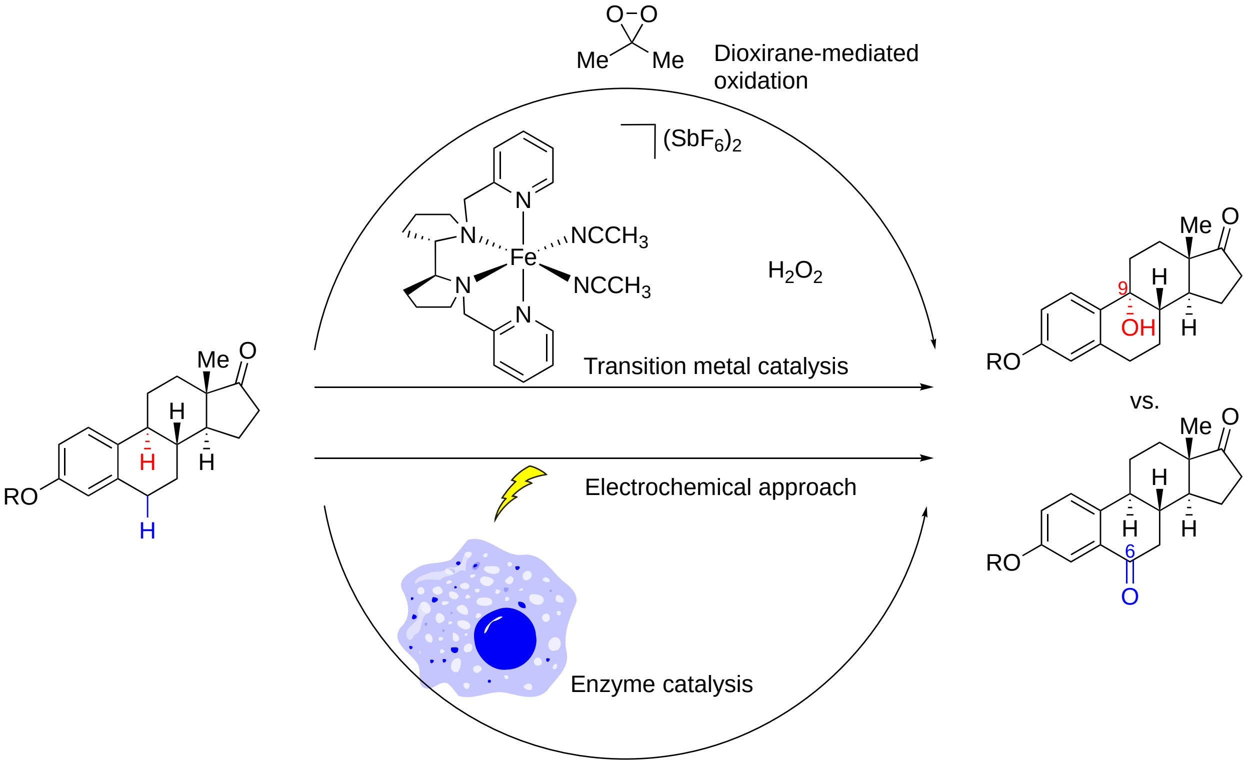


Figure 1-1 Different approaches of site-selective C–H hydroxylation

Dimethyldioxirane (DMDO) is a potent and easy-to-use organic oxidant capable of performing selective C–H hydroxylation. Previous experimental studies have showed that DMDO can selectively oxidize one or two C–H bonds in a complex substrate.[11–13] This makes DMDO-mediated oxidation a useful tool for drug diversification. DMDO can be prepared from acetone and $KHSO_5$, which are inexpensive starting materials.[14] Also, the only side product from DMDO-mediated C–H hydroxylation is acetone, which makes the disposal after experiments quite easy.

DMDO-mediated C–H hydroxylation often favors C–H bonds with low bond dissociation energies (BDEs), including tertiary alkyl C–H bonds (BDE $\approx$ 96 kcal mol^{-1}),[15] ether α C–H bonds (BDE $\approx$ 96 kcal mol^{-1}),[16] and acetal α C–H bonds (BDE $\approx$ 92 kcal mol^{-1}).[17] However, other factors alter the selectivity. In DMDO-mediated oxidation of tetrahydrofuran (Figure 1-2a), the ether α C–H bonds were selectively oxidized in both major and minor products. The minor product is the result of C–H hydroxylation while the major product is the result of overoxidation of the ɑ-C–H hydroxylation product.[18] Here, the site-selectivity can be explained by BDE alone. The BDE of ether α C–H bonds in tetrahydrofuran is lower than that of typical alkyl C–H bonds due to the stabilization of the radical center by the lone-pair electrons of the neighboring oxygen atom in the radical intermediate formed after the homolysis of an ether α C–H bond. In the DMDO-mediated oxidation of 3-methyltetrahydropyran (Figure 1-2b),[18] no C–H hydroxylation product was observed at the tertiary C3 site. The major product is a lactone formed via overoxidation at the C6 site. The minor product is formed via overoxidation at the C2 site. Among the two electronically activated ether α C–H bonds, the C6–H bond is favored over C2–H possibly because of less steric clash with the C3-methyl group. DMDO-mediated oxidation of 5β-androstan-3α-17β-diacetoxy (Figure 1-2c)[19] favors the C5–H hydroxylation. Here, C5–H is less sterically hindered because the substrate is a 5β form of steroid, which might contribute to the site-selectivity. By contrast, in the DMDO-mediated oxidation of a stereoisomer of the steroid, 5α-androstan-3β-17β-diacetoxy (Figure 1-2d), only C14–H hydroxylation occurred.[20] In the DMDO-mediated oxidation of estrone acetate (Figure 1-2e),[4] the substrate underwent selective C9–H hydroxylation followed by dehydration, which led to a $\Delta^{9,11}$ unsaturated derivative. Here, the C9–H, a tertiary benzylic C–H bond, is more electronically activated than C6–H, a secondary benzylic C–H bond, and other tertiary alkyl C–H bonds in the same substrate. In the DMDO-mediated oxidation of a cholestane derivative (Figure 1-2f), only C25–H hydroxylation occurred.[21] There are six tertiary C–H bonds with similar electronic properties in this substrate, while only one of them was oxidized. This example demonstrated that it can be difficult to predict the selectivity prior to experiments. When tigogenin acetate was submitted to DMDO (Figure 1-2g), only the tertiary ether α C–H bond, C16–H, is hydroxylated.[22] When a bryostatin analogue was submitted to DMDO (Figure 1-2h), only C9–H hydroxylation occurred.[23] The macrocyclic substrate can adopt multiple conformations in solution, making it even more challenging to understand the steric properties of different C–H bonds. The selectivity of this conformationally substrate will be investigated computationally in Chapter 4. The reactive C9–H is one of the four tertiary ether α C–H bonds. C15–H, a tertiary acetal α C–H bond, which is also an allylic C–H bond, is expected to be even more electronically activated because its BDE is expected to be significantly lower than that of C9–H. However, C15–H was not oxidized in the experiment, which indicated that not only electronic effects but also steric effects play an important role in DMDO-mediated C–H hydroxylation.

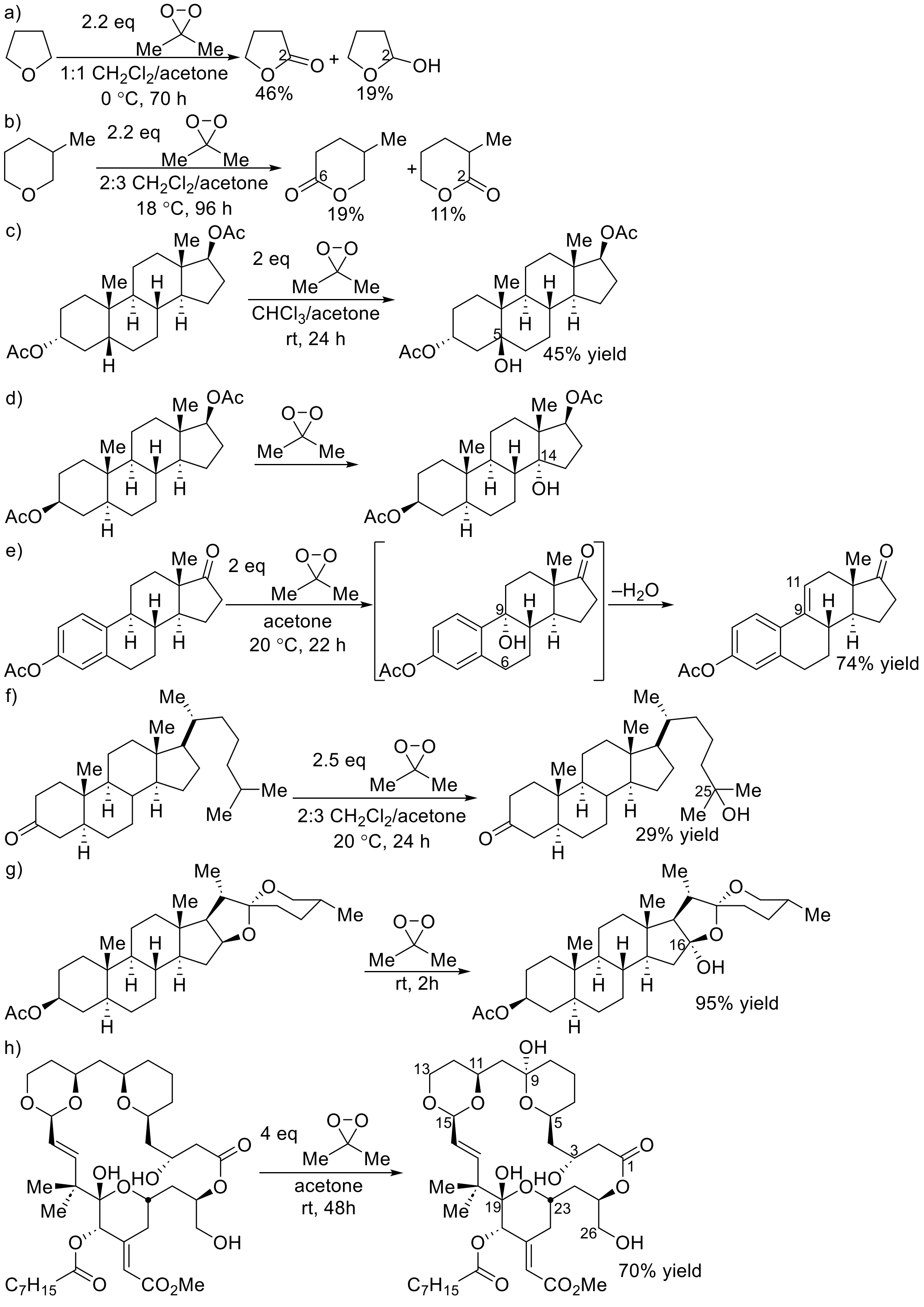


Figure 1-2 DMDO-mediated site-selective C–H hydroxylation

Methyl(trifluoromethyl)dioxirane (TFDO) is an alternative oxidant used in C–H hydroxylation and often shares similar site-selectivity as in DMDO-mediated reactions. In the TFDO-mediated oxidation of tetrahydrofuran (Figure 1-3a),[18] ether α C–H bonds were oxidized in both major and minor products, yielding similar site-selectivity as in the DMDO-mediated C–H hydroxylation. In the TFDO-mediated oxidation of 3-methyltetrahydropyran (Figure 1-3b),[18] the major product is formed via overoxidation at the C6 site, whereas the minor product is formed via overoxidation at the C2 site. The site-selectivity favoring the less sterically hindered C-H bond is similar to that of DMDO. In the TFDO-mediated oxidation of estrone triflate (Figure 1-3c),[9] the substrate initially underwent selective C9–H hydroxylation, the same site in the DMDO-mediate reaction, followed by dehydration to afford a $\Delta^{9,11}$ unsaturated derivative. With the more reactive TFDO, the $\Delta^{9,11}$ unsaturated derivative subsequently underwent alkene epoxidation and 1,2-hydride shift to get the major product with a C=O double bond installed on the C11 position. In the TFDO-mediated oxidation of a cholestane derivative (Figure 1-3d), only C25–H hydroxylation occurred, which is the same site that was hydroxylated with DMDO.[21] Similarly, in the TFDO-mediated oxidation of tetraacetyl-brassinolide (Figure 1-3e), only the exocyclic tertiary C–H bond (C25–H) is hydroxylated.[24] The TFDO-mediated selective C–H hydroxylation has been applied in a synthetic route towards (+)-phorbol (Figure 1-3f). A fused tetracyclic intermediate was submitted to TFDO and only a secondary C12–H hydroxylation occurred,[25] which is vicinal to the cyclopropane ring. Here, strain-release effect might come into play because the ring strain in the fused [6,3]-cyclic system may be partially released in the transiently produced radical intermediate.

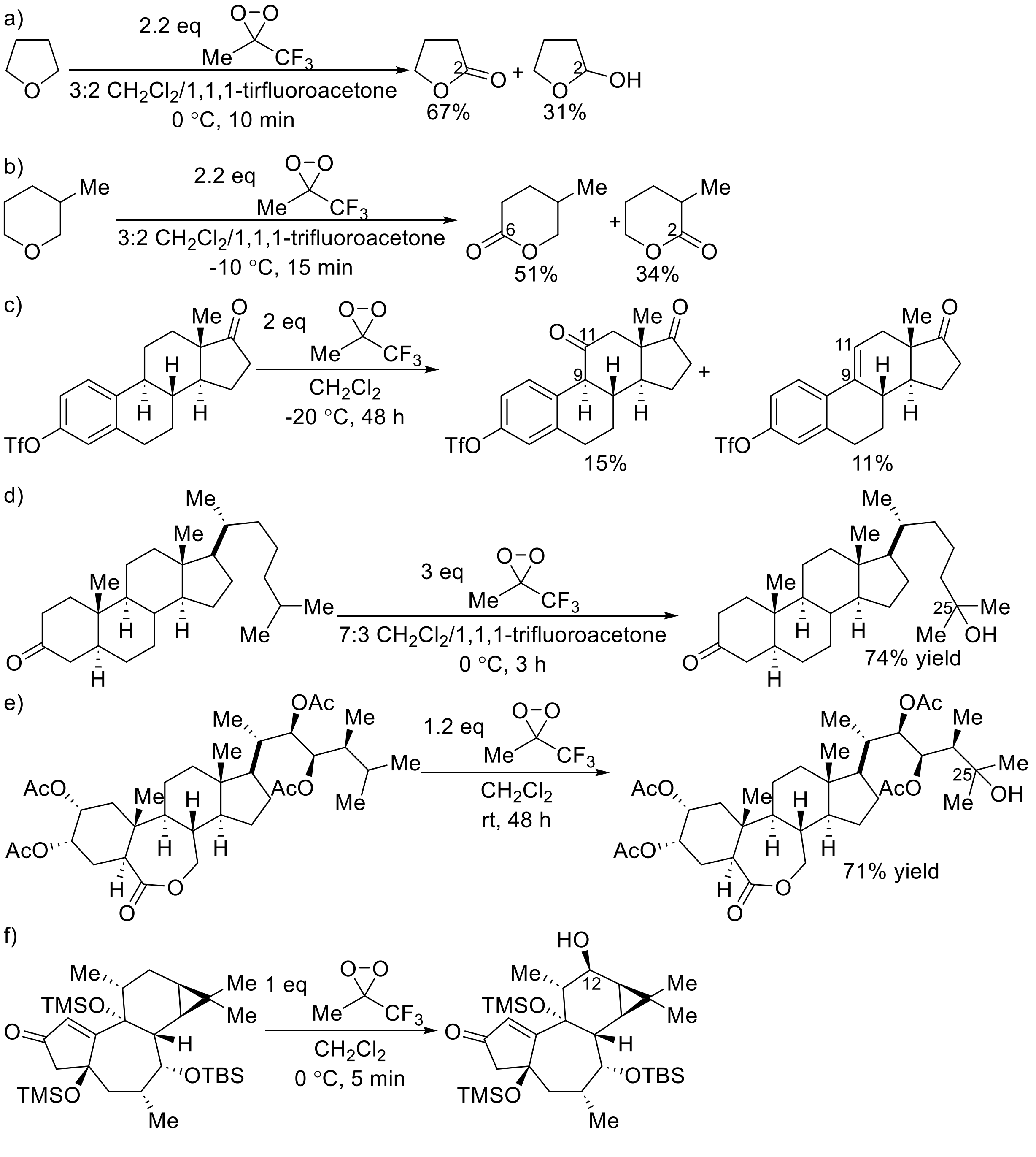


Figure 1-3 Experimental results of TFDO-mediated C–H hydroxylation

Taken together, the previous experimental results indicated that DMDO and TFDO gave similar site-selectivity in C–H hydroxylation of various structurally complex scaffolds, although

TFDO is often more reactive, leading to higher yields, and occasionally, more byproducts via alkene epoxidation. The site-selectivity in both DMDO- and TFDO-mediated reactions appears to be controlled by multiple factors, including electronic effects, strain-release effects, and steric effects.

1.2 Mechanistic Insights from Experimental and Computational Studies on DMDO-Mediated C–H Hydroxylation

Previous experimental results indicated that DMDO-mediated C–H hydroxylation is an electrophilic process. Isotope labeling experiments were performed by preparing ^{17}O-enriched DMDO using ^{17}O-labeled acetone.[26] The ^{17}O chemical shift was 302 ppm, which is the result of large deshielding. The NMR results suggested that there is a low-lying unoccupied σ^*(O–O) orbital as LUMO and a high-lying occupied π^*(O–O) orbital as HOMO.[27] The π^*(O–O) orbital is higher in energy because of coplanarity of lone pairs on the oxygen atoms enforced by the three-membered ring. The three-membered ring also leads to depleted electron density in the direction of the O–O bond, which weakens the O–O bond and promotes the reaction with electron-rich C–H bonds. The reaction rates of DMDO-mediated reactions with a series of *para*-substituted cumenes were measured.[28] The Hammett plot gave a large negative ρ value of –2.76, which further confirmed the electrophilic nature of the DMDO-mediated C–H hydroxylation.

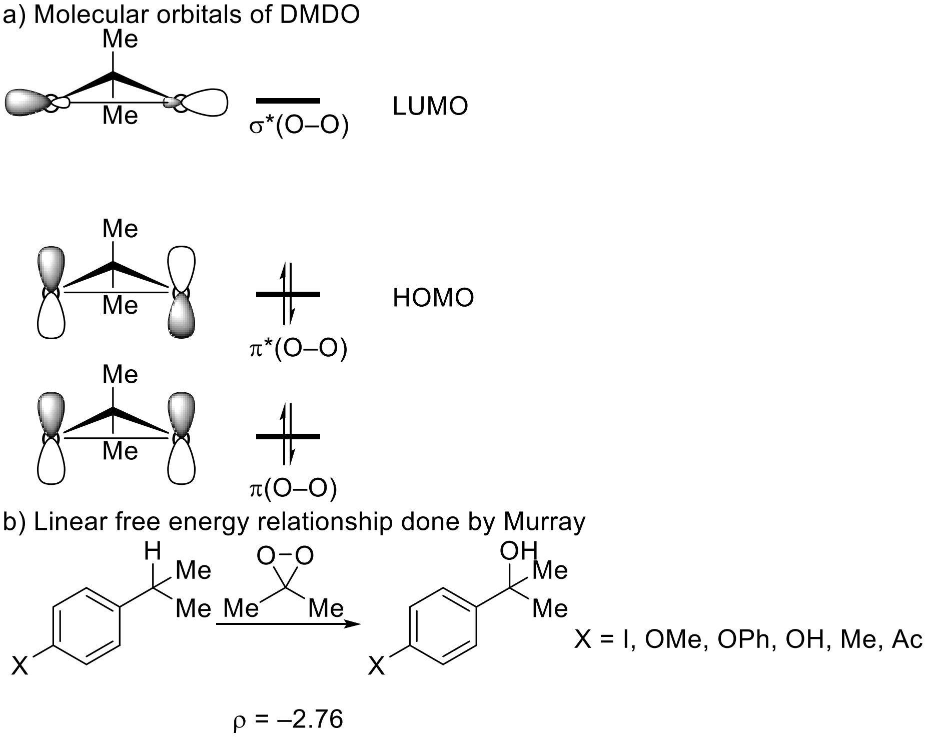

a) Molecular orbitals of DMDO
Me
Me
σ*(O–O)
LUMO
Me
Me
π*(O–O)
HOMO
Me
Me
π(O–O)
b) Linear free energy relationship done by Murray
H
Me
Me
X
O–O
Me
Me
OH
Me
Me
X
X = I, OMe, OPh, OH, Me, Ac
ρ = –2.76

Figure 1-4 DMDO-mediated C–H hydroxylation is electrophilic.

Some experimental results of DMDO-mediated C–H hydroxylation can exclude the mechanism involving long-lived free radicals. In the DMDO-mediated hydroxylation of enantioenriched (*R*)-2-phenylbutane (Figure 1-5), (*S*)-2-phenylbutan-2-ol was obtained with complete stereorentention.[29] The rate constant of the racemization of radicals derived from enantioenriched chiral substrates was estimated to be greater than or equal to 10^{12} s^{-1}, which is already very fast. If DMDO-mediated C–H hydroxylation proceeded via a mechanism involving long-lived free radicals, loss of enantiomeric excess (*ee*) should have been observed. The stereoretention shown in Figure 1-5 is consistent with a mechanism involving an intimate radical pair, which quickly collapses once generated. Another plausible mechanism is a concerted C–H insertion process without radical intermediates. DMDO-mediated C–H hydroxylation on cyclododecane and deuterated cyclododecane showed $k_H/k_D = 4.97$. The primary kinetic isotope effect indicated that the cleavage of C–H bond happens during the rate-determining step of this reaction.[30]

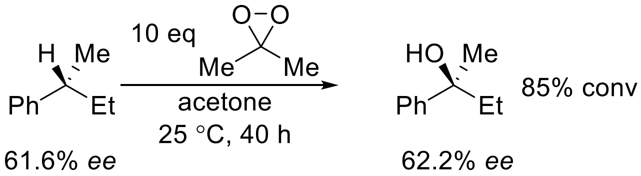


Figure 1-5 DMDO-mediated C–H hydroxylation of (*R*)-2-phenylbutane

Previous computational studies from density functional theory (DFT) calculations and multiconfiguration methods, such as CASPT2, support a mechanism involving intimate radical pairs.[31,32] In the commonly accepted mechanism of DMDO-mediated C–H hydroxylation, the first step is hydrogen atom transfer (HAT), which involves a concerted process of homolytic cleavage of both the C–H bond in the substrate and the O–O bond in DMDO with an O–H bond formation, leading to an intimate radical pair of an alkyl radical and an oxygen-centered radical. This is followed by a second, oxygen rebound step, where the C–OH bond in DMDO is homolytically cleaved and the hydroxyl radical is transferred to the alkyl radical derived from the substrate, leading to the C–H hydroxylation product and an acetone molecule as byproduct. The second step (oxygen rebound) is expected to be barrierless in solution.[33] The rate- and selectivity-determining step of DMDO-mediated C–H hydroxylation is HAT. The HAT transition state is an open-shell singlet, which has significant diradical character, making it challenging to compute, especially for structurally complex substrates.

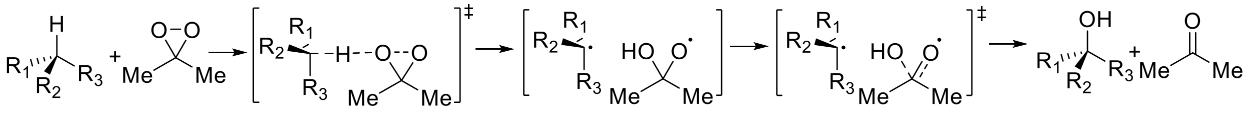


Figure 1-6 Mechanism of DMDO-mediated C–H hydroxylation

To understand the substrate effects on reactivity, the Houk group used DFT to calculate the enthalpies of activation ($\Delta H^{\ddagger}$) for the HAT step in DMDO-mediated C–H hydroxylation of a library of relatively small model substrates (Figure 1-7).[34] Bimodal linear relationships between the computed enthalpy of activation and the BDE of C–H bond were observed.[34] The 26 substrates tested were divided into two groups based on whether there is resonance stabilization of the radical intermediates generated after HAT. Good correlations between the enthalpy of activation and BDE were found within each group, which confirmed that BDE plays an important role in determining reactivity and site-selectivity of DMDO-mediated C–H hydroxylation. However, steric effects were not considered in this study, because the relatively small substrates would not reflect how the steric environment of C–H bonds contributes to the reactivity and site-selectivity in DMDO-mediated hydroxylation.

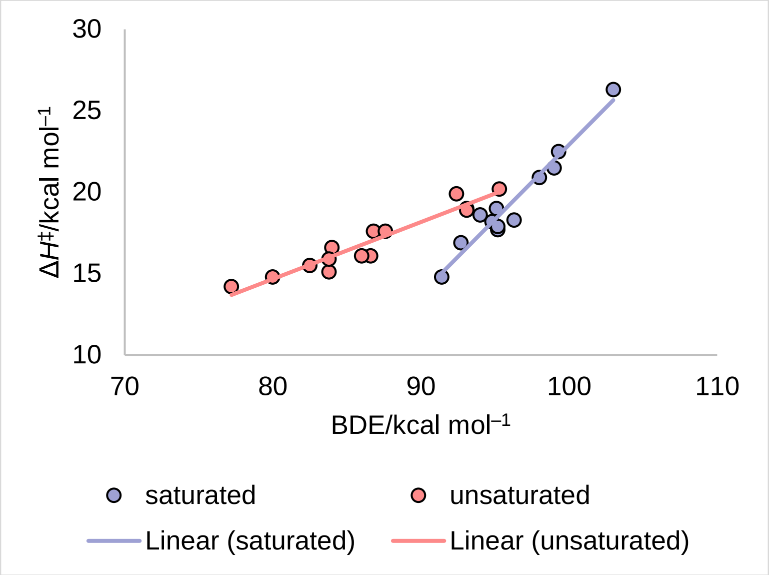


Figure 1-7 Bimodal correlations between Δ*H*‡ and BDE in DMDO-mediated C–H hydroxylation of relatively small substrates. "Saturated" C–H bonds are alkyl C–H bonds not adjacent to any multiple bond or benzene ring. "Unsaturated" C–H bonds are those adjacent to C=C double bond, C=O double bond, C≡N triple bond, or benzene ring.

Besides electronic effects, other factors including steric effects and strain-release effects in DMDO-mediated C–H hydroxylation have only been qualitatively studied. In Figure 1-8a, the methyl ester acetate derivative of lithocholic acid was submitted to DMDO and only 5β-hydroxy product was obtained,[35] which is similar to the site-selectivity of DMDO-mediated C–H hydroxylation of certain 5β steroids.[19] The A and B rings in a 5β steroid are like a *cis*-decalin, where the C5–H is less sterically hindered. In Figure 1-8b, the methyl ester diacetate derivative of chenodeoxycholic acid was submitted to DMDO, in addition to the 5β-hydroxy product, 17α-hydroxy product was also obtained.[35] The acetoxy group on C7 deactivates C5–H via inductive effects while allows competitive C–H hydroxylation on other sterically accessible C–H bonds. In Figure 1-8c, the methyl ester diacetate derivative of ursodeoxycholic acid was submitted to

DMDO, in addition to 5β- and 17α-hydroxy products, 14α-hydroxy product was also obtained.[35] Here, the C14–H in the ursodeoxycholic acid derivative is expected to be less sterically hindered compared to the C14–H in the chenodeoxycholic acid derivative. A previous computational study indicated the computed enthalpies of activation for DMDO-mediated C–H hydroxylation of equatorial tertiary C–H bonds of a series of cyclohexane derivatives decreased with substrates containing more axial methyl substituents (Figures 1-9).[31] This can be explained by the strain-release effect: when the intimate radical pair is formed, the 1,3-diaxial interactions are partially released because the radical center undergoes planarization. The triaxial 1,3,5-trimethylcyclohexane has the greatest 1,3-diaxial interactions, which accounts for its lowest enthalpy of activation.

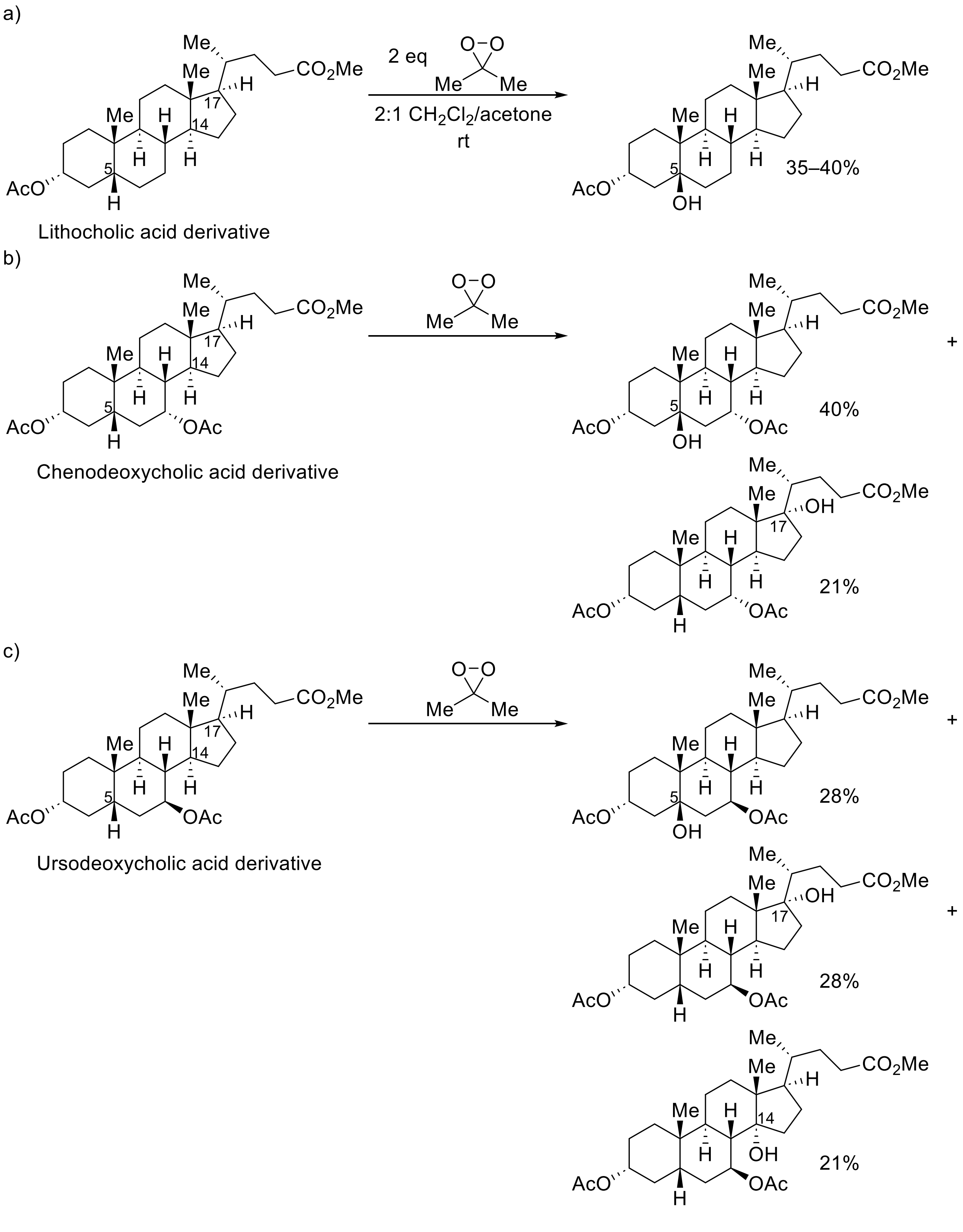

a)
Lithocholic acid derivative
2 eq
Me
Me
2:1 CH_2Cl_2/acetone
rt
35–40%
b)
Chenodeoxycholic acid derivative
40%
21%
c)
Ursodeoxycholic acid derivative
28%
28%
21%

Figure 1-8 Steric effects in DMDO-mediated C–H hydroxylation on derivatives of bile acids

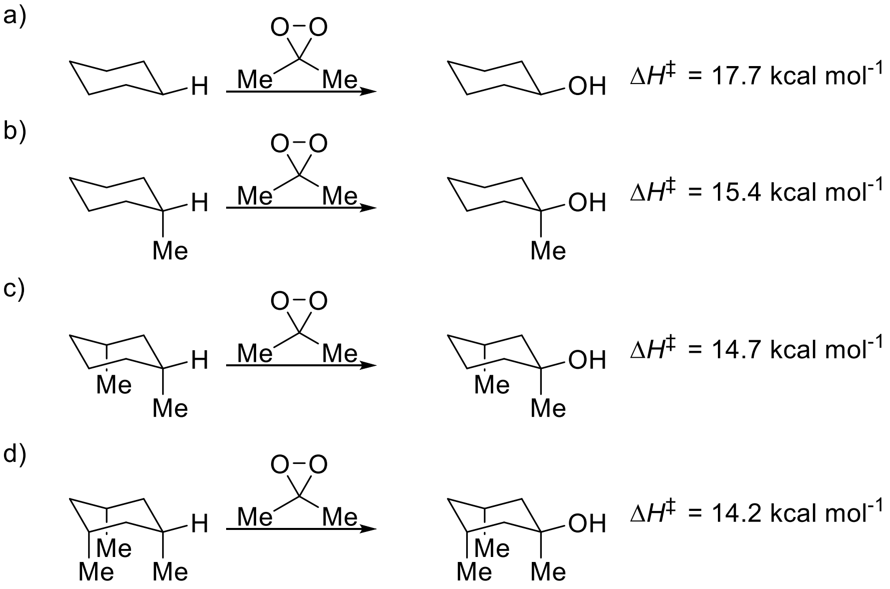
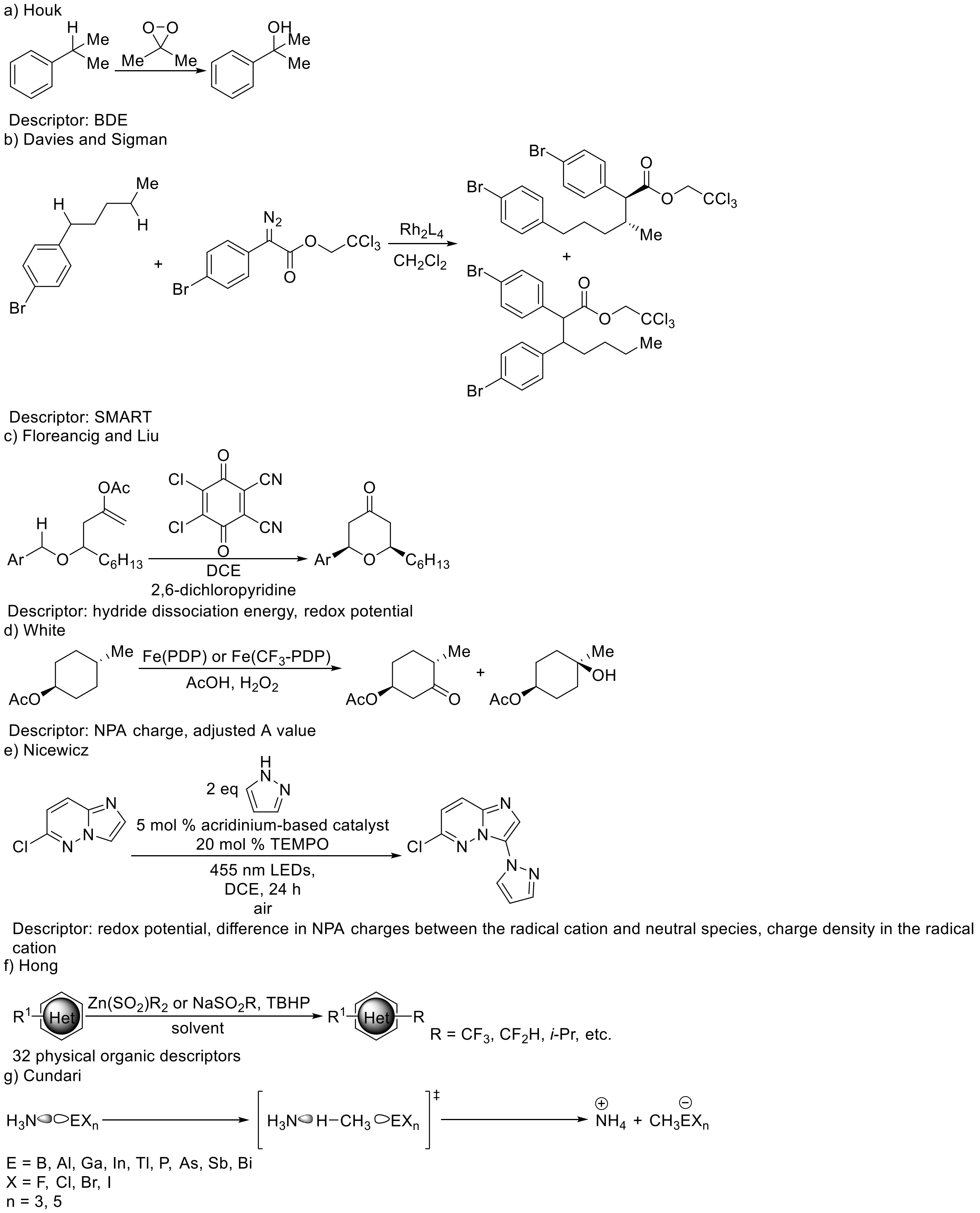


Figure 1-9 Strain-release effects promote the DMDO-mediated C–H hydroxylation of cyclohexane derivatives with axial substituents.

1.3 Predictive Models of Reactivity and Selectivity of Chemical Reactions

Various regression models have been developed to predict reactivity and site-selectivity of different C–H functionalization reactions, where free energy of activation is predicted by relatively easy-to-obtain substrate descriptors.[7,34,36–38] Different regression models, including multivariate linear regression,[36] gaussian process regression,[39] kernel ridge regression,[40–42] and random forest,[43,44] have been used. These models have been applied to various types of reactions, including methane activation by frustrated Lewis pairs,[45] as well as other reaction types, such as nucleophilic aromatic substitution[46] and nucleophilic additions to covalent drugs.[47]

Various linear regression models have been reported for C–H bond functionalization. In these studies, different types of substrate descriptors are used. Houk reported correlations between BDE of C–H bonds and enthalpy of activation for DMDO-mediated C–H hydroxylation in some small substrates (Figure 1-10a).[34] Davies and Sigman developed a SMART (Spatial Molding for Approachable Rigid Targets) descriptor to quantify spatial constraint in C–H functionalization of 1-bromo-4-pentylbenzene via different dirhodium catalysts (Figure 1-10b).[36] They utilized multivariate linear regression to predict the relative free energy of activation using multiple descriptors including the SMART descriptor. Floreancig and Liu found that carbocation stability and electrostatic attraction are two important factors on the reactivity of oxidative C–H functionalization with 2,3-dichloro-5,6-dicyano-1,4-benzoquinone (DDQ).[37] They used substrate hydride dissociation energy to quantify the contribution of carbocation stability and substrate redox potential to quantify the contribution of electrostatic attraction (Figure 1-10c). They used multivariate linear regression to develop a predictive model for reactivity of DDQ-mediated C–H functionalization. White studied the C–H hydroxylation using non-heme iron catalysts (Figure 1-10d).[7] Her group used natural population analysis (NPA) charge, adjusted A value, and higher order terms to model the relative free energy of activation. Nicewicz studied the site-selectivity of photoredox-mediated aryl and heteroaryl C–H functionalization (Figure 1-10e).[38] His group employed redox potential, the difference in NPA charges between the radical cation and neutral species, and charge density in the radical cation to predict site-selectivity and successfully summarized the rules they found in a knowledge-based flowchart. Hong developed a random forest model with 32 physical organic descriptors to predict site-selectivity of radical C–H functionalization of heterocycles (Figure 1-10f).[48] Cundari used neural network to predict the free energy of activation for methane activation by frustrated Lewis pairs (Figure 1-10g).[45] Regression models have also been widely applied to study other types of reactions besides C–H functionalization.[49] Buttar used gaussian process regression to predict the free energy of activation of nucleophilic aromatic substitution (S_NAr) reactions based on some descriptors (Figure 1-11a).[46] Lilienfeld utilized kernel ridge regression to predict free energy of activation for S_N2 reaction of non-aromatic substrates (Figure 1-11b).[50] Researchers from Boehringer Ingelheim utilized the extremely randomized trees algorithm (Figure 1-11c),[51] which is similar to random forest but performs random splits and does not require bootstrap, to predict the free energy of activation for the reaction of acrylamides and 2-chloroacetamides with methylthiolate anion (CH_3S^-).[47] Here, methylthiolate anion was used as a mimic of glutathione in biological systems.



a) Houk

Descriptor: BDE

b) Davies and Sigman

Descriptor: SMART

c) Floreancig and Liu

Descriptor: hydride dissociation energy, redox potential

d) White

Descriptor: NPA charge, adjusted A value

e) Nicewicz

Descriptor: redox potential, difference in NPA charges between the radical cation and neutral species, charge density in the radical cation

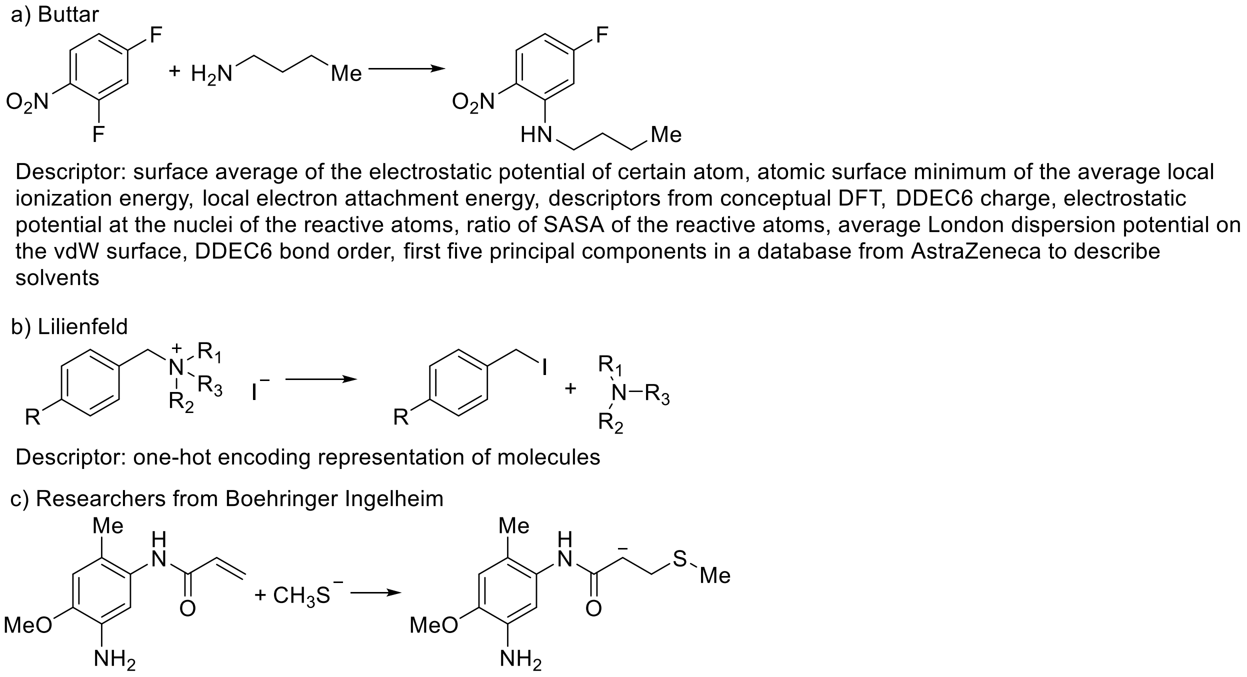
f) Hong

32 physical organic descriptors

g) Cundari

E = B, Al, Ga, In, Tl, P, As, Sb, Bi
X = F, Cl, Br, I
n = 3, 5

Figure 1-10 Recent applications of predictive models for C–H functionalization



Descriptor: surface average of the electrostatic potential of certain atom, atomic surface minimum of the average local ionization energy, local electron attachment energy, descriptors from conceptual DFT, DDEC6 charge, electrostatic potential at the nuclei of the reactive atoms, ratio of SASA of the reactive atoms, average London dispersion potential on the vdW surface, DDEC6 bond order, first five principal components in a database from AstraZeneca to describe solvents

Descriptor: one-hot encoding representation of molecules

Figure 1-11 Recent applications of predictive models for other reactions

Descriptor selection is often a major challenge when developing robust and transferrable predictive models. Employing a smaller number of descriptors in a model is often more desirable because it not only reduces the number of necessary data points to train the model but also improves the interpretability of the model. The Sigman group studied the site-selectivity of oxidative addition in Suzuki reaction where 38 phosphine ligands were tested.[52] The initially established model was complex and difficult to interpret. The initial model was established by using stepwise linear regression and applying certain criterion associated with the number of descriptors. Then the initial model was refined by adding and removing terms manually. This led to a simpler model that was more interpretable. The Doyle group used a random forest model to study palladium-catalyzed Buchwald–Hartwig cross-coupling of 4-methylaniline with aryl halides.[53] They tried to predict reaction yield based on substrate and reagent descriptors. Principal component analysis (PCA) was used to reduce the number of descriptors. The Yu group used a modified ant colony optimization (ACO) algorithm to perform descriptor selection for quantitative structure–activity relationship studies of cyclooxygenase inhibitors.[54]

Reaction-specific descriptor has been shown to improve the robustness and reliability of predictive models and is expected to better describe steric and stereoelectronic effects in some reactions. Although a number of phosphine ligand parameters have been already calculated and incorporated into databases,[55] the existing transferable descriptors alone are often insufficient to predict transition metal-catalyzed reactions. The Schoenebeck group developed reaction-specific descriptors, such as Pd–I–I–Pd dihedral angle and Wiberg bond order between both Pd centers in a complex, and employed them in machine learning algorithms to predict new phosphine ligands that can generate dinuclear $Pd^{(I)}$ species.[56] When studying the selectivity of Pd-catalyzed amination of 3,2- and 5,2- Br/Cl-pyridines, Sigman and Tan found that none of the descriptors from a ligand descriptor database correlate well with the experimentally measured difference in free energy of activation.[57] However, the distance between Pd and Cl in computed structures of L_2PdCl_2 complexes had a positive correlation with the experimental selectivity. Including this rarely used Pd–Cl distance descriptor led to a predictive model of site-selectivity and facilitated the discovery of a new ligand to further improve the selectivity. To analyze electronic and steric effects in first-[58] and second-generation Grubbs olefin metathesis catalysts,[59] Suresh developed descriptors based on molecular electrostatic potential[60] and successfully quantified the relative importance of electronic effects and steric effects in Grubbs olefin metathesis catalysts with his reaction-specific descriptor. Taken together, these previous examples suggest that the development of new substrate descriptors is an important strategy to develop robust and transferrable predictive regression models.

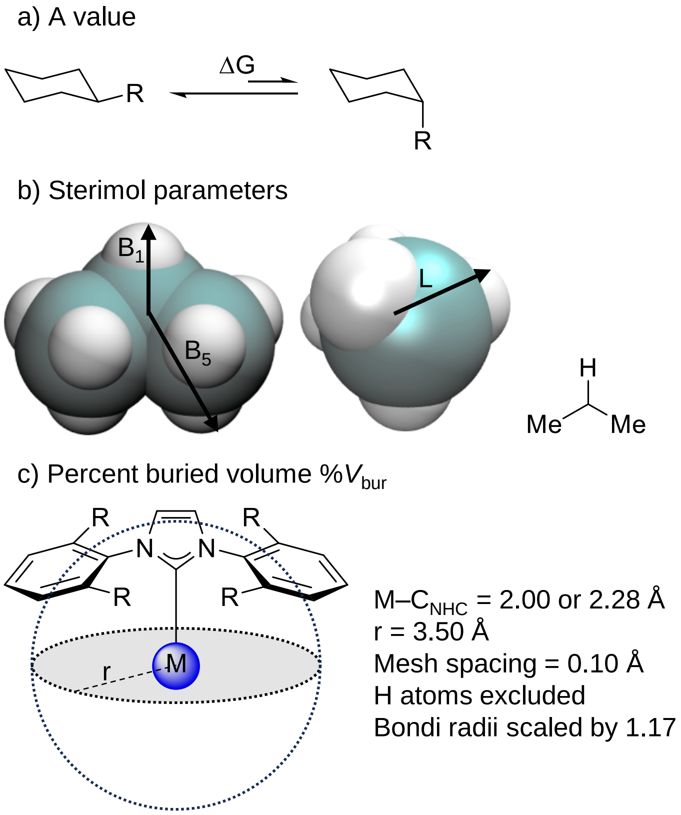


Figure 1-12 Commonly used steric descriptors in predictive models

Sometimes a customized steric descriptor for a specific reaction can significantly improve the quality of the predictive model under development. Some commonly used steric descriptors are A value,[61] Sterimol parameters,[62] and percent buried volume (Figure 1-12).[63] In the previously mentioned collaboration between the Davies group and Sigman group (Figure 1-10b),[36] the researchers initially found that Sterimol parameters,[62] percent buried volume,[63] and some descriptor related to the solid angle[64] showed poor correlations with the experimentally observed site-selectivity since the dirhodium catalysts under investigation featured bowl-shaped pockets. The researchers had to develop steric descriptors tailored to the bowl-shaped pockets since commonly used steric descriptors were not suitable for such catalysts. They docked a macrocyclic thioether molecule to the cavity of the dirhodium catalysts and sampled its conformations. They calculated V_{CAVITY}, which is the volume of the surface enclosing all conformations of the macrocyclic thioether molecule. They obtained N_{PROBE}, which is the number of conformations. They computed entry surface area (ESA) as well. They named the three descriptors SMART descriptors, which stands for Spatial Molding for Approachable Rigid Targets. In their final model after forward selection, which is a form of stepwise regression, the maximum value of V_{CAVITY} was retained along with two electronic descriptors. This study demonstrates that a customized steric descriptor can be valuable in the development of predictive models for certain reactions.

We aim to develop a predictive model of the site-selectivity of DMDO-mediated C–H hydroxylation with some descriptors that are easy to calculate. We prioritize the interpretability of this model. In this way, chemical intuition can guide the development of this model. This model can help the usage of DMDO in organic synthesis. The protocol established during the development of this model is expected to be generalizable to other reactions.

2.0 Computational Methods

2.1 Computational Details of Generating Descriptors and Computing Activation Free Energies for Substrates in the Data Set

All DFT calculations were performed with the Gaussian 09 software.[65] For each substrate in the data set, geometry optimization was performed with the B3LYP[66] functional and 6-31G(d) basis set.[67–76] After geometry optimization, vibrational frequency calculation was performed to make sure that the optimized structures are local minima on the potential energy surface. The Compliance software was used to retrieve the relaxed force constants of C–H bonds from the Hessian matrix in the formatted checkpoint file.[77] Sterimol parameters[62] and solvent accessible surface area (SASA)[78] were calculated from the DFT-optimized geometry using Sterimol.py and VMD[79] respectively. A probe radius of 1.4 Å was used to calculate solvent accessible surface area (SASA). Single-point energy was then calculated with the B3LYP functional[66] and 6-311++G(d,p) basis set in the gas phase in Gaussian.[73,74,80–86] Natural population analysis (NPA) charge,[87] CHELPG charge,[88] and Laplacian bond order[89] were calculated at the same level of theory as the single point energy calculations. BDEs were calculated by the reaction enthalpy of the homolytic cleavage reaction of C–H bond at the B3LYP/6-311++G(d,p)//B3LYP/6-31G(d) level of theory. The free energies of activation of the hydrogen atom transfer step of DMDO-mediated C–H hydroxylation were computed at the B3LYP/6-311++G(d,p)/SMD(acetone)//B3LYP/6-31G(d) level of theory, with geometry optimization of the transition state performed with broken-symmetry DFT with the UB3LYP functional and 6-31G(d) basis set. Vibrational frequency calculations were performed to ensure that the optimized transition state structures are first-order saddle points on the potential energy surface and to get the thermal corrections. Wavefunction stability test was performed for all open-shell singlet transition state structures with the “stable=opt” keyword in Gaussian. Because DMDO-mediated C–H hydroxylation is usually conducted in acetone solvent, when calculating the hydrogen atom transfer barriers, single-point energies were calculated with the SMD solvation model[90] in acetone. The UB3LYP functional and 6-311++G(d,p) basis set were used in the single-point energy calculations.

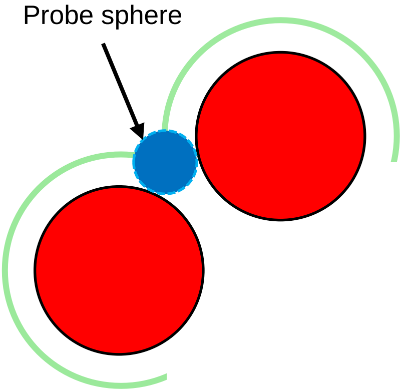


Figure 2-1 Illustration of solvent accessible surface area (SASA) calculations. A probe with radius of 1.4 Å was used in the calculations of the present study.

3.0 Developing a Predictive Model for DMDO-Mediated C–H Hydroxylation

3.1 Data Set Generation and Preliminary Analysis

We established a library of substrates with diverse steric and electronic properties to train and validate the predictive model. The substrates include those used in previous experimental studies of DMDO-mediated C–H hydroxylation as well as some relatively small model substrates. In total, 31 substrates are included in the data set, and in many cases, multiple C–H bonds in one substrate were included. There are 79 C–H bonds in total, including 31 tertiary alkyl C–H bonds, 28 secondary ether α C–H bonds, 14 tertiary ether α C–H bonds, and six acetal α C–H bonds in the data set. C–H bonds known to be unreactive with DMDO due to their high BDEs were not included in the data set. The C–H bonds included in the data set are highlighted in red in Figure 3-1.

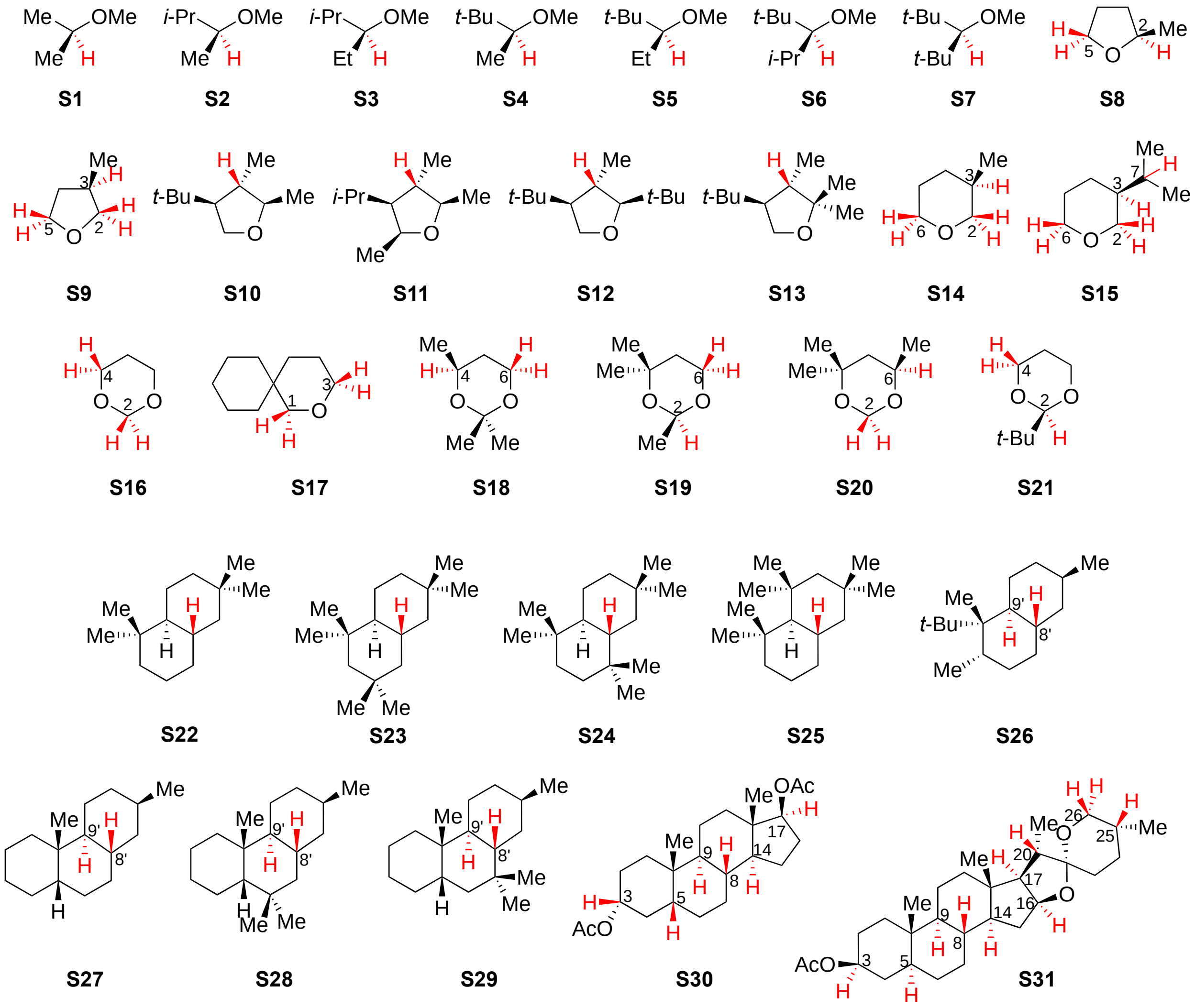


Figure 3-1 Data set for developing a reactivity and selectivity model of DMDO-mediated C–H hydroxylation

For each targeted C–H bond in the data set, we calculated several steric and electronic descriptors (Table 3-1), as well as the free energy of activation for the HAT step in DMDO-mediated C–H hydroxylation. The solvent accessible surface area (SASA) of a C–H bond describes the area of the surface area of the C–H bond that is exposed to a solvent molecule. The SASA of a C–H bond was calculated from the DFT-optimized geometry of the substrate using the "measure" command in VMD on Windows.[79] Both the carbon atom and the hydrogen atom in the

C–H bond were specified. The DFT-optimized geometry of the substrate was also used to calculate Sterimol parameters L, B_1, and B_5 using Sterimol.py.[62] After defining the C–H bond as a primary axis of attachment, L is the total distance following the primary axis of attachment. B_1 is the shortest distance perpendicular from the primary axis of attachment. B_5 is the longest distance perpendicular from the primary axis of attachment. Although SASA and Sterimol parameters both describe the steric properties of a C–H bond, the Sterimol parameters describe the steric property of the alkyl group attached to the hydrogen (*e.g.*, length, width), whereas the SASA describe the degree of exposure of the bond itself. We expect that the subtle differences between SASA and Sterimol parameters may lead to different effectiveness in describing the steric effects in different types of reactions. In the DMDO-mediated C–H hydroxylation, because the O–O bond of the DMDO molecule is nearly collinear with the C–H bond of the substrate in the HAT transition state, we surmised that the SASA of the C–H, may be a more suitable steric parameter than Sterimol parameters.

Several electronic descriptors of the substrate were calculated using single point energies at the B3LYP/6-311++G(d,p) level of theory, including the natural population analysis (NPA) charge[87] and the CHELPG charge[88] of the hydrogen atom in C–H bonds, and the Laplacian bond order of the C–H bond.[89] Natural population analysis is usually linked with rehybridization. NPA charge is derived based on natural atomic orbitals, which are localized. CHELPG charge is derived via fitting to the electrostatic potential at points selected according to certain scheme. Laplacian bond order is the integral of the negative part of the Laplacian of the electron density in fuzzy overlap space multiplied by –10. The C–H bond BDE were also calculated at the B3LYP/6-311++G(d,p)//B3LYP/6-31G(d) level of theory from the reaction enthalpy of the C–H bond homolysis.

Relaxed force constant is a somewhat special descriptor because it describes both electronic and steric properties of a C–H bond.[77] The relaxed force constants were calculated from geometry optimization and frequency calculation of the substrates at the B3LYP/6-31G(d) level of theory, which give the Hessian matrix. A software named Compliance was used to calculate the relaxed force constant from the Hessian matrix.[77]

Table 3-1 List of substrate descriptors

Descriptors for electronic properties	BDE, NPA charge, CHELPG charge, Laplacian bond order
Descriptors for steric properties	SASA, L, B_1, B_5
Descriptors for both eletronic and steric properties	Relaxed force constant

The HAT transition states were computed at the B3LYP/6-311++G(d,p)/SMD(acetone)// B3LYP/6-31G(d) level of theory. Because the HAT transition state is an open-shell singlet, broken-symmetry DFT calculations (UB3LYP) were used in both geometry optimization and single point energy calculations. The initial guess of the wavefunction was generated by setting one fragment composed of the DMDO molecule and the hydrogen atom to be transferred as a doublet and the other fragment composed of remaining atoms as a doublet using "guess(fragment=2)" keyword. The stability of the wavefunction was checked using the "stable=opt" keyword to ensure the most stable open-shell singlet wavefunction was used in both geometry optimization and single point energy calculations.

Next, we analyzed the computed descriptors of the C–H bonds in the data set. First, we constructed a correlation matrix of the descriptors to evaluate if there are any highly correlated descriptors in the data set. The existence of highly correlated descriptors can be detrimental to the performance of the model, which, for example, can be caused by multicollinearity of regression models. Highly correlated descriptors also negatively impact the interpretability of the model being developed. Pearson correlation coefficients between the computed descriptors were used as the elements of the correlation matrix (Figure 3-2). Because the Pearson correlation coefficient does not depend on the order of descriptors in calculation, correlation matrix is always a symmetric matrix, half of which is shown in Figure 3-2. Pearson correlation coefficients range from –1 to 1. If the Pearson correlation coefficient of variables X and Y is –1, the data points lie exactly on a line with a negative slope. If the Pearson correlation coefficient of variables X and Y is 1, the data points lie exactly on a line with a positive slope. If the Pearson correlation coefficient of variables X and Y is 0, there is no linear dependency between X and Y. From the correlation matrix, a relatively strong negative correlation between two steric descriptors B_1 and SASA (r = –0.86) was observed. A moderate positive correlation between SASA and Laplacian bond order (LBO) (r = 0.74) was observed as well. Besides these two correlations, other descriptors in the data set are not strongly correlated. Based on the correlation matrix analysis, all computed descriptors were kept in the subsequent analysis and model development steps.

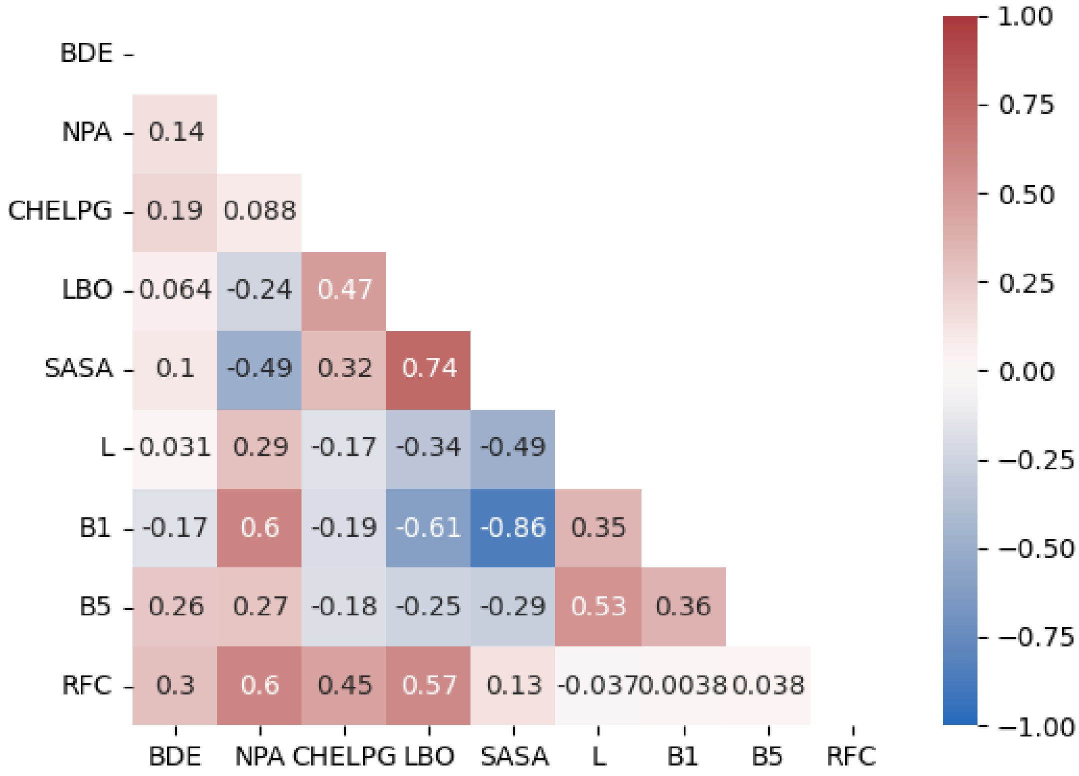


Figure 3-2 Correlation matrix of descriptors from the data set

Next, we analyzed the chemical space of the data set to make sure the electronic and steric properties of substrates in the data set are diverse enough to represent a large chemical space, where the data points are not clustered that may cause bias in the resulting model. The computed values of electronic descriptors exhibit relatively broad ranges for the natural population analysis (NPA) charges of the hydrogen atom in the C–H bond (0.123~0.211) and the BDE of C–H bonds (84.2~96.8 kcal mol^{-1}). This suggests the C–H bonds in the data set represent reasonable ranges of charge density and bond strength. In terms of steric properties, the SASAs of the C–H bonds cover a very broad range from 0.29 Å^2 to 33.17 Å^2, indicating C–H bonds with diverse steric properties, from those are almost completely blocked to completely unhindered, have been included in the data set. We examined the distribution of NPA charge versus SASA and the distribution of BDE versus SASA for all data points (Figure 3-3). The scattered distributions in these plots suggest that the chosen data points are not clustered in the steric and electronic space. Based on these analyses, we consider the range and distribution of the data point satisfactory for subsequent model developments.

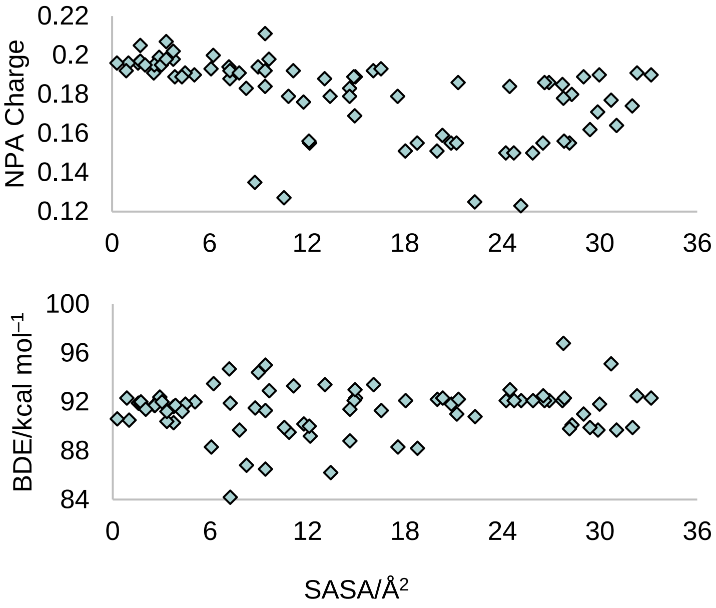


Figure 3-3 Chemical space of the data set

3.2 Single-Descriptor Predictive Models

We explored several different types of predictive models for C–H bond reactivity in DMDO-mediated hydroxylation. First, we considered single-descriptor models, using univariate linear regression that uses only one descriptor to predict the activation free energy. Similar approach was used by Houk in a previous study with a data set of mostly sterically unhindered substrates, where BDE was used as the sole descriptor to develop a predictive model for activation enthalpy.[34]

Before any linear regression is performed, we split the data set into a training set and a validation set. Because our data set is not very large, method for training/validation split may impact the performance of the model. We employed the Kennard–Stone algorithm to split the data set.[91] The Kennard–Stone algorithm can provide a subset that is more representative of the entire data set than a subset from random split. In our data set of 79 C–H bonds, we employed the Kennard–Stone algorithm to extract a training set composed of 59 C–H bonds, which is about 75% of the entire data set. The remaining 20 C–H bonds were used as the validation set.

Next, we performed least squares fitting between each descriptor and the computed free energy of activation ($\Delta G^{\ddagger}_{\mathrm{DFT}}$) of C–H bonds in the training set. The results are plotted in Figures 3-4, 3-5, and 3-6. The performance of each univariate linear regression model was analyzed based on the coefficients of determination (R^2) values shown in the figures. In addition, the slopes of the linear relationships were analyzed because they represent the sensitivity to each descriptor when a correlation with activation free energy is present.

In general, none of the univariate linear regression models gave strong correlation with the $\Delta G^{\ddagger}_{\mathrm{DFT}}$ values. The NPA charge gave the highest coefficient of determination ($R^2 = 0.461$) among all the electronic descriptors. Interestingly, although the NPA charge gave a reasonable correlation, no correlation was observed with the CHELPG charge ($R^2 = 0.0006$), suggesting that this molecular electrostatic potential (MESP)-based charge scheme is not sufficient to describe the electron density effects on reactivity with DMDO. The drastically different performances of NPA and CHELPG charges imply that the electronic effects of DMDO-mediated C–H hydroxylation manifest via charge transfer from the C–H bond to DMDO, rather than affected by non-covalent electrostatic interactions between the C–H bond and the DMDO, because the NPA charge can better describe the charge density of C–H bond and its tendency as electron donor, whereas CHELPG charge better describes non-covalent electrostatic interactions with other polar molecules or point charges.

BDE gave a low coefficient of determination of R^2 = 0.0827, which is in contrast to previous report[34] that suggests BDE alone can serve as a sufficient descriptor for reactivity prediction of relatively small molecules. This result highlighted the significance of including other parameters, especially steric descriptors, in predictive reactivity model for structurally complex substrates.

Although the correlations of the univariate regression models are modest or poor, the slopes of these relationships still provide useful insights into how these individual factors contribute to the reactivity. The positive slope in the fitted equation between BDE and free energy of activation agrees with the previous study by Houk.[34] However, a smaller slope of 0.661 was obtained from the present data set, compared to a slope of 0.91 with “saturated” C–H bonds in the previous study. It should be noted that the “saturate” C–H bonds in the previous study are mostly primary, secondary, and tertiary alkyl C–H bonds, whereas the C–H bonds in the present work include a substantial percentage of ether α C–H bonds. The difference in sensitivity to BDE may be attributed to the different types of C–H bonds included in the data sets, or the inclusion of sterically distinct substrate slightly decreasing the sensitivity to electronic effects. The positive slope in the fitted equation between NPA charge and free energy of activation is expected, because it is consistent with the fact that DMDO-mediated C–H hydroxylation is an electrophilic process where more electron-rich C–H bonds react faster. The slope of this correlation (154) is comparable to those in other electrophilic C–H bond functionalization reactions. For example, in a recent computational study from our group, slopes of 123 to 302 were observed for the correlations between $\Delta G^{‡}$ and NPA charges of hydrogen atoms in the Ag-catalyzed benzylic C–H amination with different types of ligands and nitrene precursors.[92]

The negative slope in the fitted equation between SASA and free energy of activation is also expected, because it is consistent with the fact that more sterically hindered C–H bonds undergo slower reactions. SASA gives the highest coefficient of determination (R^2 = 0.443) among all the steric descriptors. Among the steric parameters, B_1 gives a better correlation with $\Delta G^{‡}_{\mathrm{DFT}}$ than B_5 and L. Because B_1 can be viewed as the minimum width of the alkyl substituent attached to the C–H bond, this result suggests that this reaction is more sensitive to the local steric environment around the C–H bond, rather than longer-range steric repulsions, which can be better described by B_5, which describes the maximum width of the substituent, and L, which describes the length of the substituent. We surmised that the better correlation of SASA compared to Sterimol parameters can also be attributed to the fact that the DMDO-mediated hydroxylation is more sensitive to *local* steric environment around the hydrogen atom, because SASA describes the steric hindrance of the hydrogen atom itself, rather than alkyl substituent.

Among the other descriptors considered Laplacian bond order gives a relatively poor coefficient of determination (R^2 = 0.179), whereas no correlation was observed with between relaxed force constant and free energy of activation (R^2 = 0.0598).

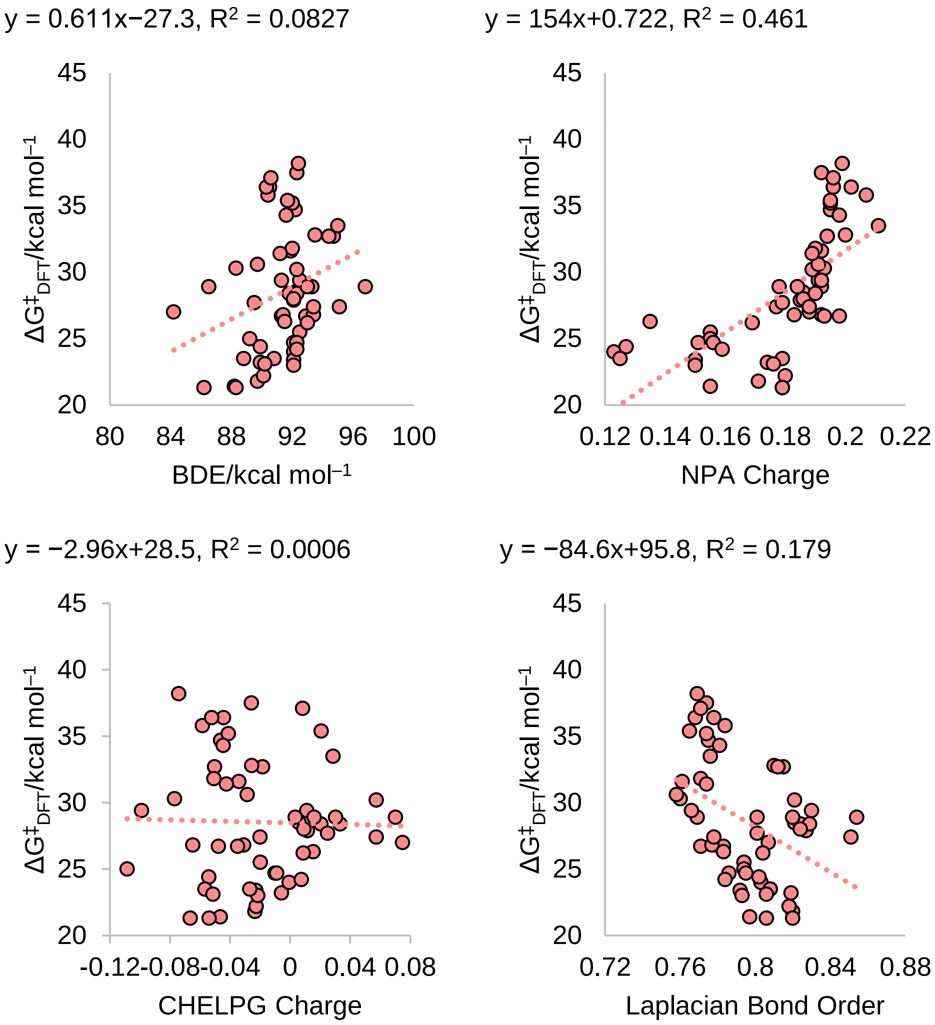


Figure 3-4 Univariate linear regression between electronic descriptor and free energy of activation

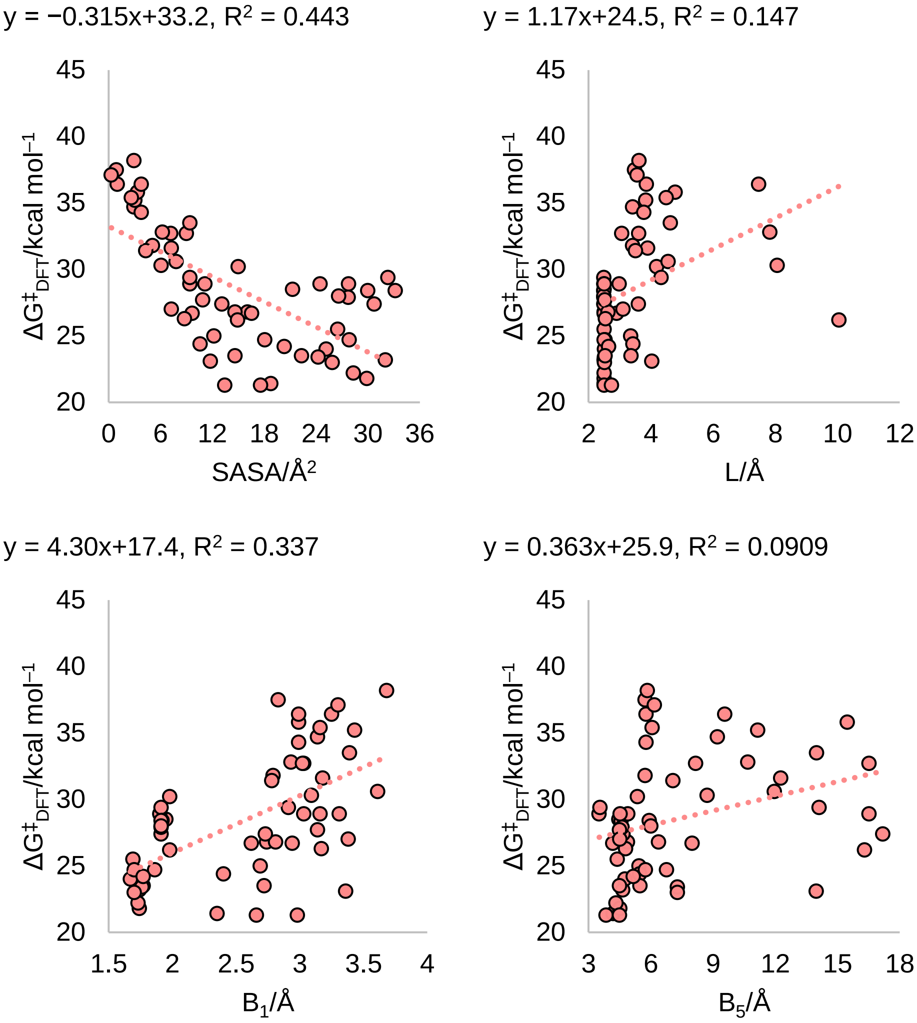


Figure 3-5 Univariate linear regression between steric descriptor and free energy of activation

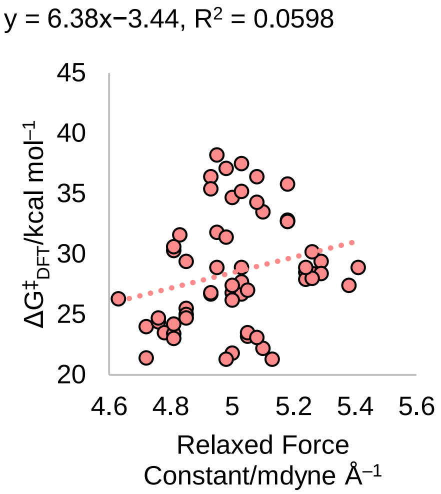


Figure 3-6 Univariate linear regression between relaxed force constant and free energy of activation

Taken together, several useful conclusions can be drawn from the above analysis. First, none of the univariate models provides sufficient accuracy to predict free energy of activation. This strongly suggests the reactivity of the compounds in the data set are affected by more than one factors, and thus a multivariate approach with both electronic and steric descriptors are needed to develop a more accurate model. Second, NPA charge and SASA show the best correlations with free energy of activation, among all electronic and steric descriptors, respectively. The slopes from these correlations are consistent with our chemical intuition.

3.3 Multivariate Linear Regression Models

Next, we explored strategies to develop multivariate linear regression models that uses both electronic and steric descriptors for reactivity prediction. We performed stepwise regression on the training set to select appropriate descriptors from those included in the data set (Table 3-1) and to identify the minimum number of descriptors required for the multivariate regression model. Stepwise regression uses an iterative, step-by-step approach to construct the regression model by evaluating how adding or removing a potential descriptor affects the performance of the model, based on certain criterion. We used the commonly used Akaike information criterion (AIC) with the form of AIC = $2k+n\ ln(RSS/n)–2C$ for model evaluation.[93] In our case, k is equal to the number of descriptors plus one and n is the number of C–H bonds in the training set (n = 59). RSS is the residual sum of squares in least squares fitting. The constant C does not affect the evaluation results and thus is often ignored in model comparisons. Models with smaller AIC values are preferred. For models with comparable AIC values, those with fewer number of descriptors are preferred because they are easier to interpret and may require a small number of training set data points.

We performed forward selection of stepwise regression on the training set using Akaike information criterion (Table 3-2). In the first step, the NPA charge descriptor was selected, giving an AIC of 147.64. In the second step, the SASA descriptor was added while the AIC of the model decreases by 20.84. This suggests an improved performance of the two-descriptor model compared with the one-descriptor model. In the third step, the BDE descriptor was added and the AIC of the model further decreased by 17.63, indicating further improvement over the two-descriptor model with NPA and SASA. In the fourth step, the CHELPG charge descriptor was added. However, the AIC of the model only moderately decreased by 1.22. After that, the algorithm could not lower the AIC of the model by adding another descriptor and stopped. Because the four-descriptor model did not substantially improve the AIC value, we decided to use the three-descriptor model with NPA charge, SASA, and BDE as descriptors. It should be noted that the three chosen descriptors are among those with the best correlations with $\Delta G^{\ddagger}$ in the univariate regression models discussed earlier.

Table 3-2 Use Akaike information criterion in the forward selection of descriptors

Number of steps	Descriptors selected	Akaike information criterion (AIC)
1	NPA charge	147.64
2	NPA charge, SASA	126.80
3	NPA charge, SASA, BDE	109.17
4	NPA charge, SASA, BDE, CHELPG charge	107.95

To validate the descriptors chosen by the forward selection process, we developed three multivariate linear regression models using descriptors chosen in the second, third, and fourth steps of the forward selection, which have two, three, and four descriptors, respectively (Table 3-2). The resulting models from the training set and the predicted values in the validation set are shown in Figure 3-7. All three models gave much improved performances than the univariate regression models shown in Figures 3-4, 3-5, and 3-6. Based on the RMSE (root mean square error) and coefficient of determination (R^2) of the training set data points, the performance of the three-descriptor model is noticeable better than that of the two-descriptor model, whereas the four-descriptor model only shows minimal improvement from the three-descriptor model. Similar trends were observed from the RMSE and R^2 values of validation set data points, where the three-descriptor model gives a lower RMSE and a higher R^2 compared to the two-descriptor model, but the four-descriptor model essentially give the same RMSE and R^2 values as the three-descriptor model. These results further confirmed the results AIC analysis in the forward selection process indicated that the fourth descriptor, CHELPG charge, will not substantially improve the model performance. This is consistent with the poor correlation between CHELPG charge and Δ*G*‡ observed in our univariate model analysis, which is shown in Figure 3-4.

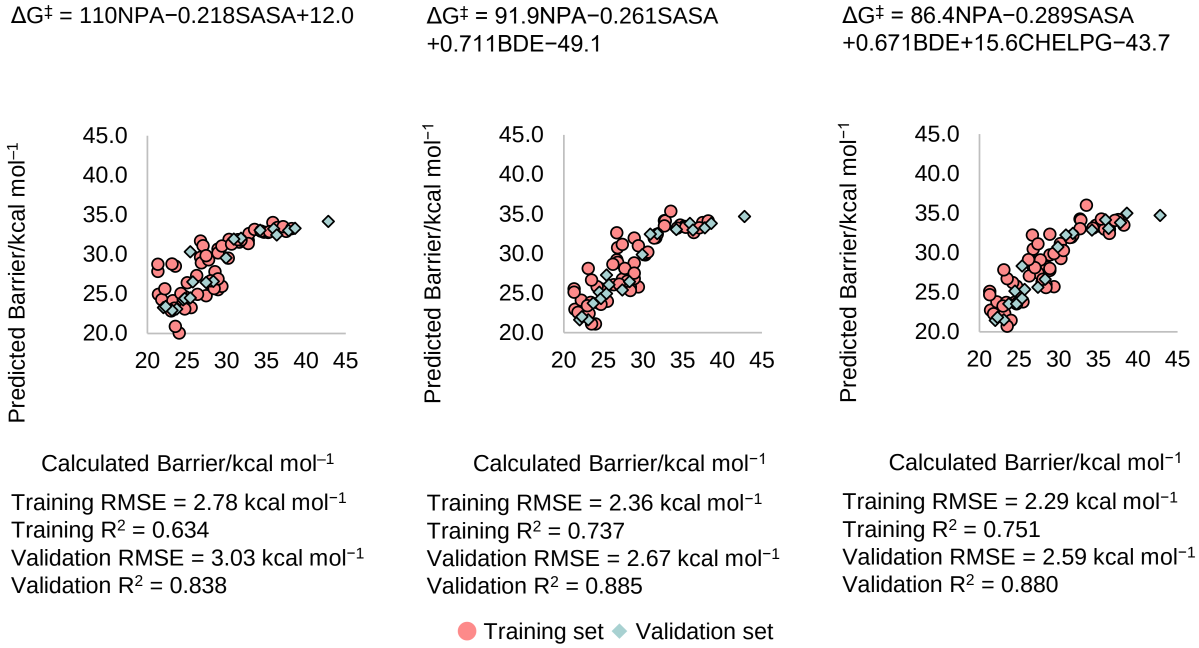


Figure 3-7 Performance of two-, three-, and four-descriptor linear regression models with descriptors chosen from forward selection stepwise regression

Although the three descriptors from the predictive model Δ*G*‡ = 91.9NPA–0.261SASA+0.711BDE–49.1 were identified solely based on statistical analysis, each of the descriptors represents an expected effect that affects the C–H hydroxylation reactivity. In particular, NPA charge describes electronic effects, BDE is affected by both bond strength and strain-release after the formation of intimate radical pair, and SASA describes steric effects. These individual factors have been previously identified or proposed as components contributing to the reactivity and site-selectivity of DMDO-mediated C–H hydroxylation. The coefficients of the three descriptors in the multivariate model are also consistent with the chemical intuition: the positive coefficient for NPA charge is consistent with the electrophilic process, the negative coefficient for SASA is consistent with the expected steric effects that C–H bonds with larger SASA should encounter less steric hindrance, and the positive coefficient for BDE is consistent with the fact that the reaction favors weaker C–H bonds as well as those can alleviate intramolecular steric strain once the planarized radical center is formed.

The identification of SASA as a more effective steric descriptor than the more commonly used Sterimol parameters is also worth noting. We propose that SASA is a particularly effective steric descriptor for the C–H hydroxylation because the C–H hydroxylation involves an *outer-sphere* C–H cleavage mechanism where the DMDO reagent does not form a bond with the carbon atom in the rate- and selectivity-determining HAT transition state (Figure 3-8). This is distinct from many known transition metal-catalyzed C–H activation mechanisms that involve an *inner-sphere* pathway, where a metal–carbon bond is being formed in the transition state. Therefore, we surmise that the SASA of the C–H bond, can be more effective to describe steric effects in outer-sphere C–H activation mechanisms (*e.g.*, in HAT) whereas Sterimol parameters (L, B_1, and B_5), which describe the steric bulk of the alkyl group of the C–H bond, can be generally more effective for inner-sphere C–H activation mechanisms.

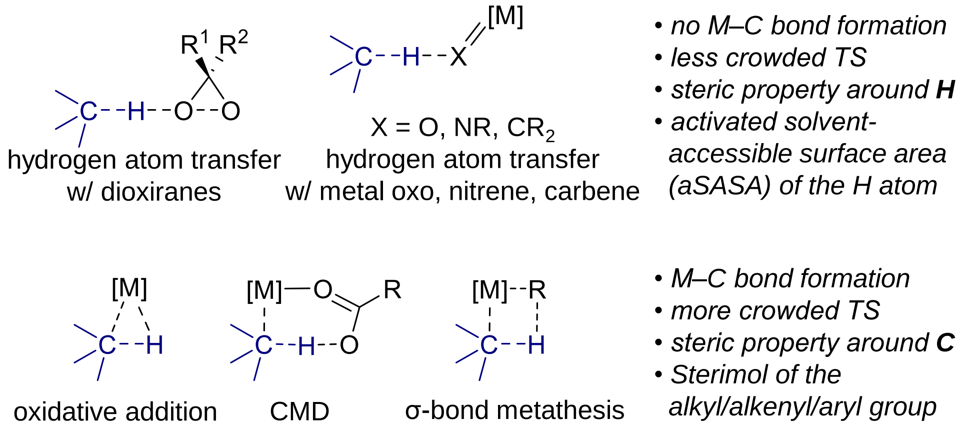

R1 R2
C--H--O--O
hydrogen atom transfer w/ dioxiranes
[M]
C--H--X
X = O, NR, CR2
hydrogen atom transfer w/ metal oxo, nitrene, carbene
• no M–C bond formation
• less crowded TS
• steric property around H
• activated solvent-accessible surface area (aSASA) of the H atom
[M]
C--H
oxidative addition
[M]—O R
C--H--O
CMD
[M]--R
C--H
σ-bond metathesis
• M–C bond formation
• more crowded TS
• steric property around C
• Sterimol of the alkyl/alkenyl/aryl group

Figure 3-8 Characteristics of the transition state in HAT step of DMDO-mediated C–H hydroxylation

In summary, based on the forward selection stepwise regression results, we have identified NPA, SASA, and BDE as three descriptors for multivariate linear regression models. Although these descriptors were chosen based on the statistical analysis of their performances in stepwise regression, they are in the same time chemically meaningful to describe different factors affecting the HAT transition state stability. In subsequent sections, we turned our attention to further refinements of the three-descriptor model based on these three descriptors.

3.4 Using an Activation Function to Improve the Performance of Solvent Accessible Surface Area as a Reaction-Specific Steric Descriptor in Multivariate Regression

Although the three-descriptor model described in the previous section provides reasonable agreement with the DFT-calculated free energies of activation, the RMSE is still relatively large (2.36 and 2.67 kcal/mol for training and validation sets, respectively), which translates to up to *c.a.* 100-fold error in terms of rate constants. We surmised that the challenges of accurate reactivity prediction are partially attributed to the fact that C–H bonds with very diverse steric properties were included in the data set. In particular, we questioned whether the free energy of activation truly maintains a linear relationship with SASA in a relatively broad range (from 0.29 $Å^2$ to 33.17 $Å^2$, see Figure 3-3). We postulate that because of the relatively small size of DMDO, substrates with small to medium degrees of steric hinderance may be less sensitive to steric effect. This hypothesis is supported by previous studies by Houk which indicated electronic parameters alone (*i.e.*, BDE) is sufficient to predict reactivity of C–H bonds in relatively less hindered substrates.[34]

The hypothesis that less hindered and more hindered C–H bonds have different degrees of sensitivities to steric effects is supported by the $\Delta G^{\ddagger}_{\mathrm{DFT}}$ versus SASA plot shown in Figure 3-9. Here, two regimes were observed: a reasonable correlation between $\Delta G^{\ddagger}_{\mathrm{DFT}}$ and SASA was observed for C–H bonds with relatively small SASA values (red points in Figure 3-9), indicating steric effects being a major factor affecting their reactivity, whereas no correlation was observed for C–H bonds with relatively large SASA values (yellow points in Figure 3-9).

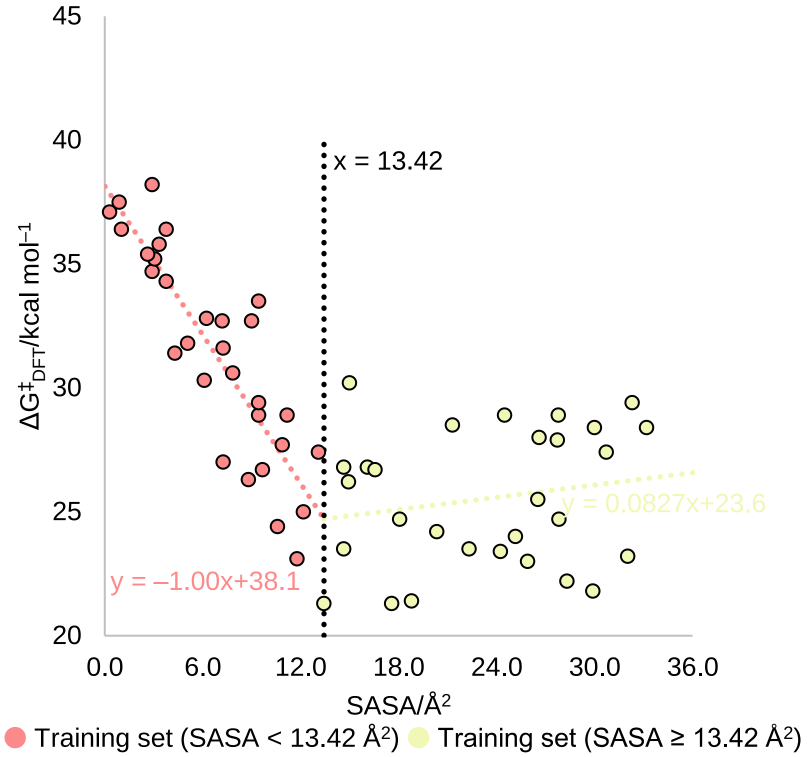


Figure 3-9 Location of dividing line determined by piecewise linear regression between SASA and free energy of activation

The nonlinear relationship between $\Delta G^{\ddagger}_{\mathrm{DFT}}$ and SASA is caused by the different sensitivities to steric effects for more hindered and less hindered C–H bonds,[90] leading to a "growth regime" for more hindered C–H bonds and a "plateau regime" for less hindered C–H bonds. We propose to use an activation function to address this nonlinear relationship. Various forms of activation functions have been used to develop artificial neural networks, but their applications in physical organic chemistry free energy relationships are still limited. Activation functions, such as the sigmoid function (Figure 3-11), can convert a linear input to a nonlinear output. Here, we propose to use the sigmoid function because it provides a growth regime preceding the plateau regime, which matches the type of nonlinearity for SASA, where smaller SASA leads to a growth regime (*i.e.*, more sensitive) and larger SASA leads to a plateau regime (*i.e.*, less sensitive). The sigmoid function has already been used in modeling autocatalytic reactions,[94] modeling Brønsted coefficient of certain reaction,[95] and neural networks.[96,97] The genuine sigmoid function has the form of

$$sigmoid(x) = \frac{1}{1 + e^{-x}}$$

The curve of sigmoid function has an inflection point at (0, 0.5). When the input value is smaller than zero, the slope of sigmoid function increases as the input value increases. When the input value is greater than zero, the slope of sigmoid function decreases as the input value increases. The tangent to the curve of sigmoid function at the inflection point (0, 0.5) intersects one asymptote $y=1$ at the point (2, 1). Thus, $x=2$ acts as a dividing line. More generally, the tangent to the curve of 1/(1+exp(–x/(0.5d))) at the inflection point (0, 0.5) intersects one asymptote $y=1$ at the point (d, 1). In this case, $x=d$ acts as a dividing line (Figure 3-10). This dividing line separates the growth and plateau regimes of the sigmoid function. Therefore, when designing a sigmoid function for SASA, a reasonable position of the dividing line (d) separating the growth and plateau regimes, needs to be chosen. We used the data in the training set to locate the dividing line (d). From the nonlinear relationship observed in the scatter plot of $\Delta G^{\ddagger}_{\mathrm{DFT}}$ and SASA of C–H bonds (Figure 3-9), we performed piecewise linear regression between $\Delta G^{\ddagger}_{\mathrm{DFT}}$ and SASA with a *R* package named *segmented* to detect the change point.[98–102] Piecewise linear regression has been used to model the relationship between biological activity and certain substituent constants of 2-furylethylene compounds.[103] The change point (*i.e.*, dividing line, d) is calculated to be 13.42 Å^2 with the standard error being 1.179 Å^2 in our case. The multiple coefficient of determination is 0.735. Using d = 13.42 Å^2 to separate the growth (red) and plateau (yellow) regimes, we obtained very different slopes before and after the change point (Figure 3-9). For all C–H bonds with SASA smaller than 13.42 Å^2 (*i.e.*, in the growth regime), the slope is –1. For those with SASA larger than 13.42 Å^2 (*i.e.*, in the plateau regime), the slope is close to zero.

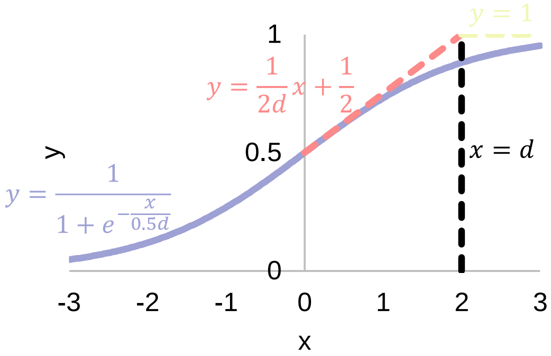


Figure 3-10 Curve of the sigmoid function and location of the dividing line. Here, *d* is equal to two. The red line is the tangent at the inflection point. The yellow line is an asymptote.

Using the change point detected via piecewise linear regression, we propose a sigmoid activation function for SASA, named aSASA, as a new descriptor for steric effects (Figure 3-11):

$$aSASA = sigmoid\left(\frac{SASA}{0.5d}\right) = \frac{1}{1+e^{-\frac{SASA}{0.5d}}}$$

where *d* (the dividing line) is 13.42 Å^2. We expect that the aSASA descriptor should have a more linear relationship with reactivity, because it has different sensitivities with smaller and larger SASA values. When SASA is smaller than *d*, the aSASA descriptor is quite sensitive towards changes in SASA, which corresponds to the growth regime. When SASA is larger than *d*, the aSASA descriptor is not as sensitive towards changes in SASA, which corresponds to the plateau regime.

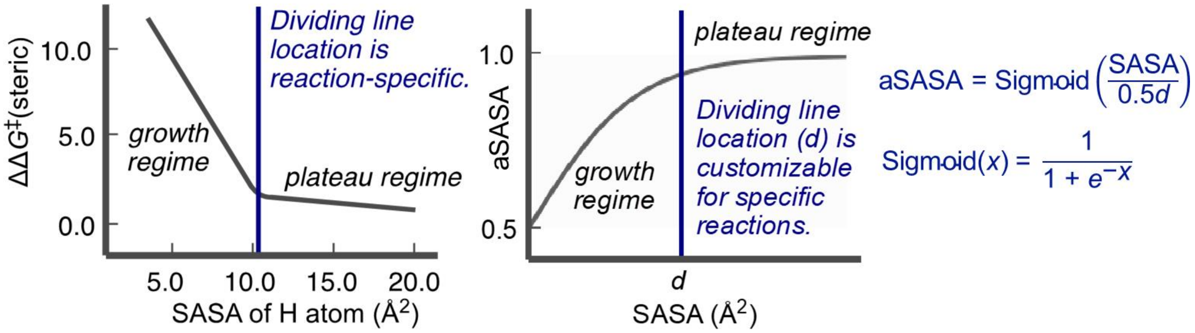


Figure 3-11 Activated solvent accessible surface area (aSASA) descriptor

To explore whether the use of the activated aSASA descriptor further improves the performance of the three-descriptor multivariate linear regression model, we replaced the SASA descriptor in the three-descriptor model with sigmoid(SASA/6.71), which corresponds to a dividing line of d = 13.42 Å^2. We compared the performance of the previous three-descriptor model with the new model using aSASA descriptor in Figure 3-12. A noticeable improvement was observed with the new model using aSASA (Figure 3-12, right). A smaller RMSE was obtained for both training and validation sets, whereas the coefficients of determination (R^2) improved for both training and validation sets. Further examination of these results indicates that although the improvements were observed for both the growth (SASA < 13.42 Å^2) and plateau (SASA ⩾ 13.42 Å^2) regimes, the improvements are more noticeable for the growth regime where steric effects are more important for reactivity.

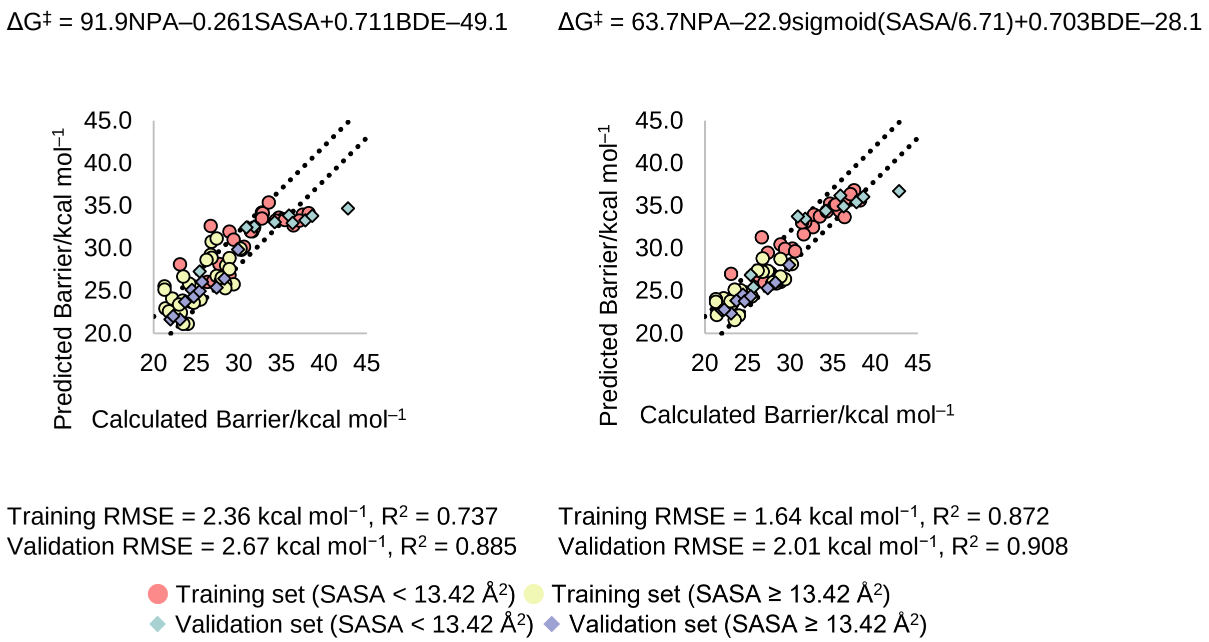


Figure 3-12 Comparison of the performance of three-descriptor linear regression models with the standard SASA descriptor and with the aSASA descriptor using sigmoid activation function. The three-descriptor linear regression model with the standard SASA descriptor is shown on the left. The three-descriptor linear regression model with the aSASA descriptor using sigmoid activation function is shown on the right.

Next, we performed tests with different sigmoid functions to validate whether our approach to identify the dividing line (*d*) is optimal for the performance of the aSASA descriptor. We used sigmoid(SASA) and sigmoid (SASA/18) in place of sigmoid(SASA/6.71) to train the three-descriptor linear regression model (Figure 3-13). These new sigmoid functions correspond to a dividing line at 2 Å^2 and 36 Å^2, respectively. The generated models show clearly inferior performance compared to the model with dividing line set at 13.42 Å^2. Therefore, these results suggest that a careful and rational choice of the dividing line (*d*) for the sigmoid function is important to obtain optimal performance using the activated aSASA descriptors.

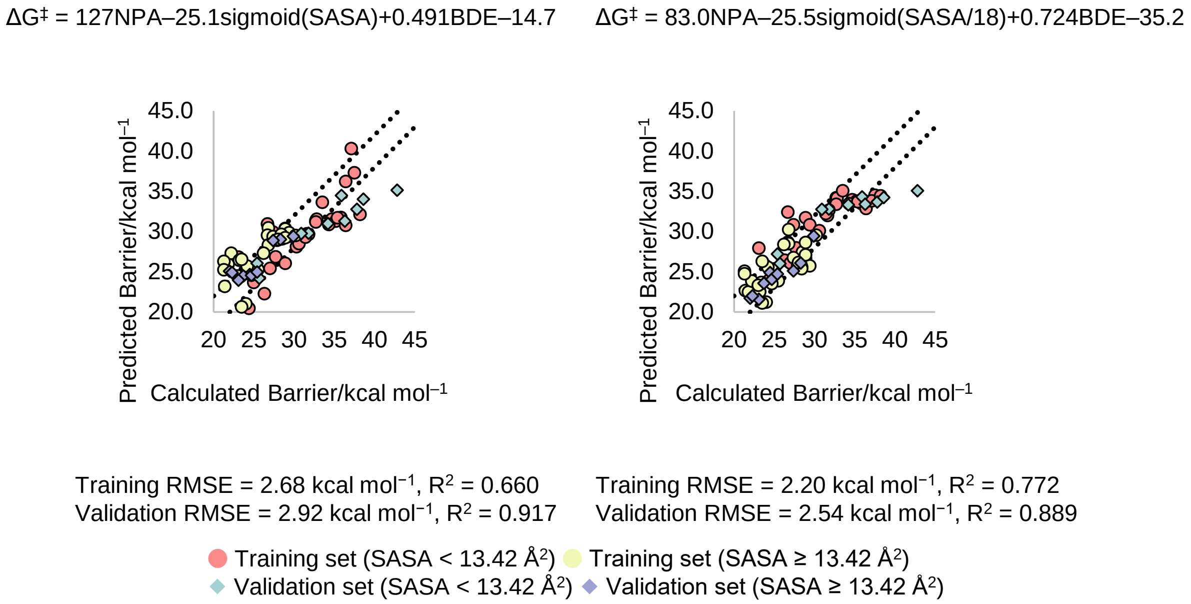


Figure 3-13 Performance of two models with different dividing lines

4.0 Further Applications of the Predictive Model to Structurally Complex Molecules

4.1 Challenges of Site-Selectivity Prediction for Conformationally Flexible Macrocyclic Molecules

Predicting site-selectivity for late-stage C–H functionalization of macrocyclic molecules is often more challenging because of their conformational flexibility and multiple potentially reactive C–H bonds with similar steric and electronic properties. Here, we attempt to apply the predictive model developed in Chapter 3 to the DMDO-mediated C–H hydroxylation of a bryostatin analogue. Bryostatin is a group of macrocyclic lactones isolated from marine bryozoan *Bugula neritina*, which serve as drug candidates.[104–113] Among these natural products, bryostatin 1 is the most studied member and shows antineoplastic activity.[104] The mechanism may involve the binding of bryostatin 1 towards the first conserved (C1) domain of protein kinase C.[114,115] Bryostatin can potentially be used to treat Alzheimer's disease[116] and HIV infection.[117] Bryostatin 1 can inhibit SARS-CoV-2 BA.1 as well.[118] The key building blocks in biosynthesis of bryostatin 1 are acetate, *S*-adenosyl methionine, and glycerol.[119] There are many synthetic attempts towards this group of macrocyclic lactones in literature.[120–127] Paul A. Wender is one of the pioneers in developing synthetic routes for bryostatin analogues.[128–133] Wender and his collaborators found that the (*R*)-configuration of the C3–OH, the free hydroxyl group at the C26 position, and the C ring are essential for the biological activity of bryostatin, while A and B rings are less important. Wender replaced the tetrahydropyran in the B ring with an acetal linkage to reduce the burdens in synthesis of bryostatin analogues. It was proposed that the hydroxyl group at the C9 position in the A ring is not responsible for the binding of bryostatin 1 towards protein kinase C and three analogues of bryostatin 1 were synthesized to verify this hypothesis. Here, we computationally evaluate the site-selectivity of DMDO-mediated C–H hydroxylation of one of these analogues, which is shown in the right half of Figure 4-1. Experimentally, this analogue underwent selective C9–H hydroxylation with DMDO (Figure 4-2a).[23] The DMDO-mediated reaction with the analogue 2a shown in Figure 4-2b gave a mixture of three products, including the C9–H hydroxylation and oxidation of the free hydroxyl group at the C26 position. The additional methyl substituent at the C26 position of this analogue might have enhanced the reactivity of the tertiary C–H at this position. The analogue 3a shown in Figure 4-2c also underwent selective C9–H hydroxylation with DMDO. Although the C26–H in the third analogue is also a tertiary C–H bond, substituting the hydroxyl group at C26 with an acetoxy group is apparently sufficient to protect the hydroxyl group in the third analogue. The late-stage diversification shown in Figure 4-2 did not significantly alter their binding affinities, which supported the hypothesis proposed by Wender.[23] Later the C ring was identified to be responsible for the binding of bryostatin and the C1 domain of isoforms of protein kinase C.[134]

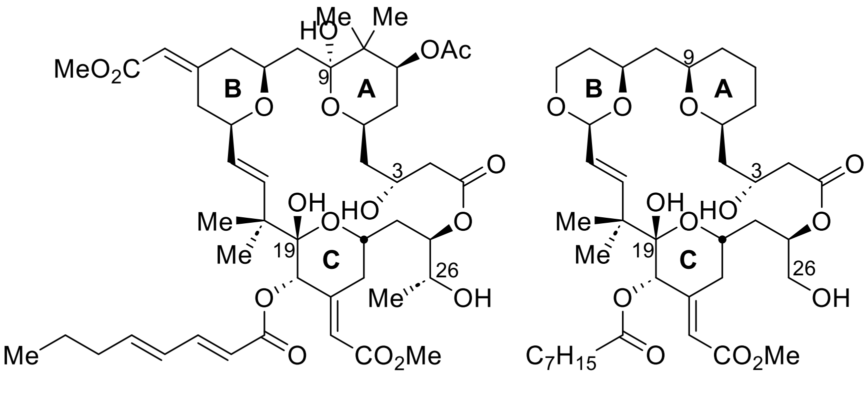


Figure 4-1 Bryostatin 1 and one of its analogues under study

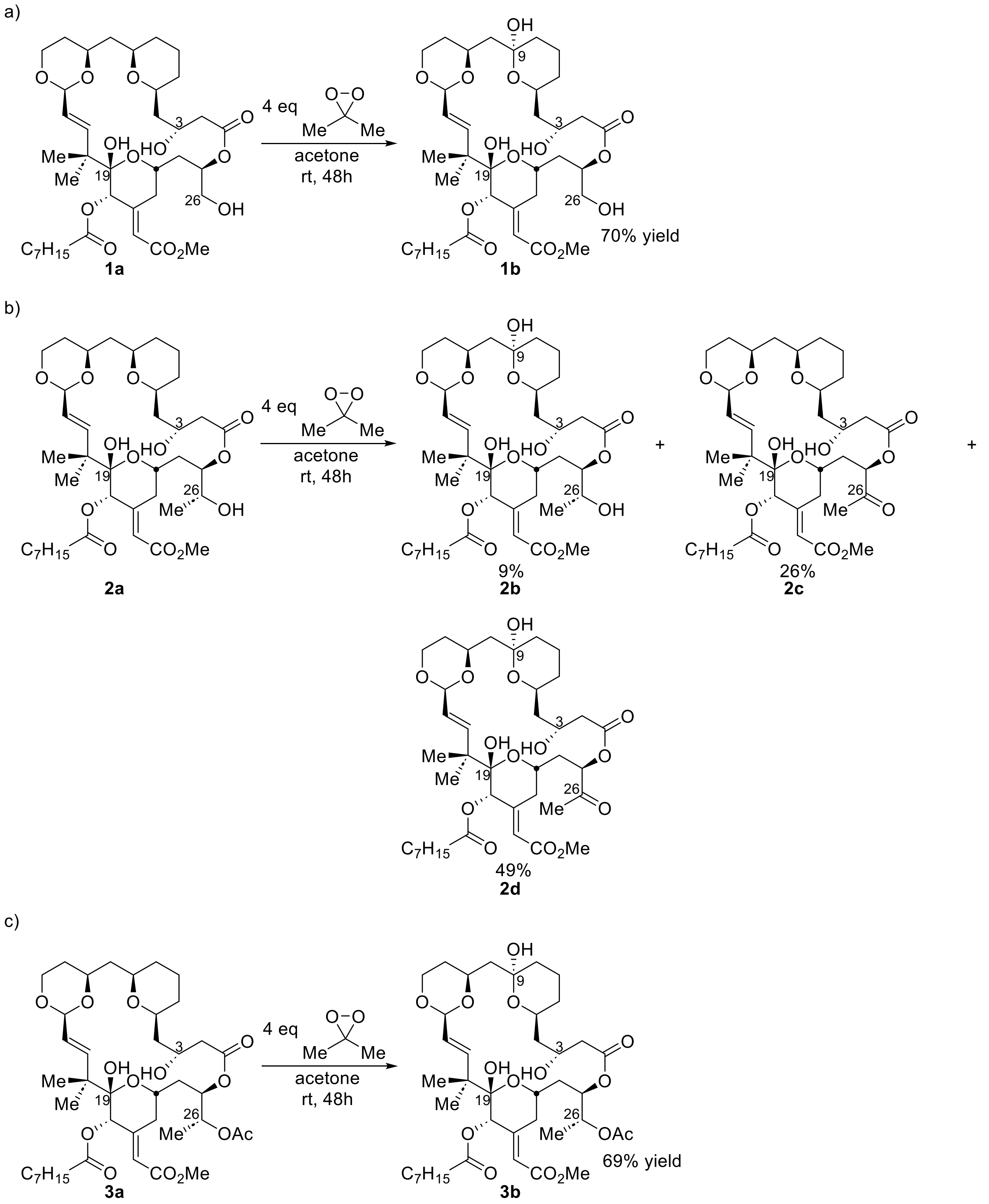


Figure 4-2 DMDO-mediated C–H hydroxylation on analogues of bryostatin 1

Intramolecular hydrogen bonds are observed in the crystal structure of bryostatin 1. The hydrogen bonds may be crucial for its biological activity[104] as well as conformational flexibility. There is a hydrogen bond between C19–OH as the hydrogen bond donor and the oxygen atom in C3–OH as hydrogen bond acceptor. In addition, a bifurcated hydrogen bond[135]was observed, involving C3–OH as the donor, and the oxygen atom in the A ring and the oxygen atom in the B ring both serving as the acceptors. Bryostatin 10 in $CDCl_3$ also contains this hydrogen bond network, which was revealed by ROESY spectrum and temperature-dependent coupling patterns in NMR.[136,137] In bryostatin analogues, if C3–OH is changed to the (*S*)-configuration, the hydrogen bond between C19–OH and the oxygen atom in C3–OH still exists.[131] However, the hydrogen bond between C3–OH and the oxygen atom in the B ring disappears and the biological activity of the analogue substantially decreases.

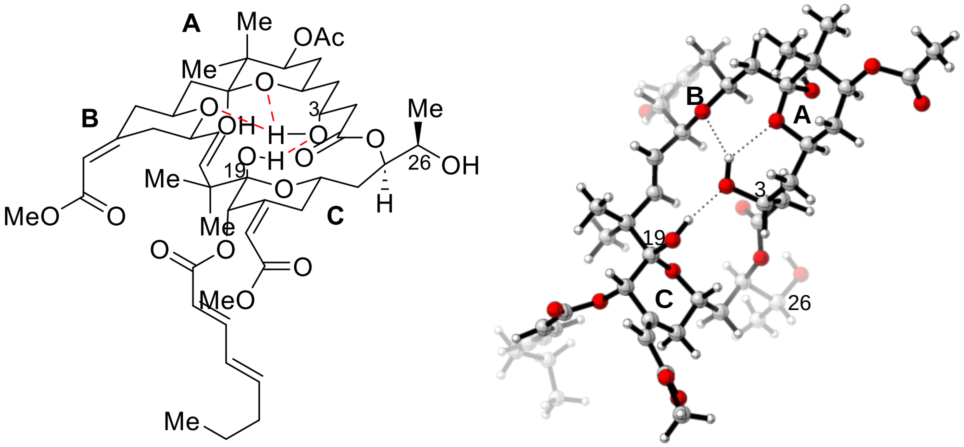


Figure 4-3 Intramolecular hydrogen bonds in the crystal structure of bryostatin 1

The conformational landscape of macrocycles such as bryostatin can be quite complex[138] and may be affected by interactions with solvent molecules.[139] Through rotational echo double resonance NMR and molecular dynamics, one bryostatin analogue was observed to adopt multiple conformations in a membrane environment, which are different from the crystal structure.[140] The conformational flexibility may complicate the application of the model developed in Chapter 3 to bryostatin analogues.

4.2 Application of the Predictive Model to a Simplified Model Based on the Crystal Structure

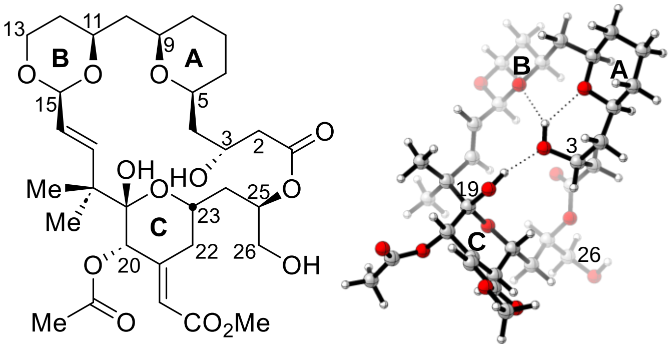


Figure 4-4 A simplified model of the bryostatin analogue 1a used in computational study

Because the crystal structure of bryostatin 1 has already been reported,[104,126] we used the crystal structure to create a slightly simplified model substrate of the bryostatin analogue 1a by replacing the long alkyl chain on the ester group with a methyl group to reduce computational costs (Figure 4-4). We calculated the NPA charge, SASA, and BDE descriptors for this molecule. Here, the NPA charge and BDE were calculated at the B3LYP/6-311++G(d,p)//B3LYP/6-31G(d) level of theory, which is the same as the methods employed in the previous chapter. SASA values were calculated using optimized structure of the crystal structure. We then applied the model developed in Chapter 3

$$\Delta G^{\ddagger}_{\text{predicted}} = 63.7\text{NPA}–22.9\text{sigmoid}(\text{SASA}/6.71)+0.703\text{BDE}–28.1$$

to predict the free energy of activation for each potentially reactive C–H bond in the simplified model of the bryostatin analogue 1a. The computed NPA, BDE, SASA descriptors, and the predicted free energies of activation are provided in Table 4-1. To validate the predicted values, we also used DFT to calculate the activation free energies for each C–H bond at the B3LYP/6-311++G(d,p)/SMD(acetone)//B3LYP/6-31G(d) level of theory. The $\Delta G^{\ddagger}_{\text{DFT}}$ values are the free energy difference between the computed transition state structure and the reactants.

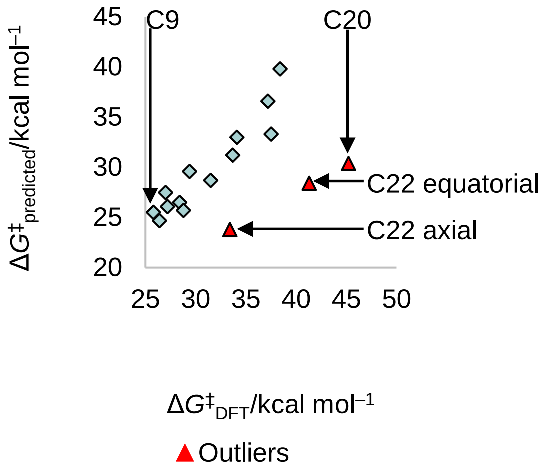
Table 4-1 Results of the crystal structure

Site	NPA	SASA/Å^2	BDE/kcal mol^{-1}	$\Delta G^{\ddagger}_{\text{DFT}}$/kcal mol^{-1}	$\Delta G^{\ddagger}_{\text{predicted}}$/kcal mol^{-1}
C3	0.190	5.45	92.2	34.1	33.0
C5	0.156	7.72	89.7	27.0	27.5
C9	0.152	12.40	90.6	25.8	25.5
C11	0.161	6.76	90.0	31.5	28.7
C15	0.160	1.79	86.0	29.4	29.6
C20	0.222	2.03	81.9	45.2	30.4
C23	0.183	3.19	90.9	37.5	33.3
C25	0.221	2.65	96.0	38.4	39.8
C2 H_R	0.223	17.14	94.4	33.7	31.2

C2 H_S	0.221	8.30	97.2	37.2	36.6
C13 axial	0.161	27.91	92.6	26.4	24.7
C13 equatorial	0.190	28.05	92.6	28.4	26.5
C22 axial	0.222	11.54	81.3	33.4	23.8
C22 equatorial	0.256	7.05	81.3	41.3	28.4
C26 H_R	0.170	20.71	92.3	28.8	25.7
C26 H_S	0.185	23.32	91.9	27.2	26.1

The comparison of $\Delta G^{\ddagger}_{\text{predicted}}$ values predicted using the three-descriptor regression model and DFT-computed $\Delta G^{\ddagger}_{\text{DFT}}$ values are shown in Figure 4-1. Three outliers are noticeable: at the C22 axial site, the $\Delta G^{\ddagger}_{\text{predicted}}$(C22 axial) is 9.6 kcal mol^{-1} lower than the DFT-computed $\Delta G^{\ddagger}_{\text{DFT}}$(C22 axial), at the C22 equatorial site, $\Delta G^{\ddagger}_{\text{predicted}}$(C22 equatorial) is 12.9 kcal mol^{-1} lower than $\Delta G^{\ddagger}_{\text{DFT}}$(C22 equatorial), at the C20 site, $\Delta G^{\ddagger}_{\text{predicted}}$(C20) is 14.8 kcal mol^{-1} lower than $\Delta G^{\ddagger}_{\text{DFT}}$(C20). Considering the three outliers are all allylic C–H bonds, the poor performance of the regression model for these C–H bonds is not completely unexpected, because we did not include any allylic or benzylic C–H bonds in the training set. Because previous work from Houk revealed that allylic and benzylic C–H bonds have a very different dependence on BDE,[34] we surmised that our model, which was trained using non-allylic and non-benzylic C–H bonds, cannot be applied to allylic and benzylic C–H bonds. Therefore, this test case revealed an important limitation to the regression model developed in Chapter 3, and suggested that future work is needed to include allylic and benzylic C–H bonds in the training set to obtain a more general model.

$\Delta G^{\ddagger}_{\text{predicted}}$ = 63.7NPA–22.9sigmoid(SASA/6.71)+0.703BDE–28.1



R^2 = 0.349
RMSE = 5.76 kcal mol^{-1}
After removing outliers, R^2 = 0.886
After removing outliers, RMSE = 2.02 kcal mol^{-1}

Figure 4-5 Performance of the three-descriptor regression model for site-selectivity of DMDO-mediated C–H hydroxylation of a bryostatin analogue

When the three allylic C–H bonds (C20, C22 axial, and C22 equatorial) were excluded, the performance of the predictive model substantially improved. A coefficient of determination (R^2) of 0.886 and RMSE of 2.02 kcal mol^{-1} were obtained, which are both similar to the performance of the model using the validation set (see Chapter 3). This data suggested that predictive model can provide reliable prediction for alkyl and α ether C–H bonds in a structurally complex macrocyclic molecule.

When excluding the allylic C–H bonds, which cannot be accurately predicted by the model, we identified five most reactive sites that are within 2 kcal mol^{-1} of the lowest $\Delta G^{\ddagger}_{\text{predicted}}$ (*i.e.*, $\Delta G^{\ddagger}_{\text{predicted}}$ = 24.7~26.7 kcal mol^{-1}): C9–H ($\Delta G^{\ddagger}_{\text{predicted}}$ = 25.5 kcal mol^{-1}), C13$_{\text{axial}}$–H ($\Delta G^{\ddagger}_{\text{predicted}}$ = 24.7 kcal mol^{-1}), C13$_{\text{equatorial}}$–H ($\Delta G^{\ddagger}_{\text{predicted}}$ = 26.5 kcal mol^{-1}), C26$_{\text{R}}$–H ($\Delta G^{\ddagger}_{\text{predicted}}$ = 25.7 kcal mol^{-1}), and C26$_{S}$–H ($\Delta G^{\ddagger}_{\text{predicted}}$ = 26.1 kcal mol^{-1}). Because these $\Delta G^{\ddagger}_{\text{predicted}}$ values are within the expected error range (2.01 kcal mol^{-1}) of the regression model, we do not expect our predictive model can conclusively predict which of the five sites is the most reactive with DMDO. Although the multiple C–H bonds with comparable reactivity complicate the site-selectivity prediction, the experimentally observed reactive site (C9) was successfully identified by the regression model as one of the most reactive site with a $\Delta G^{\ddagger}_{\text{predicted}}$ of 25.5 kcal mol^{-1}, which is the second lowest after C13$_{\text{axial}}$–H. Interestingly, the DFT-computed $\Delta G^{\ddagger}_{\text{DFT}}$ values correctly identified the most reactive site for hydroxylation: the $\Delta G^{\ddagger}_{\text{DFT}}$ for C9 (25.8 kcal mol^{-1}) is the lowest, although the $\Delta G^{\ddagger}_{\text{DFT}}$ for C13 axial (26.4 kcal mol^{-1}) is only slightly higher, further revealing the similar reactivities between these two sites.

It should be noted that the regression model successfully predicted the higher reactivity at C9 compared to other sites with comparable electronic properties, highlighting the importance of considering substrate steric effects in the predictive model. For example, C3, C5, C11, and C15 are all successfully predicted to be significantly less reactive than C9, which is consistent with the experimental results. Because the SASA values for these C–H bonds, along with C9–H, are all within the growth regime of the sigmoid function (SASA < 13.42 Å^{2}), the activation function would clearly predict the greater sensitivity to steric effects for these C–H bonds.

For the two tertiary α ether C–H bonds in the A ring (C5 and C9), the contribution from the aSASA term (*i.e.*, –22.9(sigmoid(SASA$_{\text{C5}}$/6.71)–sigmoid(SASA$_{\text{C9}}$/6.71))) to the $\Delta G^{\ddagger}_{\text{predicted}}$ difference is 2.4 kcal mol^{-1} based on the SASA values for C5–H (7.72 Å^{2}) and C9–H (12.40 Å^{2}). This steric effect makes a dominant contribution to the predicted reactivity difference between these sites ($\Delta G^{\ddagger}_{\text{predicted}}$(C5)–$\Delta G^{\ddagger}_{\text{predicted}}$(C9) = 2.0 kcal mol^{-1}). This specific example further highlights the importance of using activated steric descriptor, aSASA, because although the SASA values of these two C–H sites only differ 4.7 Å^2, because both are in the growth regime of the sigmoid activation function, this subtle SASA difference is amplified in the linear regression model to correctly predict the steric effects on site-selectivity.

In the future development of the predictive model, the first step should be including more allylic and benzylic C–H bonds like C6–H in estrone derivatives into the data set. After that, some tuning should be performed like adjusting the dividing line in the aSASA descriptor. Based on our previous experience, chemical intuition should still be taken into consideration if the interpretability of the predictive model needs conserving. Our experience of designing a reaction-specific descriptor may be generalizable towards other reactions.

4.3 Conclusions

We developed a computational protocol tackling the site-selectivity of DMDO-mediated C–H hydroxylation via predictive models. Chemical insights obtained by detailed analysis on previous experimental and computational results and statistical tools were integrated together to generate a predictive model for the reaction under investigation. A customized steric descriptor based on an activation function named aSASA was proposed during the development of this predictive model, which is expected to be generalizable to other C–H functionalization reactions. We later examined the applicability of the model via application to a macrocyclic lactone substrate.

Multivariate linear regression was extensively used in this work, which is easy to use and can be clearly linked to chemical intuitions. Lack of allylic and benzylic C–H bonds in the data set inherently limited the usage of this predictive model, which was revealed in the application to a macrocyclic lactone substrate containing allylic C–H bonds. However, due to rare appearance of allylic and benzylic C–H bonds in previous experimental results of DMDO-mediated C–H hydroxylation, this limitation cannot be exaggerated. The interpretability brought by linear regression made it easier to improve the model. The chemical insights helped us decide whether the slope in univariate linear regression, which corresponds to sensitivity, should be positive or negative if the descriptor is suitable for certain reaction. The model generated from linear regression can be formally decomposed when performing comparisons between two C–H bonds since the intercept remains unchanged, which is shown in the application of the model to a macrocyclic lactone substrate and can help us better understand the controlling factors of site-selectivity in this reaction.

The aSASA descriptor developed in this work is tunable and expected to be applicable to other C–H functionalization reactions. The interval covered by values of certain descriptor can be divided into a growth regime and a plateau regime while the same changes in the descriptor within different regimes can lead to significantly different changes in the property researchers are interested in, which in this case is the free energy of activation. This nonlinear behavior can be approximated by an activation function where a dividing line should be set at an appropriate position based on expertise. We resorted to piecewise linear regression to detect the change point and then set the dividing line at the change point, which in practice was time-efficient and could lead to a result that was near optimal. This might add to the toolbox of predictive models used by chemists.

This interpretability-prioritized protocol built on top of previous experimental and computational results and exploration of reaction-specific descriptor might be insightful to chemists hoping to use predictive models for other C–H functionalization reactions and even more complex transformations. Our aSASA descriptor is a contribution to the field of steric descriptors. The predictive model for site-selectivity of DMDO-mediated C–H hydroxylation developed in this work can be valuable to synthetic organic chemists attempting to use this reaction.

Appendix A Code for performing quality threshold clustering

Attached is a computer code written by me to perform quality threshold clustering[141,142] on a symmetric multiprocessing computing node. In the generated clusters, no pair of members will have a similarity value greater than a threshold specified by the user. The similarity value should have no dependence on the order of two members and is supposed to be calculated before using this program. A typical example of the similarity value is root-mean-square deviation of atomic positions (RMSD) between frames in the output of a molecular dynamics simulation. This program is written in C programming language with OpenMP. It uses features from the C99 standard. This program has been tested with the C compiler from GNU Compiler Collection (gcc version 10.2.0). A compile-time flag "-fopenmp" should be used when compiling the code. The most up-to-date version of the code can be obtained from GitHub at https://github.com/yimin-chen-at-sdf/parallel-quality-threshold-clustering or from Codeberg at https://codeberg.org/yimin-chen-at-sdf/parallel-quality-threshold-clustering.

```
#include <omp.h>

#include <stdio.h>

#include <stddef.h>

#include <stdlib.h>

#include <string.h>

#include <unistd.h>

#include <getopt.h>
```

```
void PrintUsage()
{
    printf("To use this program, edit the first number in the command
in the following line according to the size of your rmsd.dat file or
equivalently the number of frames in your trajectory generated by
molecular dynamics. ");
    printf("Next, edit the second number in the command in the
following line according to your own needs and pay attention to the
dimension or unit of the data stored in your rmsd.dat file. ");
    printf("Then, edit the name of the file which will store the
output after calculation. ");
    printf("Finally, set the OMP_NUM_THREADS environment variable
before typing the edited command.\n");
    printf("./parallelqt --number 1000 --threshold 0.42 --output
clusteringresult.txt\n");
}

void* AllocIntArray (int rows, int cols)
{
    return malloc( sizeof(int[rows][cols]) );
}

void* AllocFloatArray (int rows, int cols)
{
    return malloc( sizeof(float[rows][cols]) );
```

```
}

void ReadFloatArray (int rows, int cols, float array[rows][cols])
{
    FILE *data;
    data=fopen("rmsd.dat", "rb");
    if (data == NULL)
    {
        fprintf(stderr, "rmsd.dat does not exist!\n");
        exit(1);
    }
    fread(array, sizeof(float[rows][cols]), 1, data);
    fclose(data);
}

int **ConvertIntMatrix(int *a, int nrow, int ncol)
{
    int **m;
    m = (int **) malloc((size_t) ((nrow)*sizeof(int*)));
    if (m == NULL)
    {
        fprintf(stderr, "Memory allocation failure in
ConvertIntMatrix().\n");
        exit(1);
    }
    m[0] = a;
    for (int i=1; i<nrow; i++)
        m[i] = m[i-1] + ncol;
    return m;
}

float **ConvertFloatMatrix(float *a, int nrow, int ncol)
{
    float **m;
    m = (float **) malloc((size_t) ((nrow)*sizeof(float*)));
    if (m == NULL)
    {
        fprintf(stderr, "Memory allocation failure in
ConvertFloatMatrix().\n");
        exit(1);
    }
    m[0] = a;
    for (int i=1; i<nrow; i++)
        m[i] = m[i-1] + ncol;
    return m;
}

void FreeConvertIntMatrix(int **b)
{
    free((char*) b);
```

```
}

void FreeConvertFloatMatrix(float **b)
{
    free((char*) b);
}

void SetBit (unsigned A[], int o)
{
    A[o/sizeof(unsigned)] |= 1 << (o%sizeof(unsigned));
}

int TestBit (unsigned A[], int o)
{
    return ( (A[o/sizeof(unsigned)] & (1 << (o%sizeof(unsigned)) )) !=
0 );
}

void MultiClustering (int **frame, float **rmsd, int **localframe,
float **restrict localrmsd, int *index, int stack, int start, int
remainder, unsigned *restrict clusterbit, int clusterbitsize, int del,
int lmax, int *restrict outputnumber)
{
    int cardinality, anchor, cancer, flag;
    float nominee;
    for (int l=0; l<stack; l++)
    {
        for (int m=0; m<clusterbitsize; m++)
        {
            clusterbit[m] = 0;
        }
        SetBit(clusterbit, index[l+start]);
        cardinality = 1;
        while (cardinality < remainder)
        {
            nominee = -1.0;
            for (int m=0; m<lmax; m++)
            {
                if (localframe[l][m] < 0)
                    break;
                if (TestBit(clusterbit, localframe[l][m]) == 0 &&
localrmsd[l][m] > 0.0)
                {
                    if (nominee < 0.0)
                    {
                        anchor = m;
                        nominee = localrmsd[l][m];
                    }
                    else
                    {
                        if (localrmsd[l][m] < nominee)
                        {
                            anchor = m;
                            nominee = localrmsd[l][m];
                        }
                    }
                }
            }
            if (nominee < 0.0)
                break;
            cardinality++;
            anchor = localframe[l][anchor];
            SetBit(clusterbit, anchor);
            if (anchor < index[l+start] || anchor > index[start+stack-
1])
            {
                for (int m=0; m<lmax; m++)
                {
                    if (localframe[l][m] < 0)
                        break;
                    if (TestBit(clusterbit, localframe[l][m]) != 0 ||
localrmsd[l][m] < 0.0)
                        continue;
                    cancer = -1;
                    for (int n=0; n<del; n++)
                    {
                        if (frame[anchor][n] < 0 || frame[anchor][n] >
localframe[l][m])
                            break;
                        if (frame[anchor][n] == localframe[l][m])
                        {
                            cancer = n;
                            break;
                        }
                    }
                    if (cancer < 0)
                        localrmsd[l][m] = -1.0;
                    else
                    {
                        if (localrmsd[l][m] < rmsd[anchor][cancer])
                            localrmsd[l][m] = rmsd[anchor][cancer];
                    }
                }
            }
            else
            {
                for (int m=l+1; m<stack; m++)
                {
                    if (index[m+start] == anchor)
                    {
                        flag = m;
                        break;
                    }
                }
                for (int m=0; m<lmax; m++)
                {
                    if (localframe[l][m] < 0)
                        break;
                    if (TestBit(clusterbit, localframe[l][m]) != 0 ||
localrmsd[l][m] < 0.0)
                        continue;
                    cancer = -1;
                    for (int n=0; n<lmax; n++)
                    {
                        if (localframe[flag][n] < 0 ||
localframe[flag][n] > localframe[l][m])
                            break;
                        if (localframe[flag][n] == localframe[l][m])
                        {
                            cancer = n;
                            break;
                        }
                    }
                    if (cancer < 0)
                        localrmsd[l][m] = -1.0;
                    else
                    {
                        if (localrmsd[l][m] < localrmsd[flag][cancer])
                            localrmsd[l][m] = localrmsd[flag][cancer];
                    }
                }
            }
        }
        outputnumber[l+start] = cardinality;
    }
}

int EndingClustering (int **frame, float **rmsd, int *shortidlist,
float *restrict diameter, int seqseed, int remainder, unsigned
*restrict clusterbit, int clusterbitsize, int del, int lmax)
{
    int cardinality, anchor, cancer;
    float nominee;
    for (int l=0; l<clusterbitsize; l++)
    {
        clusterbit[l] = 0;
    }
    SetBit(clusterbit, seqseed);
    cardinality = 1;
    while (cardinality < remainder)
    {
        nominee = -1.0;
        for (int l=0; l<lmax; l++)
        {
            if (TestBit(clusterbit, shortidlist[l]) == 0 &&
diameter[l] > 0.0)
            {
                if (nominee < 0.0)
                {
                    anchor = l;
                    nominee = diameter[l];
                }
                else
                {
                    if (diameter[l] < nominee)
                    {
                        anchor = l;
                        nominee = diameter[l];
                    }
                }
            }
        }
        if (nominee < 0.0)
            break;
        cardinality++;
        anchor = shortidlist[anchor];
        SetBit(clusterbit, anchor);
        for (int l=0; l<lmax; l++)
        {
            if (TestBit(clusterbit, shortidlist[l]) != 0 ||
diameter[l] < 0.0)
                continue;
            cancer = -1;
            for (int m=0; m<del; m++)
            {
                if (frame[anchor][m] < 0 || frame[anchor][m] >
shortidlist[l])
                break;
                if (frame[anchor][m] == shortidlist[l])
                {
                    cancer = m;
                    break;
                }
            }
            if (cancer < 0)
                diameter[l] = -1.0;
            else
            {
                if (diameter[l] < rmsd[anchor][cancer])
                    diameter[l] = rmsd[anchor][cancer];
            }
        }
    }
    return cardinality;
}

void MonoClustering (int **frame, float **rmsd, float *restrict
diameter, int seqseed, unsigned *indexbit, int remainder, unsigned
*restrict clusterbit, int clusterbitsize, int del, float toc)
{
    int cardinality, anchor, cancer;
    float nominee;
    #pragma omp parallel for
    for (int l=0; l<clusterbitsize; l++)
    {
        clusterbit[l] = 0;
    }
    SetBit(clusterbit, seqseed);
    #pragma omp parallel for
    for (int l=0; l<del; l++)
    {
        if (TestBit(indexbit, frame[seqseed][l]) == 0)
            diameter[l] = rmsd[seqseed][l];
        else
            diameter[l] = -1.0;
```

```
}
cardinality = 1;
while (cardinality < remainder)
{
    nominee = -1.0;
    #pragma omp parallel for reduction(max:nominee)
    for (int l=0; l<del; l++)
    {
        if (diameter[l] > 0.0)
            nominee = diameter[l];
    }
    if (nominee < 0.0)
        break;
    nominee = toc + 1.0;
    #pragma omp parallel for reduction(min:nominee)
    for (int l=0; l<del; l++)
    {
        if (diameter[l] > 0.0 && diameter[l] < nominee)
            nominee = diameter[l];
    }
    anchor = del;
    #pragma omp parallel for reduction(min:anchor)
    for (int l=0; l<del; l++)
    {
        if (diameter[l] == nominee)
            anchor = l;
    }
    diameter[anchor] = -1.0;
    cardinality++;
    anchor = frame[seqseed][anchor];
    SetBit(clusterbit, anchor);
    #pragma omp parallel for private(cancer)
    for (int l=0; l<del; l++)
    {
        if (diameter[l] < 0.0)
            continue;
        cancer = -1;
        for (int m=0; m<del; m++)
        {
            if (frame[anchor][m] < 0)
                break;
            if (frame[anchor][m] == frame[seqseed][l])
            {
                cancer = m;
                break;
            }
        }
        if (cancer < 0)
            diameter[l] = -1.0;
        else
```

```
            {
                if (diameter[l] < rmsd[anchor][cancer])
                    diameter[l] = rmsd[anchor][cancer];
            }
        }
    }
}

void PrintCitation()
{
    printf("If you want to use this program in any of your published
work, please cite the following papers:\n");
    printf("1. Laurie J. Heyer, Semyon Kruglyak, and Shibu Yooseph,
Exploring Expression Data: Identification and Analysis of Coexpressed
Genes, Genome Research, 1999, 9, 1106-1105.
DOI:10.1101/gr.9.11.1106.\n");
    printf("2. Anthony Danalis, Collin McCurdy, and Jeffrey S. Vetter,
Efficient Quality Threshold Clustering for Parallel Architectures,
2012 IEEE 26th International Parallel and Distributed Processing
Symposium, Shanghai, China, 2012, pp. 1068-1079. DOI:
10.1109/IPDPS.2012.99.\n");
}

int main (int argc, char *argv[])
{
    int numberofelements = -1;
    float thresholdofclustering = -1.0;
    char *outputtext;
    int opt = 0;
    static struct option long_options[] =
    {
        {"number",    required_argument, NULL, 'n'},
        {"threshold", required_argument, NULL, 't'},
        {"output",    required_argument, NULL, 'o'}
    };
    int long_index = 0;
    while ((opt = getopt_long(argc, argv, "n:t:o:", long_options,
&long_index)) != -1)
    {
        switch (opt)
        {
            case 'n':
                numberofelements = atoi(optarg);
                break;
            case 't':
                thresholdofclustering = atof(optarg);
                break;
            case 'o':
                outputtext = (char
*)malloc((strlen(optarg)+1)*sizeof(char));
```

```
                if (outputtext == NULL)
                {
                    fprintf(stderr, "Memory allocation failure in
obtaining the name of the output file.\n");
                    exit(1);
                }
                outputtext[strlen(optarg)] = '\0';
                strcpy(outputtext, optarg);
                break;
            default:
                PrintUsage();
                exit(1);
        }
    }
    if (numberofelements == -1 || thresholdofclustering == -1.0)
    {
        PrintUsage();
        exit(1);
    }
    int rows = numberofelements;
    int cols = numberofelements;
    int order, counter, delta, cancer, available;
    float d = thresholdofclustering;
    float (*R)[cols] = AllocFloatArray(rows, cols);
    if (R == NULL)
    {
        fprintf(stderr, "Memory allocation failure in creating matrix
R.\n");
        exit(1);
    }
    ReadFloatArray(rows, cols, R);
    int *interaction;
    interaction = (int *)malloc(rows*sizeof(int));
    if (interaction == NULL)
    {
        fprintf(stderr, "Memory allocation failure in creating array
interaction.\n");
        exit(1);
    }
    #pragma omp parallel for private(counter)
    for (int i=0; i<numberofelements; i++)
    {
        counter = 0;
        for (int j=0; j<numberofelements; j++)
        {
            if (R[i][j] <= d)
                counter++;
        }
        interaction[i] = counter - 1;
    }
    delta = 0;
    #pragma omp parallel for reduction(max:delta)
    for (int i=0; i<numberofelements; i++)
    {
        if (interaction[i] > delta)
            delta = interaction[i];
    }
    if (delta == 0)
    {
        printf("The threshold specified by the user is smaller than
the minimum value of pairwise distance in rmsd.dat file. Each element
can form a one-membered cluster by itself while the result has no
practical value.\n");
        free(interaction);
        free(R);
        exit(1);
    }
    rows = numberofelements;
    cols = delta;
    int (*B1)[cols] = AllocIntArray(rows, cols);
    if (B1 == NULL)
    {
        fprintf(stderr, "Memory allocation failure in creating matrix
B1.\n");
        exit(1);
    }
    float (*B2)[cols] = AllocFloatArray(rows, cols);
    if (B2 == NULL)
    {
        fprintf(stderr, "Memory allocation failure in creating matrix
B2.\n");
        exit(1);
    }
    #pragma omp parallel for private(counter)
    for (int i=0; i<numberofelements; i++)
    {
        counter = 0;
        for (int j=0; j<numberofelements; j++)
        {
            if (j == i)
                continue;
            else
            {
                if (R[i][j] <= d)
                {
                    B1[i][counter] = j;
                    counter++;
                }
            }
        }
        for (int j=counter; j<delta; j++)
        {
            B1[i][j] = -1;
        }
    }
    #pragma omp parallel for private(counter)
    for (int i=0; i<numberofelements; i++)
    {
        counter = 0;
        for (int j=0; j<numberofelements; j++)
        {
            if (j == i)
                continue;
            else
            {
                if (R[i][j] <= d)
                {
                    B2[i][counter] = R[i][j];
                    counter++;
                }
            }
        }
        for (int j=counter; j<delta; j++)
        {
            B2[i][j] = -1.0;
        }
    }
    free(R);
    int *I, *work, *capacity;
    unsigned *S;
    I = (int *)malloc(rows*sizeof(int));
    if (I == NULL)
    {
        fprintf(stderr, "Memory allocation failure in creating array
I.\n");
        exit(1);
    }
    available = omp_get_max_threads();
    work = (int *)malloc(2*available*sizeof(int));
    if (work == NULL)
    {
        fprintf(stderr, "Memory allocation failure in creating array
work.\n");
        exit(1);
    }
    capacity = (int *)malloc(rows*sizeof(int));
    if (capacity == NULL)
    {
        fprintf(stderr, "Memory allocation failure in creating array
capacity.\n");
```

```
        exit(1);
    }
    if (rows%sizeof(unsigned) == 0)
        order = rows / sizeof(unsigned);
    else
        order = rows / sizeof(unsigned) + 1;
    S = (unsigned *)malloc(order*sizeof(unsigned));
    if (S == NULL)
    {
        fprintf(stderr, "Memory allocation failure in creating array
S.\n");
        exit(1);
    }
    #pragma omp parallel for
    for (int i=0; i<numberofelements; i++)
    {
        I[i] = i;
    }
    #pragma omp parallel for
    for (int i=0; i<order; i++)
    {
        S[i] = 0;
    }
    int unclustered = numberofelements;
    int maxcardinality, sum, ideal, temp, fragment, record;
```

```
float nominee;

unsigned *SP;

int *icache;

float *dcache;

int **aB1;

float **aB2;

aB1 = ConvertIntMatrix(&B1[0][0], rows, cols);

aB2 = ConvertFloatMatrix(&B2[0][0], rows, cols);

FILE *fp;

fp = fopen(outputtext, "w");

while (unclustered > 0)

{

    if (unclustered == 1)

    {

        SetBit(S, I[0]);

        fprintf(fp, "%d\n", I[0]);

        break;

    }

    else

    {

        if (unclustered > available && available > 1)

        {

            sum = 0;

            for (int i=0; i<unclustered; i++)

            {
```

```
    sum += interaction[I[i]];
}
sum += unclustered;
ideal = sum / available;
work[0] = 0;
#pragma omp parallel for
for (int i=1; i<2*available; i++)
{
    work[i] = -1;
}
counter = available;
if (sum % available == 0)
{
    for (int i=0; i<available; i++)
    {
        if (work[2*i] >= unclustered)
        {
            work[2*i] = -1;
            counter = i;
            break;
        }
        temp = interaction[I[work[2*i]]] + 1;
        if (temp >= ideal)
        {
            work[2*i+1] = work[2*i];
```

```
                            if (i < available-1)
                                work[2*i+2] = work[2*i] + 1;
                        }
                        else
                        {
                            for (int j=1; temp<ideal; j++)
                            {
                                if (work[2*i]+j >= unclustered)
                                {
                                    record = j - 1;
                                    break;
                                }
                                temp = temp +
interaction[I[work[2*i]+j]] + 1;
                                record = j;
                            }
                            work[2*i+1] = work[2*i] + record;
                            if (i < available-1)
                                work[2*i+2] = work[2*i] + record + 1;
                        }
                    }
                }
                else
                {
                    for (int i=0; i<available; i++)
                    {
                        if (work[2*i] >= unclustered)
                        {
                            work[2*i] = -1;
                            counter = i;
                            break;
                        }
                        temp = interaction[I[work[2*i]]] + 1;
                        if (temp > ideal)
                        {
                            work[2*i+1] = work[2*i];
                            if (i < available-1)
                                work[2*i+2] = work[2*i] + 1;
                        }
                        else
                        {
                            for (int j=1; temp<=ideal; j++)
                            {
                                if (work[2*i]+j >= unclustered)
                                {
                                    record = j - 1;
                                    break;
                                }
                                temp = temp +
interaction[I[work[2*i]+j]] + 1;
```

```
                record = j;
            }
            work[2*i+1] = work[2*i] + record;
            if (i < available-1)
                work[2*i+2] = work[2*i] + record + 1;
        }
     }
}
while (counter < available)
{
    record = -1;
    fragment = 0;
    for (int i=0; i<counter; i++)
    {
        temp = 0;
        for(int j=work[2*i]; j<=work[2*i+1]; j++)
            temp = temp + interaction[I[j]] + 1;
        if (temp > fragment)
        {
            record = i;
            fragment = temp;
        }
    }
    if (record == -1)
    {
```

```
                        fprintf(stderr, "Error in allocating per-
thread starting and ending indices, which indicates the user might
have forgotten to set the OMP_NUM_THREADS environment variable.\n");
                        exit(1);
                    }
                    if (work[2*record+1] == work[2*record])
                    {
                        record = -1;
                        fragment = 0;
                        for (int i=0; i<counter; i++)
                        {
                            if (work[2*i+1]-work[2*i]+1 > fragment)
                            {
                                record = i;
                                fragment = work[2*i+1] - work[2*i] +
1;
                            }
                        }
                        if (record == -1)
                        {
                            fprintf(stderr, "Error in allocating per-
thread starting and ending indices.\n");
                            exit(1);
                        }
                        if (fragment == 1)
```

```
                            break;
                    }
                    if (work[2*counter] != -1)
                    {
                        fprintf(stderr, "Error in allocating per-
thread starting and ending indices.\n");
                        exit(1);
                    }
                    work[2*counter] = work[2*counter-1];
                    work[2*counter+1] = work[2*counter-1];
                    for (int i=record+1; i<counter; i++)
                    {
                        work[2*i] -= 1;
                        work[2*i+1] -= 1;
                    }
                    work[2*record+1] -= 1;
                    counter++;
                }
                #pragma omp parallel for private(SP, temp)
schedule(static)
                for (int i=0; i<counter; i++)
                {
                    SP = (unsigned *)malloc(order*sizeof(unsigned));
                    if (SP == NULL)
                    {
                        fprintf(stderr, "Memory allocation failure in
creating array SP to be used by MultiClustering().\n");
                        exit(1);
                    }
                    int localrow, localmax;
                    localrow = work[2*i+1] - work[2*i] + 1;
                    localmax = 0;
                    for (int j=0; j<localrow; j++)
                    {
                        temp = 0;
                        for (int k=0; k<delta; k++)
                        {
                            if (B1[I[work[2*i]+j]][k] < 0)
                                break;
                            if (TestBit(S, B1[I[work[2*i]+j]][k]) ==
0)
                                temp++;
                            if (temp == interaction[I[work[2*i]+j]])
                                break;
                        }
                        interaction[I[work[2*i]+j]] = temp;
                        if (temp > localmax)
                            localmax = temp;
                    }
                    int (*multiicache)[localmax] =
AllocIntArray(localrow, localmax);
                    if (multiicache == NULL)
                    {
                        fprintf(stderr, "Memory allocation failure in
creating matrix multiicache.\n");
                        exit(1);
                    }
                    float (*multidcache)[localmax] =
AllocFloatArray(localrow, localmax);
                    if (multidcache == NULL)
                    {
                        fprintf(stderr, "Memory allocation failure in
creating matrix multidcache.\n");
                        exit(1);
                    }
                    for (int j=0; j<localrow; j++)
                    {
                        temp = 0;
                        for (int k=0; k<delta; k++)
                        {
                            if (B1[I[work[2*i]+j]][k] < 0)
                                break;
                            if (TestBit(S, B1[I[work[2*i]+j]][k]) ==
0)
```

```
                            {
                                multiicache[j][temp] =
B1[I[work[2*i]+j]][k];
                                multidcache[j][temp] =
B2[I[work[2*i]+j]][k];
                                temp++;
                            }
                        }
                        for (int k=temp; k<localmax; k++)
                        {
                            multiicache[j][k] = -1;
                        }
                        for (int k=temp; k<localmax; k++)
                        {
                            multidcache[j][k] = -1.0;
                        }
                    }
                    int **amultiicache;
                    float **amultidcache;
                    amultiicache =
ConvertIntMatrix(&multiicache[0][0], localrow, localmax);
                    amultidcache =
ConvertFloatMatrix(&multidcache[0][0], localrow, localmax);
```

```
                    MultiClustering(aB1, aB2, amultiicache,
amultidcache, I, localrow, work[2*i], unclustered, SP, order, delta,
localmax, capacity);
                    free(SP);
                    SP = NULL;
                    FreeConvertIntMatrix(amultiicache);
                    FreeConvertFloatMatrix(amultidcache);
                    free(multiicache);
                    free(multidcache);
                    multiicache = NULL;
                    multidcache = NULL;
                    amultiicache = NULL;
                    amultidcache = NULL;
                }
            }
            else
            {
                #pragma omp parallel for private(SP, icache, dcache,
temp) schedule(static)
                for (int i=0; i<unclustered; i++)
                {
                    SP = (unsigned *)malloc(order*sizeof(unsigned));
                    if (SP == NULL)
                    {
                        fprintf(stderr, "Memory allocation failure in
creating array SP to be used by EndingClustering().\n");
                        exit(1);
                    }
                    temp = 0;
                    for (int k=0; k<delta; k++)
                    {
                        if (B1[I[i]][k] < 0)
                            break;
                        if (TestBit(S, B1[I[i]][k]) == 0)
                            temp++;
                        if (temp == interaction[I[i]])
                            break;
                    }
                    interaction[I[i]] = temp;
                    icache = (int *)malloc(temp*sizeof(int));
                    if (icache == NULL)
                    {
                        fprintf(stderr, "Memory allocation failure in
creating array icache to be used by EndingClustering().\n");
                        exit(1);
                    }
                    dcache = (float *)malloc(temp*sizeof(float));
                    if (dcache == NULL)
                    {
                        fprintf(stderr, "Memory allocation failure in
creating array dcache to be used by EndingClustering().\n");
                        exit(1);
                    }
                    temp = 0;
                    for (int k=0; k<delta; k++)
                    {
                        if (B1[I[i]][k] < 0)
                            break;
                        if (TestBit(S, B1[I[i]][k]) == 0)
                        {
                            icache[temp] = B1[I[i]][k];
                            dcache[temp] = B2[I[i]][k];
                            temp++;
                        }
                        if (temp == interaction[I[i]])
                            break;
                    }
                    capacity[i] = EndingClustering(aB1, aB2, icache,
dcache, I[i], unclustered, SP, order, delta, temp);
                    free(SP);
                    free(dcache);
                    SP = NULL;
                    dcache = NULL;
                }
```

```
            }
            maxcardinality = 0;
            #pragma omp parallel for reduction(max:maxcardinality)
            for (int i=0; i<unclustered; i++)
            {
                if (capacity[i] > maxcardinality)
                    maxcardinality = capacity[i];
            }
            for (int i=0; i<unclustered; i++)
            {
                if (capacity[i] == maxcardinality)
                {
                    counter = i;
                    break;
                }
            }
            SP = (unsigned *)malloc(order*sizeof(unsigned));
            if (SP == NULL)
            {
                fprintf(stderr, "Memory allocation failure in creating
array SP to be used by MonoClustering().\n");
                exit(1);
            }
            dcache = (float *)malloc(delta*sizeof(float));
            if (dcache == NULL)
            {
                fprintf(stderr, "Memory allocation failure in creating
array dcache to be used by MonoClustering().\n");
                exit(1);
            }
            MonoClustering(aB1, aB2, dcache, I[counter], S,
unclustered, SP, order, delta, thresholdofclustering);
            free(dcache);
            dcache = NULL;
            fprintf(fp, "%d", I[counter]);
            for (int i=0; i<rows; i++)
            {
                if (TestBit(SP, i) != 0 && i != I[counter])
                    fprintf(fp, ",%d", i);
            }
            fprintf(fp, "\n");
            #pragma omp parallel for
            for (int i=0; i<order; i++)
            {
                S[i] += SP[i];
            }
            for (int i=0; i<rows; i++)
            {
                if (TestBit(SP, i) != 0)
                {
```

```
                for (int j=0; j<unclustered; j++)
                {
                    if (I[j] == i)
                    {
                        for (int k=j; k<unclustered-1; k++)
                            I[k] = I[k+1];
                        I[unclustered-1] = -1;
                        unclustered--;
                        break;
                    }
                }
            }
        }
        free(SP);
        SP = NULL;
    }
}
fclose(fp);
free(I);
free(S);
FreeConvertIntMatrix(aB1);
FreeConvertFloatMatrix(aB2);
free(B1);
free(B2);
free(interaction);
```

```
    free(work);

    free(capacity);

    PrintCitation();

    return 0;

}
```

Appendix B Data Set

Appendix Table 1 Training and validation sets splitting results using the Kennard–Stone algorithm

ID	Compound	Label	Site	Category	Split Result
1	5β-androstan-3α-17β-diacetoxy	S30	C3	ether α tertiary	Training
2			C5	tertiary alkyl	Training
3			C8	tertiary alkyl	Training
4			C9	tertiary alkyl	Training
5			C14	tertiary alkyl	Training
6			C17	ether α tertiary	Training
7	3-methyltetrahydropyran	S14	C2 axial	ether α secondary	Validation
8			C2 equatorial	ether α secondary	Training
9			C3	tertiary alkyl	Training

10			C6 axial	ether α secondary	Validation
11			C6 equatorial	ether α secondary	Validation
12	2-methyltetrahydrofuran	S8	C2	ether α tertiary	Training
13			C5 cis to methyl	ether α secondary	Validation
14			C5 trans to methyl	ether α secondary	Training
15	3-methyltetrahydrofuran	S9	C2 cis to methyl	ether α secondary	Validation
16			C2 trans to methyl	ether α secondary	Training
17			C3	tertiary alkyl	Training
18			C5 cis to methyl	ether α secondary	Training
19			C5 trans to methyl	ether α secondary	Validation
20	1,3-dioxane	S16	C2 axial	acetal α	Training

21			C2 equatorial	acetal α	Training
22			C4 axial	ether α secondary	Training
23			C4 equatorial	ether α secondary	Training
24	2,2,4-trimethyl-1,3-dioxane	S18	C4	ether α tertiary	Training
25			C6 axial	ether α secondary	Validation
26			C6 equatorial	ether α secondary	Validation
27	2,4,4-trimethyl-1,3-dioxane	S19	C2	acetal α	Training
28			C6 axial	ether α secondary	Validation
29			C6 equatorial	ether α secondary	Training
30	4,4,6-trimethyl-1,3-dioxane	S20	C2 axial	acetal α	Training
31			C2 equatorial	acetal α	Training
32			C6	ether α tertiary	Validation

33	3-isopropyltetrahydro-*2H*-pyran	S15	C2 axial	ether α secondary	Training
34			C2 equatorial	ether α secondary	Validation
35			C3	tertiary alkyl	Training
36			C6 axial	ether α secondary	Training
37			C6 equatorial	ether α secondary	Training
38			C7	tertiary alkyl	Training
39	2-(*tert*-butyl)-1,3-dioxane	S21	C2	acetal α	Training
40			C4 axial	ether α secondary	Training
41			C4 equatorial	ether α secondary	Training
42	2-oxaspiro[5.5]undecane	S17	C1 axial	ether α secondary	Training
43			C1 equatorial	ether α secondary	Training

44			C3 axial	ether α secondary	Training
45			C3 equatorial	ether α secondary	Training
46	acyclic	S1	1	ether α tertiary	Training
47	acyclic	S2	4	ether α tertiary	Training
48	acyclic	S3	5	ether α tertiary	Training
49	acyclic	S4	7	ether α tertiary	Training
50	acyclic	S5	8	ether α tertiary	Validation
51	acyclic	S6	9	ether α tertiary	Training
52	acyclic	S7	10	ether α tertiary	Training
53	tigogenin acetate	S31	C3	ether α tertiary	Training
54			C5	tertiary alkyl	Training

55			C8	tertiary alkyl	Training
56			C9	tertiary alkyl	Training
57			C14	tertiary alkyl	Training
58			C16	ether α tertiary	Training
59			C17	tertiary alkyl	Training
60			C20	tertiary alkyl	Training
61			C25	tertiary alkyl	Training
62			C26 axial	ether α secondary	Training
63			C26 equatorial	ether α secondary	Training
64	4-(*tert*-butyl)-2,3-dimethyltetrahydrofuran	S10	C3	tertiary alkyl	Training
65	3-isopropyl-2,4,5-trimethyltetrahydrofuran	S11	C4	tertiary alkyl	Validation

66	2,4-di-*tert*-butyl-3-methyltetrahydrofuran	S12	C3	tertiary alkyl	Validation
67	4-(*tert*-butyl)-2,2,3-trimethyltetrahydrofuran	S13	C3	tertiary alkyl	Training
68	1,1,6,6-tetramethyldecahydronaphthalene	S22	C4a	tertiary alkyl	Training
69	1,1,3,3,6,6-hexamethyldecahydronaphthalene	S23	C4a	tertiary alkyl	Training
70	1,1,4,4,6,6-hexamethyldecahydronaphthalene	S24	C4a	tertiary alkyl	Validation
71	1,1,3,3,8,8-hexamethyldecahydronaphthalene	S25	C4a	tertiary alkyl	Validation
72	2,4b-dimethyltetradecahydrophenanthrene	S27	C10a	tertiary alkyl	Validation
73	1-(*tert*-butyl)-1,2,6-trimethyldecahydronaphthalene	S26	C4a	tertiary alkyl	Training
74	2,4b,9,9-tetramethyltetradecahydrophenanthrene	S28	C10a	tertiary alkyl	Validation
75	2,4b,10,10-tetramethyltetradecahydrophenanthrene	S29	C10a	tertiary alkyl	Training
76	2,4b-dimethyltetradecahydrophenanthrene	S27	C4a	tertiary alkyl	Validation

77	1-(*tert*-butyl)-1,2,6-trimethyldecahydronaphthalene	S26	C8a	tertiary alkyl	Training
78	2,4b,9,9-tetramethyltetradecahydrophenanthrene	S28	C4a	tertiary alkyl	Training
79	2,4b,10,10-tetramethyltetradecahydrophenanthrene	S29	C4a	tertiary alkyl	Validation

Appendix Table 2 Computed electronic descriptors

ID	BDE/kcal mol^{-1}	NPA charge	CHELPG charge	Laplacian bond order
1	94.7	0.194	-0.0183	0.815
2	92.9	0.198	-0.0477	0.771
3	92.2	0.195	-0.0463	0.775
4	90.5	0.196	-0.0444	0.768
5	88.3	0.193	-0.077	0.76
6	93.5	0.2	-0.0255	0.81
7	92.2	0.151	-0.0198	0.787
8	92.2	0.186	0.006	0.821
9	93.4	0.192	-0.0309	0.779
10	92.1	0.15	-0.0239	0.792
11	92.1	0.186	0.0162	0.824
12	88.2	0.155	-0.0464	0.797

13	89.7	0.164	-0.0161	0.816
14	89.7	0.171	-0.0235	0.82
15	89.8	0.155	-0.0199	0.802
16	90.1	0.18	-0.0226	0.818
17	91.3	0.193	-0.0349	0.783
18	89.9	0.174	-0.0057	0.819
19	89.9	0.162	-0.0201	0.811
20	92.1	0.123	-0.0006	0.803
21	96.8	0.178	0.0703	0.854
22	92.5	0.155	-0.0197	0.794
23	92.5	0.191	0.0112	0.83
24	89.2	0.155	-0.1085	0.794
25	91	0.155	-0.0695	0.801
26	91	0.189	0.0199	0.824
27	89.9	0.127	-0.0541	0.802
28	91.8	0.155	-0.0659	0.797
29	91.8	0.19	0.0205	0.824
30	90.8	0.125	-0.0566	0.808
31	95.1	0.177	0.0574	0.851
32	90	0.156	-0.0837	0.79
33	92.1	0.151	-0.01	0.786
34	92.1	0.189	0.0384	0.828

35	93.3	0.192	0.0034	0.769
36	92.1	0.15	-0.0227	0.792
37	92.1	0.185	0.0115	0.827
38	91.4	0.183	-0.0648	0.777
39	91.5	0.135	0.0154	0.783
40	92.3	0.156	-0.0089	0.795
41	92.3	0.19	0.0334	0.829
42	92.3	0.159	0.0075	0.784
43	92.3	0.189	0.0574	0.821
44	92.1	0.15	-0.0215	0.793
45	92.1	0.186	0.0092	0.824
46	88.3	0.179	-0.054	0.82
47	88.8	0.179	-0.0269	0.808
48	89.5	0.179	0.0249	0.801
49	86.2	0.179	-0.0664	0.806
50	86.8	0.183	0.0221	0.803
51	86.5	0.184	0.016	0.801
52	84.2	0.188	0.075	0.807
53	94.4	0.194	-0.0502	0.812
54	91.3	0.192	-0.0988	0.766
55	92	0.195	-0.0411	0.774
56	91.9	0.192	-0.034	0.761

57	89.7	0.191	-0.0286	0.758
58	90.2	0.176	-0.0513	0.806
59	95	0.211	0.0287	0.776
60	90.4	0.207	-0.0584	0.784
61	93.4	0.188	-0.02	0.778
62	93	0.169	0.0089	0.804
63	93	0.184	0.0303	0.82
64	91.6	0.198	-0.0446	0.781
65	91.2	0.198	-0.0499	0.782
66	92	0.205	-0.0352	0.781
67	90.3	0.202	-0.0522	0.778
68	92	0.19	-0.0507	0.771
69	92.3	0.192	-0.0258	0.774
70	91.9	0.196	-0.0196	0.768
71	91.8	0.191	-0.0433	0.775
72	91.7	0.189	-0.0535	0.772
73	91.2	0.189	-0.0423	0.774
74	91.7	0.191	-0.0349	0.774
75	90.6	0.196	0.00841	0.771
76	91.4	0.195	-0.00643	0.763
77	92.4	0.199	-0.0741	0.769
78	91.7	0.195	0.0208	0.765

79	91.8	0.197	0.0368	0.767

Appendix Table 3 Computed steric descriptors

ID	SASA/Å^2	L/Å	B_1/Å	B_5/Å
1	7.19	3.07	3.03	8.16
2	9.65	2.9	2.94	8
3	2.9	3.42	3.14	9.21
4	1.01	7.47	3.25	9.56
5	6.08	8.06	3.09	8.72
6	6.22	7.83	2.93	10.67
7	19.99	2.52	1.7	4.85
8	21.29	2.5	1.95	4.46
9	16.07	2.5	2.74	4.88
10	24.72	2.51	1.76	5.93
11	26.89	2.49	1.91	3.6
12	18.77	2.5	2.35	4.19
13	31.04	2.5	1.74	5.14
14	29.88	2.5	1.74	4.51
15	28.14	2.71	1.67	4.52
16	28.29	2.5	1.73	4.33
17	16.55	2.5	2.62	4.17
18	32.01	2.5	1.74	4.64

19	29.4	2.5	1.68	5.19
20	25.15	2.51	1.67	4.76
21	27.76	2.49	1.9	3.51
22	26.51	2.5	1.69	4.39
23	32.3	2.49	1.91	3.55
24	12.16	3.36	2.69	5.44
25	21.19	3.27	1.8	5.6
26	29.01	2.49	1.91	4.71
27	10.57	3.43	2.4	5.46
28	20.86	3.3	1.79	5.56
29	29.98	2.49	1.92	4.57
30	22.31	3.37	1.77	5.48
31	30.7	2.49	1.91	4.67
32	12.11	3.39	2.43	5.62
33	18.05	2.53	1.7	5.74
34	14.87	3.27	1.96	5.72
35	11.15	2.5	3.16	4.9
36	24.23	2.51	1.76	7.29
37	27.71	2.49	1.91	4.61
38	14.62	2.63	2.81	6.37
39	8.78	2.54	3.17	4.79
40	27.81	2.5	1.86	6.76

41	33.17	2.49	1.91	5.93
42	20.32	2.65	1.77	5.16
43	14.97	4.19	1.98	5.35
44	25.88	2.51	1.7	7.28
45	26.6	2.49	1.91	6
46	17.57	2.5	2.66	3.84
47	14.62	2.53	2.72	4.49
48	10.86	2.52	3.14	4.5
49	13.42	2.74	2.98	4.49
50	8.25	2.94	3.12	4.51
51	9.41	2.99	3.31	4.53
52	7.24	3.11	3.38	4.51
53	8.98	3.61	3.02	16.53
54	9.41	4.33	2.91	14.11
55	3.04	3.84	3.43	11.15
56	7.24	3.9	3.18	12.26
57	7.82	4.56	3.61	11.96
58	11.77	4.03	3.36	13.98
59	9.41	4.63	3.39	13.99
60	3.33	4.78	2.99	15.48
61	13.08	3.6	2.73	17.18
62	14.92	10.05	1.98	16.31

63	24.47	2.5	3.03	16.53
64	3.76	3.78	2.99	5.77
65	3.33	3.69	2.99	5
66	1.74	3.75	3	5.83
67	3.76	3.86	2.99	5.77
68	5.07	3.42	2.79	5.73
69	0.87	3.48	2.83	5.73
70	1.59	3.49	3.62	5.76
71	4.49	3.41	2.84	5.71
72	3.86	3.48	2.77	6.02
73	4.29	3.51	2.78	7.07
74	2.56	3.48	2.8	6.07
75	0.29	3.56	3.3	6.18
76	2.03	4.53	3.16	5.24
77	2.9	3.62	3.68	5.83
78	2.61	4.5	3.16	6.06
79	1.74	4.47	3.18	5.71

Appendix Table 4 Relaxed force constant, aSASA descriptor, and free energy of activation calculated by DFT

ID	Relaxed force constant/mdyne Å^{-1}	aSASA = sigmoid(SASA/6.71)	$\Delta G^{\ddagger}_{\mathrm{DFT}}$/kcal mol^{-1}
1	5.18	0.745	32.7

2	4.93	0.808	26.7
3	5	0.606	34.7
4	4.93	0.538	36.4
5	4.81	0.712	30.3
6	5.18	0.716	32.8
7	4.76	0.952	24.5
8	5.24	0.960	28.5
9	5	0.916	26.8
10	4.81	0.975	23.7
11	5.26	0.982	28.3
12	4.72	0.943	21.4
13	4.78	0.990	22
14	5	0.988	21.8
15	4.74	0.985	23.1
16	5.1	0.985	22.2
17	5.03	0.922	26.7
18	5.05	0.992	23.2
19	4.81	0.988	22.3
20	4.72	0.977	24
21	5.41	0.984	28.9
22	4.85	0.981	25.5
23	5.29	0.992	29.4

24	4.85	0.860	25
25	4.9	0.959	24.7
26	5.26	0.987	27.4
27	4.76	0.829	24.4
28	4.88	0.957	25.4
29	5.26	0.989	28.4
30	4.78	0.965	23.5
31	5.38	0.990	27.4
32	4.81	0.859	25.7
33	4.76	0.936	24.7
34	5.32	0.902	29.9
35	4.95	0.840	28.9
36	4.81	0.974	23.4
37	5.24	0.984	27.9
38	4.93	0.898	26.8
39	4.63	0.787	26.3
40	4.85	0.984	24.7
41	5.29	0.993	28.4
42	4.81	0.954	24.2
43	5.26	0.903	30.2
44	4.81	0.979	23
45	5.26	0.981	28

46	5.13	0.932	21.3
47	5.05	0.898	23.5
48	5.03	0.835	27.7
49	4.98	0.881	21.3
50	5.03	0.774	25.4
51	5.03	0.803	28.9
52	5.05	0.746	27
53	5.18	0.792	32.7
54	4.85	0.803	29.4
55	5.03	0.611	35.2
56	4.83	0.746	31.6
57	4.81	0.762	30.6
58	5.08	0.852	23.1
59	5.1	0.803	33.5
60	5.18	0.622	35.8
61	5	0.875	27.4
62	5	0.902	26.2
63	5.24	0.975	28.9
64	5.08	0.637	34.3
65	5.13	0.622	34.2
66	5.15	0.564	42.8
67	5.08	0.637	36.4

68	4.95	0.680	31.8
69	5.03	0.532	37.5
70	4.95	0.559	35.9
71	5	0.661	31.8
72	4.98	0.640	30.9
73	4.98	0.655	31.4
74	5	0.594	36.3
75	4.98	0.511	37.1
76	4.93	0.575	37.8
77	4.95	0.606	38.2
78	4.93	0.596	35.4
79	4.93	0.564	38.6

Bibliography

(1) Castellino, N. J.; Montgomery, A. P.; Danon, J. J.; Kassiou, M. Late-stage Functionalization for Improving Drug-like Molecular Properties. *Chem. Rev.* **2023.** DOI: 10.1021/acs.chemrev.2c00797.

(2) Niwa, T.; Murayama, N.; Imagawa, Y.; Yamazaki, H. Regioselective hydroxylation of steroid hormones by human cytochromes P450. *Drug Metab. Rev.* **2015**, *47* (2), 89–110. DOI: 10.3109/03602532.2015.1011658.

(3) Michaudel, Q.; Journot, G.; Regueiro-Ren, A.; Goswami, A.; Guo, Z.; Tully, T. P.; Zou, L.; Ramabhadran, R. O.; Houk, K. N.; Baran, P. S. Improving Physical Properties via C–H Oxidation: Chemical and Enzymatic Approaches. *Angew. Chem. Int. Ed.* **2014**, *53* (45), 12091–12096. DOI: 10.1002/anie.201407016.

(4) Bovicelli, P.; Lupattelli, P.; Mincione, E.; Prencipe, T.; Curci, R. Oxidation of Natural Targets by Dioxiranes. Oxyfunctionalization of Steroids. *J. Org. Chem.* **1992**, *57* (7), 2182–2184. DOI: 10.1021/jo00033a053.

(5) Adams, A. M.; Du Bois, J. Organocatalytic C–H hydroxylation with Oxone® enabled by an aqueous fluoroalcohol solvent system. *Chem. Sci.* **2014**, *5* (2), 656–659. DOI: 10.1039/C3SC52649F.

(6) Arnone, A.; Foletto, S.; Metrangolo, P.; Pregnolato, M.; Resnati, G. Highly Enantiospecific Oxyfunctionalization of Nonactivated Hydrocarbon Sites by Perfluoro-*cis*-2-*n*-butyl-3-*n*-propyloxaziridine. *Org. Lett.* **1999**, *1* (2), 281–284. DOI: 10.1021/ol990594e.

(7) Gormisky, P. E.; White, M. C. Catalyst-Controlled Aliphatic C–H Oxidations with a Predictive Model for Site-Selectivity. *J. Am. Chem. Soc.* **2013**, *135* (38), 14052–14055. DOI: 10.1021/ja407388y.

(8) Wein, L. A.; Wurst, K.; Angyal, P.; Weisheit, L.; Magauer, T. Synthesis of (–)-Mitrephorone A via a Bioinspired Late Stage C–H Oxidation of (–)-Mitrephorone B. *J. Am. Chem. Soc.* **2019**, *141* (50), 19589–19593. DOI: 10.1021/jacs.9b11646.

(9) Saito, M.; Kawamata, Y.; Meanwell, M.; Navratil, R.; Chiodi, D.; Carlson, E.; Hu, P.; Chen, L.; Udyavara, S.; Kingston, C.; Tanwar, M.; Tyagi, S.; McKillican, B. P.; Gichinga, M. G.; Schmidt, M. A.; Eastgate, M. D.; Lamberto, M.; He, C.; Tang, T.; Malapit, C. A.; Sigman, M. S.; Minteer, S. D.; Neurock, M.; Baran, P. S. *N*-Ammonium Ylide Mediators for Electrochemical C–H Oxidation. *J. Am. Chem. Soc.* **2021**, *143* (20), 7859–7867. DOI: 10.1021/jacs.1c03780.

(10) Li, F.; Deng, H.; Renata, H. Remote B-Ring Oxidation of Sclareol with an Engineered P450 Facilitates Divergent Access to Complex Terpenoids. *J. Am. Chem. Soc.* **2022**, *144* (17), 7616–7621. DOI: 10.1021/jacs.2c02958.

(11) Horiguchi, T.; Cheng, Q.; Oritani, T. Highly regio- and stereospecific hydroxylation of C-1 position of 2-deacetoxytaxinine J derivative with DMDO. *Tetrahedron Lett.* **2000**, *41* (20), 3907–3910. DOI: 10.1016/S0040-4039(00)00514-1.

(12) Iida, T.; Yamaguchi, T.; Nakamori, R.; Hikosaka, M.; Mano, N.; Goto, J.; Nambara, T. A highly efficient, stereoselective oxyfunctionalization of unactivated carbons in steroids with dimethyldioxirane. *J. Chem. Soc., Perkin Trans. 1* **2001** (18), 2229–2236. DOI: 10.1039/b104938k.

(13) Curci, R.; D'Accolti, L.; Fusco, C. A Novel Approach to the Efficient Oxygenation of Hydrocarbons under Mild Conditions. Superior Oxo Transfer Selectivity Using Dioxiranes. *Acc. Chem. Res.* **2006**, *39* (1), 1–9. DOI: 10.1021/ar050163y.

(14) Adam, W.; Curci, R.; Edwards, J. O. Dioxiranes: A New Class of Powerful Oxidants. *Acc. Chem. Res.* **1989**, *22* (6), 205–211. DOI: 10.1021/ar00162a002.

(15) Blanksby, S. J.; Ellison, G. B. Bond dissociation energies of organic molecules. *Acc. Chem. Res.* **2003**, *36* (4), 255–263. DOI: 10.1021/ar020230d.

(16) Hioe, J.; Zipse, H. Radical stability and its role in synthesis and catalysis. *Org. Biomol. Chem.* **2010**, *8* (16), 3609–3617. DOI: 10.1039/c004166a.

(17) Gorelik, D. J.; Turner, J. A.; Virk, T. S.; Foucher, D. A.; Taylor, M. S. Site- and Stereoselective C–H Alkylations of Carbohydrates Enabled by Cooperative Photoredox, Hydrogen Atom Transfer, and Organotin Catalysis. *Org. Lett.* **2021**, *23* (13), 5180–5185. DOI: 10.1021/acs.orglett.1c01718.

(18) Curci, R.; D'Accolti, L.; Fiorentino, M.; Fusco, C.; Adam, W.; González-Nuñez, M. E.; Mello, R. Oxidation of acetals, an orthoester, and ethers by dioxiranes through α-CH insertion. *Tetrahedron Lett.* **1992**, *33* (29), 4225–4228. DOI: 10.1016/S0040-4039(00)74695-8.

(19) Bovicelli, P.; Gambacorta, A.; Lupattelli, P.; Mincione, E. A Highly Regio- and Stereoselective C_5 Oxyfunctionalization of Coprostane Steroids by Dioxiranes: an Improved Access to Progestogen and Androgen Hormones. *Tetrahedron Lett.* **1992**, *33* (48), 7411–7412. DOI: 10.1016/S0040-4039(00)60202-2.

(20) Bovicelli, P.; Lupattelli, P.; Fiorini, V.; Mincione, E. Oxyfunctionalization of Steroids by Dioxiranes: Site and Stereoselective C_{14} and C_{17} Hydroxylation of Pregnane and Androstane Steroids. *Tetrahedron Lett.* **1993**, *34* (38), 6103–6104. DOI: 10.1016/S0040-4039(00)61739-2.

(21) Bovicelli, P.; Lupattelli, P.; Mincione, E.; Prencipe, T.; Curci, R. Oxidation of Natural Targets by Dioxiranes. 2. Direct Hydroxylation at the Side-Chain C-25 of Cholestane Derivatives and of Vitamin D_3 Windaus–Grundmann Ketone. *J. Org. Chem.* **1992**, *57* (19), 5052–5054. DOI: 10.1021/jo00045a004.

(22) Bovicelli, P.; Lupattelli, P.; Fracassi, D.; Mincione, E. Sapogenins and Dimethyldioxirane: a New Entry to Cholestanes Functionalized at the Side Chain. *Tetrahedron Lett.* **1994**, *35* (6), 935–938. DOI: 10.1016/S0040-4039(00)76004-7.

(23) Wender, P. A.; Hilinski, M. K.; Mayweg, A. V. W. Late-Stage Intermolecular CH Activation for Lead Diversification: A Highly Chemoselective Oxyfunctionalization of the C-9 Position of Potent Bryostatin Analogues. *Org. Lett.* **2005**, *7* (1), 79–82. DOI: 10.1021/ol047859w.

(24) Voigt, B.; Porzel, A.; Golsch, D.; Adam, W.; Adam, G. Regioselective Oxyfunctionalization of Brassinosteroids by Methyl(trifluoromethyl)dioxirane: Synthesis of 25-Hydroxy-brassinolide and 25-Hydroxy-24-epibrassinolide by Direct C–H Insertion. *Tetrahedron* **1996**, *52* (32), 10653–10658. DOI: 10.1016/0040-4020(96)00587-X.

(25) Kawamura, S.; Chu, H.; Felding, J.; Baran, P. S. Nineteen-step total synthesis of (+)-phorbol. *Nature* **2016**, *532* (7597), 90–93. DOI: 10.1038/nature17153.

(26) Cassidei, L.; Fiorentino, M.; Mello, R.; Sciacovelli, O.; Curci, R. Oxygen-17 and Carbon-13 Identification of the Dimethyldioxirane Intermediate Arising in the Reaction of Potassium Caroate with Acetone. *J. Org. Chem.* **1987**, *52* (4), 699–700. DOI: 10.1021/jo00380a045.

(27) Ehinger, C.; Gordon, C. P.; Copéret, C. Oxygen transfer in electrophilic epoxidation probed by ^{17}O NMR: differentiating between oxidants and role of spectator metal oxo. *Chem. Sci.* **2019**, *10* (6), 1786–1795. DOI: 10.1039/c8sc04868a.

(28) Murray, R. W.; Gu, H. Linear Free Energy Relationship Studies of the Dimethyldioxirane C–H Bond Insertion Reaction. *J. Org. Chem.* **1995**, *60* (17), 5673–5677.

(29) Adam, W.; Curci, R.; D'Accolti, L.; Dinoi, A.; Fusco, C.; Gasparrini, F.; Kluge, R.; Paredes, R.; Schulz, M.; Smerz, A. K.; Angela Veloza, L.; Weinkötz, S.; Winde, R. Epoxidation and Oxygen Insertion into Alkane CH Bonds by Dioxirane Do Not Involve Detectable Radical Pathways. *Chem. Eur. J.* **1997**, *3* (1), 105–109. DOI: 10.1002/chem.19970030117.

(30) Murray, R. W.; Jeyaraman, R.; Mohan, L. Chemistry of Dioxiranes. 4. Oxygen Atom Insertion into Carbon-Hydrogen Bonds by Dimethyldioxirane. *J. Am. Chem. Soc.* **1986**, *108* (9), 2470–2472. DOI: 10.1021/ja00269a069.

(31) Zou, L.; Paton, R. S.; Eschenmoser, A.; Newhouse, T. R.; Baran, P. S.; Houk, K. N. Enhanced Reactivity in Dioxirane C-H Oxidations via Strain Release: A Computational and Experimental Study. *J. Org. Chem.* **2013**, *78* (8), 4037–4048. DOI: 10.1021/jo400350v.

(32) Bach, R. D. The DMDO Hydroxylation of Hydrocarbons via the Oxygen Rebound Mechanism. *J. Phys. Chem. A* **2016**, *120* (5), 840–850. DOI: 10.1021/acs.jpca.5b12086.

(33) Yang, Z.; Yu, P.; Houk, K. N. Molecular Dynamics of Dimethyldioxirane C-H Oxidation. *J. Am. Chem. Soc.* **2016**, *138* (12), 4237–4242. DOI: 10.1021/jacs.6b01028.

(34) Liu, F.; Yang, Z.; Yu, Y.; Mei, Y.; Houk, K. N. Bimodal Evans-Polanyi Relationships in Dioxirane Oxidations of sp^3 C-H: Non-perfect Synchronization in Generation of Delocalized Radical Intermediates. *J. Am. Chem. Soc.* **2017**, *139* (46), 16650–16656. DOI: 10.1021/jacs.7b07988.

(35) Cerrè, C.; Hofmann, A. F.; Schteingart, C. D.; Jia, W.; Maltby, D. Oxyfunctionalization of (5β)-Bile Acids by Dimethyldioxirane: Hydroxylation at C-5, C-14, and C-17. *Tetrahedron* **1997**, *53* (2), 435–446. DOI: 10.1016/S0040-4020(96)01062-9.

(36) Cammarota, R. C.; Liu, W.; Bacsa, J.; Davies, H. M. L.; Sigman, M. S. Mechanistically Guided Workflow for Relating Complex Reactive Site Topologies to Catalyst Performance in C–H Functionalization Reactions. *J. Am. Chem. Soc.* **2022**, *144* (4), 1881–1898. DOI: 10.1021/jacs.1c12198.

(37) Morales-Rivera, C. A.; Floreancig, P. E.; Liu, P. Predictive Model for Oxidative C-H Bond Functionalization Reactivity with 2,3-Dichloro-5,6-dicyano-1,4-benzoquinone. *J. Am. Chem. Soc.* **2017**, *139* (49), 17935–17944. DOI: 10.1021/jacs.7b08902.

(38) Margrey, K. A.; McManus, J. B.; Bonazzi, S.; Zecri, F.; Nicewicz, D. A. Predictive Model for Site-Selective Aryl and Heteroaryl C–H Functionalization via Organic Photoredox Catalysis. *J. Am. Chem. Soc.* **2017**, *139* (32), 11288–11299. DOI: 10.1021/jacs.7b06715.

(39) Deringer, V. L.; Bartók, A. P.; Bernstein, N.; Wilkins, D. M.; Ceriotti, M.; Csányi, G. Gaussian Process Regression for Materials and Molecules. *Chem. Rev.* **2021**, *121* (16), 10073–10141. DOI: 10.1021/acs.chemrev.1c00022.

(40) Exterkate, P. Model selection in kernel ridge regression. *Comput Stat Data Anal* **2013**, *68*, 1–16. DOI: 10.1016/j.csda.2013.06.006.

(41) Vu, K.; Snyder, J. C.; Li, L.; Rupp, M.; Chen, B. F.; Khelif, T.; Müller, K.-R.; Burke, K. Understanding Kernel Ridge Regression: Common Behaviors from Simple Functions to Density Functionals. *Int J Quantum Chem* **2015**, *115* (16), 1115–1128. DOI: 10.1002/qua.24939.

(42) Stuke, A.; Todorović, M.; Rupp, M.; Kunkel, C.; Ghosh, K.; Himanen, L.; Rinke, P. Chemical diversity in molecular orbital energy predictions with kernel ridge regression. *J. Chem. Phys.* **2019**, *150* (20), 204121. DOI: 10.1063/1.5086105.

(43) Sheridan, R. P. Using Random Forest to Model the Domain Applicability of Another Random Forest Model. *J. Chem. Inf. Model.* **2013**, *53* (11), 2837–2850. DOI: 10.1021/ci400482e.

(44) Lovatti, B. P.; Nascimento, M. H.; Neto, Á. C.; Castro, E. V.; Filgueiras, P. R. Use of Random forest in the identification of important variables. *Microchem. J.* **2019**, *145*, 1129–1134. DOI: 10.1016/j.microc.2018.12.028.

(45) Migliaro, I.; Cundari, T. R. Density Functional Study of Methane Activation by Frustrated Lewis Pairs with Group 13 Trihalides and Group 15 Pentahalides and a Machine Learning Analysis of Their Barrier Heights. *J. Chem. Inf. Model.* **2020**, *60* (10), 4958–4966. DOI: 10.1021/acs.jcim.0c00862.

(46) Jorner, K.; Brinck, T.; Norrby, P.-O.; Buttar, D. Machine learning meets mechanistic modelling for accurate prediction of experimental activation energies. *Chem. Sci.* **2021**, *12* (3), 1163–1175. DOI: 10.1039/d0sc04896h.

(47) Palazzesi, F.; Hermann, M. R.; Grundl, M. A.; Pautsch, A.; Seeliger, D.; Tautermann, C. S.; Weber, A. BIreactive: A Machine-Learning Model to Estimate Covalent Warhead Reactivity. *J. Chem. Inf. Model.* **2020**, *60* (6), 2915–2923. DOI: 10.1021/acs.jcim.9b01058.

(48) Li, X.; Zhang, S.-Q.; Xu, L.-C.; Hong, X. Predicting Regioselectivity in Radical C−H Functionalization of Heterocycles through Machine Learning. *Angew. Chem. Int. Ed.* **2020**, *59* (32), 13252–13259. DOI: 10.1002/anie.202000959.

(49) Santiago, C. B.; Guo, J.-Y.; Sigman, M. S. Predictive and mechanistic multivariate linear regression models for reaction development. *Chem. Sci.* **2018**, *9* (9), 2398–2412. DOI: 10.1039/c7sc04679k.

(50) Bragato, M.; Rudorff, G. F. von; Lilienfeld, O. A. von. Data enhanced Hammett-equation: reaction barriers in chemical space. *Chem. Sci.* **2020**, *11* (43), 11859–11868. DOI: 10.1039/d0sc04235h.

(51) Geurts, P.; Ernst, D.; Wehenkel, L. Extremely randomized trees. *Mach Learn* **2006**, *63* (1), 3–42. DOI: 10.1007/s10994-006-6226-1.

(52) Niemeyer, Z. L.; Milo, A.; Hickey, D. P.; Sigman, M. S. Parameterization of phosphine ligands reveals mechanistic pathways and predicts reaction outcomes. *Nat. Chem.* **2016**, *8*, 610–617. DOI: 10.1038/nchem.2501.

(53) Ahneman, D. T.; Estrada, J. G.; Lin, S.; Dreher, S. D.; Doyle, A. G. Predicting reaction performance in C–N cross-coupling using machine learning. *Science* **2018**, *360* (6385), 186–190. DOI: 10.1126/science.aar5169.

(54) Shen, Q.; Jiang, J.-H.; Tao, J.-C.; Shen, G.-L.; Yu, R.-Q. Modified Ant Colony Optimization Algorithm for Variable Selection in QSAR Modeling: QSAR Studies of Cyclooxygenase Inhibitors. *J. Chem. Inf. Model.* **2005**, *45* (4), 1024–1029. DOI: 10.1021/ci049610z.

(55) Gensch, T.; Dos Passos Gomes, G.; Friederich, P.; Peters, E.; Gaudin, T.; Pollice, R.; Jorner, K.; Nigam, A.; Lindner-D'Addario, M.; Sigman, M. S.; Aspuru-Guzik, A. A Comprehensive Discovery Platform for Organophosphorus Ligands for Catalysis. *J. Am. Chem. Soc.* **2022**, *144* (3), 1205–1217. DOI: 10.1021/jacs.1c09718.

(56) Hueffel, J. A.; Sperger, T.; Funes-Ardoiz, I.; Ward, J. S.; Rissanen, K.; Schoenebeck, F. Accelerated dinuclear palladium catalyst identification through unsupervised machine learning. *Science* **2021**, *374* (6571), 1134–1140. DOI: 10.1126/science.abj0999.

(57) Keylor, M. H.; Niemeyer, Z. L.; Sigman, M. S.; Tan, K. L. Inverting Conventional Chemoselectivity in Pd-Catalyzed Amine Arylations with Multiply Halogenated Pyridines. *J. Am. Chem. Soc.* **2017**, *139* (31), 10613–10616. DOI: 10.1021/jacs.7b05409.

(58) Mathew, J.; Suresh, C. H. Assessment of Stereoelectronic Effects in Grubbs First-Generation Olefin Metathesis Catalysis Using Molecular Electrostatic Potential. *Organometallics* **2011**, *30* (6), 1438–1444. DOI: 10.1021/om101034a.

(59) Mathew, J.; Suresh, C. H. Assessment of Steric and Electronic Effects of N-Heterocyclic Carbenes in Grubbs Olefin Metathesis Using Molecular Electrostatic Potential. *Organometallics* **2011**, *30* (11), 3106–3112. DOI: 10.1021/om200196u.

(60) Suresh, C. H.; Anila, S. Molecular Electrostatic Potential Topology Analysis of Noncovalent Interactions. *Acc. Chem. Res.* **2023**, *56* (13), 1884–1895. DOI: 10.1021/acs.accounts.3c00193.

(61) Winstein, S.; Holness, N. J. Neighboring Carbon and Hydrogen. XIX. *t*-Butylcyclohexyl Derivatives. Quantitative Conformational Analysis. *J. Am. Chem. Soc.* **1955**, *77* (21), 5562–5578. DOI: 10.1021/ja01626a037.

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

(63) Clavier, H.; Nolan, S. P. Percent buried volume for phosphine and *N*-heterocyclic carbene ligands: steric properties in organometallic chemistry. *ChemComm* **2010**, *46* (6), 841–861. DOI: 10.1039/b922984a.

(64) Guzei, I. A.; Wendt, M. An improved method for the computation of ligand steric effects based on solid angles. *Dalton Trans.* **2006**, *128* (33), 3991–3999. DOI: 10.1039/b605102b.

(65) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Mennucci, B.; Petersson, G. A.; Nakatsuji, H.; Caricato, M.; Li, X.; Hratchian, H. P.; Izmaylov, A. F.; Bloino, J.; Zheng, G.; Sonnenberg, J. L.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Vreven, T.; Montgomery, J. A., Jr.; Peralta, J. E.; Ogliaro, F.; Bearpark, M.; Heyd, J. J.; Brothers, E.; Kudin, K. N.; Staroverov, V. N.; Kobayashi, R.; Normand, J.; Raghavachari, K.; Rendell, A.; Burant, J. C.; Iyengar, S. S.; Tomasi, J.; Cossi, M.; Rega, N.; Millam, N. J.; Klene, M.; Knox, J. E.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Martin, R. L.; Morokuma, K.; Zakrzewski, V. G.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Dapprich, S.; Daniels, A. D.; Farkas, O.; Foresman, J. B.; Ortiz, J. V.; Cioslowski, J.; Fox, D. J. Gaussian 09, Revision D.01:Gaussian, Inc., Wallingford CT, 2013. In.

(66) 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. *J. Phys. Chem.* **1994**, *98* (45), 11623–11627.

(67) Ditchfield, R.; Hehre, W. J.; Pople, J. A. Self-Consistent Molecular-Orbital Methods. IX. An Extended Gaussian-Type Basis for Molecular-Orbital Studies of Organic Molecules. *J. Chem. Phys.* **1971**, *54* (2), 724–728. DOI: 10.1063/1.1674902.

(68) Hehre, W. J.; Ditchfield, R.; Pople, J. A. Self—Consistent Molecular Orbital Methods. XII. Further Extensions of Gaussian—Type Basis Sets for Use in Molecular Orbital Studies of Organic Molecules. *J. Chem. Phys.* **1972**, *56* (5), 2257–2261. DOI: 10.1063/1.1677527.

(69) Hariharan, P. C.; Pople, J. A. The Influence of Polarization Functions on Molecular Orbital Hydrogenation Energies. *Theor. Chim. Acta* **1973**, *28* (213–222). DOI: 10.1007/BF00533485.

(70) Hariharan, P. C.; Pople, J. A. Accuracy of AH_n equilibrium geometries by single determinant molecular orbital theory. *Mol Phys* **1974**, *27* (1), 209–214. DOI: 10.1080/00268977400100171.

(71) Gordon, M. S. The isomers of silacyclopropane. *Chem. Phys. Lett.* **1980**, *76* (1), 163–168. DOI: 10.1016/0009-2614(80)80628-2.

(72) 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. *J. Chem. Phys.* **1982**, *77* (7), 3654–3665. DOI: 10.1063/1.444267.

(73) Binning, R. C.; Curtiss, L. A. Compact Contracted Basis Sets for Third-Row Atoms: Ga-Kr. *J. Comput. Chem.* **1990**, *11* (10), 1206–1216. DOI: 10.1002/jcc.540111013.

(74) Blaudeau, J.-P.; McGrath, M. P.; Curtiss, L. A.; Radom, L. Extension of Gaussian-2 (G2) theory to molecules containing third-row atoms K and Ca. *J. Chem. Phys.* **1997**, *107* (13), 5016–5021. DOI: 10.1063/1.474865.

(75) Rassolov, V. A.; Pople, J. A.; Ratner, M. A.; Windus, T. L. 6-31G* basis set for atoms K through Zn. *J. Chem. Phys.* **1998**, *109* (4), 1223–1229. DOI: 10.1063/1.476673.

(76) Rassolov, V. A.; Ratner, M. A.; Pople, J. A.; Redfern, P. C.; Curtiss, L. A. 6-31G* Basis Set for Third-Row Atoms. *J. Comput. Chem.* **2001**, *22* (9), 976–984. DOI: 10.1002/jcc.1058.

(77) Brandhorst, K.; Grunenberg, J. How strong is it? The interpretation of force and compliance constants as bond strength descriptors. *Chem. Soc. Rev.* **2008**, *37* (8), 1558–1567. DOI: 10.1039/b717781j.

(78) Lee, B.; Richards, F. M. The Interpretation of Protein Structures: Estimation of Static Accessibility. *J. Mol. Biol.* **1971**, *55* (3), 379–400. DOI: 10.1016/0022-2836(71)90324-X.

(79) Humphrey, W.; Dalke, A.; Schulten, K. VMD: Visual Molecular Dynamics. *J. Molec. Graphics* **1996**, *14*, 33–38.

(80) McLean, A. D.; Chandler, G. S. Contracted Gaussian basis sets for molecular calculations. I. Second row atoms, Z =11–18. *J. Chem. Phys.* **1980**, *72* (10), 5639–5648. DOI: 10.1063/1.438980.

(81) Krishnan, R.; Binkley, J. S.; Seeger, R.; Pople, J. A. Self‐consistent molecular orbital methods. XX. A basis set for correlated wave functions. *J. Chem. Phys.* **1980**, *72* (1), 650–654. DOI: 10.1063/1.438955.

(82) Wachters, A. J. H. Gaussian Basis Set for Molecular Wavefunctions Containing Third-Row Atoms. *J. Chem. Phys.* **1970**, *52* (3), 1033–1036. DOI: 10.1063/1.1673095.

(83) Hay, P. J. Gaussian basis sets for molecular calculations. The representation of 3*d* orbitals in transition‐metal atoms. *J. Chem. Phys.* **1977**, *66* (10), 4377–4384. DOI: 10.1063/1.433731.

(84) Raghavachari, K.; Trucks, G. W. Highly correlated systems. Excitation energies of first row transition metals Sc–Cu. *J. Chem. Phys.* **1989**, *91* (2), 1062–1065. DOI: 10.1063/1.457230.

(85) McGrath, M. P.; Radom, L. Extension of Gaussian-1 (G1) theory to bromine-containing molecules. *J. Chem. Phys.* **1991**, *94* (1), 511–516. DOI: 10.1063/1.460367.

(86) Curtiss, L. A.; McGrath, M. P.; Blaudeau, J.-P.; Davis, N. E.; Binning, R. C.; Radom, L. Extension of Gaussian-2 theory to molecules containing third-row atoms Ga–Kr. *J. Chem. Phys.* **1995**, *103* (14), 6104–6113. DOI: 10.1063/1.470438.

(87) Reed, A. E.; Weinstock, R. B.; Weinhold, F. Natural population analysis. *J. Chem. Phys.* **1985**, *83* (2), 735–746. DOI: 10.1063/1.449486.

(88) Breneman, C. M.; Wiberg, K. B. Determining Atom-Centered Monopoles from Molecular Electrostatic Potentials. The Need for High Sampling Density in Formamide Conformational Analysis. *J. Comput. Chem.* **1990**, *11* (3), 361–373. DOI: 10.1002/jcc.540110311.

(89) Lu, T.; Chen, F. Bond Order Analysis Based on the Laplacian of Electron Density in Fuzzy Overlap Space. *J. Phys. Chem. A* **2013**, *117* (14), 3100–3108. DOI: 10.1021/jp4010345.

(90) 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. *J. Phys. Chem. B* **2009**, *113* (18), 6378–6396. DOI: 10.1021/jp810292n.

(91) Kennard, R. W.; Stone, L. A. Computer Aided Design of Experiments. *Technometrics* **1969**, *11* (1), 137–148. DOI: 10.1080/00401706.1969.10490666.

(92) Fu, Y.; Zerull, E. E.; Schomaker, J. M.; Liu, P. Origins of Catalyst-Controlled Selectivity in Ag-Catalyzed Regiodivergent C–H Amination. *J. Am. Chem. Soc.* **2022**, *144* (6), 2735–2746. DOI: 10.1021/jacs.1c12111.

(93) Akaike, H. A New Look at the Statistical Model Identification. *IEEE Trans. Automat. Contr.* **1974**, *19* (6), 716–723. DOI: 10.1109/TAC.1974.1100705.

(94) Reed, L. J.; Berkson, J. The Application of the Logistic Function to Experimental Data. *J. Phys. Chem.* **1929**, *33* (5), 760–779. DOI: 10.1021/j150299a014.

(95) Levine, R. D. Free Energy of Activation. Definition, Properties, and Dependent Variables with Special Reference to "Linear" Free Energy Relations. *J. Phys. Chem.* **1979**, *83* (1), 159–170. DOI: 10.1021/j100464a023.

(96) DeStefano, J. J. Logistic Regression and the Boltzmann Machine. In *1990 IJCNN International Joint Conference on Neural Networks*; 199-204. DOI: 10.1109/IJCNN.1990.137845.

(97) Kilian, J.; Siegelmann, H. T. The Dynamic Universality of Sigmoidal Neural Networks. *Inf Comput* **1996**, *128* (1), 48–56. DOI: 10.1006/inco.1996.0062.

(98) Muggeo, V. M. R. Estimating regression models with unknown break-points. *Stat Med* **2003**, *22* (19), 3055–3071. DOI: 10.1002/sim.1545.

(99) Muggeo, V. M. R. segmented: an R Package to Fit Regression Models with Broken-Line Relationships. *R News* **2008**, *8* (1), 20–25.

(100) Muggeo, V.; Atkins, D. C.; Gallop, R. J.; Dimidjian, S. Segmented mixed models with random changepoints: a maximum likelihood approach with application to treatment for depression study. *Stat Modelling* **2014**, *14* (4), 293–313. DOI: 10.1177/1471082X13504721.

(101) Muggeo, V. M. R. Testing with a nuisance parameter present only under the alternative: a score-based approach with application to segmented modelling. *J Stat Comput Simul* **2016**, *86* (15), 3059–3067. DOI: 10.1080/00949655.2016.1149855.

(102) Muggeo, V. M. Interval estimation for the breakpoint in segmented regression: a smoothed score-based approach. *Aust N Z J Stat* **2017**, *59* (3), 311–322. DOI: 10.1111/anzs.12200.

(103) Molina, E.; Estrada, E.; Nodarse, D.; Torres, L. A.; González, H.; Uriarte, E. Quantitative Structure-Antibacterial Activity Relationship Modeling Using a Combination of Piecewise Linear Regression-Discriminant Analysis (I): Quantum Chemical, Topographic, and Topological Descriptors. *Int J Quantum Chem* **2008**, *108* (10), 1856–1871. DOI: 10.1002/qua.21702.

(104) Pettit, G. R.; Herald, C. L.; Doubek, D. L.; Herald, D. L.; Arnold, E.; Clardy, J. Isolation and Structure of Bryostatin 1. *J. Am. Chem. Soc.* **1982**, *104* (24), 6846–6848. DOI: 10.1021/ja00388a092.

(105) Pettit, G. R.; Herald, C. L.; Kamano, Y. Structure of the *Bugula neritina* (Narine Bryozoa) Antineoplastic Component Bryostatin 3. *J. Org. Chem.* **1983**, *48* (26), 5354–5356. DOI: 10.1021/jo00174a037.

(106) Pettit, G. R.; Kamano, Y.; Herald, C. L.; Tozawa, M. Structure of Bryostatin 4. An Important Antineoplastic Constituent of Geographically Diverse *Bugula neritina* (Bryozoa). *J. Am. Chem. Soc.* **1984**, *106* (22), 6768–6771. DOI: 10.1021/ja00334a050.

(107) Pettit, G. R.; Kamano, Y.; Herald, C. L.; Tozawa, M. Isolation and structure of bryostatins 5–7. *Can J Chem* **1985**, *63* (6), 1204–1208. DOI: 10.1139/v85-205.

(108) Pettit, G. R.; Kamano, Y.; Aoyagi, R.; Herald, C. L.; Doubek, D. L.; Schmidt, J. M.; Rudloe, J. J. Antineoplastic agents 100: The marine bryozoan *Amathia convoluta*. *Tetrahedron* **1985**, *41* (6), 985–994. DOI: 10.1016/S0040-4020(01)96466-X.

(109) Pettit, G. R.; Kamano, Y.; Herald, C. L. Isolation and Structure of Bryostatins 10 and 11. *J. Org. Chem.* **1987**, *52* (13), 2848–2854. DOI: 10.1021/jo00389a036.

(110) Pettit, G. R.; Leet, J. E.; Herald, C. L.; Kamano, Y.; Boettner, F. E.; Baczynskyj, L.; Nieman, R. A. Isolation and Structure of Bryostatins 12 and 13. *J. Org. Chem.* **1987**, *52* (13), 2854–2860. DOI: 10.1021/jo00389a037.

(111) Pettit, G. R.; Gao, F.; Sengupta, D.; Coll, J. C.; Herald, C. L.; Doubek, D. L.; Schmidt, J. M.; van Camp, J. R.; Rudloe, J. J.; Nieman, R. A. Isolation and structure of bryostatins 14 and 15. *Tetrahedron* **1991**, *47* (22), 3601–3610. DOI: 10.1016/S0040-4020(01)80873-5.

(112) Pettit, G. R.; Kamano, Y.; Herald, C. L.; Schmidt, J. M.; Zubrod, C. G. Relationship of *Bugula neritina* (Bryozoa) antineoplastic constituents to the yellow sponge *Lissodendoryx isodictyalis*. *Pure Appl. Chem.* **1986**, *58* (3), 415–421. DOI: 10.1351/pac198658030415.

(113) Pettit, G. R.; Gao, F.; Blumberg, P. M.; Herald, C. L.; Coll, J. C.; Kamano, Y.; Lewin, N. E.; Schmidt, J. M.; Chapuis, J.-C. Antineoplastic Agents. 340. Isolation and Structural Elucidation of Bryostatins 16–18. *J. Nat. Prod.* **1996**, *59* (3), 286–289. DOI: 10.1021/np960100b.

(114) Mochly-Rosen, D.; Das, K.; Grimes, K. V. Protein kinase C, an elusive therapeutic target? *Nat. Rev. Drug Discov.* **2012**, *11* (12), 937–957. DOI: 10.1038/nrd3871.

(115) Berkow, R. L.; Kraft, A. S. Bryostatin, a non-phorbol macrocyclic lactone, activates intact human polymorphonuclear leukocytes and binds to the phorbol ester receptor. *Biochem. Biophys. Res. Commun.* **1985**, *131* (3), 1109–1116. DOI: 10.1016/0006-291X(85)90205-0.

(116) Hongpaisan, J.; Alkon, D. L. A structural basis for enhancement of long-term associative memory in single dendritic spines regulated by PKC. *Proc. Natl. Acad. Sci. U.S.A.* **2007**, *104* (49), 19571–19576. DOI: 10.1073/pnas.0709311104.

(117) DeChristopher, B. A.; Loy, B. A.; Marsden, M. D.; Schrier, A. J.; Zack, J. A.; Wender, P. A. Designed, synthetically accessible bryostatin analogues potently induce activation of latent HIV reservoirs *in vitro*. *Nat. Chem.* **2012**, *4* (9), 705–710. DOI: 10.1038/nchem.1395.

(118) Pérez-Vargas, J.; Shapira, T.; Olmstead, A. D.; Villanueva, I.; Thompson, C. A. H.; Ennis, S.; Gao, G.; Guzman, J. de; Williams, D. E.; Wang, M.; Chin, A.; Bautista-Sánchez, D.; Agafitei, O.; Levett, P.; Xie, X.; Nuzzo, G.; Freire, V. F.; Quintana-Bulla, J. I.; Bernardi, D. I.; Gubiani, J. R.; Suthiphasilp, V.; Raksat, A.; Meesakul, P.; Polbuppha, I.; Cheenpracha, S.; Jaidee, W.; Kanokmedhakul, K.; Yenjai, C.; Chaiyosang, B.; Teles, H. L.; Manzo, E.; Fontana, A.; Leduc,

R.; Boudreault, P.-L.; Berlinck, R. G. S.; Laphookhieo, S.; Kanokmedhakul, S.; Tietjen, I.; Cherkasov, A.; Krajden, M.; Nabi, I. R.; Niikura, M.; Shi, P.-Y.; Andersen, R. J.; Jean, F. Discovery of lead natural products for developing pan-SARS-CoV-2 therapeutics. *Antiviral Res.* **2023**, *209*, 105484. DOI: 10.1016/j.antiviral.2022.105484.

(119) Kerr, R. G.; Lawry, J.; Gush, K. A. *In Vitro* Biosynthetic Studies of the Bryostatins, Anti-Cancer Agents from the Marine Bryozoan *Bugula neritina*. *Tetrahedron Lett.* **1996**, *37* (46), 8305–8308. DOI: 10.1016/0040-4039(96)01943-0.

(120) Masamune, S. Asymmetric synthesis and its applications: towards the synthesis of bryostatin 1. *Pure Appl. Chem.* **1988**, *60* (11), 1587–1596. DOI: 10.1351/pac198860111587.

(121) Hale, K. J.; Hummersone, M. G.; Manaviazar, S.; Frigerio, M. The chemistry and biology of the bryostatin antitumour macrolides. *Nat. Prod. Rep.* **2002**, *19* (4), 413–453. DOI: 10.1039/b009211h.

(122) Trost, B. M.; Dong, G. Total synthesis of bryostatin 16 using atom-economical and chemoselective approaches. *Nature* **2008**, *456* (7221), 485–488. DOI: 10.1038/nature07543.

(123) Trost, B. M.; Dong, G. Total Synthesis of Bryostatin 16 Using a Pd-Catalyzed Diyne Coupling as Macrocyclization Method and Synthesis of C20-*epi*-Bryostatin 7 as a Potent Anticancer Agent. *J. Am. Chem. Soc.* **2010**, *132* (46), 16403–16416. DOI: 10.1021/ja105129p.

(124) Keck, G. E.; Poudel, Y. B.; Cummins, T. J.; Rudra, A.; Covel, J. A. Total Synthesis of Bryostatin 1. *J. Am. Chem. Soc.* **2011**, *133* (4), 744–747. DOI: 10.1021/ja110198y.

(125) Lu, Y.; Woo, S. K.; Krische, M. J. Total Synthesis of Bryostatin 7 *via* C–C Bond-Forming Hydrogenation. *J. Am. Chem. Soc.* **2011**, *133* (35), 13876–13879. DOI: 10.1021/ja205673e.

(126) Wender, P. A.; Hardman, C. T.; Ho, S.; Jeffreys, M. S.; Maclaren, J. K.; Quiroz, R. V.; Ryckbosch, S. M.; Shimizu, A. J.; Sloane, J. L.; Stevens, M. C. Scalable synthesis of bryostatin 1

and analogs, adjuvant leads against latent HIV. *Science* **2017**, *358* (6360), 218–223. DOI: 10.1126/science.aan7969.

(127) Zhang, Y.; Guo, Q.; Sun, X.; Lu, J.; Cao, Y.; Pu, Q.; Chu, Z.; Gao, L.; Song, Z. Total Synthesis of Bryostatin 8 Using an Organosilane-Based Strategy. *Angew. Chem. Int. Ed.* **2018**, *57* (4), 942–946. DOI: 10.1002/anie.201711452.

(128) Wender, P. A.; Cribbs, C. M.; Koehler, K. F.; Sharkey, N. A.; Herald, C. L.; Kamano, Y.; Pettit, G. R.; Blumberg, P. M. Modeling of the bryostatins to the phorbol ester pharmacophore on protein kinase C. *Proc. Natl. Acad. Sci. U.S.A.* **1988**, *85* (19), 7197–7201. DOI: 10.1073/pnas.85.19.7197.

(129) Wender, P. A.; Brabander, J. de; Harran, P. G.; Jimenez, J.-M.; Koehler, M. F. T.; Lippa, B.; Park, C.-M.; Shiozaki, M. Synthesis of the First Members of a New Class of Biologically Active Bryostatin Analogues. *J. Am. Chem. Soc.* **1998**, *120* (18), 4534–4535. DOI: 10.1021/ja9727631.

(130) Wender, P. A.; Brabander, J. de; Harran, P. G.; Hinkle, K. W.; Lippa, B.; Pettit, G. R. Synthesis and Biological Evaluation of Fully Synthetic Bryostatin Analogues. *Tetrahedron Lett.* **1998**, *39* (47), 8625–8628. DOI: 10.1016/S0040-4039(98)01955-8.

(131) Wender, P. A.; Hinkle, K. W.; Koehler, M. F. T.; Lippa, B. The Rational Design of Potential Chemotherapeutic Agents: Synthesis of Bryostatin Analogues. *Med Res Rev* **1999**, *19* (5), 388–407. DOI: 10.1002/(SICI)1098-1128(199909)19:5<388:AID-MED6>3.0.CO;2-H.

(132) Wender, P. A.; Lippa, B. Synthesis and biological evaluation of bryostatin analogues: the role of the A-ring. *Tetrahedron Lett.* **2000**, *41* (7), 1007–1011. DOI: 10.1016/S0040-4039(99)02195-4.

(133) Wender, P. A.; Verma, V. A.; Paxton, T. J.; Pillow, T. H. Function-oriented synthesis, step economy, and drug design. *Acc. Chem. Res.* **2008**, *41* (1), 40–49. DOI: 10.1021/ar700155p.

(134) Ketcham, J. M.; Volchkov, I.; Chen, T.-Y.; Blumberg, P. M.; Kedei, N.; Lewin, N. E.; Krische, M. J. Evaluation of Chromane-Based Bryostatin Analogues Prepared via Hydrogen-Mediated C–C Bond Formation: Potency Does Not Confer Bryostatin-like Biology. *J. Am. Chem. Soc.* **2016**, *138* (40), 13415–13423. DOI: 10.1021/jacs.6b08695.

(135) Rajendran, N. D.; Mookan, N.; Samuel, I.; Mookan, S. B. Experimental validation of bifurcated hydrogen bond of 2,5-lutidinium bromanilate and its charge density distribution. *Chem. Pap.* **2020**, *74* (8), 2689–2699. DOI: 10.1007/s11696-020-01107-3.

(136) Kamano, Y.; Zhang, H.; Hino, A.; Yoshida, M.; Pettit, G. R.; Herald, C. L.; Itokawa, H. An Improved Source of Bryostatin 10, *Bugula neritina* from the Gulf of Aomori, Japan. *J. Nat. Prod.* **1995**, *58* (12), 1868–1875. DOI: 10.1021/np50126a009.

(137) Kamano, Y.; Zhang, H.; Morita, H.; Itokawa, H.; Shirota, O.; Pettit, G. R.; Herald, D. L.; Herald, C. L. Conformational Analysis of a Marine Antineoplastic Macrolide, Bryostatin 10. *Tetrahedron* **1996**, *52* (7), 2369–2376. DOI: 10.1016/0040-4020(95)01080-7.

(138) Zivanovic, S.; Colizzi, F.; Moreno, D.; Hospital, A.; Soliva, R.; Orozco, M. Exploring the Conformational Landscape of Bioactive Small Molecules. *J. Chem. Theory Comput.* **2020**, *16* (10), 6575–6585. DOI: 10.1021/acs.jctc.0c00304.

(139) Witek, J.; Keller, B. G.; Blatter, M.; Meissner, A.; Wagner, T.; Riniker, S. Kinetic Models of Cyclosporin A in Polar and Apolar Environments Reveal Multiple Congruent Conformational States. *J. Chem. Inf. Model.* **2016**, *56* (8), 1547–1562. DOI: 10.1021/acs.jcim.6b00251.

(140) Yang, H.; Staveness, D.; Ryckbosch, S. M.; Axtman, A. D.; Loy, B. A.; Barnes, A. B.; Pande, V. S.; Schaefer, J.; Wender, P. A.; Cegelski, L. REDOR NMR Reveals Multiple

Conformers for a Protein Kinase C Ligand in a Membrane Environment. *ACS Cent. Sci.* **2018**, *4* (1), 89–96. DOI: 10.1021/acscentsci.7b00475.

(141) Heyer, L. J.; Kruglyak, S.; Yooseph, S. Exploring Expression Data: Identification and Analysis of Coexpressed Genes. *Genome Res.* **1999**, *9*, 1106–1115. DOI: 10.1101/gr.9.11.1106.

(142) Danalis, A.; McCurdy, C.; Vetter, J. S. Efficient Quality Threshold Clustering for Parallel Architectures. In *2012 IEEE 26th International Parallel and Distributed Processing Symposium*; IEEE, 2012; pp 1068–1079. DOI: 10.1109/IPDPS.2012.99.