DISSERTATION

DIGITAL MOLECULAR REPRESENTATIONS FOR REACTION PREDICTION AND OPTIMIZATION

Submitted by:

Guilian W. Luchini

Department of Chemistry

In partial fulfillment of the requirements

For the Degree of Doctor of Philosophy

Colorado State University

Fort Collins, Colorado

Summer 2023

Doctoral Committee:

Advisor: Robert S. Paton

Anthony K. Rappé
Jeffrey S. Bandar
Patrick D. Shipman

ABSTRACT

DIGITAL MOLECULAR REPRESENTATIONS FOR

REACTION PREDICTION AND OPTIMIZATION

The properties of molecules can be related to measurable outcomes from chemical reactions, for example, reaction yield, selectivity, or rate. The unique ways we represent molecules computationally can provide insight into how a particular molecular property influences reaction outcome. This dissertation discusses different ways a molecule can be digitally represented, and properties that can be measured from these molecular representations. The first two chapters provide context and a landscape of the current field of digital molecular representations and properties. Chapter three focuses on a specific type of molecular property, the steric properties of molecules, describing the size and shape that molecules occupy in space. Existing steric parameters have proven useful in how the size or steric bulk of a molecule can influence reaction outcome. We address an opportunity in the literature currently unaccounted for by describing the concept of steric proximity. We quantify how near to a reactive site the steric bulk of a molecule lies in two novel steric parameter sets, Sterimol2vec and vol2vec. Chapter four focuses on another class of molecular properties, electronic properties. Commonly used in computational studies, the partial atomic charge is often used in mechanistic studies to justify reactivity at atomic sites, by providing a conceptual medium for the buildup of electronic charge across a molecule, partitioned into its atoms. This study benchmarks and compares different methods for computing partial atomic charge through comparisons with experimentally tabulated Hammett parameters. We find that the choice of method and the atomic position for which the charge is measured is important in relating to the reactivity of the system. Many computational

studies rely on programming and computational workflows for data collection and analysis. Chapter five is used to summarize open-source Python tools resulting from this and additional work relating to the collection and analysis of molecular properties. Three programs are summarized. GoodVibes is used to compute and apply corrections to thermochemistry data (entropy, enthalpy, and Gibbs free energy), while automating tasks for computing and visualizing relative thermochemistry potential energy surfaces. DBSTEP is used for computing novel steric parameters described in chapter three, along with existing parameters, Sterimol and percent buried volumes. The final program discussed is Py-X Struct, a program designed to query molecules and substructures in X-ray crystal structures from the Cambridge Structural Database, measuring geometric information like bond distances, angles, and dihedrals between user specified atoms. The final chapter summarizes results and potential future directions for these projects.

ACKNOWLEDGEMENTS

I would like to thank Robert Paton, who has been a wonderful mentor and advisor during my graduate studies. I always felt supported in all my work and through his knowledge and direction I have been able to accomplish more than I would have imagined when I began graduate school. He has provided me with many opportunities during my degree to perform to my best ability, allowing me to meet and interact with so many great people all invested in advancing science.

I would also like to thank members of the Paton Lab and the Theory Suite. I will miss doing potlucks, barbecues, board games and going to the Marmot with all of you. Special thanks to Juan V. Alegre-Requena, Heidi Klem, Liliana Gallegos, Louis de Lescure, Shree Sowndarya, Brandon Portela, Yeonjoon Kim, Mihai Popescu, Raúl Pérez-Soto, Yingzi Li, Sreenithya Avadakkam, and Santeri Aikonen, all who made my Ph.D. a bit more fun.

Lastly, I'd like to thank my family. My parents, sister, and brother have all been immensely supportive and encouraging during my time in graduate school and I feel incredibly lucky to always have them a phone call away.

TABLE OF CONTENTS

CHAPTER 1: INTRODUCTION

1.1: Chemical Reaction Prediction

Chemical reaction development and optimization is not a trivial task. Chemists seek to optimize reaction components to achieve the best reactive outcomes, such as yield, selectivity, or reaction rate. Reaction optimization is an important step across broad areas of chemistry, including small molecule organic synthesis, organometallic chemistry, pharmaceutical and agricultural chemistry and materials chemistry.[1] Apart from better reaction outcomes, reaction optimization can involve making reactions greener or more sustainable, or even optimize cost to save on materials.[2]

Reaction optimization efforts often involve the use of predictive modeling. Predictive models developed for reaction optimization often rely on linear relationships between reaction inputs and the reactivity outputs. These "linear free energy relationships" have been studied since the early 20th century by chemists and relate properties of reactants to their reactivity. In general, these linear relationships are constructed from univariate correlations of relative reaction rates to tabulated parameter values. Early examples include Hammett relationships, which relate electronic properties of aryl substituents to reaction rates.[3] Taft was able to construct similar relationships analyzing the steric contributions of substituents on reaction rates.[4] Identifying these relationships aids in pinpointing which components and properties of a molecule influence the reaction outcome, providing insights into reaction mechanism and aiding in optimizing a desired reactivity. This process is visualized in Figure 1.1, where each molecule is "embedded" into a single property or collection of properties, quantitatively representing each molecule as a unique fingerprint that can be related to reaction outcome through simple comparisons or linear regression. In some cases, a reaction may not simply be related to a single variable property, and so more intensive statistical modeling efforts using multivariate linear regression or machine learning algorithms may be more suitable for predicting chemical behavior.[5]

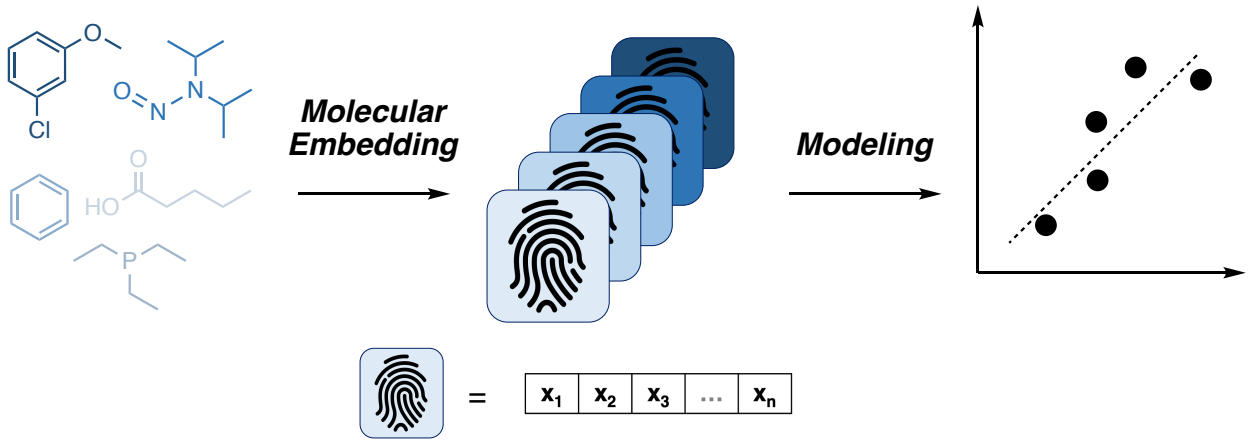


Figure 1.1. *Each molecule in a study can be quantified into a property or series of properties through the "molecular embedding" process. These embeddings, much like a unique fingerprint for a molecule, is then related to chemical reactivity through statistical modeling.*

Molecular properties can be measured experimentally with a "top-down" approach, capturing macroscopic observables from molecules (i.e., from spectroscopic techniques), or with a "bottom-up" approach, by building up the chemical system computationally. Top-down methods rely on experimental measurements, and molecular properties come from spectra, reaction rates, yield, or selectivity measurements. Factors contributing to these outcomes can vary based on reaction conditions. Reaction conditions can be fine-tuned based on the type of reaction being performed, for example, in the cross-coupling reaction shown in Figure 1.2, several conditions can be altered to influence the reaction outcomes. In this example, the desired reaction takes place at the C5 position on the imidazole reactant, however, it is possible that the C2 product is also formed. Additionally, other undesired reaction pathways which include difunctionalization on both C2 and C5 positions, as well as the hydrolysis of the pyrrole ester are also possible. There are several components used to optimize the desired yield and selectivity of the reaction, including the ligand, solvent, or base used, temperature and reaction time.[6,7] Reaction components can increase number of possible reactions rapidly, with each alteration to a component possibly influencing the reaction outcome (Figure 1.2). Typically, conditions are chosen and vetted based on the experimental chemist's prior knowledge, however, recently developed computer-assisted

methods for reaction optimization can aid in expanding and exploring the chemical space of a reaction efficiently.[7]

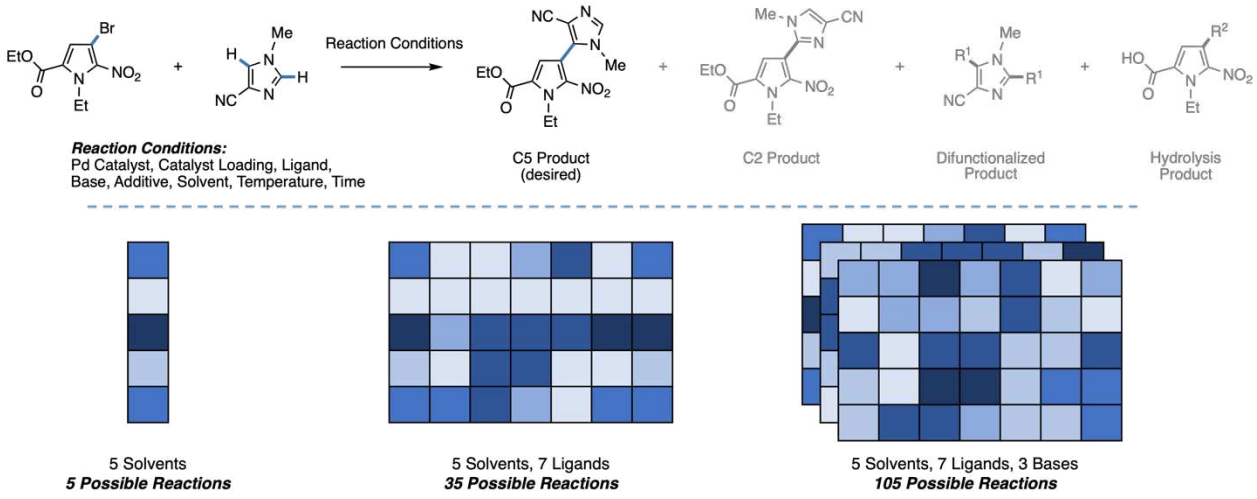


Figure 1.2. *(Top) Cross-coupling reaction between a pyrrole and imidazole, showing desired and potentially undesired product formation. Several reaction conditions can be altered to tune the reactivity to optimize yield and selectivity for the desired product formation. (Bottom) The number of possible reactions rapidly increases as reaction conditions are altered.*

Alternative to an experimental approach, bottom-up computational methods can also aid in reaction optimization. With this approach, computational modeling is used to capture the conformational space and relevant molecular properties to describe the chemical system.[8] Accurate high-throughput simulations using Density Functional Theory (DFT) can subvert the cost and time of running an experiment, and the scope of molecules studied is easily expanded. Computational studies can be performed on a small scale, modeling individual reactants, intermediates, and transition state structures to study a reaction mechanism through a potential energy surface from computed Gibbs free energies. In these studies, three dimensional structures are produced and experimental reaction components like solvent, temperature, and species concentrations can be accounted for.[9] From three dimensional structures, additional molecular properties can be computed, quantifying the steric and electronic contributions of a species.

Chemical reactivity can also be studied on a larger scale. The area of cheminformatics deals with describing trends in chemical properties from large datasets of molecules (thousands

to millions of datapoints).[10] To reduce computational cost, molecules are often represented by much simpler properties than properties obtained from DFT structures. Relevant properties here include count descriptors (number of carbons, number of aromatic rings), or connectivity descriptors, which include topological properties, describing how atoms are bonded in a molecule, encoding which unique substructures are present in each molecule. Since these studies involve larger datasets, they are suitable for more intensive machine learning and deep learning statistical models.[11,12]

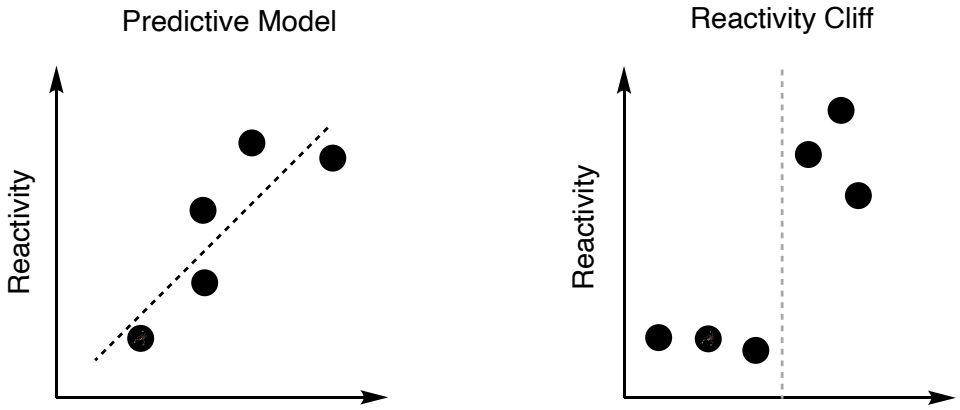


Figure 1.3. *(Left) A generic predictive model showing a linear trend with reactivity. (Right) A reactivity cliff, showing a jump in reactivity where a linear trend in reactivity is not obvious with the analysis.*

Statistical models developed from either experimental or computational properties aid in chemical understanding. In general, reaction prediction operates under the assumption that similar molecules will exhibit similar properties, so molecules that differ in small structural changes will have similar properties, and similar reactivities. For simple univariate models or multivariate linear regression, the reaction outcome generally follows a continuous trend, and it can be easy to interpret the influence of a specific property on reaction outcome (Figure 1.3). From this the reaction specifications can be tuned to further optimize reactivity. Deep learning models can be more “black-box” and offer little interpretability for how molecular properties can influence the predicted outcome, but efforts and calls for meaningful interpretations of these models are on the rise.[13] In some cases, the chemical reactivity is not captured by the specific properties measured

for a chemical system. This would cause an anomaly in reaction prediction called a “reactivity cliff,” shown in Figure 1.3, where the reactivity behavior exhibited by the chemical system is not described continuously across the reactivity axis.[14] These reactivity cliffs make modeling the chemical system a challenge, and in these cases, there may be more appropriate, unutilized properties to describe the chemical system better. This can also result in the development of novel molecular properties, in the case that existing measurable properties are unable to capture the reactivity of the system.

1.2: Document Overview

This work’s primary focus on the study of computed molecular properties (“molecular properties” here is used interchangeably with “molecular descriptors” and “molecular features” in this work). In this dissertation, I will focus on the current landscape of molecular properties used in the literature. I will showcase how I have worked to develop novel computational molecular properties, as well as explore how properties can be used and related to experimental reactivity properties. I have also worked to develop accessible Python packages for chemists to collect molecular properties for their own research.

1.2.1: Digitizing Molecules: Molecular Representations

This dissertation provides insight into existing methods to represent molecules and their properties of digitally. Chapter 2 will present the numerous methods we describe what a molecule is on a computer, and which properties we can obtain from each of those representations. This translation of the concept of a molecule into quantitative values that computers can interpret is important in the identification of patterns. These patterns help us construct and study reactivity relationships for new chemical systems, driving discovery. This chapter also explores specific molecular properties relevant to the field of organic catalysis, describing ground and transition structure properties, and how conformational studies of molecules are influential in describing molecular reactivity.

In Chapter 3, a specific class of properties describing size and shape of molecules, steric parameters are discussed, and novel parameters are presented. An identified gap in the literature involving the idea of how near steric bulk lies in relation to a chemical center of interest (i.e., a reactive center) is addressed. The proximity of steric bulk can influence reaction outcome by blocking or providing access to the chemical center of interest, and in existing parameter measurements, this proximity information is not clearly captured. New parameter sets which expand up on existing steric parameters are introduced. Several examples are presented outlining their use and applicability in different areas of chemistry, comparing to where proximal or distal steric information is relevant in unique studies.

In Chapter 4, a different class of properties, electronic parameters, are discussed. An early example of experimental electronic parameters is the Hammett parameter, useful in describing reactivity at meta and para positions in chemical reactions involving aromatic ring systems. A specific type of computed electronic parameter, the partial atomic charge, is often used in computational studies to justify how the molecule's charge distributions across its atoms give rise to reactivity at specific sites. Since these computed charges are easily obtained and do not rely on experiment, many studies have involved utilizing partial atomic charges to predict experimental Hammett parameters. Many computational methods exist to compute partial atomic charges. The work of this chapter compares different methods for computing partial atomic charge directly with the experimental Hammett values, providing recommendations and best practices for studies comparing these values.

1.2.2: Python Tools for Chemists

Chapter 5 of this document switches the perspective to programming efforts that went into works described throughout the document. Software development is a large part of research processes in computational chemistry. The development of tools and workflows to obtain the data presented in a manuscript may be applicable for subsequent studies, or even in studies in unrelated areas.

In agreement with FAIR data principles,[15] which includes guidelines for improving Findability, Accessibility, Interoperability and Reuse of data, we publish useful Python packages and workflows as open-source software, creating GitHub repositories when possible. We aim to design accessible tools for chemists to utilize in their research. Part of the software development process for these packages involves writing documentation, usually in the form of a ReadMe page, describing program options, providing examples and theoretical background. Additionally, the software development process is iterative. There is always room for improving methods, optimizing how functions are called, keeping Python software and packages dependencies up to date with current versions. GitHub allows for user interaction through Issue Reports or Pull Requests, which allows users to submit issues for bugs or errors in the code they observe, and if users wish to edit the code themselves, we welcome contributions that help make packages better.

This chapter contains work from three separate Python packages, *GoodVibes*, *DBSTEP*, and *Py-X Struct*. These works find use in many different areas of chemistry. Perhaps most broadly, GoodVibes allows for the computation and correction of thermochemical values from quantum mechanical software, so that reaction thermochemistry may be studied. This allows for comparisons of thermochemically relevant conformations of molecules, along with comparisons of relative thermochemical values from potential energy surface calculations. DBSTEP is designed for obtaining steric descriptors from three-dimensional molecular structures, providing access for researchers to obtain parameters in high-throughput scripts. Finally, Py-X Struct is used for searching and returning geometric data from experimental X-Ray crystal structures contained in the Cambridge Structural Database. These programs are all freely available from our laboratory GitHub repository and are easily installed.

1.3: References

1. Taylor, C. J.; Pomberger, A.; Felton, K. C.; Grainger, R.; Barecka, M.; Chamberlain, T. W.; Bourne, R. A.; Johnson, C. N.; Lapkin, A. A., *Chem. Rev.* **2023,** *123*, 3089-3126.
2. Bryan, M. C.; Dillon, B.; Hamann, L. G.; Hughes, G. J.; Kopach, M. E.; Peterson, E. A.; Pourashraf, M.; Raheem, I.; Richardson, P.; Richter, D.; Sneddon, H. F., *J. Med. Chem.* **2013,** *56*, 6007-6021.
3. Hammett, L. P., *J. Am. Chem. Soc.* **1937,** *59*, 96-103.
4. (a) Taft Jr, R. W., *J. Am. Chem. Soc.* **1952,** *74*, 3120-3128; (b) Taft, R. W., *J. Am. Chem. Soc.* **1952,** *74*, 2729-2732; (c) Taft Jr, R. W., *J. Am. Chem. Soc.* **1953,** *75*, 4538-4539.
5. Santiago, C. B.; Guo, J. Y.; Sigman, M. S., *Chem. Sci.* **2018,** *9*, 2398-2412.
6. Fox, R. J.; Cuniere, N. L.; Bakrania, L.; Wei, C.; Strotman, N. A.; Hay, M.; Fanfair, D.; Regens, C.; Beutner, G. L.; Lawler, M.; Lobben, P.; Soumeillant, M. C.; Cohen, B.; Zhu, K.; Skliar, D.; Rosner, T.; Markwalter, C. E.; Hsiao, Y.; Tran, K.; Eastgate, M. D., *J. Org. Chem.* **2019,** *84*, 4661-4669.
7. Torres, J. A. G.; Lau, S. H.; Anchuri, P.; Stevens, J. M.; Tabora, J. E.; Li, J.; Borovika, A.; Adams, R. P.; Doyle, A. G., *J. Am. Chem. Soc.* **2022,** *144*, 19999-20007.
8. Ahn, S.; Hong, M.; Sundararajan, M.; Ess, D. H.; Baik, M.-H., *Chem. Rev.* **2019,** *119*, 6509-6560.
9. Gallegos, L. C.; Luchini, G.; St. John, P. C.; Kim, S.; Paton, R. S., *Acc. Chem. Res.* **2021,** *54*, 827-836.
10. Chen, W. L., *J. Chem. Inf. Model.* **2006,** *46*, 2230-2255.
11. (a) Meuwly, M., *Chem. Rev.* **2021,** *121*, 10218-10239; (b) Ramakrishnan, R.; von Lilienfeld, O. A., Machine Learning, Quantum Chemistry, and Chemical Space. In *Rev. Comput. Chem.*, 2017; pp 225-256.
12. (a) Chen, H.; Kogej, T.; Engkvist, O., *Mol. Inform.* **2018,** *37*, 1800041; (b) Mater, A. C.; Coote, M. L., *J. Chem. Inf. Model.* **2019,** *59*, 2545-2559.
13. Bender, A.; Schneider, N.; Segler, M.; Patrick Walters, W.; Engkvist, O.; Rodrigues, T., *Nat. Rev. Chem.* **2022,** *6*, 428-442.
14. Newman-Stonebraker, S. H.; Smith, S. R.; Borowski, J. E.; Peters, E.; Gensch, T.; Johnson, H. C.; Sigman, M. S.; Doyle, A. G., *Science* **2021,** *374*, 301-308.
15. Wilkinson, M. D.; Dumontier, M.; Aalbersberg, I. J.; Appleton, G.; Axton, M.; Baak, A.; Blomberg, N.; Boiten, J.-W.; Da Silva Santos, L. B.; Bourne, P. E.; Bouwman, J.; Brookes, A. J.; Clark, T.; Crosas, M.; Dillo, I.; Dumon, O.; Edmunds, S.; Evelo, C. T.; Finkers, R.; Gonzalez-Beltran, A.; Gray, A. J. G.; Groth, P.; Goble, C.; Grethe, J. S.; Heringa, J.; 'T Hoen, P. A. C.; Hooft, R.; Kuhn, T.; Kok, R.; Kok, J.; Lusher, S. J.; Martone, M. E.; Mons, A.; Packer, A. L.; Persson, B.; Rocca-Serra, P.; Roos, M.; Van Schaik, R.; Sansone, S.-A.; Schultes, E.; Sengstag, T.; Slater, T.; Strawn, G.; Swertz, M. A.; Thompson, M.; Van Der Lei, J.; Van Mulligen, E.; Velterop, J.; Waagmeester, A.; Wittenburg, P.; Wolstencroft, K.; Zhao, J.; Mons, B., *Sci. Data* **2016,** *3*, 160018.

CHAPTER 2: MOLECULAR REPRESENTATIONS: COMPUTATIONALLY QUANTIFYING MOLECULAR PROPERTIES

2.1: Chapter Overview

Computed molecular properties are represented as quantified numbers or collections of numbers so that computational algorithms can identify patterns in how properties may influence chemical reactivity. For a computational chemist, deciding the correct property to measure from a molecule to relate to the chemical reactivity of the system is an important step. The choice of property should capture important features of the underlying mechanism for the system, under the assumption that properties of unique molecules will influence the outcome of the reaction in unique ways. This chapter presents an overview of the current field of computational molecular descriptors. This chapter contains work from two manuscripts, separated in sections 2.2 and 2.3. In section 2.2, an introduction and summary of methods to computationally represent molecular properties, is presented. This section contains a portion of a manuscript that was published in *Accounts of Chemical Research* entitled “Importance of Engineered and Learned Molecular Representations in Predicting Organic Reactivity, Selectivity and Chemical Properties.” This work was done in collaboration with scientists at the National Renewable Energy Lab, Dr. Seonah Kim (now at Colorado State University) and Dr. Peter C. St. John (now at NVIDIA), along with Liliana C. Gallegos. I have included sections relevant to this chapter that I specifically made contributions on, the *Introduction* (Section 2.2.1) and the following section, *Molecular Representations: From One to Four Dimensions* (Section 2.2.2), which describes the types of features you can obtain from different molecular representations, from simple connectivity information, to considering molecular flexibility in three dimensions. I worked on the *Introduction* section in collaboration with Robert Paton and Peter St. John, and wrote the following section, *Molecular Representations: From One to Four Dimensions,* separately.

Section 2.3 presents more detailed information on specific molecular properties available for a chemist to use in predictive modeling. This summary presents an overview of the current landscape of descriptors obtained from molecules for applications designed for organic catalysis. The molecular properties described in this section are primarily obtained from Density Functional Theory calculations, as accurate descriptions of molecular structure and energies is necessary for the properties resulting from the molecules discussed. This section discusses existing steric and electronic properties that may influence reaction outcome, providing examples where they have found use. In mechanistic studies, researchers may optimize both ground state and transition state structures to compare a reaction potential energy surface. These transition structures and the properties they hold are important in learning how a reaction mechanism may proceed, and so properties from the transition structure can provide valuable insight into predicting reactivity. Additionally, if molecules in the ground or transition state can take on multiple unique geometric conformations, it is important to capture the conformational space of the system, this work provides an overview of how and when conformational studies can be used. This work is only a portion of a manuscript written in collaboration with Liliana C. Gallegos, however, I have only included sections that I contributed to. The work in this section is under preparation to be submitted to *ACS Catalysis*.

2.2: Importance of Engineered and Learned Molecular Representations in Predicting Organic Reactivity, Selectivity and Chemical Properties

2.2.1: Introduction

Data-driven chemistry is propelled by innovations in the generation and curation of chemical data, the machine learning algorithms used for regression and classification, and how molecules are represented.[1] Machine-readable chemical structure representations were originally introduced to create the first searchable computational databases of molecules and reactions in the 1960s.[2] They are now a central element of chemical machine learning (ML). For decades, the development of predictive quantitative structure–activity and structure–property relationships

(QSAR and QSPR) directly from chemical structure has been an area of active research in which the construction of expressive molecular feature representations that inform the physical nature of the input–output mapping is a central task.[3] Feature vectors encode information about molecular structure, in most cases, by combining a series of physically meaningful molecular descriptors that describe spatial, electronic, and energetic properties. The use of "expert crafted" descriptors provides an opportunity to incorporate chemical knowledge, domain expertise, and physical constraints into any given machine- learning approach while also potentially offering greater interpretability to the chemist as a result.[4]

The featurization or embedding of discrete molecular structures into a continuous vector space (i.e., as feature vectors) is a critical phase undertaken before model selection. Attempts to predict specific reaction outcomes such as reactivity or selectivity are routinely faced with small data sets on the order of tens to hundreds of examples. In these cases, manual approaches to feature engineering that rely upon specialized domain knowledge, such as a structural or mechanistic hypothesis, and physicochemical descriptors tend to achieve better results than more generalizable representations. Feature vectors derived from physical-organic parameters that describe a molecule's or substituent's electronic (e.g., HOMO/LUMO energies, atomic charges, Fukui[5] coefficients) and steric (e.g., Tolman cone angle,[6] Sterimol,[7] buried volume[8]) influence have been used in ML models to predict the yields[9] and diastereo- and enantioselectivities[10] of organic and organometallic reactions by Sigman,[11] Doyle,[12] and others including ourselves. The continued development of physically motivated descriptors that succinctly and transparently capture the subtleties of molecular stereochemistry, conformation, and electronic effects is central to data-driven approaches for organic reaction prediction, as the ability to link a quantitative predictive model back to interpretable descriptors can be used to derive new understanding and mechanistic inferences.[13]

While manually engineered features may focus on describing a specific type of molecule or reaction, more flexible, general- purpose molecular representations such as attributed molecular graphs[14] can be used in combination with ML approaches to learn the complex relationship between a structure and prediction target. Deep learning approaches have proven to be particularly well-suited to the representation of organic structures, automatically learning "rich" features and improving the accuracy of chemical property and reactivity prediction over traditional hand-coded or molecular fingerprint representations.[15] In particular, the rise of graph neural networks (GNNs)[16] in modeling chemical properties has enabled "end-to-end" learning on molecular structure: an ML strategy where traditional feature engineering is replaced by a learned molecular representation derived from an attributed molecular graph. These approaches have led to best-in-class prediction accuracies on a range of applications from total energies, interatomic forces,[17] and bond strengths,[18] especially as the amount of available training data grows.[19]

In this Account, we present an overview of how distinct featurization strategies can be applied to organic and organometallic chemistry, to predict reactivity, stereoselectivity, and chemical properties. In the case of small (<100) reaction data sets, hand-crafted physicochemical descriptors are shown to yield interpretable models such as multivariate linear regressions (MLR) of catalytic enantio- and diastereoselectivities. With larger data sets, such as those obtained from high-throughput quantum chemical data sets, learned representations with flexible GNNs can be trained to produce excellent quantitative predictions of atomic or molecular properties at low computational cost. This is illustrated for organic bond dissociation enthalpies. Finally, we describe how GNN predictions can be incorporated into mechanistically informed statistical models of chemical reactivity and selectivity. Once trained, this approach avoids the expensive computational overhead associated with QM calculations and maintains chemical interpretability.

2.2.2: Molecular Representations: From One to Four Dimensions

While QSAR/QSPR models have found large success in areas of medicinal chemistry and drug discovery, attempts to relate structure with catalytic activity and selectivity have emerged more recently. Traditional cheminformatics representations largely focus on 2D or topological molecular representations that define the connectivity and bonding types of atoms in a molecule. 2D molecular descriptors, such as topological fingerprints, are simple to define and can be obtained without geometry optimization. However, a number of features of mechanistic relevance to reactivity and selectivity depend upon a molecule's 3D structure and conformation, including electronic properties such as atomic and molecular charge distributions, as well as other structure-dependent features that capture a molecule's steric influence, chirality, volume or surface area. Quantum mechanically (QM) optimized molecular coordinates (e.g., using density functional theory, DFT) can be used to obtain such descriptors, in addition to other QM-computed properties such as thermochemical values, molecular orbital energies, vibrational frequencies, and noncovalent interaction energies. In QSAR vernacular, 3D-descriptors generally refer to those that map molecular interactions to a prealigned grid of points, as in Comparative Molecular Field Analysis[20] (COMFA) and the more recently developed Average Steric Occupancy (ASO) from Denmark.[21] More generally, properties dependent upon 3D-structure, and especially spatial/steric occupancies, have been described by scalar parameters such as buried volume and higher dimensional objects such as topological maps or multidimensional Sterimol parameters, approaches pioneered by the Cavallo[8] and Sigman[22] groups, respectively. This hierarchy of molecular representations used across organic chemistry is shown in Figure 2.1.

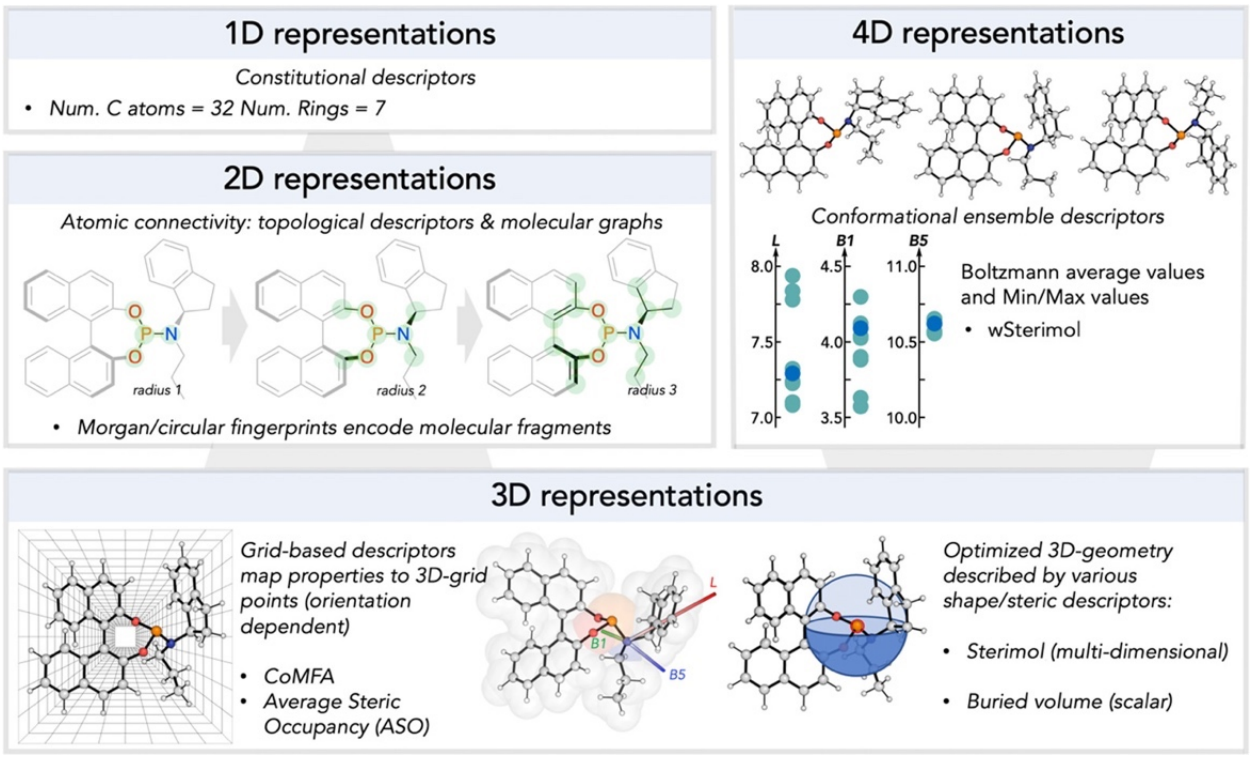


Figure 2.1. *Hierarchy of molecular representations used to encode organic and organometallic structures.*

The dependence of 3D or DFT-derived descriptors upon the molecular geometry means that unlike 2D representations, conformational dynamics are important to consider. Indeed, the presence of multiple conformations may itself be an important descriptor relating to catalytic proficiency.[23] Thus, in addition to featurization of the most stable conformer, consideration of the full conformational ensemble may be necessary to quantitatively encode macroscopic behavior. Ensemble approaches have been referred to as 4D-QSAR.[24] Spatial parameters may be sensitive to conformational behavior and to the level of theory employed to generate an ensemble, which led us to develop a software tool, wSterimol, to automate conformer-sampling and featurization.[25] With this, we have been able to include estimates of parameter uncertainty into regression models of stereoselectivity.

2.3: Molecular Descriptors for Data-Driven Catalysis: "From-the-Molecule" Descriptors

2.3.1: Shape and Electronic Descriptors

The shape and electronic properties of catalyst molecules influence molecular behavior and reaction outcome. These catalyst properties can be tuned and optimized by swapping out key functional groups or larger structural moieties to gain or lose steric bulk or adjust electron donating and withdrawing character of the catalyst to achieve an optimal reaction outcome, such as yield, rate, or chemical selectivity. The field of catalysis has benefited from quantitative descriptions of these electronic and steric properties. Atomic or molecular descriptors initially designed for applications in other areas of chemistry can be utilized in the field catalysis by providing information previously inaccessible or unconsidered but relevant to the reaction mechanism of study.

Different types of descriptors can describe molecules in varying levels of detail. Cheminformatic descriptors are usually quick to obtain, describing general information about a molecule (1D, Count, and topological descriptors). The inexpensive nature of obtaining these descriptors allows for computing features for large datasets. Cheminformatic studies have commonly been utilized in drug discovery processes,[26] relying on simple molecular descriptors to predict important structural moieties for early drug candidates. Simple descriptors can also be fed into more complex machine learning models. More expensive molecular descriptors can also be generated, relying on three-dimensional molecule representations. Ground state electronic structures can be generated using a variety of techniques (FF, semi-empirical, DFT). Collecting properties based on shape and electronic properties of modeled molecular structures, discussed further in sections 2.3.2 and 2.3.3, respectively, can provide additional insight into reaction mechanism that simpler descriptors are unable to capture.

Molecular descriptors can come in many forms. Descriptors aim to quantitatively capture specific aspects of chemical phenomena, from varying computational molecular

representations.[27] Often cheminformatic-type descriptors are single-valued, they can describe aspects of molecular structure like count-descriptors, for example, the number of carbon atoms or the number of aromatic rings, or general properties of molecules such as molecular weight. More complex steric and electronic descriptors, discussed below, can also be single-valued, describing geometric bond lengths, angles, or torsions, as well as partial atomic charge. Descriptors can also be multi-dimensional, containing multiple measurements as a part of a set of descriptors.[22] These parameter sets can consist of a few measurements, for example the Sterimol set of parameters consists of three different measurements of steric bulk on a molecule. This can also be taken to an even greater extreme; descriptors can also be made up of hundreds or even thousands of data points measured per molecule in 3D-QSAR studies, which rely on molecules being aligned in a cubic grid, with properties of each molecule measured at evenly spaced grid points or voxels, seen in COMFA analyses[20] or Average Steric Occupancy[28] 3D-descriptor sets.

2.3.2: Shape-Based Descriptors

Molecular size and shape can alter reaction outcome. Describing the shape of molecules in catalysis has proven useful in predicting the outcome where varying steric bulk of a molecule can either provide or hinder access to a reactive site. Steric parameters designed for catalytic applications can include the Tolman cone angle,[6] which measures the angle of a cone encapsulating a phosphine ligand centered at a metal. The percent buried volume,[8] visualized in Figure 2.2A, was designed to describe steric bulk of a ligand within a sphere of 3.5Å of a metal center, capturing proximal metal-ligand steric interactions in the first coordination sphere of a metal.

Descriptors initially designed for use in other areas of chemistry have also proven useful in catalysis. The Sterimol descriptors, also shown in Figure 2.2A, are an anisotropic class of three descriptors describing the length (L), minimum width (B_{min}) and maximum width (B_{max}) of a

molecule. Originally, these descriptors were designed showing their usefulness in predicting reactivity of pesticide molecules,[29] and in the years following quickly found use in insecticide, herbicide, and pharmaceutical studies.[7] These parameters have since been revisited and have found use in asymmetric catalysis, proving useful to describe how the steric bulk of ligand molecules and substituents influence reaction outcome.[22,30]

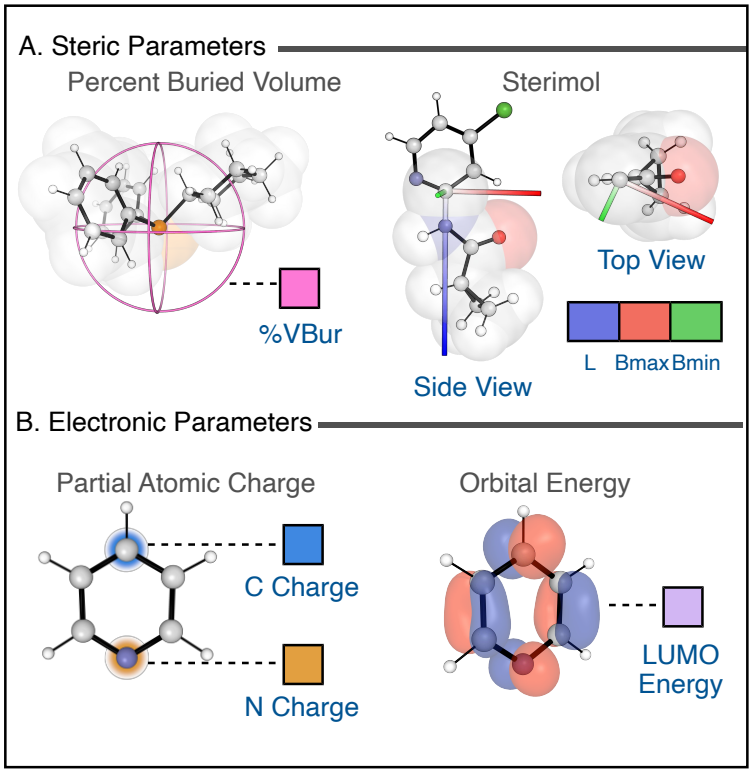


Figure 2.2. *Visualizations of steric parameters percent buried volume (%V_{bur}) and Sterimol parameters (A) and electronic parameters including partial atomic charge and orbital energies (B).*

2.3.3: Electronic Descriptors

Describing differences in electronic properties of a molecule can be important in determining an appropriate substrate scope for a reaction. Varying degrees of electron withdrawing or donating character can contribute to variations in chemical reactivity. Quantifying electronic effects is possible in a variety of methods. Many computational methods exist to attempt to model experimental observables including NMR shifts, various vibrational spectroscopies, and UV-Vis, providing values directly comparable to experiment.[31]

Other classes of computational electronic descriptors exist which attempt to quantify changes and differences in the distribution of electron density across a molecule, or at specific atom centers. Formulations for partial atomic charges involve partitioning the charge of a molecule into individual atoms and can be derived from a molecule's electrostatic potential, atomic orbitals, or electron density.[32] The Natural Population Analysis (NPA) is a method to compute atomic partial charges from molecular orbital populations, providing quantitative values for atomic charges, as well as orbital energies.[33] Qualitative methods of visualizing molecular orbitals, like the LUMO orbital shown in Figure 2.2B, obtained through these analyses have also been used to provide mechanistic insight into a reaction. Recently, Modak and others utilized NPA to aid in understanding regioselectivity differences in substituents for homologation reactions of aryl halides.[34] Using key σ^* molecular orbital energies obtained with NPA, authors constructed a classification model identifying an energetic threshold separating highly regioselective substrates against lower selectivities.

2.3.4: Transition State Descriptors

It is possible to compute molecular descriptors on transition state (TS) structures to relate to reaction outcome as well, like ground state electronic structures. However, optimizing TS structures using DFT is often more complex than ground state counter parts. For this reason, many traditional cheminformatic studies opt to exclude TS features, as the computational expense and manual input that goes into obtaining these structures can vary. To address this challenge, many tools exist to aid in obtaining these structures. Covered more extensively in recent work by Jensen,[35] tools like CatVS[36] and Q2MM[37] can be used in combination to develop TS force fields for a system to probe and screen TS conformational space, while programs like ANI[38] and xTB[39] can be used to optimize TS structures at low computational expense. Comprehensive workflow programs like AARON[40] or AQME[41] can be used to generate TS conformations and optimize

structures with DFT with considerably less user input and manipulation than manual optimization.

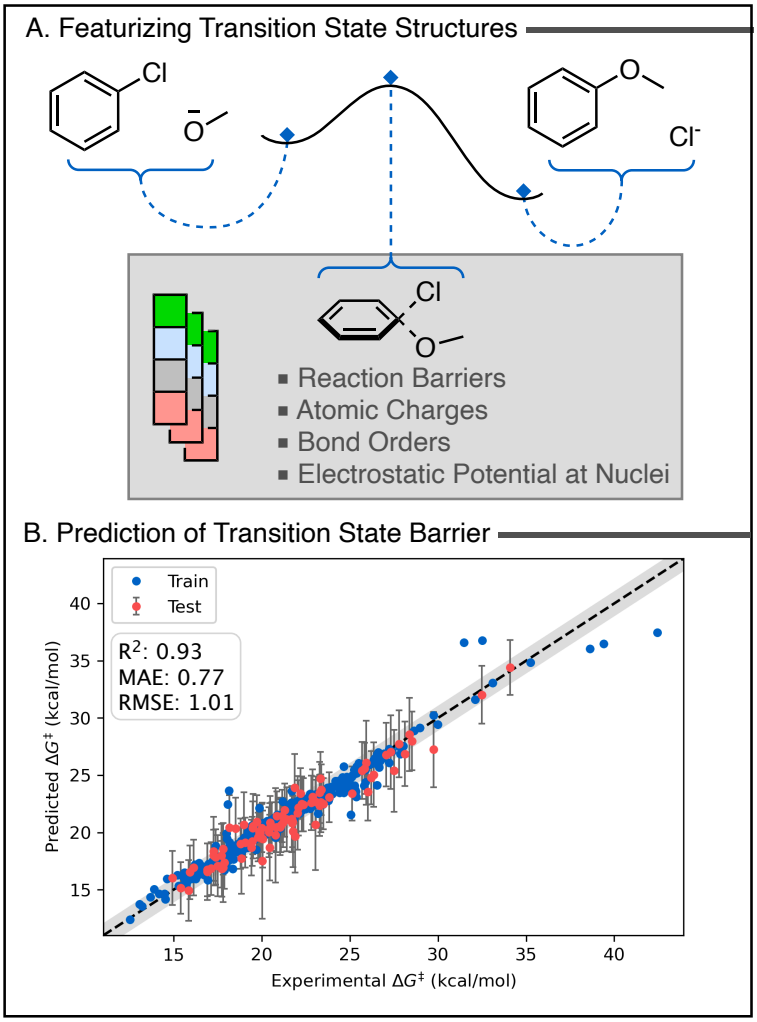


Figure 2.3. *(A) Features were generated from DFT-optimized transition structures including reaction barriers, atomic charge, bond order, and electrostatic potential at the nuclei of transition state structures. (B) Gaussian process regression modeling was performed to predict the energy of the transition state barrier from features computed from the optimized ground and transition state structures.*

2.3.5: Relating Transition State Features to Reaction Outcome

Transition state structures can be utilized to make reaction predictions and have been found to be useful in prediction activation energies. In recent work by Jorner and coworkers,[42] authors modeled reactants, TS structures, and products (Figure 2.3A) for 336 unique S_NAr reactions using an automated computational workflow to obtain DFT-optimized structures. Steric and electronic descriptors were collected from both ground state and transition state structures. The collected

TS features included atomic charges, bond orders, electrostatic potentials, and computed activation energies. Different kinds of predictive models were trained using varying combinations of training and test set ratios, resulting in a gaussian process regression model with the best performance obtaining a mean absolute error (MAE) of 0.77 kcal mol^{-1} and a Pearson R^2 value of 0.93, shown in Figure 2.3B. When performing feature selection to assess performance of predictive models, the models that used features from the TS structures, notably the DFT computed activation energy, $\Delta G^{\ddagger}_{DFT}$, performed consistently better in predicting experimental activation energies than models that excluded TS features.

2.3.6: Conformational Effects

Effective conformational sampling should be performed to accurately describe the system. Molecular conformation influences steric and electronic properties that a molecule will exhibit. Experimentally observable properties result from an accumulation of contributions from individual molecular conformations. Describing an accurate range of conformers appropriate for the system will help in accurately modeling a chemical process when comparing to experiment.

Several low-cost tools and algorithms exist to search for a molecule's conformers. Systematically scanning dihedrals of all rotatable bonds in set intervals for a molecule is possible by manipulating molecular coordinates using tools like RDKit.[43] However, this approach becomes more tedious and can generate many unstable conformations with greater molecular flexibility. It is also possible to generate conformers through a process called embedding with RDKit using the ETKDG algorithm,[44] which generates conformers by sampling dihedral angles using experimentally parameterized datasets. CREST is a program from Grimme used to sample the conformational space of molecules using semiempirical methods.[45]

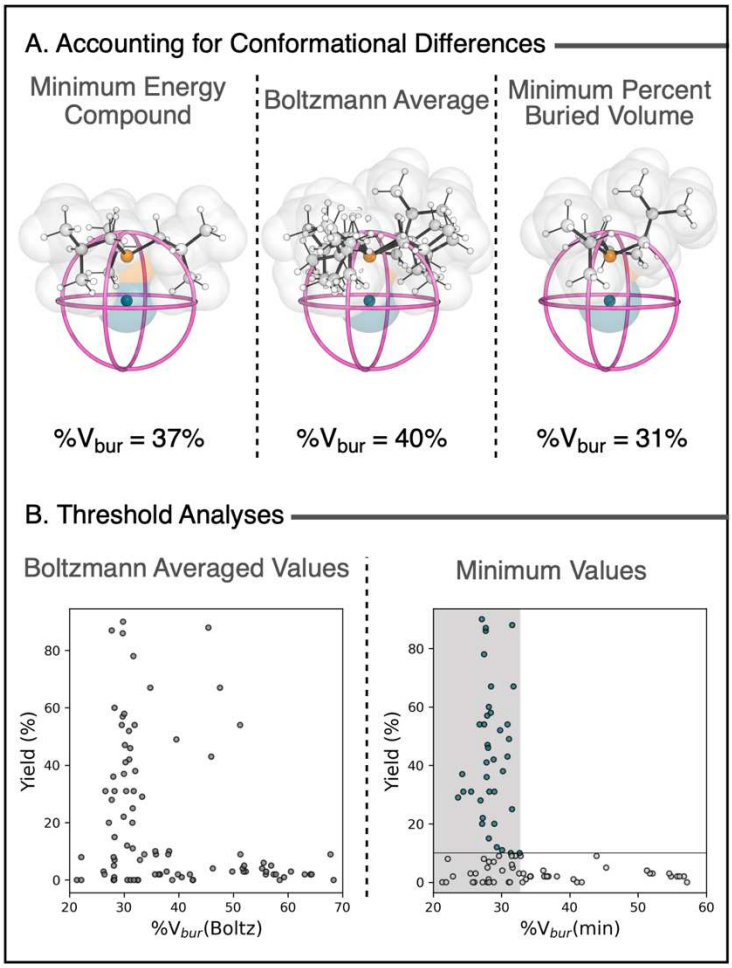


Figure 2.4. *(A) Differences in percent buried volume measurements for conformations of P(i-Bu)3 bound to Pd comparing the lowest energy conformer, Boltzmann average, and minimum percent buried volume measurement across all identified conformers. (B) Threshold analyses comparing reaction yield with percent buried volume measurements attempting to identify a univariate threshold to explain differences in reactivity. Graphs shown display the Boltzmann averaged values (left) and the minimum values (right) measured for phosphine ligands.*

There may be slight differences in the lowest energy conformer properties versus the Boltzmann-averaged properties, especially if multiple low energy conformers are energetically accessible for a molecule. Often, using the Boltzmann-averaged value is an appropriate method to account for individual contributions of the conformational ensemble. In wSterimol, the Boltzmann weighted Sterimol values compares best to reactivity.[25] However, contributions from individual conformers can influence reactivity more than others. For example, it has been shown that the conformer with the smallest value for percent buried volume across all conformations of

a molecule, $\%V_{bur}$(min)—not necessarily the lowest energy conformer—can give the best predictive reliability for reactivity.[46] Figure 2.4A shows comparisons between conformations of triisobutylphosphine bound to nickel, comparing the lowest energy conformer with a visualization of the Boltzmann averaged value and the conformer which gives the smallest percent buried volume value. This study utilized the Kraken database,[47] a large virtual collection of phosphine ligands with DFT-level descriptors. The conformational space for each ligand was obtained by sampling both free and metal-bound ligands in CREST, and then further optimized using DFT. Values obtained from Kraken contain information on individual conformers, giving ranges for descriptor values as the conformational space varies. By providing descriptors on a conformational basis, statistical techniques can be used to assess the impact of individual conformations of molecules on the desired reaction outcome. Figure 2.4B shows how a univariate threshold comparing $\%V_{bur}$ values to percent yield was determined using the $\%V_{bur}$(min), which was not as pronounced when using Boltzmann averaged values.

2.4: References

1. Haghighatlari, M.; Li, J.; Heidar-Zadeh, F.; Liu, Y.; Guan, X.; Head-Gordon, T., *Chem* **2020,** *6*, 1527-1542.
2. Chen, W. L., *J. Chem. Inf. Model.* **2006,** *46*, 2230-2255.
3. Dudek, A. Z.; Arodz, T.; Gálvez, J., *Comb. Chem. High Throughput Screen.* **2006,** *9*, 213-228.
4. Todeschini, R.; Consonni, V., *Handbook of molecular descriptors*. John Wiley & Sons: 2008.
5. Fukui, K.; Yonezawa, T.; Nagata, C.; Shingu, H., *J. Chem. Phys.* **1954,** *22*, 1433-1442.
6. Tolman, C. A., *Chem. Rev.* **1977,** *77*, 313-348.
7. Verloop, A., *Pesticide Chemistry: Human Welfare and Environment* **1983**, 339-344.
8. (a) Poater, A.; Cosenza, B.; Correa, A.; Giudice, S.; Ragone, F.; Scarano, V.; Cavallo, L., *Eur. J. Inorg. Chem.* **2009,** *2009*, 1759-1766; (b) Falivene, L.; Credendino, R.; Poater, A.; Petta, A.; Serra, L.; Oliva, R.; Scarano, V.; Cavallo, L., *Organometallics* **2016,** *35*, 2286-2293.
9. Yada, A.; Nagata, K.; Ando, Y.; Matsumura, T.; Ichinoseki, S.; Sato, K., *Chem. Lett.* **2018,** *47*, 284-287.
10. Reid, J. P.; Sigman, M. S., *Nature* **2019,** *571*, 343-348.
11. Santiago, C. B.; Guo, J. Y.; Sigman, M. S., *Chem. Sci.* **2018,** *9*, 2398-2412.
12. Nielsen, M. K.; Ahneman, D. T.; Riera, O.; Doyle, A. G., *J. Am. Chem. Soc.* **2018,** *140*, 5004-5008.
13. (a) Niemeyer, Z. L.; Milo, A.; Hickey, D. P.; Sigman, M. S., *Nat. Chem.* **2016,** *8*, 610-617; (b) Wu, K.; Doyle, A. G., *Nat. Chem.* **2017,** *9*, 779-784; (c) Amar, Y.; Schweidtmann, A. M.; Deutsch, P.; Cao, L.; Lapkin, A., *Chem. Sci.* **2019,** *10*, 6697-6706.
14. Yang, K.; Swanson, K.; Jin, W.; Coley, C.; Eiden, P.; Gao, H.; Guzman-Perez, A.; Hopper, T.; Kelley, B.; Mathea, M., *J. Chem. Inf. Model.* **2019,** *59*, 3370-3388.
15. Mater, A. C.; Coote, M. L., *J. Chem. Inf. Model.* **2019,** *59*, 2545-2559.
16. Battaglia, P. W.; Hamrick, J. B.; Bapst, V.; Sanchez-Gonzalez, A.; Zambaldi, V.; Malinowski, M.; Tacchetti, A.; Raposo, D.; Santoro, A.; Faulkner, R., *arXiv preprint arXiv:1806.01261* **2018**.
17. Schütt, K. T.; Sauceda, H. E.; Kindermans, P.-J.; Tkatchenko, A.; Müller, K.-R., *J. Chem. Phys.* **2018,** *148*, 241722.
18. St. John, P. C.; Guan, Y.; Kim, Y.; Kim, S.; Paton, R. S., *Nat. Commun.* **2020,** *11*.
19. (a) Faber, F. A.; Hutchison, L.; Huang, B.; Gilmer, J.; Schoenholz, S. S.; Dahl, G. E.; Vinyals, O.; Kearnes, S.; Riley, P. F.; Von Lilienfeld, O. A., *J. Chem. Theory Comput.* **2017,** *13*, 5255-5264; (b) Feinberg, E.; Sheridan, R.; Joshi, E.; Pande, V.; Cheng, A., Step Change Improvement in ADMET Prediction with PotentialNet Deep Featurization. arXiv. org. March: 2019; (c) Coley, C. W.; Barzilay, R.; Green, W. H.; Jaakkola, T. S.; Jensen, K. F., *J. Chem. Inf. Model.* **2017,** *57*, 1757-1772.
20. (a) Lipkowitz, K. B.; Pradhan, M., *J. Org. Chem.* **2003,** *68*, 4648-4656; (b) Ianni, J. C.; Annamalai, V.; Phuan, P.-W.; Panda, M.; Kozlowski, M. C., *Angew. Chem. Int. Ed.* **2006,** *45*, 5502-5505.
21. Zahrt, A. F.; Athavale, S. V.; Denmark, S. E., *Chem. Rev.* **2019,** *120*, 1620-1689.
22. Harper, K. C.; Bess, E. N.; Sigman, M. S., *Nat. Chem.* **2012,** *4*, 366-74.
23. Crawford, J. M.; Sigman, M. S., *Synthesis* **2019,** *51*, 1021-1036.

24. Andrade, C. H.; Pasqualoto, K. F.; Ferreira, E. I.; Hopfinger, A. J., *Molecules* **2010,** *15*, 3281-94.
25. Brethomé, A. V.; Fletcher, S. P.; Paton, R. S., *ACS Catal.* **2019,** *9*, 2313-2323.
26. (a) Kubinyi, H., *Drug Discovery Today* **1997,** *2*, 457-467; (b) Kubinyi, H., *Drug Discovery Today* **1997,** *2*, 538-546.
27. Gallegos, L. C.; Luchini, G.; St. John, P. C.; Kim, S.; Paton, R. S., *Acc. Chem. Res.* **2021,** *54*, 827-836.
28. Zahrt, A. F.; Henle, J. J.; Rose, B. T.; Wang, Y.; Darrow, W. T.; Denmark, S. E., *Science* **2019,** *363*, eaau5631.
29. Verloop, A., Drug Design. Ariens, E. J., Ed. Academic Press: New York,, 1976; Vol. III.
30. Piou, T.; Romanov-Michailidis, F.; Romanova-Michaelides, M.; Jackson, K. E.; Semakul, N.; Taggart, T. D.; Newell, B. S.; Rithner, C. D.; Paton, R. S.; Rovis, T., *J Am Chem Soc* **2017,** *139*, 1296-1310.
31. Barone, V.; Alessandrini, S.; Biczysko, M.; Cheeseman, J. R.; Clary, D. C.; McCoy, A. B.; DiRisio, R. J.; Neese, F.; Melosso, M.; Puzzarini, C., *Nature Rev. Methods Primers* **2021,** *1*, 38.
32. Gonthier, J. F.; Steinmann, S. N.; Wodrich, M. D.; Corminboeuf, C., *Chem. Soc. Rev.* **2012,** *41*, 4671.
33. Reed, A. E.; Weinstock, R. B.; Weinhold, F., *J. Chem. Phys.* **1985,** *83*, 735-746.
34. Modak, A.; Alegre-Requena, J. V.; de Lescure, L.; Rynders, K. J.; Paton, R. S.; Race, N. J., *J. Am. Chem. Soc.* **2022,** *144*, 86-92.
35. Foscato, M.; Jensen, V. R., *ACS Catal.* **2020,** *10*, 2354-2377.
36. Rosales, A. R.; Wahlers, J.; Limé, E.; Meadows, R. E.; Leslie, K. W.; Savin, R.; Bell, F.; Hansen, E.; Helquist, P.; Munday, R. H.; Wiest, O.; Norrby, P.-O., *Nat. Catal.* **2019,** *2*, 41-45.
37. Hansen, E.; Rosales, A. R.; Tutkowski, B.; Norrby, P.-O.; Wiest, O., *Acc. Chem. Res.* **2016,** *49*, 996-1005.
38. (a) Devereux, C.; Smith, J. S.; Huddleston, K. K.; Barros, K.; Zubatyuk, R.; Isayev, O.; Roitberg, A. E., *J. Chem. Theory Comput.* **2020,** *16*, 4192-4202; (b) Smith, J. S.; Isayev, O.; Roitberg, A. E., *Chem. Sci.* **2017,** *8*, 3192-3203; (c) Smith, J. S.; Isayev, O.; Roitberg, A. E., *Sci. Data* **2017,** *4*, 170193; (d) Smith, J. S.; Nebgen, B.; Lubbers, N.; Isayev, O.; Roitberg, A. E., *J. Chem. Phys.* **2018,** *148*, 241733.
39. (a) Christoph Bannwarth; Sebastian Ehlert; Stefan Grimme, *J. Chem. Theory Comput.* **2019,** *15*, 1652-1671; (b) Bannwarth, C.; Caldeweyher, E.; Ehlert, S.; Hansen, A.; Pracht, P.; Seibert, J.; Spicher, S.; Grimme, S., *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2021,** *11*.
40. Guan, Y.; Ingman, V. M.; Rooks, B. J.; Wheeler, S. E., *J. Chem. Theory Comput.* **2018,** *14*, 5249-5261.
41. Alegre-Requena, J. V.; Sowndarya S. V., S.; Pérez-Soto, R.; Alturaifi, T. M.; Paton, R. S., *Wiley Interdiscip. Rev. Comput. Mol. Sci. n/a*, e1663.
42. Jorner, K.; Brinck, T.; Norrby, P.-O.; Buttar, D., *Chem. Sci.* **2021,** *12*, 1163-1175.
43. RDKit: Open-source cheminformatics.
44. Riniker, S.; Landrum, G. A., *J. Chem. Inf. Model.* **2015,** *55*, 2562-2574.
45. Pracht, P.; Bohle, F.; Grimme, S., *Phys. Chem. Chem. Phys.* **2020,** *22*, 7169-7192.
46. Newman-Stonebraker, S. H.; Smith, S. R.; Borowski, J. E.; Peters, E.; Gensch, T.; Johnson, H. C.; Sigman, M. S.; Doyle, A. G., *Science* **2021,** *374*, 301-308.
47. 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., *J. Am. Chem. Soc.* **2022**.

CHAPTER 3: STERIC DESCRIPTORS: CAPTURING STERIC PROXIMITY WITH DATA-RICH VECTORS

3.1: Chapter Overview

Steric descriptors, as discussed in Chapter 2, Section 2.3.2, are properties of molecules describing the size and shape a molecule occupies in space. Arising from the repulsive interaction properties of nonbonding electrons, the steric bulk of a molecule can influence reaction outcome by helping to provide or block access to the reactive center over the course of a chemical reaction. In this work, we address a gap in the literature for existing steric parameters, namely the idea of proximity. Simple steric parameters taken from three-dimensional structures are often measured using an atom or bond in a molecule as a reference for where a measurement starts, and measurements are taken for the entire molecule, often represented as a single number. These parameters lack information on how close (proximal) or far (distal) the steric bulk of the molecule captured lies in relation to the chemical point of interest. In our work, we expand on existing steric parameters by making use of three-dimensional information to "slice" a molecule in set intervals, linearly or radially, to take discrete measurements capturing how the steric bulk changes ranging from proximal to distal sites on the molecule relative to the reference site chosen. While the Python package resulting from this work is already freely available (https://github.com/patonlab/DBSTEP), the work in this chapter is under preparation to be submitted. This work was done in collaboration with Tobin Patterson, whose contributions aided in formatting and standardizing the source code of the open-source Python package DBSTEP, as well as benchmarking and testing the package.

3.2: Introduction

Size and shape are important aspects of catalyst design, influencing reaction outcomes such as yield, selectivity, and rate.[1] By understanding how steric influence contributes to reaction

outcome, steric contributions can be tuned for a desirable outcome, more efficient and economic reactions.[2] Quantitative methods of capturing of molecular shape allow chemists to use statistical measures to make predictions to enhance reactivity or discover new molecules by optimizing molecule shape.[3] While existing experimental and computational steric measures find applicability in broad areas of chemistry, further development of these parameters can enhance the understanding and role of molecular shape in reaction outcome.

Steric effects manifest as a repulsive influence by adjacent atoms or functional groups on a molecule, or as intermolecular interactions, where certain molecular shapes influence reaction outcome.[4] Steric parameters (used interchangeably with features or descriptors in this text) are a class of descriptors that aim to quantitatively capture the physical space a molecule occupies. In general, these geometric descriptors are distinct from electronic effects, but in the case for some chemical substituents it has been troublesome to separate the two.[5] Historically, these effects have been described by experimental observations. Early notions of steric effects are summarized by Taft and Charton, relating reaction rates to substituent size in a linear Hammett fashion.[6] Tolman's cone angle is still commonly used to provide information on the breadth of a ligand in reference to a metal site and assess impact on chemical reactivity.[7,8] Various computational steric parameters exist for a chemist to choose,[9] and an increasing access to digital representations of molecules makes the process of obtaining and statistically vetting the appropriate steric parameter to use for experimental prediction simpler.[1, 10]

Digital representations of molecules can store unique information about assorted molecular properties.[11] 2D representations of molecules, such as topological fingerprints can summarize general connectivity information, which can be used to obtain general, parameterized steric information, like volumes described by McGowan, or steric measurements like Crippen molar refractivity.[12] Features may also be obtained from 3D molecular structures using methods ranging from semi-empirical to density functional theory (DFT) to accurately describe the

electronic structure of a molecule. From these representations, steric information can be analytically obtained by using optimized molecular coordinates or electron density surfaces to define molecular volumes to quantify steric effects.

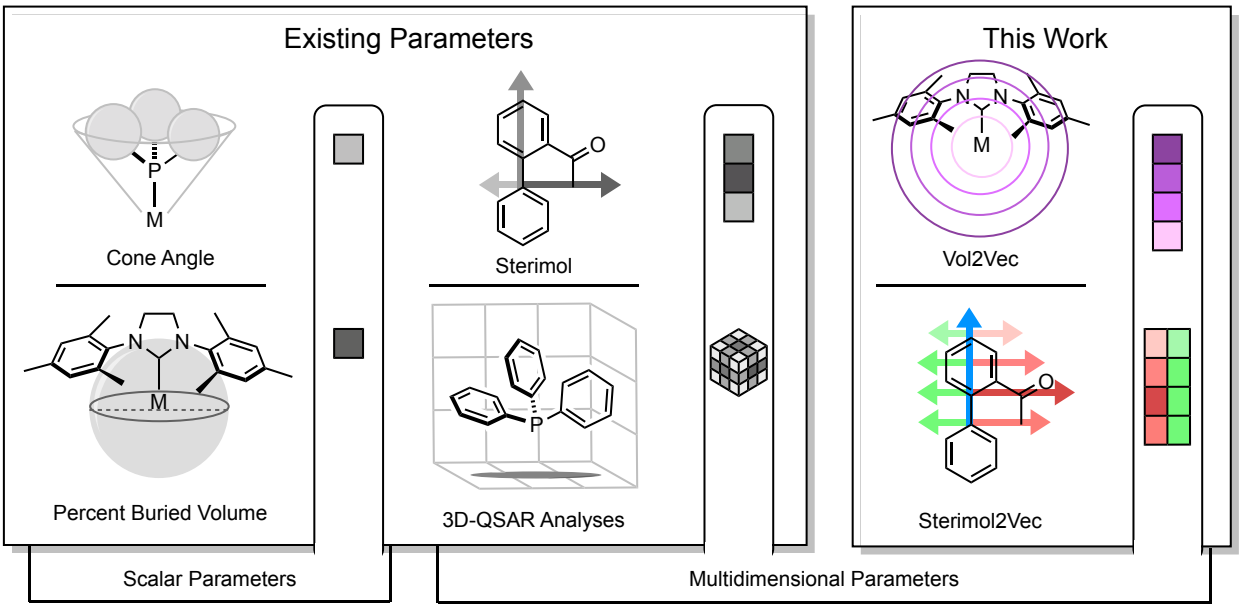


Figure 3.1. *Existing steric parameters commonly used for predictions of reaction outcomes, including single-valued parameters cone angle and percent buried volume, multivariate parameters Sterimol and 3D-QSAR parameter sets, and parameters described by this work, vol2vec and Sterimol2vec.*

We and other research groups have taken interest in the Sterimol set of parameters, originally conceived in the 1970s by Verloop,[13] seeing resurgence due to their usefulness in aiding predictions in asymmetric catalysis.[14,15] Another computational parameter, percent buried volume ($\%V_{bur}$), can account for the amount of ligand occupying the space within the first coordination sphere in a metal-ligand complex.[16] An increase in parameter complexity can be used to discover precise catalytic features.[15a] Use of multidimensional parameters like Sterimol can provide more specific information into which aspects of the physical space a molecule is occupying in order to influence the reaction outcome. Other, more data-rich descriptor sets utilized in "3D-QSAR" studies like Comparative Molecular Field Analysis[17] or Average Steric Occupancy[18] can provide insight into reactivity by mapping molecular interactions to a grid of points. However, preparation

in the form of alignment of molecules is necessary to obtain these parameters, which can become complex on a large scale or with more flexible molecules.

Information that individual steric parameters can capture can be insufficient for reaction prediction. Being able to quantify proximal and distal steric bulk in an interpretable fashion is challenging with available steric parameters. In some cases, combinations of steric parameters have been utilized to capture different aspects of molecular shape. Doyle's study on the steric effect of phosphine ligands relied on specific aspects of percent buried volume and cone angles to showcase how distal steric bulk can influence reaction yield.[19] In this work, we introduce parameter sets that quantitatively capture a set range from proximal to distal steric information relative to a chemical point of interest.

This work describes a collection of steric parameters to tackle unique problems in reaction prediction by quantitatively capturing how near or far from a reaction center steric influence is greatest. Borrowing the "2vec" notation,[20] by translating input molecular structures into a vector or collection of numbers, we introduce two 3-dimensional steric parameters, vol2vec and Sterimol2vec, visualized on the right column of Figure 3.1. These parameters adapt and expand upon existing steric features, percent buried volume and Sterimol, respectively, and are each made up of a combination of individual steric features, ranging from proximal to distal to the reference chemical point of interest. These parameter sets share a common goal of capturing information about the proximity of steric influence to a reactive site, analogous to a radial distribution of a molecule relative to an interesting chemical site. We also introduce 2D-atomic layer-based parameters to collect proximal and distal steric information from a molecular graph, without conformational considerations. These parameters aim to capture detailed information about steric bulk proximal or distal to a reactive site. This work showcases the generalizability of these parameters through their use in predictive statistical models in various applications, which can help support mechanistic insight and aid in the design of new catalysts.

3.3: Methods

Sterimol parameters are constructed from three sub-parameters: L, the length of the molecule, and the minimum (B_{min}) and maximum (B_{max}) widths of the molecule on the axis perpendicular to L.[13] The Sterimol2vec parameter set extends Sterimol parameters by taking new measurements of the width parameters B_{min} and B_{max} in set intervals along the L direction. Visualized in Figure 3.2, the Sterimol2vec parameter set can differentiate between molecules where traditional Sterimol parameters cannot. In the example shown, Sterimol parameters measure the same B_{max} value of 5.6 Å for both ortho and meta methoxy substituents, ignoring proximity of the steric bulk to the reference bond vector chosen in this measurement. With Sterimol2vec, however, we can measure quantitative differences in the collection of B_{max} values for each molecule, as they range from near to far from the chosen reference point, with ortho-methoxy having more proximal steric bulk and meta-methoxy having greater distal steric bulk.

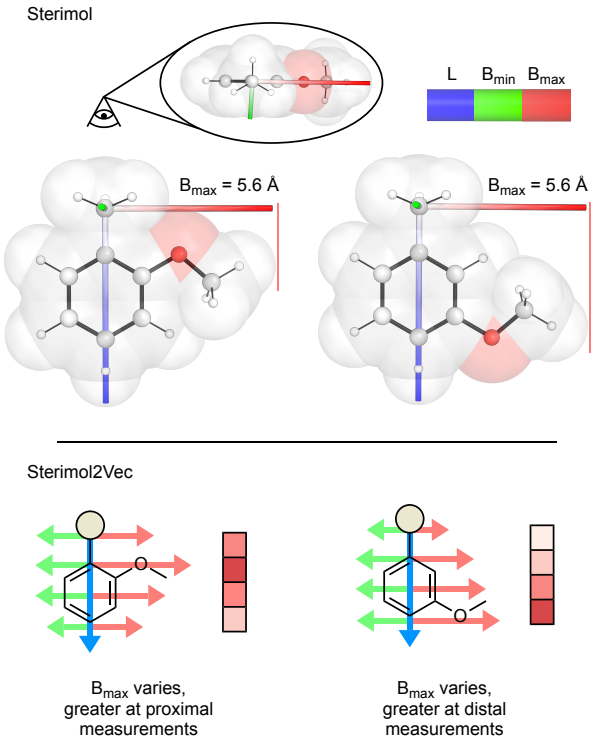


Figure 3.2. *Comparison of Sterimol measurement with Sterimol2vec measurement.*

The percent buried volume[16b] is measured by placing a sphere, typically with a radius of 3.5Å, centered at the metal in a metal-ligand complex, and measuring the percent occupancy of the sphere by the ligand. This provides information on proximal steric bulk of the complex, typically within the first coordination sphere of the metal. We have also augmented this parameter, allowing the radius of the sphere to expand in set intervals. We measure the percent buried volume for each increasing sphere, subtracting out smaller interval measurements from the sphere centers, creating hollow sphere measurements. By measuring the percent buried volume increasing hollow spheres, we can compare radial differences in percent occupancy as measurements move further from the reference center. Figure 3.3 shows how this parameter can identify more precisely where the bulk of a molecule lies in reference to a metal center.

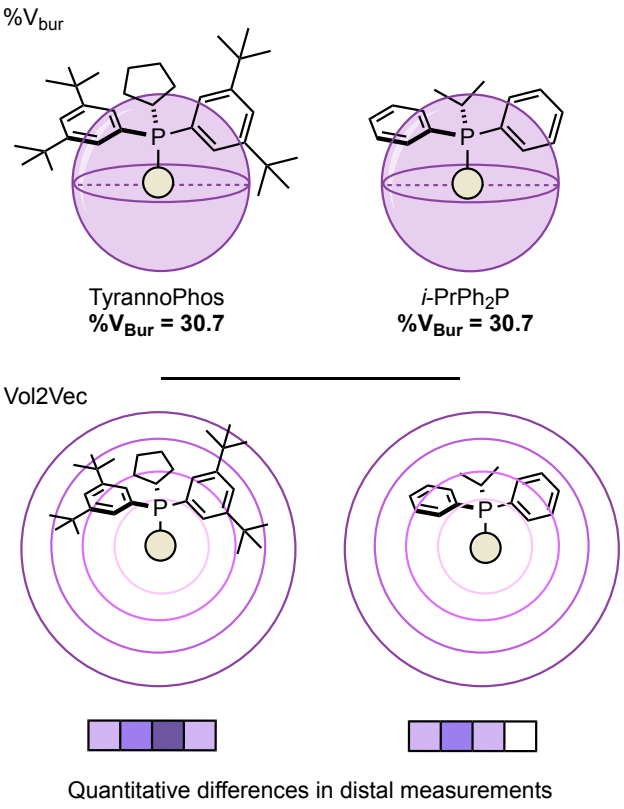


Figure 3.3. *Percent buried volume compared to vol2vec parameter measurement.*

More detailed workflows for how these parameters are collected are detailed in Figure 3.4. For Sterimol2vec measurements, two atoms should be chosen, defining the direction of the vector

that measures L, the length of the molecule. The molecule will then be sliced in even increments along this vector. For the top example shown in Figure 3.4, the molecule was sliced in 1.0Å intervals, from 0.0 to 4.0Å. The minimum and maximum width parameters B_{min} and B_{max} are then measured for each slice. The collection of these widths along the L vector details changes in steric bulk for the input molecule. For vol2vec measurements, one reference atom is needed to identify the center of the spherical measurements. In the step-by-step example in the bottom of Figure 3.4, measurements are made from 0.0-8.0Å, from hollow spheres with a thickness of 2.0Å. Volumetric percent occupancy of each shell is then computed, resulting in a collection of measurements that can be compared to show how steric bulk changes radially relative to the measurement center.

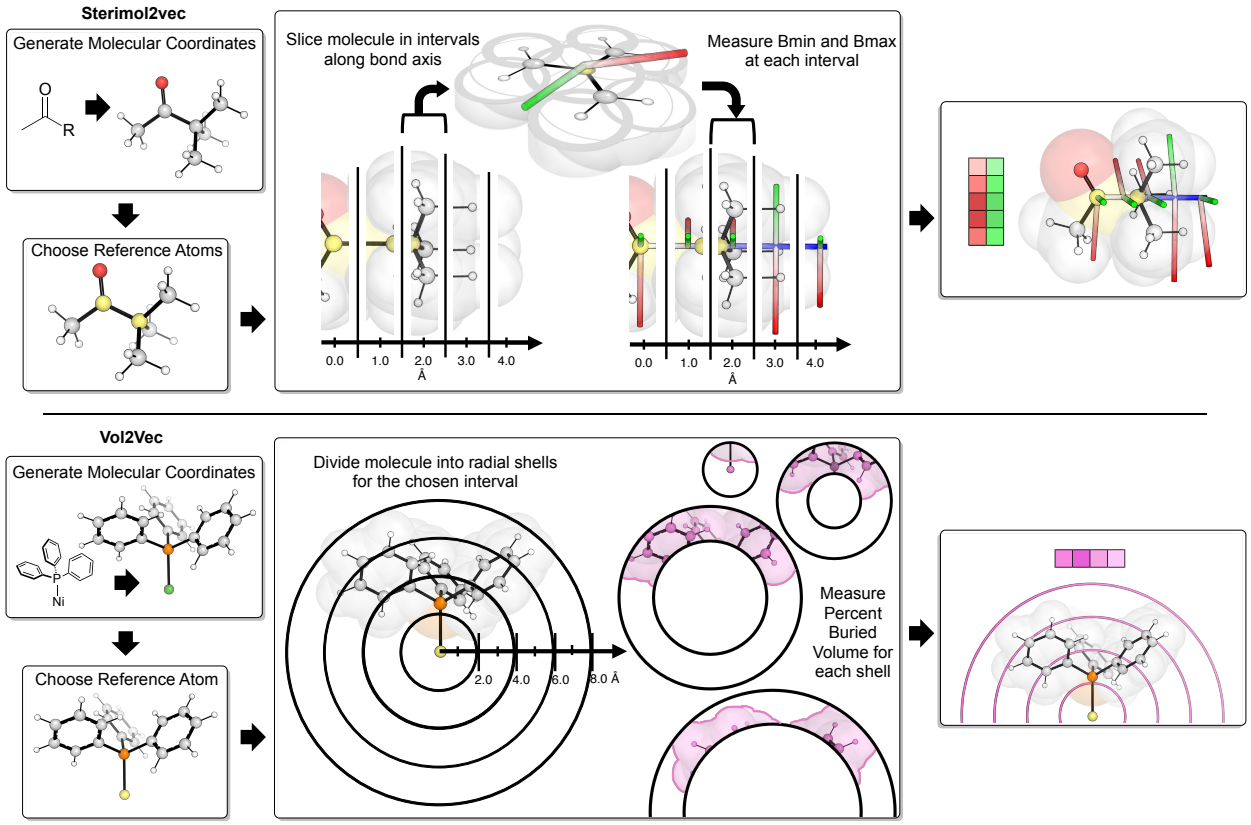


Figure 3.4. *Workflow on parameter collection for Sterimol2vec parameters (top) and vol2vec parameters (bottom).*

Measuring each of these parameters in set intervals allows for the discretization of these descriptors into respective feature vectors. With these steric vectors, statistical models can help identify patterns and importance of what regions of steric influence have the greatest effect on reaction outcome. When considering molecular flexibility, conformational ensemble information can be summarized using these vectors as a minimum-maximum range or Boltzmann weighted ensemble.[21]

This work utilizes Bondi radii[22] to approximate atomic radius. These parameters are measured by overlaying a three-dimensional grid on molecular coordinates, relying on atomic radii to define the space occupied by each atom in the molecule. In some cases when hard spheres meet on the outer surfaces of molecules, unrealistic creases or crevasses may form on the molecular surface. To avoid this, parameters may also be measured from electron density molecular surfaces, which can be constructed from quantum chemistry programs, measuring steric bulk consistent with the curvature of a molecule's electron cloud.

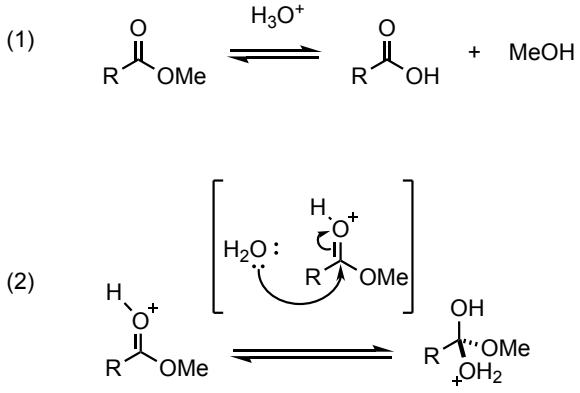
3.4: Applications

With the increased detail that vectorized steric parameters can capture, these descriptor sets are able to generalize information about molecular shape, describing how that shape changes in relation to a specific point on the molecule. With the aid of visual and statistical methods, this section addresses how vectorized steric parameters can relate proximal and distal steric information to reaction outcome, providing insight into where the bounds of steric influence lie.

3.4.1: Modeling Reaction Rates with Vol2Vec

The empirical Taft steric parameter relates the bulkiness of a substituent to the rate of ester hydrolysis shown in Scheme 3.1, with bulkier groups contributing to slower reaction rates, reflected by larger Taft values. A key step where sterics play a role in this reaction happens in the formation of a tetrahedral intermediate (Scheme 3.1, 2), in which the carbonyl carbon undergoes

a nucleophilic attack by a water or hydroxide. Bulkier substituents on this carbonyl can hinder this attack.



Scheme 3.1. *Ester hydrolysis to form the carboxylic acid (1) and the nucleophilic addition step to form the tetrahedral intermediate (2).*

Charton related Taft's findings to the Van der Waals radii of substituents, opening the door to estimation and theoretical prediction of these parameters. In this example, we have collected experimental Taft values for 72 alkyl-substituents,[23] and computationally modeled the carboxylic acid product with DFT, with the B3LYP[24] functional with a D3(Becke-Johnson)[25] dispersion correction at the def2-TZVP[26] basis. We collected vol2vec parameter sets in 1.5Å intervals, beginning 2.5Å away from the carbonyl carbon center, with volumetric shells capturing occupancy by each substituent. A principal component analysis (PCA) shown in Figure 3.5 shows a dimensionality reduction of the four measurements made in two dimensions, each point colored by the molecule's experimental Taft value. The loadings of these principal components aid in the interpretation of how variables relate to individual components and are shown in the sub-table in Figure 3.5. Major contributions to the first component PC1 are the first two proximal measurements done at 2.5 and 4.0Å, and the larger magnitude contributions to PC2 are the two distal measurements at 5.5 and 7.0Å.

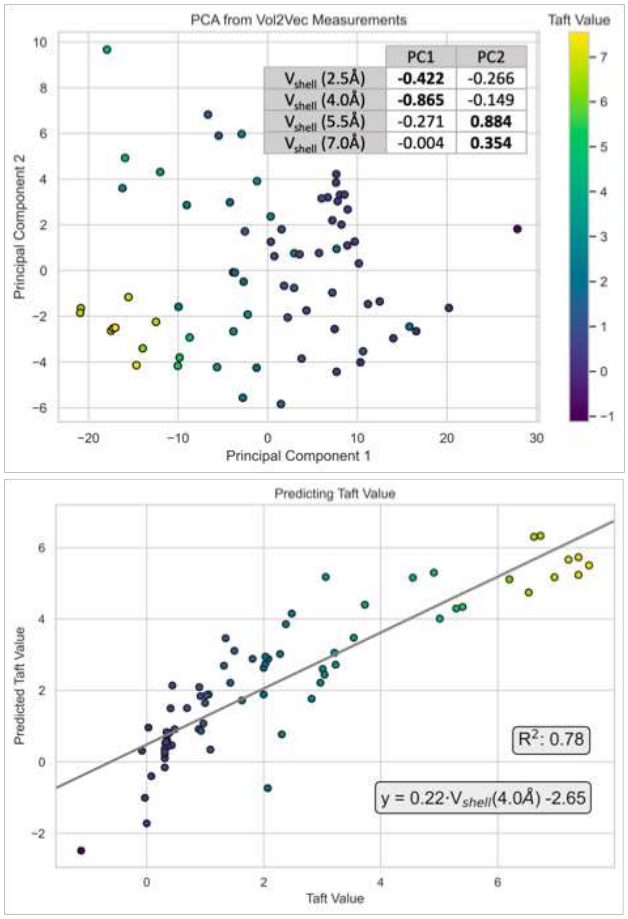


Figure 3.5. *A PCA dimensionality reduction (top) using vol2vec V_{shell} measurements as inputs, with corresponding loadings. A univariate prediction for the experimental Taft value and using the vol2vec measurement at 4.0Å away from the carbonyl carbon of the carboxylic acid (bottom).*

We observed a transition of large to small Taft values along the x-axis of the PCA and investigated the two proximal measurements in assessing their predictivity for the experimental value. A univariate correlation was observed between the Taft value and the proximal V_{shell} measurement made at 4.0Å, suggesting steric contributions 3.25-4.75Å away from the carbonyl impact the rate of this reaction, with larger steric measurements contributing to larger Taft values and slower reaction rates.

3.4.2: Atropisomer Rotational Barriers with Sterimol2vec

Another experimentally measured steric parameter, the interference value, measures the half-life of racemization for atropisomers, biaryl molecules in which ortho substituents hinder rotation about the connecting biaryl C-C bond. In a simplified model system, shown in Figure 3.6, we have

computed the rotational barriers for 16 biaryl systems with varying ortho-substituents using DFT by optimizing the ground and transition structures using B3LYP[24] functional with a D3(Becke-Johnson)[25] dispersion correction and 6-31+G(d)[27] basis and computing the relative free energies between them, with the ground state as a reference. We measured Sterimol2vec values using the biaryl C-C bond as a reference point.

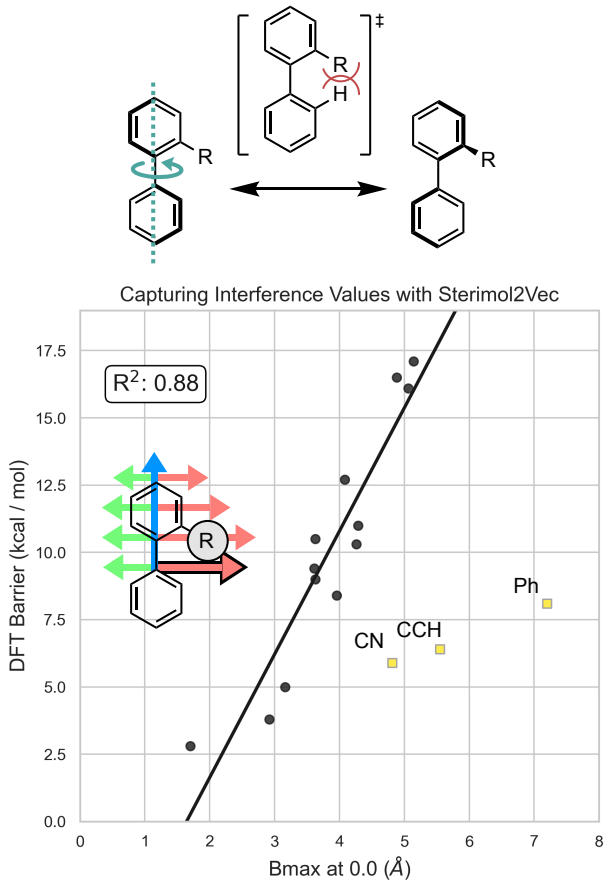


Figure 3.6. *A depiction of the rotational barrier measured by DFT (top) and a univariate comparison between the B_{max} measured at 0.0 Å. Highlighted points were not included in the calculation of the displayed R^2.*

When comparing steric contributions along this bond vector for all points, R^2 correlation value of 0.19 was observed between the rotational barrier and the maximum width, Bmax, at 0.0Å, however, the three yellow highlighted points in Figure 3.6, more linear or linear-profiled substituents, did not seem to follow the same trend as the rest of the dataset. When these points

were removed, a strong R^2 of 0.88 was observed. This trend is greater than using the normal overall Sterimol values, in which case an R^2 of 0.55 was observed for the same points (see Appendix A for further information), showing that differentiating between distal and proximal steric influence is important can help model experimental values for prediction of molecular behavior.

3.4.3: Vol2Vec Applied to Phosphine Ligands

These features can be used to capture both proximal and distal sterics in systems where certain steric criteria need to be met for each case. In a 2017 study, Doyle and Wu observed that when featurizing the ligand of a Ni-catalyzed cross-coupling reaction, a combination of low proximal steric contributions and larger distal steric influence was able to explain ligand effects in reactivity quite well, building a linear correlation using percent buried volume to represent proximal sterics and using the cone angle to capture distal steric effects, along with an electronic term, the minimum electrostatic potential of the ligand, V_{min}. While it was not the initial intent of either of these steric parameters to be used in tandem, the vol2vec parameter sets were designed to capture both proximal and distal steric information.

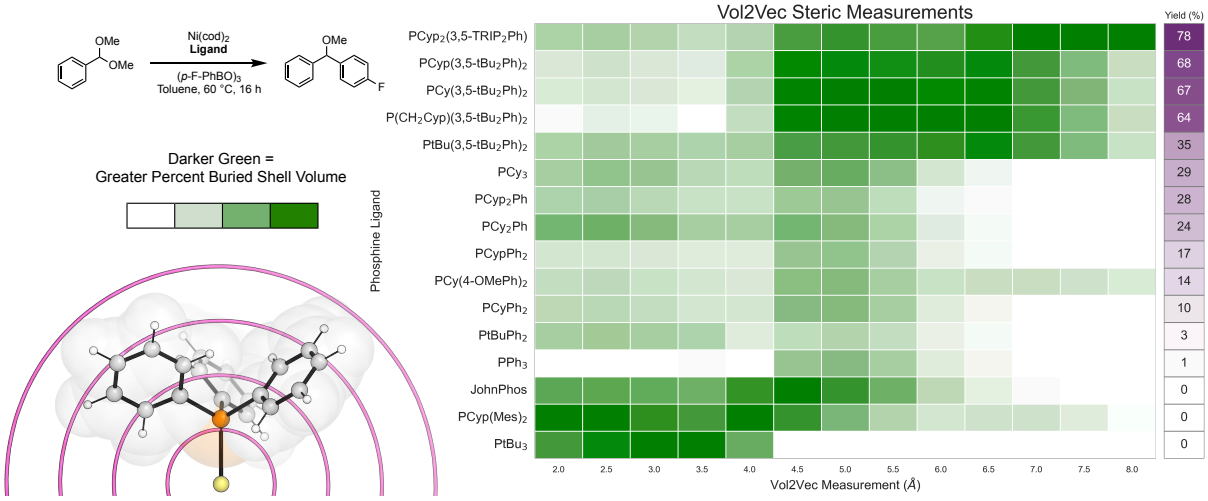


Figure 3.7. *Vol2vec parameters measured from monophosphine ligand structures used in Ni catalyzed cross coupling reactions. Relative vol2vec measurement values are plotted, with higher values in darker green, plotted alongside yield, shown in purple.*

We have measured vol2vec parameters for each phosphine ligand used in this study in 0.5Å intervals from 2.0-8.0Å. A visualization of these features is shown in Figure 3.7. Structures

were generated and optimized using the xTB[28] program, using the semi-empirical method GFN2-xTB to optimize geometries. Relative shell occupancies for each measurement interval are shown, normalized for each column, with ligands sorted by the corresponding reaction yield. In agreement with the original findings of Doyle, ligands with lower proximal steric contributions (2.0-4.0Å), but larger distal contributions (4.0-8.0Å away) result in the best performing reactions, shown in the first few rows of Figure 3.7. Ligands with greater proximal yield, depicted by darker green occupancy in the lower left of Figure 3.7, and low distal yield do not allow this reaction to progress, indicated by low or zero yield.

3.4.4: Rapid Steric Proximity Categorization to Explore Chemical Space

Fragment based drug discovery is a technique that can be used in early drug-discovery processes to assess chemical fragments as key structural motifs or building blocks for a more potent drug molecule.[29] This process involves assessing the ability of a large number of fragments to bind or inhibit a target protein of interest. Fragments can then be incorporated and optimized to form a lead molecule with stronger desirable pharmaceutical properties. High throughput screening of many molecules is possible using NMR or X-ray crystallography to characterize fragment-binding affinity, increasing efficiency of determining hit fragments, however, high throughput virtual screening through the form of molecular docking simulations or building quantitative structure activity relationships with relevant chemical properties can aid in identifying hit fragment molecules before experimental trials, saving time and resources.[30-31]

The Carboxylic Acid Fragment Library, supplied by Enamine is a fragment-based library consisting of 4000 small diverse carboxylic acid molecules, made available virtually and experimentally. From the SMILES strings of these molecules, we have collected 2D-layered steric parameters using the carboxylic acid as the fragment of interest, measuring Crippen molar refractivity values for each layer, from 1 to 8 bonds away from the carboxylic acid group. These parameter sets for each molecule are quickly obtained and allow for quantitative differentiation

between different fragments based on how their shape changes ranging near to far from the carboxylic acid group. One method of dimensionality reduction, UMAP, was employed to depict the chemical space of these molecules, using 2D-layered steric parameters as an input, from layers 1-8, reduced to two dimensions, depicted in Figure 3.8.

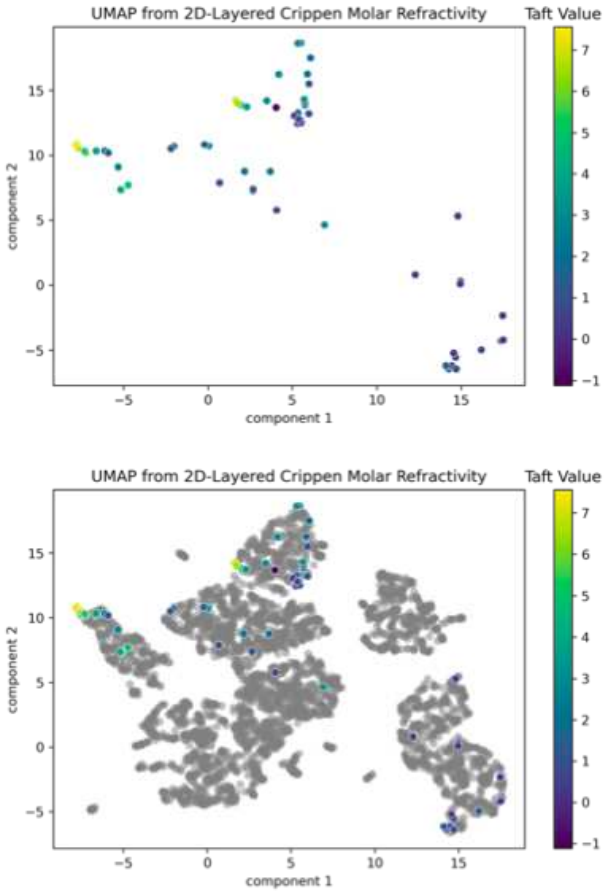


Figure 3.8. *Visualization of UMAP dimensionality reduction for 2D graph-based parameters measured Crippen molar refractivity values for 72 alkyl-substituted carboxylic acids (top, colored by Taft value) overlaid on 4000 carboxylic acid fragments (bottom, gray).*

When comparing the initial Taft dataset to the fragment library, we find areas of unexplored chemical space, containing molecules with shapes that diverge from the original set of 72 alkanes. Manual or algorithmic methods can be employed to explore this chemical space further to fill gaps in knowledge about reactivity in these unexplored areas.[32] The 2D-layered steric parameters were

rapidly obtained and provide a method of quantifying and exploring diverse structural differences in molecules.

3.5: Computing Parameter Sets

We have developed an open-source Python program, DFT-Based Steric Parameters (DBSTEP)[33] to collect the described vectorized steric parameters, including vol2vec, Sterimol2vec, and 2D layer-based parameters, along with percent buried volume, and Sterimol. The program will accept and extract coordinates from a variety of computational chemistry output files, along with other common chemical file formats (.xyz, .sdf, .pdb), and SMILES strings for 2D graph-based parameters. For ease of use, the program has been designed to be used on the command line or in a Python script or Jupyter notebook. Comparisons of parameter accuracy compared to original Sterimol formulations, as well as comparisons with DBSTEP computed volumes with experimental molecular volume measurements can be found in Appendix A.

3.6: Conclusions

This paper introduces three multidimensional steric parameters, vol2vec, Sterimol2vec and 2D graph-based steric parameters, each of which quantifies steric influence from a range proximally or distally to a chemical point of interest. These parameters are generalizable, capturing aspects of existing experimentally obtained steric values, such as Taft and interference values. When conformational effects are considered, parameters can describe energetic minimum and maximum structures, as well as a Boltzmann-averaged feature set. 2D graph-based steric parameters can be used in rapidly determining the chemical space of a dataset, without relying on computing or optimizing individual molecular conformations. We have also introduced and have made available an open-source Python package, DBSTEP, to collect these parameter sets from 3D coordinates or SMILES strings in the case for 2D steric parameters.

3.7: References

1. J. Durand, D.; Natalie Fey, *Chem. Rev.* **2019,** *119*, 6561-6594.
2. Anastas, P. T.; Kirchhoff, M. M.; Williamson, T. C., *Appl. Catal. A-Gen* **2001,** *221*, 3-13.
3. Santiago, C. B.; Guo, J.-Y.; Sigman, M. S., *Chem. Sci.* **2018,** *9*, 2398-2412.
4. (a) Kier, L. B., *Med. Res. Rev.* **1987,** *7*, 417-440; (b) Joyce, J. P.; Billman, M. M.; Chandorkar, S.; Rappé, A. K., Sterics, the core of intermolecular interactions. In *Intra- and Intermolecular Interactions Between Non-covalently Bonded Species*, Elsevier: 2021; pp 1-38.
5. Solel, E.; Ruth, M.; Schreiner, P. R., *J. Am. Chem. Soc.* **2021,** *143*, 20837-20848.
6. (a) Taft, R. W., *J. Am. Chem. Soc.* **1952,** *74*, 2729-2732; (b) Taft Jr, R. W., *J. Am. Chem. Soc.* **1952,** *74*, 3120-3128; (c) Taft Jr, R. W., *J. Am. Chem. Soc.* **1953,** *75*, 4538-4539; (d) Charton, M., *J. Am. Chem. Soc.* **1975,** *97*, 1552-1556; (e) Charton, M., *J. Org. Chem.* **1976,** *41*, 2217-2220.
7. Tolman, C. A., *Chem. Rev.* **1977,** *77*, 313-348.
8. Mcford, A. W.; Butts, C. P.; Fey, N.; Alder, R. W., *Journal of the American Chemical Society* **2021,** *143*, 13573-13578.
9. Falivene, L.; Cao, Z.; Petta, A.; Serra, L.; Poater, A.; Oliva, R.; Scarano, V.; Cavallo, L., *Nat. Chem.* **2019,** *11*, 872-879.
10. Natalie Fey, *Dalton Trans.* **2010,** *39*, 296-310.
11. Gallegos, L. C.; Luchini, G.; St. John, P. C.; Kim, S.; Paton, R. S., *Acc. Chem. Res.* **2021,** *54*, 827-836.
12. (a) Abraham, M. H.; Mcgowan, J. C., *Chromatographia* **1987,** *23*, 243-246; (b) Wildman, S. A.; Crippen, G. M., *J. Chem. Inf. Comput. Sci.* **1999,** *39*, 868-873.
13. Verloop, A., Drug Design. Ariens, E. J., Ed. Academic Press: New York,, 1976; Vol. III.
14. Piou, T.; Romanov-Michailidis, F.; Romanova-Michaelides, M.; Jackson, K. E.; Semakul, N.; Taggart, T. D.; Newell, B. S.; Rithner, C. D.; Paton, R. S.; Rovis, T., *Journal of the American Chemical Society* **2017,** *139*, 1296-1310.
15. (a) Harper, K. C.; Bess, E. N.; Sigman, M. S., *Nat. Chem.* **2012,** *4*, 366-374; (b) Ardkhean, R.; Mortimore, M.; Paton, R. S.; Fletcher, S. P., *Chemical Science* **2018,** *9*, 2628-2632.
16. (a) Falivene, L.; Credendino, R.; Poater, A.; Petta, A.; Serra, L.; Oliva, R.; Scarano, V.; Cavallo, L., *Organometallics* **2016,** *35*, 2286-2293; (b) Hillier, A. C.; Sommer, W. J.; Yong, B. S.; Petersen, J. L.; Cavallo, L.; Nolan, S. P., *Organometallics* **2003,** *22*, 4322-4326; (c) Poater, A.; Cosenza, B.; Correa, A.; Giudice, S.; Ragone, F.; Scarano, V.; Cavallo, L., *Eur. J. Inorg. Chem.* **2009,** *2009*, 1759-1766.
17. (a) Lipkowitz, K. B.; Pradhan, M., *J. Org. Chem.* **2003,** *68*, 4648-4656; (b) Ianni, J. C.; Annamalai, V.; Phuan, P.-W.; Panda, M.; Kozlowski, M. C., *Angew. Chem. Int. Ed.* **2006,** *45*, 5502-5505.
18. Zahrt, A. F.; Henle, J. J.; Rose, B. T.; Wang, Y.; Darrow, W. T.; Denmark, S. E., *Science* **2019,** *363*, eaau5631.
19. Wu, K.; Doyle, A. G., *Nat. Chem.* **2017,** *9*, 779-784.
20. Mikolov, T.; Chen, K.; Corrado, G.; Dean, J., **2013**.
21. Brethomé, A. V.; Fletcher, S. P.; Paton, R. S., *ACS Catalysis* **2019,** *9*, 2313-2323.
22. (a) Bondi, A., *J. Phys. Chem.* **1964,** *68*, 441-451; (b) Mantina, M.; Chamberlin, A. C.; Valero, R.; Cramer, C. J.; Truhlar, D. G., *J. Phys. Chem. A* **2009,** *113*, 5806-5812.
23. Macphee, J. A.; Panaye, A.; Dubois, J.-E., *Tetrahedron* **1978,** *34*, 3553-3562.

24. Stephens, P. J.; Devlin, F. J.; Chabalowski, C. F.; Frisch, M. J., *J. Phys. Chem.* **1994,** *98*, 11623-11627.
25. Grimme, S.; Ehrlich, S.; Goerigk, L., *J. Comput. Chem.* **2011,** *32*, 1456-1465.
26. (a) Weigend, F.; Ahlrichs, R., *Phys. Chem. Chem. Phys.* **2005,** *7*, 3297-3305; (b) Weigend, F., *Phys. Chem. Chem. Phys.* **2006,** *8*, 1057-1065.
27. (a) Clark, T.; Chandrasekhar, J.; Spitznagel, G. W.; Schleyer, P. V. R., *J. Comput. Chem.* **1983,** *4*, 294-301; (b) Petersson, G. A.; Bennett, A.; Tensfeldt, T. G.; Al-Laham, M. A.; Shirley, W. A.; Mantzaris, J., *J. Chem. Phys.* **1988,** *89*, 2193-2218; (c) Petersson, G. A.; Al-Laham, M. A., *J. Chem. Phys.* **1991,** *94*, 6081-6090.
28. Christoph Bannwarth; Sebastian Ehlert; Stefan Grimme, *J. Chem. Theory Comput.* **2019,** *15*, 1652-1671.
29. Erlanson, D. A.; Mcdowell, R. S.; O'Brien, T., *J. Med. Chem.* **2004,** *47*, 3463-3482.
30. Congreve, M.; Chessari, G.; Tisi, D.; Woodhead, A. J., *J. Med. Chem.* **2008,** *51*, 3661-3680.
31. Salum, L. B.; Andricopulo, A. D., *Molecular Diversity* **2009,** *13*, 277-285.
32. Shields, B. J.; Stevens, J.; Li, J.; Parasram, M.; Damani, F.; Alvarado, J. I. M.; Janey, J. M.; Adams, R. P.; Doyle, A. G., *Nature* **2021,** *590*, 89-96.
33. Luchini, G.; Patterson, T.; Paton, R. S., DBSTEP: DFT-Based Steric Parameters. 2021.

CHAPTER 4: ELECTRONIC DESCRIPTORS: EVALUATING COMPUTED ELECTRONIC PROPERTIES IN PREDICTING HAMMETT CONSTANTS

4.1: Chapter Overview

Continuing the discussion on computed molecular properties, this chapter examines how well computed electronic properties relate to tabulated experimental values. The Hammett parameters, σ_m, and σ_p, are monumental properties in reaction prediction and optimization. Experimentally determined from reaction rates, they directly capture the electronic influence (i.e., electron donating or withdrawing character) that aryl substituents have on reaction outcome through linear relationships. Hammett relationships provide insight on the underlying reaction mechanism, quantitatively capturing how the aryl system is influenced by varying substituents. Since these relationships are useful in describing chemical reactivity, there have been several ventures to predict these parameters computationally. A relatively simple method to describe electronic effect of the aryl system computationally is by measuring the partial atomic charge for atoms in the aryl ring. The partial atomic charge captures how the overall molecular charge is partitioned by individual atomic contributions. Partial atomic charge does not correspond to a physically observed measurement, and so many methods have arisen with unique ways to partition molecular charges. This work compares and benchmarks different methods for computing atomic charge, along with computed NMR values, comparing to how well each method relates to the experimental Hammett value. The work in this chapter is in preparation to be submitted, and calculations and analysis were all performed by G. Luchini.

4.2: Introduction

Empirical linear free energy relationships can be insightful in providing mechanistic interpretation for a chemical reaction.[1] Parameters and descriptors derived from these relationships are also often highly transferrable to new chemical systems to describe steric or electronic effects, for

example. These “top-down” descriptors capture information about a chemical system’s behavior in a flask, describing empirically observable properties of molecules. These properties are specific to the system being studied, for example, a reaction rate. Contrarily, “bottom-up” descriptors, obtained computationally from electronic structure calculations, rely on building up a chemical ensemble from unique accessible conformations of a molecule, from which a key conformer’s computed property or an ensemble-averaged value can provide insight into a reaction mechanism. An advantage of these bottom-up descriptors is that they do not rely on the time and resources of performing an experiment, and the scope of molecules is easily expanded, however, appropriate and adequate conformational sampling should still be conducted to best represent and capture the behavior of each molecule studied. Bottom-up descriptors are in general more broadly applicable. For example, atomic charge can be computed and compared across a variety of chemical systems.

Hammett constants are a key example of top-down descriptors. Experimentally determined, these constants are commonly represented by two parameters, σ_m, and σ_p. They were initially measured from ionization constants of benzoic acids, comparing relative rates of ionization when substituents are placed in the meta (σ_m) or para (σ_p) position, relative to the acid,[2] but have been found to be generally applicable across a multitude of reactions involving aryl substituents as a method of determining the sensitivity of a reaction to electronic effects. In a chemical reaction, bonds are formed and broken at the transition state (TS). This higher energy species can determine the rate of the reaction. Electron withdrawing substituents on the aryl system will have a positive σ value, and generally help to stabilize the buildup of negative charge at the reaction center in the TS. Electron donating substituents have a negative σ value and help to stabilize the buildup of positive charge in the TS. The magnitude of σ values reflect the strength of the effect. The ability to relate substituent electronic effects to chemical reactivity has historically proven useful in a variety of applications.[3] Given the utility of these parameters, there

have been several ventures to predict these parameters computationally, using bottom-up approaches.[4] Because these systems rely on characterizing an aryl substituent's ability to stabilize the buildup of charge in the TS, partial atomic charges are often used in modeling these parameters Although partial atomic charges do not correspond directly to a physically observable quantity, they are able to provide an intuitive conceptual medium for how electronic charge is partitioned in a molecule. Many unique charge models exist in the literature and are made available in a variety of quantum or semi-empirical chemistry programs.

Charge models have been described more extensively in other articles and reviews,[5] and much work has gone into the comparison of the variety of different methods for computing atomic charge.[6] There seems to be a consensus in the literature for the categorization of different charge models into different classes ranging from I-IV, justified by Cramer and Truhlar[7] and further described by Martin:[5b] with charge values (I) derived from experimentally measured properties, such as deformation densities or dipole moments; (II) computed from molecular orbital or electron density information, with examples including Mulliken,[8] Natural Population Analysis (NPA),[9] or Hirshfeld[10] charges; (III) computed from wave function or electron density information and fitting to a physical observable, (such as dipole moment or electrostatic potential) seen in the CHarges from ELectrostatic Potentials using a Grid-based method (CHELPG),[11] Merz-Kollman charges using Universal force field radii (MKUFF),[12] Hu, Lu, Yang charges with Gaussian 16's standard atomic densities (HLYGat)[13] and Atomic Polar Tensor (APT)[14] method; or (IV) Based on semiempirical adjustments to Class II or III methods, such as in Charge Model 5 (CM5).[15] Since partial atomic charge is not an experimental observable, there is no one ground truth method to produce atomic charge. Each charge method operates on its own scale, and all find value in a variety of applications.[16] In this work, we explore charge models available in the quantum mechanical (QM) software package Gaussian 16, which includes charge values from Classes II-IV.[17]

NMR chemical shifts have also been explored in their relationship to Hammett parameters, appealing due to their ability to capture local electronic effects with a direct experimental observable.[18] Perhaps due to more complex nuclear coupling relationships captured in this experimental measurement, correlations to partial atomic charge values may not be as high.[19] Still, successful studies have utilized NMR shifts in their relationship to Hammett parameters and partial atomic charges.[20] A variety of methods also exist to compute NMR shifts using QM software, including Gauge-Independent Atomic Orbital (GIAO),[21] Individual Gauges for Atoms in Molecules (IGAIM),[22] and the Continuous Set of Gauge Transformations (CSGT)[23] methods.

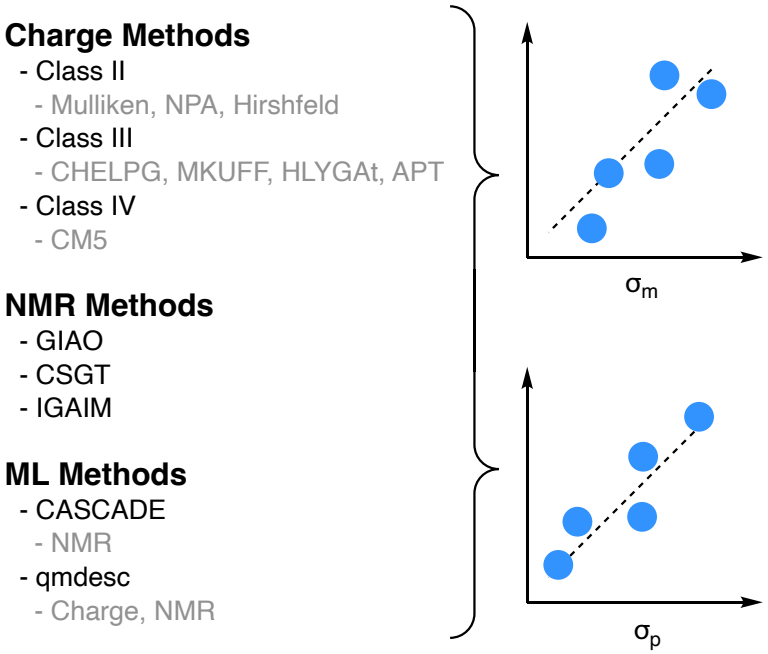


Figure 4.1. *Computed charge and NMR shift methods to be compared with experimental Hammett parameters.*

Machine learning (ML) models have been trained to compute partial atomic charge and NMR shift. When utilized within their domain of applicability, these models give a highly accurate prediction of QM or experimental properties, while being much faster than traditional QM calculations. These models typically rely on simple molecular representations, such as a SMILES string, as input. The *qmdesc* Python library utilizes a model that has been trained to predict

Hirshfeld charges and GIAO NMR shift, among other QM properties.[24] The *CASCADE* Python library also can be utilized to predict experimental ^{1}H and ^{13}C NMR shift.[25]

In this work, we compare the performance of various computational methods, mainly from the QM software package Gaussian 16, along with results from the semi-empirical package xTB[26] and additional ML-predicted values, in their ability capture electronic effects in their relation to experimental Hammett values (Figure 4.1). We analyze atoms inside and outside of the aromatic ring, probing charge values at aromatic carbons in the meta and para positions relative to the substituent, as well as the attached hydrogens. Additionally, we study a more complex conjugated nitroarene system, studying the charges again at the aromatic carbon as well as the nitro-group nitrogen. This bottom-up approach utilizes conformational ensembles of structures to investigate relationships between the method to compute electronic effects and the importance of the atomic position chosen to compare with experimental values. With this, we develop suggestions and best practices when modeling Hammett parameters with computed values.

4.3: Methods

A dataset of 89 molecules, shown in Figure 4.2 along with their experimental Hammett sigma values, was curated by Ertl,[4b] who comprised the data from the 200 most common aryl substituents in the ChEMBL database,[27] of which 89 had tabulated experimental Hammett values. Molecules were converted from SMILES strings to 3D structures with hydrogens using RDKit.[28] Coordinates were then submitted to CREST[29] to generate a conformational ensemble for each molecule. Electronic structures were then optimized, and frequencies were computed using Gaussian 16 using the B3LYP functional[30] and def2TZVP basis set[31] with a Becke-Johnson damped Grimme D3-dispersion correction[32] in chloroform using SMD implicit solvation.[33] Gaussian 16 was also used to compute a variety of charges from Classes II, III, and IV, along with NMR chemical shifts at the same level of theory. Variations of charges were also computed, including Minimal-Basis Mulliken charges (MBS-Mulliken)[34] and iterative Hirshfeld and iterative

CM5 methods.[35] ML predicted Hirshfeld charges were computed by qmdesc and NMR shifts were predicted using qmdesc and CASCADE models. Additionally, xTB was used to compute Mulliken and CM5 charges, using the Density Functional Theory (DFT) optimized geometries. For species with multiple conformers, the charge and NMR values were Boltzmann averaged using the DFT-computed Gibbs free energy values.

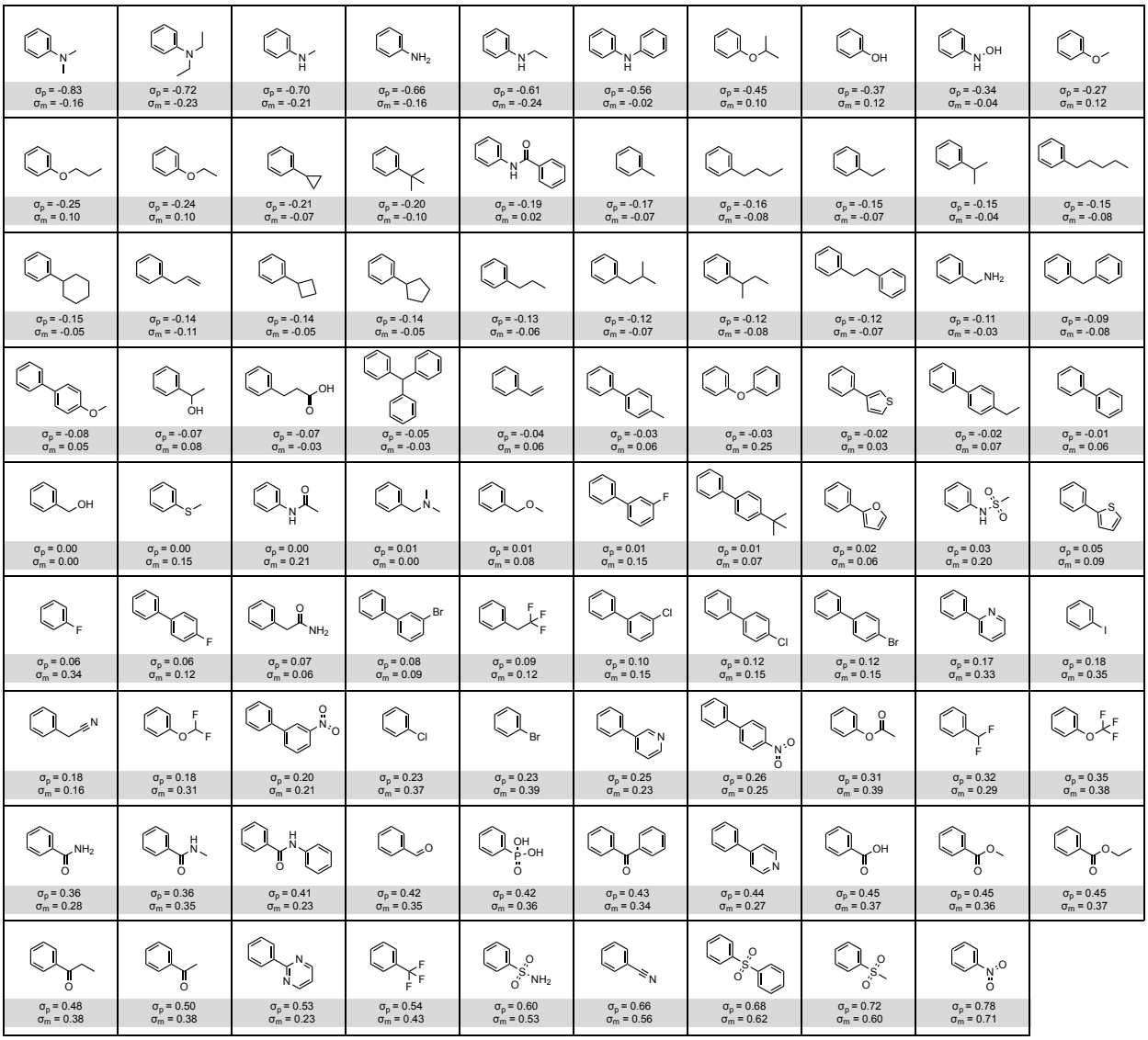


Figure 4.2. *Dataset of aryl substituents displayed along with their experimental σ_p and σ_m Hammett values.*

From the output files, computed charge and NMR shift values were parsed at the carbons in the meta and para positions, relative to the substituent point of attachment. Values obtained

from both meta carbons were averaged, though in most cases there was little difference in the values. The values were then compared on a univariate basis to the corresponding experimental σ_p and σ_m values using Pearson R^2.

4.4: Results and Discussion

A full list of charge and NMR shift models is present in Figure 4.3, which shows a correlation heatmap of Pearson R^2 values between computed values (rows) and the experimental Hammett constants (columns). In general, we observed poor correlations between computed meta values and experimental σ_m values. Apart from Hirshfeld, ML-computed Hirshfeld and CM5 charges which reach R^2 values of 0.84, 0.84 and 0.83 respectively, the charge methods range from 0.00 to 0.38, corresponding to poor correlations. Computed para values compare well, for the most part, with experimental σ_p values yielding correlations of 0.74 R^2 and above, except for the Class III charge methods, specifically CHELPG, MKUFF, and HLYGAt, which poorly correlate with experimental Hammett values. These three methods are all based on the electrostatic potential of the molecule, opposed to the other Class III method APT, which is based on dipole moment. Interestingly, NPA charge, a prominently used charge method in the literature, is also prone to differences in correlations between meta (R^2=0.12) and para (R^2=0.86) values. Even in the para case, NPA charges are outperformed in their ability to predict experimental value by MBS-Mulliken (R^2=0.90), Hirshfeld (R^2=0.92), ML-QMDESC Hirshfeld (R^2=0.89), and CM5 (R^2=0.92) charge methods.

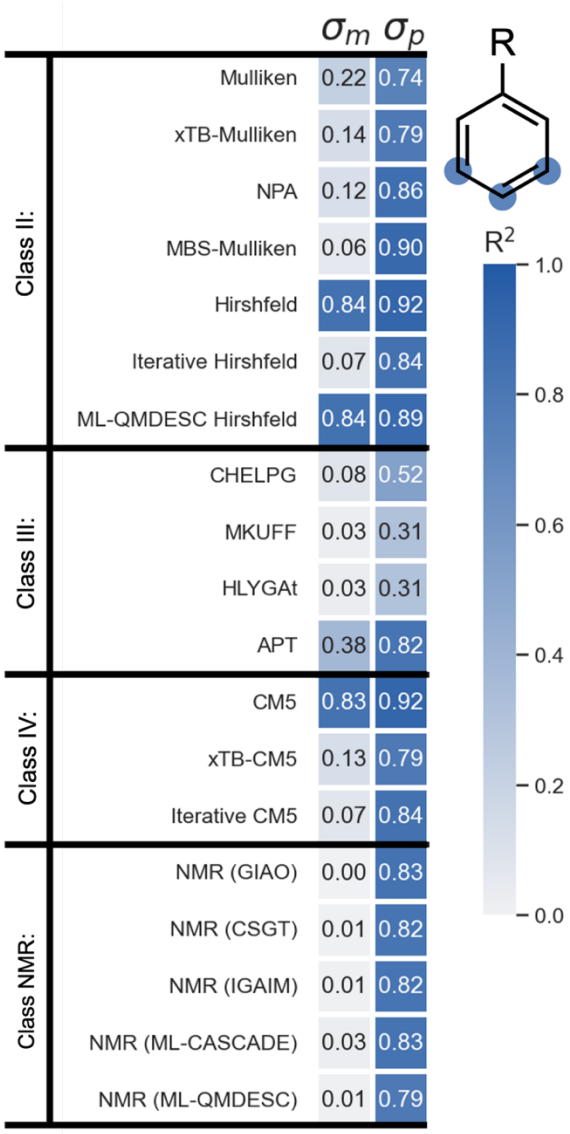


Figure 4.3. *Heatmap correlations for carbon computed charge and NMR shift values with σ_p and σ_m values.*

To further investigate discrepancies in the prediction of Hammett values between the meta and para positions, we examined computed charge and NMR shift values from hydrogens at each aryl para and meta carbon compared to their corresponding Hammett values for each substituent. This resulted in an overall improvement in R^2 correlation for all meta values apart from Class III and NMR values. Correlations for para values remained relatively consistent. Correlation heatmaps against values taken from hydrogen positions are visualized in Figure 4.4. We

hypothesize that since the carbon atoms in the meta position are less prone to changes in their electronic density by resonance and inductive substituent effects than in the para position, these small variations may not be adequately captured by many charge models. The attached hydrogen atoms outside of the aromatic system may be more sensitive to changes in the substituent, but it is less likely a through-bond effect from the meta carbon. Depending on the orientation of the substituent, the hydrogen in the meta position may experience the through-space effects of the substituent. This suggests that in some cases it might be appropriate to utilize computed charge values of atoms outside of the aromatic system to evaluate electronic substituent effects.

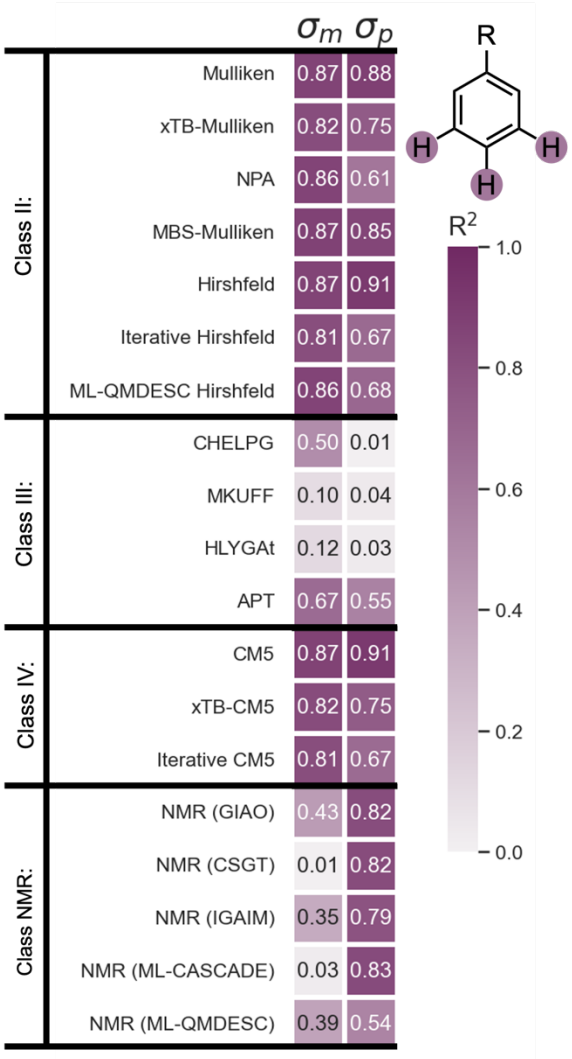
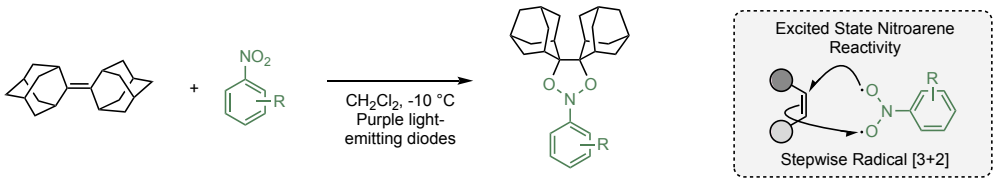


Figure 4.4. *Heatmap correlations for hydrogen computed charge and NMR shift values with σ_p and σ_m values.*

To test this hypothesis, we sought out a slightly more challenging system. In 2022, Leonori and coworkers published work showcasing the oxidative cleavage of olefins with photoexcited nitroarenes. In this work, electronic influence from the substituent(s) of the nitroarene in these reactions was explored using Hammett correlations, comparing reaction rates with tabulated sigma values.[36] Substituents utilized in this study are in a variety of combinations of meta and para positions and in the Hammett analysis it is assumed the substituent effects are additive when multiple substituents are present. A scheme of this reaction and the scope of fourteen nitroarenes used in this work is shown in Figure 4.5. The nitroarenes are shown with their reported relative rate values, with the unsubstituted nitroarene used as a reference.



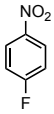
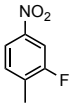
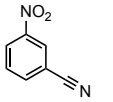
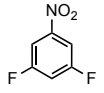
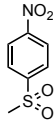
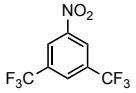
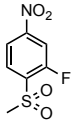
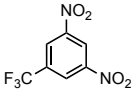
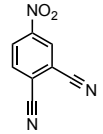
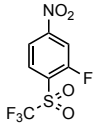
						
-0.0716	0	0.0181	0.0653	0.2513	0.3328	0.4593
						
0.4469	0.5019	0.7078	0.8351	0.8729	1.0272	1.1454

Figure 4.5. *Scheme of oxidative olefin cleavage by photoexcited nitroarenes reaction reported by Leonori (top). The scope of nitroarenes used in this study, along with their reported relative rates (bottom).*

We modeled the nitroarenes in the ground singlet state, following the same computational procedure as previously mentioned, collecting the same charge and NMR shift value. These values were obtained from both the aryl carbon attached to the nitro group as well as the nitro nitrogen for the fourteen substituents. When compared with the experimental relative rate values, shown in Figure 4.6, the values follow a similar trend as the previous dataset's meta values. The best performing methods when looking at the carbon charges were Hirshfeld (R^2=0.94), CM5

(R^2=0.86), and MBS-Mulliken (R^2=0.81), while other methods range from R^2 values of 0.11-0.42. For values computed at the nitro nitrogen, we observed an overall improvement in R^2 in most cases, again except for Class III charges. Mulliken and NPA charges still yield a poor correlation with experimental values for these values for this dataset, perhaps due to a low-data problem. These results reiterate that the method of charge calculation and the atomic position chosen for comparison influence the obtained results.

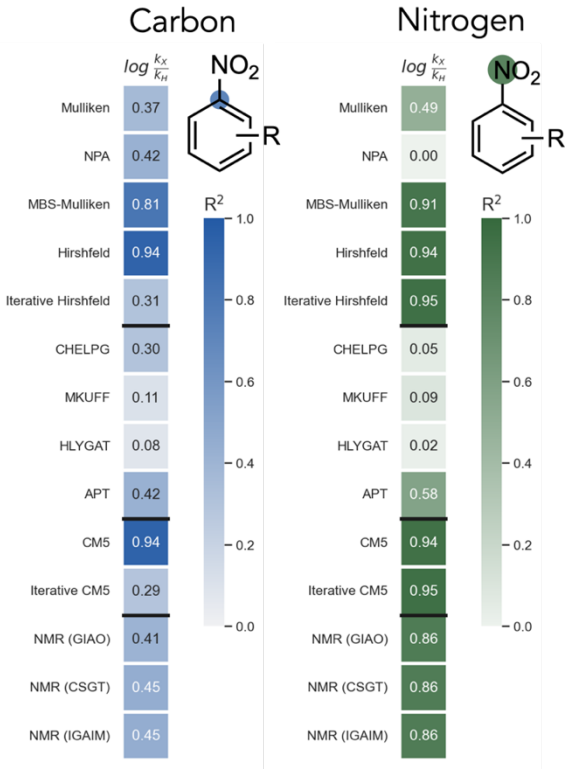


Figure 4.6. *Heatmap correlations with computed charge and NMR shift values at the carbon (shown in blue, left) and the nitro nitrogen (shown in green, right) against relative reaction rates.*

In this system, it is proposed that the cycloaddition is facilitated by excited state triplet nitroarenes. To account for this, we also modeled and computed charge values for nitroarene molecules in the triplet excited state in comparison to the singlet ground state, correlation heatmaps for the triplet state are provided in Appendix B. In general, we observe similar trends in improvement in correlations when using charge values at the nitro nitrogen, except for some

cases, specifically iterative Hirshfeld and iterative CM5, which do not perform well in the triplet state.

4.5: Conclusions

We describe comparisons between different partial atomic charge and NMR shift methods by evaluating univariate relationships with experimentally measured Hammett constants, used to explain the electronic effects of a substituent. We observe that in general, Class III charges do not correlate strongly with σ_m or σ_p values when charges are measured from atoms in or outside of the aryl ring. Charge and NMR shift correlations with σ_p remain relatively consistent if either the aryl C or the attached H value is used, however, correlations with σ_m seem to be more sensitive to the model and chosen atomic position. Hirshfeld and CM5 charge models yield strong correlations with σ_m and σ_p for both C and H values in both the simple initial dataset and the more complex nitroarene system. In studies for utilizing atomic charges to relate to experimental Hammett relationships, we suggest using charge values from outside the aromatic system for prediction. Even in more complex examples involving multiple substituents, and with a nitro group present, using an atom outside the aromatic system (the aryl H or nitro N, in our case) works best. We expect differences in predictive values for meta and para positions are less likely a through-bond effect causing differences in predictability, but more likely a through-space effect of the substituent. Finally, we observe ML predicted values of charge and NMR shifts compare well to their DFT counterparts. We see agreement in the correlation trends for the qmdesc predicted charges and NMR shift values, as well as with the CASCADE predicted NMR shifts.

4.6: References

1. Wells, P. R., *Chem. Rev.* **1963,** *63*, 171-219.
2. Hammett, L. P., *J. Am. Chem. Soc.* **1937,** *59*, 96-103.
3. Hansch, C.; Leo, A.; Taft, R., *Chem. Rev.* **1991,** *91*, 165-195.
4. (a) Charton, M., *J. Org. Chem.* **1963,** *28*, 3121-3124; (b) Ertl, P., *Chemistry-Methods* **2022,** *2*, e202200041; (c) Miranda-Quintana, R. A.; Deswal, N.; Roy, R. K., *Theor. Chem. Acc.* **2022,** *141*, 4.
5. (a) Gonthier, J. F.; Steinmann, S. N.; Wodrich, M. D.; Corminboeuf, C., *Chem. Soc. Rev.* **2012,** *41*, 4671; (b) Cho, M.; Sylvetsky, N.; Eshafi, S.; Santra, G.; Efremenko, I.; Martin, J. M., *ChemPhysChem* **2020,** *21*, 688-696.
6. (a) Wiberg, K. B.; Rablen, P. R., *J. Comput. Chem.* **1993,** *14*, 1504-1518; (b) Heidar-Zadeh, F.; Ayers, P. W.; Verstraelen, T.; Vinogradov, I.; Vöhringer-Martinez, E.; Bultinck, P., *J. Phys. Chem. A* **2017,** *122*, 4219-4245.
7. Storer, J. W.; Giesen, D. J.; Cramer, C. J.; Truhlar, D. G., *J. Comput. Aided Mol. Des.* **1995,** *9*, 87-110.
8. Mulliken, R. S., *J. Chem. Phys.* **1955,** *23*, 1833-1840.
9. Weinhold, F.; Landis, C. R., *Valency and bonding: a natural bond orbital donor-acceptor perspective*. Cambridge University Press: 2005.
10. (a) Hirshfeld, F. L., *Theor. Chim. Acta* **1977,** *44*, 129-138; (b) Hirshfeld, F., *Isr. J. Chem.* **1977,** *16*, 198-201.
11. Breneman, C. M.; Wiberg, K. B., *J. Comput. Chem.* **1990,** *11*, 361-373.
12. Besler, B. H.; Merz Jr, K. M.; Kollman, P. A., *J. Comput. Chem.* **1990,** *11*, 431-439.
13. Hu, H.; Lu, Z.; Yang, W., *J. Chem. Theory Comput.* **2007,** *3*, 1004-1013.
14. Cioslowski, J., *J. Am. Chem. Soc.* **1989,** *111*, 8333-8336.
15. Marenich, A. V.; Jerome, S. V.; Cramer, C. J.; Truhlar, D. G., *J. Chem. Theory Comput.* **2012,** *8*, 527-541.
16. Meister, J.; Schwarz, W., *J. Phys. Chem.* **1994,** *98*, 8245-8252.
17. Gaussian 16. Gaussian Inc.: Wallingford CT, 2016.
18. Ewing, D. F., Correlation of nmr Chemical Shifts with Hammett σ Values and Analogous Parameters. In *Correlation Analysis in Chemistry: Recent Advances*, Chapman, N. B.; Shorter, J., Eds. Springer US: Boston, MA, 1978; pp 357-396.
19. Erdmann, P.; Greb, L., *Angew. Chem.* **2022,** *134*, e202114550.
20. (a) Abraham, R. J.; Mobli, M., *Spectrosc. Eur.* **2004,** *16*, 16-22; (b) Schulman, E.; Christensen, K.; Grant, D. M.; Walling, C., *J. Org. Chem.* **1974,** *39*, 2686-2690.
21. Schreckenbach, G.; Ziegler, T., *J. Phys. Chem.* **1995,** *99*, 606-611.
22. Keith, T.; Bader, R., *Chem. Phys. Lett.* **1992,** *194*, 1-8.
23. Keith, T. A.; Bader, R. F., *Chem. Phys. Lett.* **1993,** *210*, 223-231.
24. Guan, Y.; Coley, C. W.; Wu, H.; Ranasinghe, D.; Heid, E.; Struble, T. J.; Pattanaik, L.; Green, W. H.; Jensen, K. F., *Chem. Sci.* **2021,** *12*, 2198-2208.
25. Guan, Y.; Shree Sowndarya, S. V.; Gallegos, L. C.; St. John, P. C.; Paton, R. S., *Chem. Sci.* **2021,** *12*, 12012-12026.
26. Christoph Bannwarth; Sebastian Ehlert; Stefan Grimme, *J. Chem. Theory Comput.* **2019,** *15*, 1652-1671.
27. Gaulton, A.; Hersey, A.; Nowotka, M.; Bento, A. P.; Chambers, J.; Mendez, D.; Mutowo, P.; Atkinson, F.; Bellis, L. J.; Cibrián-Uhalte, E.; Davies, M.; Dedman, N.; Karlsson, A.;

Magariños, M. P.; Overington, J. P.; Papadatos, G.; Smit, I.; Leach, A. R., *Nucleic Acids Res.* **2016,** *45*, D945-D954.
28. RDKit: Open-source cheminformatics.
29. Pracht, P.; Bohle, F.; Grimme, S., *Phys. Chem. Chem. Phys.* **2020,** *22*, 7169-7192.
30. Stephens, P. J.; Devlin, F. J.; Chabalowski, C. F.; Frisch, M. J., *J. Phys. Chem.* **1994,** *98*, 11623-11627.
31. Schäfer, A.; Huber, C.; Ahlrichs, R., *J. Chem. Phys.* **1994,** *100*, 5829-5835.
32. Grimme, S.; Ehrlich, S.; Goerigk, L., *J. Comput. Chem.* **2011,** *32*, 1456-1465.
33. Marenich, A. V.; Cramer, C. J.; Truhlar, D. G., *J. Phys. Chem. B* **2009,** *113*, 6378-6396.
34. Montgomery Jr, J. A.; Frisch, M. J.; Ochterski, J. W.; Petersson, G. A., *J. Chem. Phys.* **2000,** *112*, 6532-6542.
35. Vassetti, D.; Labat, F., *Int. J. Quantum Chem* **2021,** *121*, e26560.
36. Ruffoni, A.; Hampton, C.; Simonetti, M.; Leonori, D., *Nature* **2022,** *610*, 81-86.

CHAPTER 5: PYTHON TOOLS FOR CHEMISTS

5.1: Chapter Overview

Releasing open-source software is a convenient method to make the research performed in a study more reproducible and accessible to the public. In our lab, we have several Python packages that detail the collection of key molecular descriptors so that we and others can obtain parameters for research (see https://github.com/patonlab). I have contributed significantly to the development of three Python packages, detailed in this chapter. The first, GoodVibes, discussed in Section 5.2, presents work published in the open access journal *F1000 Research*. This work computes and applies corrections to reaction thermochemistry from quantum mechanical calculations. This program itself began development in 2016, before I began working on it in 2018. I contributed to the formatting and standardization of the package, as well as the automation of computing and graphing potential energy surfaces, which is showcased in the *Use Case* section of the manuscript, Section 5.2.5, which I worked on in collaboration with Dr. Juan V. Alegre-Requena.

Section 5.3 provides an overview of the DBSTEP Python package (pronounced “dubstep”). Inspired by the work presented in Chapter 3, the measurement of steric parameters can provide insight into which factors influence chemical reactivity. DBSTEP provides users with the tools to measure five different types of steric parameters, including Sterimol, percent buried volume, Sterimol2vec, vol2vec, and 2D graph-based steric parameters. This package is designed for use on the command line or in a Python script for users to obtain values in a high-throughput manner. Output visualization scripts are written for use in the molecular visualization program PyMOL to validate measurement and for use in publications. The programming work for this

program was performed by G. Luchini. Additional formatting, optimization and testing scripts were written by Tobin Patterson.

In Section 5.4, the program Py-X Struct is discussed. This Python program is used to query structures and geometric measurements (bond distances, angles and dihedrals) from experimentally determined X-ray crystal structures from the Cambridge Crystallographic Data Centre's Cambridge Structural Database. This program was published alongside a study showing use of the program in studying conformational behavior of diarylureas and diarylthioureas. This script was written by G. Luchini.

5.2: GoodVibes: Computing and Applying Corrections to Thermochemical Data

5.2.1: Introduction

Quantum chemistry software packages implement various levels of theory and basis sets, so-called model chemistries, that can be used to optimize molecular geometries and compute ground state vibrational modes. Statistical mechanical expressions using these vibrational frequencies, along with other contributing terms to the partition function, are used to obtain the enthalpy, entropy and Gibbs energy values required to understand, validate or predict experimental observations. GoodVibes was developed to address several challenges faced by practitioners of computational thermochemistry.

Firstly, the rigid-rotor harmonic oscillator (RRHO) model is routinely used to obtain vibrational entropic contributions and is the default for most electronic structure packages. However, the harmonic approximation fails to accurately describe low frequency modes and alternative models may be more appropriate.[1] Corrections to the RRHO model may be theoretically desirable, but the absence of practical and accessible tools to implement such corrections has limited their widespread adoption. Here we present and detail the use of the program GoodVibes, a Python-based project used for obtaining thermochemical values while applying corrections using quasi-harmonic approximations, alongside other corrections relevant

to the overestimation of the zero-point energy, multi-conformer ensembles, and standard concentrations.

Secondly, computational chemistry projects often combine results from different software. For example, geometries may be optimized with one program and the energies evaluated with another. Complex spreadsheets are frequently used to process these results and to prepare figures and manuscripts. However, errors in spreadsheets are commonplace and may be hard to detect. For this reason, GoodVibes allows the combination of data from multiple program outputs as part of the same project. Our group has utilized this program in informatics and organic mechanistic studies,[2] however, GoodVibes is not limited to a specific area of chemistry and has been used in published studies by more than 30 different research groups. As an example, previous studies in organic catalysis and mechanism, photocatalysis, and inorganic structure characterization have made use of this program package to process and make corrections to thermochemical data.[3] Figure 5.1 details an overview of the GoodVibes workflow from input data to the various outputs. Input options supplied via the command line enable chemists with basic experience with Python to process a large number of computational output files and to generate publication-ready data. These data (e.g. figures, tables) can be quickly reproduced by any other user with access to the raw data and the GoodVibes code. With recent and coming changes to standards of supplying full computational outputs through open repositories (such as Zenodo and ioChem-BD[4]) alongside publications, GoodVibes allows for complete transparency in how reported thermochemical values were obtained.

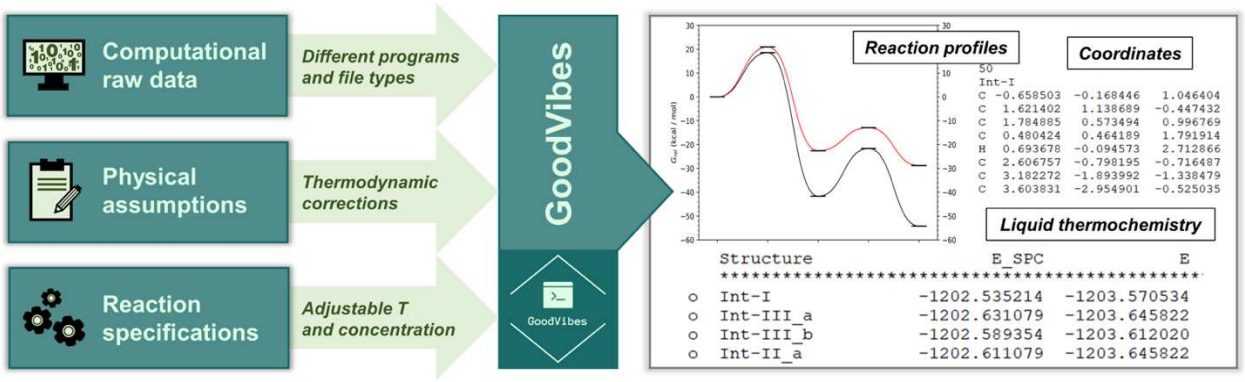


Figure 5.1. *Overview of GoodVibes*

5.2.2: Methods

By recomputing translational, rotational, vibrational and electronic partition functions from the data generated by quantum chemistry programs (such as vibrational frequencies, molecular mass, etc.), thermochemistry can be calculated in GoodVibes at any specified temperature or concentration/pressure. By default, these are set to 298.15 K and 1 atmosphere. A notable automated correction to the RRHO model is applied to low frequency vibrational modes. These low frequency modes (typically less than 100 cm-1) are not well approximated as harmonic and their entropy contributions tend to be overestimated. Methods to avoid this include nontrivial hindered- rotor calculations and computationally expensive anharmonic calculations, both of which become more infeasible with higher atom counts.[1a, 1b] Both Cramer/Truhlar[1c] and Grimme[1d] have proposed simple, more widely adopted corrections, so-called quasi-harmonic approximations, formulated specifically to obtain vibrational entropies for these low frequency modes. GoodVibes will also automatically apply empirical scaling factors to computed frequencies. These corrections arise from the tendency of electronic structure calculations (e.g. with density functional theory) to overestimate vibrational frequencies relative to experiment, and hence zero-point energies. Linear scaling factors have been collated for a number of functional and basis set combinations for sets of small organic molecules. GoodVibes accesses the scaling

factors compiled by the Truhlar group across several studies,[5] automatically detects the level of theory and basis set and scales the frequencies if there is a match.

Single point energy calculations performed at more expensive levels of theory and with larger basis sets are commonly used in combination with the thermal corrections obtained from separate calculations, often using different software packages.[6] A multitude of output files from different program packages along with correction applied by GoodVibes can be combined to construct a potential energy surface by using an easily interpretable YAML file that defines the elementary steps of the reaction, file definitions and formatting options.

5.2.3: Implementation

GoodVibes is a module implemented in Python, currently supported by 2.6-7 and all 3.x versions. The module requires the NumPy library (version 1.14.2 or greater), with optional importing of Matplotlib (version 2.2.4 or greater) to graph potential energy surfaces. To test the accuracy and upkeep of this coding package, a series of tests have been implemented in TravisCI, checking functionality and accuracy of the code when the master branch is updated on GitHub against target Python versions on Linux, macOS and Windows operating systems.

5.2.4: Operation

GoodVibes is compatible with Windows 10, Linux and macOS operating systems. This software is appropriate for use with output files produced by a wide range of calculation types, including density functional theory, wave function theory, molecular mechanics, COSMO-RS solvation calculations, and semi-empirical methods. Current supported programs include Gaussian 09,[7] Gaussian 16,[8] ORCA 4[9] single point energy calculation files, and COSMOtherm[10] COSMO-RS solvation free energy output files. GoodVibes provides an output file with tabulated thermochemical data, optionally exported as a CSV file. Cartesian coordinates of processed files can also optionally be exported in an XYZ file. Plots of energy profiles are optionally generated.

5.2.5: Use Case

In this section we show how GoodVibes can be used with a variety of input options and files to transform heterogeneous computational chemistry data into human-readable tabulated values and figures. In this example, 50 calculation output files (25 Gaussian geometry optimizations and vibrational frequency calculations with 25 corresponding ORCA single point calculations) are used to create the data in Table 5.1 and graphed in Figure 5.2. All raw data was taken from a 2018 study,[11] and is freely accessible through a Zenodo repository.[12] Each point in Figure 5.2 represents a unique conformer's Gibbs energy. The Boltzmann-weighted values are shown as dashes connected by the curved profiles. Optimizations were done with ωB97X-D/6-31+G(d)[13] implemented in Gaussian using an "ultrafine" pruned (99,590) integration grid. Considering solvent effects of the reaction, calculations were run with the SMD solvation model[14] using ethanol, and the concentration of each substance was set to 1.0 M for further thermochemical calculation. GoodVibes allows for solvent media corrections to entropy based on select solvent standard state concentration,[15] which was applied to ethanol in this case.

Thermochemistry is evaluated at a temperature of 80°C (353.15 K) in accordance with experimental conditions (the temperature assumed in the original calculations does not influence the results from GoodVibes). These values are corrected using the quasi-harmonic approximation proposed by Grimme. One conformer of intermediate II in the 'Py' pathway has a small imaginary frequency of 2.9 cm-1, which was "inverted" to a real value 2.9 cm-1, as done in previous works to small imaginary frequencies, typically under >i50 cm-1.[11, 16]

Separate single point energies are extracted from ORCA calculations performed at a coupled cluster level of theory with a (DZ/TZ) basis set extrapolation (DLPNO-CCSD(T), cc-pVDZ/cc-pVTZ),[17] then used for calculations and added to the GoodVibes output for comparison. A potential energy surface is constructed by using Boltzmann weighted averaging of all conformers at each step in the pathway. A multi-structural correction is then applied to the

resulting Gibbs free energy based on the number and energy of distinguishable conformers present for each species.[18] The Gibbs energy profile is constructed from options specified in the YAML file containing the reaction pathway steps, file definitions and plot formatting options.

Table 5.1. *Tabulated relative Boltzmann weighted thermochemical values (shown in kcal·mol-1) from the "Ph" and "Py" pathways. Including: single point calculations (ΔE_SPC), energy (ΔE), zero-point energy (ΔZPE), Gibbs free energy (ΔG) and quasi-harmonic corrected Gibbs free energy (qh-ΔG).*

Ph pathway	ΔE_SPC	ΔE	ΔZPE	ΔG	qh-ΔG
Ph-Int-I	0.00	0.00	0.00	0.00	0.00
Ph-TS-I	23.56	23.54	-1.55	21.29	22.31
Ph-Int-II	-12.29	-14.14	-0.57	-13.49	-12.86
Ph-TS-II	-8.13	-12.17	-1.34	-9.97	-9.35
Ph-Int-III	-33.59	-37.61	-0.57	-36.61	-35.11
Py pathway	**ΔE_SPC**	**ΔE**	**ΔZPE**	**ΔG**	**qh-ΔG**
Py-Int-I	0.00	0.00	0.00	0.00	0.00
Py-TS-I	15.08	16.52	-1.20	14.27	14.29
Py-Int-II	-17.45	-18.19	-0.55	-20.97	-19.21
Py-TS-II	-9.59	-13.17	-0.97	-10.96	-10.39
Py-Int-III	-29.39	-33.24	-0.70	-33.04	-31.35

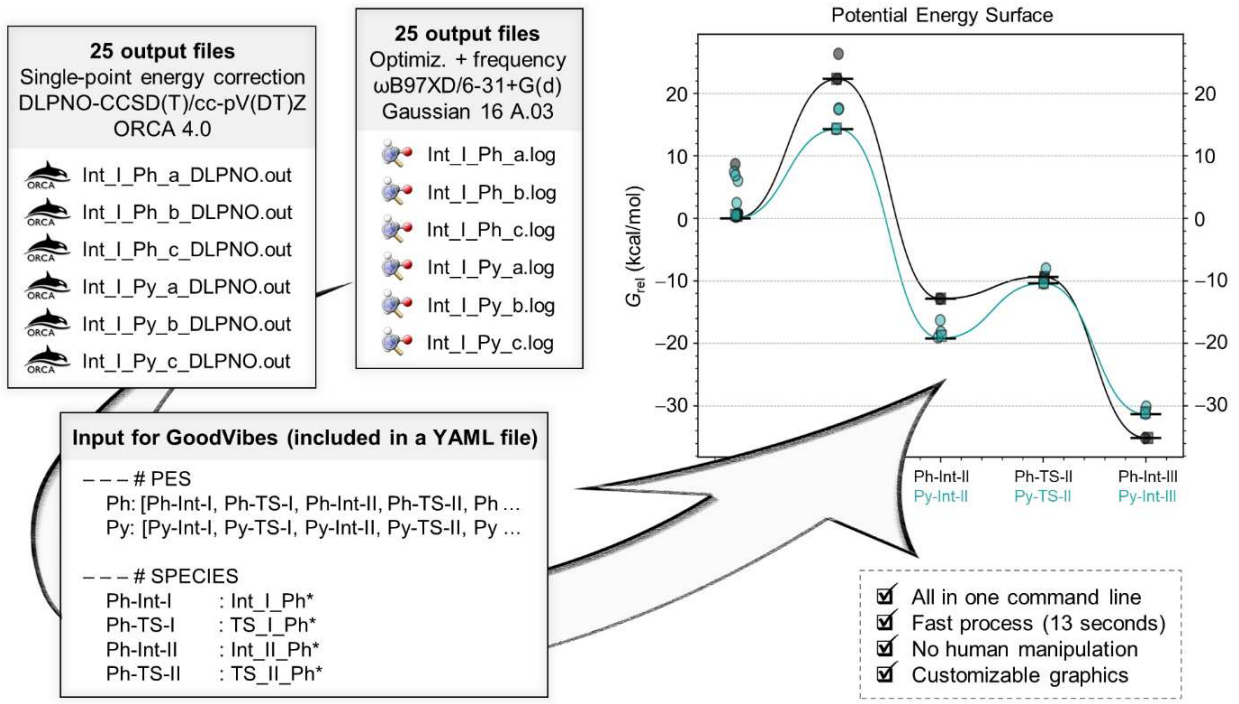


Figure 5.2. *A reaction Gibbs energy profile is produced directly from the command above and saved to PNG file. Here, two reaction pathways, "Ph" and "Py", are displayed.*

All of the output files, correction and formatting options are supplied to GoodVibes to output tabulated data and a graph of the reaction pathway from a single command:

```
python -m goodvibes *.log --spc DLPNO --pes science.yaml -graph science.yaml -t
353.15 --imag --invertifreq -5 --media ethanol -c 1
```

Additional usage examples are described at GoodVibes GitHub repository, where several features have been added in response to requests from the community of users.

5.2.6: Conclusion

GoodVibes is a Python-based tool that calculates thermochemical data from quantum mechanical calculations in a transparent and reproducible way. GoodVibes may be employed with any type of chemical structure, including organic and inorganic molecules of varying sizes as well as with single point calculations performed by differing programs. Additionally, GoodVibes contains many additional automated features that are designed to save time for researchers, allowing for the calculation of thermochemical data at any temperature or concentration, incorporating valuable and overlooked corrections to the RRHO model through quasi-harmonic and vibrational scaling

factor corrections and construction of potential energy surfaces with applied corrections accounting for the accessibility of multiple conformations. For projects involving the analysis of a large number of computational chemistry output files, GoodVibes helps to prevent human errors associated with spreadsheets, and can be used to reproduce any table or figure from the raw data.

5.2.6: Data Availability

Zenodo: Data Supporting GoodVibes: Automating and applying thermodynamic corrections to harmonic frequency calculations, https://doi.org/10.5281/zenodo.366284515.

This project contains data referenced in the use case.

Data are available under the terms of the Creative Commons Attribution 4.0 International license (CC-BY 4.0).

5.2.7: Software Availability

This code has been made accessible for chemists of various levels of computational experience and is easily installed. The most recent version of our open-sourced Python package GoodVibes v3.0.1 is freely available on GitHub at https://github.com/bobbypaton/GoodVibes. GoodVibes may be installed as a Python module from the command line using either PyPI (https://pypi.org/project/goodvibes/) or Conda (https://anaconda.org/patonlab/ goodvibes) using the commands:

```
pip install goodvibes
```

or

```
conda install -c conda-forge goodvibes
```

or, by downloading the repository from GitHub and running the following command from the extracted directory:

```
python setup.py install
```

Archived source code at the time of publication: https://doi.org/10.5281/zenodo.595246

License: MIT

5.3: DBSTEP: DFT-Based Steric Parameters

The steric bulk of a molecule can be an important factor in influencing reaction outcome. In computational studies, the measurement of steric parameters can provide quantitative insight for how the size and shape of molecules influence reaction outcome. Accompanying the work in Chapter 3, the Python package DBSTEP: DFT-Based STEric Parameters was released. Officially made available to the public in November 2020, DBSTEP is freely available to download through GitHub (https://github.com/patonlab/DBSTEP) or easily installed from the command line with pip or conda commands.[19] The DBSTEP package was designed to showcase our novel parameters Sterimol2vec and vol2vec, however the original versions of these parameters, Sterimol[20] and percent buried volume ($\%V_{bur}$),[21] are also available, along with two dimensional graph-based steric parameters.

At the time of the public release of DBSTEP, the only existing method to obtain the $\%V_{bur}$ parameter was through the SambVca web application,[22] made available from the Cavallo group, who originally developed the parameter. Since Sambvca is a web application and needs user interaction to process molecules one at a time, it is not ideal for high throughput studies, where hundreds to thousands of measurements need to be collected. DBSTEP allows for command line and in-script collection of steric parameters, making the collection of steric parameters easily integrated into a computational workflow. Today, there are a few methods to obtain $\%V_{bur}$ from freely available packages outside our own, including Morfeus[23] and SEQCROW.[24]

The DBSTEP repository is also a convenient resource for collecting Sterimol parameters. The original method to obtain Sterimol parameters was from FORTRAN 77 code, which utilized atomic coordinates and van der Waals radii. Our lab has previous, now-retired repository, “Sterimol,” that translated this code into Python to make it more accessible to a broader audience. With the release of “wSterimol” from the Paton lab in 2019,[25] it was possible to obtain Sterimol parameters from an interactive PyMOL plugin, which aided in visualization of how parameters

change across the conformational space of a molecule.[26] In DBSTEP, we chose to update the radii to Bondi van der Waals radii, which are defined for all main group elements.[27] For all other elements, an atomic radius of 2.0Å is used, consistent with similar chemistry software. The Sterimol parameter set is also now available for collection in the Morfeus and SEQCROW packages.

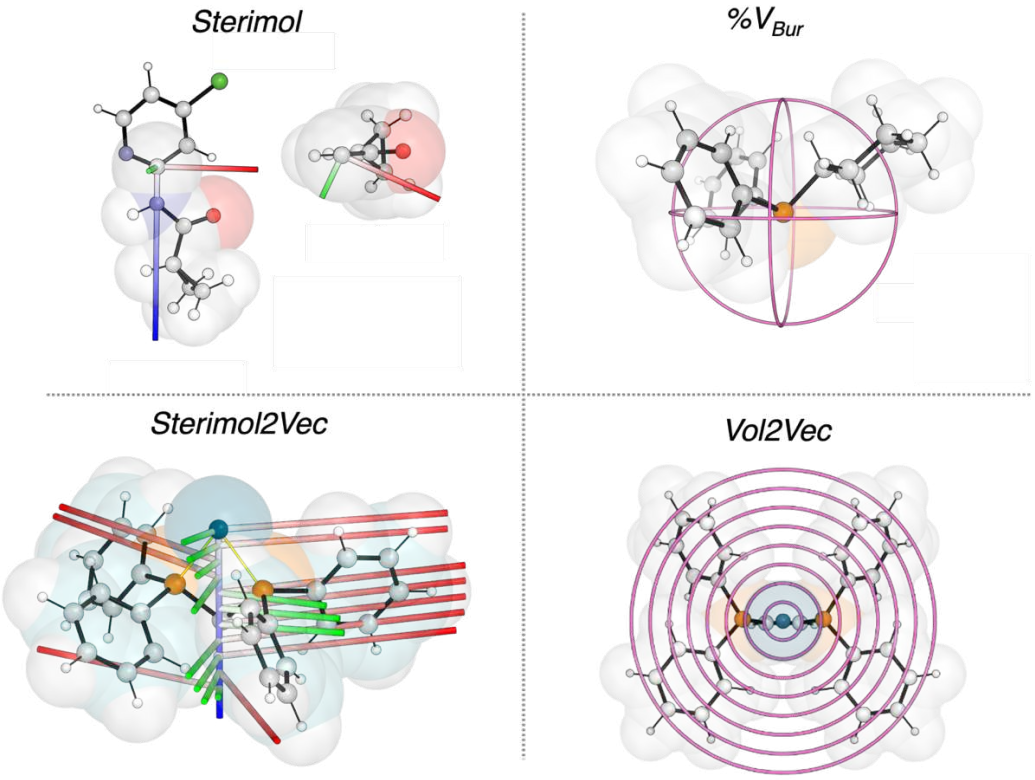


Figure 5.3. *PyMOL visualizations of measured parameters produced by DBSTEP calculations.*

DBSTEP allows for visualizations of computed parameters, shown in Figure 5.3, so that a researcher can inspect the property values overlaid with the molecule the parameter was measured with, but we have also found use for these visualizations in publications. For Sterimol and Sterimol2vec measurements, we store direction and magnitude information so that the B_{max} (red) and B_{min} (green) measurements can be overlaid with the molecule, displaying where on the molecule the maximum and minimum widths are, along with the length vector, L (blue). For $\%V_{bur}$ measurements, we overlay a wire frame with the specified measurement radius over the atomic

center chosen for measurement. Any portion of the molecule contained in the wire frame is considered for the buried volume measurements. For vol2vec visualizations, this information is represented in concentric circles centered at the chosen atom, instead of full wire frames for visualization purposes.

The DBSTEP repository has garnered a few citations since its initial release, at the time of writing this document, our citable Zenodo repository[19] has reached six citations. Of these citations, five come from outside our lab. Five of these citations use DBSTEP for the collection of $\%V_{bur}$ values, and one utilizes our Sterimol2vec parameter set. The publication utilizing Sterimol2vec comes from the Anslyn and Sigman labs, who found use in measuring proximal and distal B_{max} values for a collection of sensor molecules for circular dichroism spectroscopy in order to optimize and maximize the signal achieved in experimental measurements.[28] A resulting multivariate linear regression model utilized B_{max} values at 2.0Å and 8.5Å away from substituted pyridine reference point, citing that larger proximal sterics and smaller distal sterics were influential in maximizing the circular dichroism signal.

While main program development is largely finished, there is always room for further development in code optimization and parallelization, and bug fixes and maintenance when other package requirements update. DBSTEP measures which points in space are occupied by the molecule from a three-dimensional point cloud in space, with even grid spacing. The default grid spacing used is 0.05Å, which can be altered to smaller or larger values with an input argument (--grid). A major bottleneck on the measurement of Sterimol and volumetric properties is the iteration of points through this grid, deciding which points are occupied by the molecule and which ones are vacant in space. Currently we utilize functions from external software, scipy,[29] to perform these iterations, but we have been considering more efficient ways to measure volumetric contributions of overlapping spheres in space.

Overall, the DBSTEP program provides the means for a computational chemist to measure commonly used steric parameters for their own research. Documentation in the form of a ReadMe document is found on our GitHub page, detailing use of the program and providing examples for new users. With the release of a formal manuscript, we hope that adoption of our vectorized parameter sets, Sterimol2vec and vol2vec, become more well-known and find use in studies of systems where information on steric proximity is influential in the chemical reactivity.

5.4: Py-X Struct: Mining the Cambridge Structural Database for Geometric Data from Crystal Structures

Py-X Struct is a Python package designed to mine geometric information crystallographic X-Ray structure data from the Cambridge Structural Database (CSD).[2c,30] This program is dependent on the Cambridge Crystallographic Data Centre library, making use of the Python API to access queried structures for geometric data including bond distances, angles and dihedrals. The CSD contains over a million experimental crystal structures, making it desirable for large data analyses of molecules in unique environments. With this database, along with Py-X Struct, it is possible to study conformational behavior of molecules, querying and analyzing how molecules orient themselves in crystallization conditions.

The Py-X Struct program was designed for command line use and uses SMILES strings as input. SMILES[31] are unique ways of representing molecular connectivity in a text-based format, for example, the SMILES string “CCO” represents the structure for ethanol (two aliphatic carbons bonded together, with the second forming a single bond with oxygen). With SMILES strings, hydrogens are often implied, although they can be specified. A user of this program can query the CSD for any valid SMILES, and requested geometric measurements, and the program will return hits in a .CSV file format, returning the identifier for any structure that matches the SMILES query in the CSD, along with the requested geometric measurement.

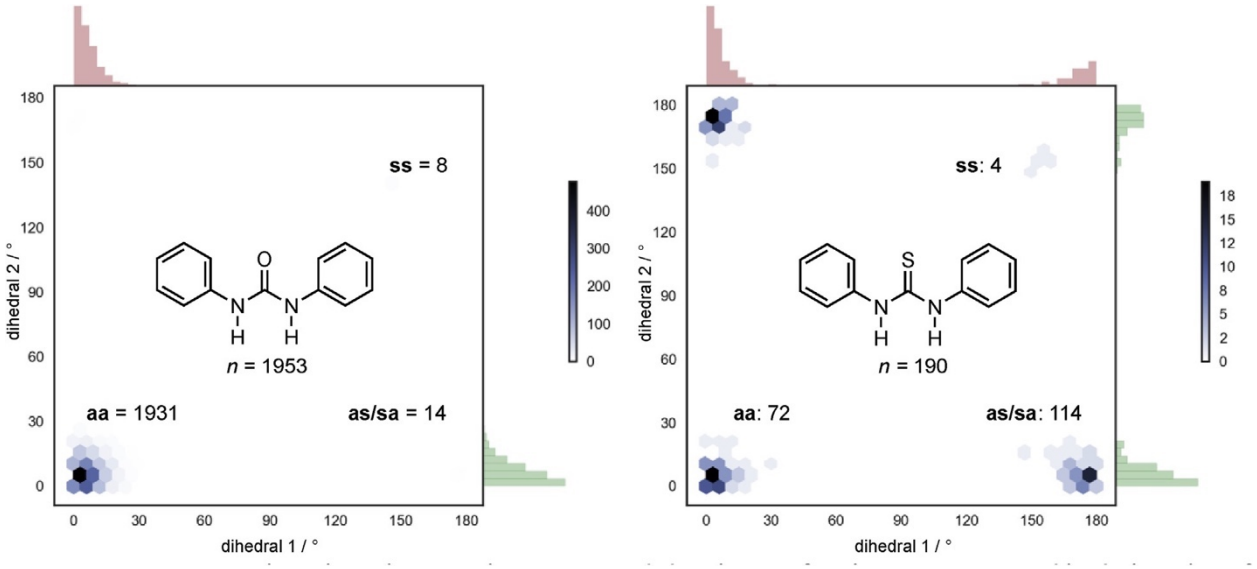


Figure 5.4. *Bivariate hexbin plots showing the counts and distribution for the two C-N-C=X dihedral angles of diphenylureas (LHS) and diphenylthioureas (RHS) found in the Cambridge Structural Database (CSD, version 5.39). Univariate histograms are also shown for each angle along the axes.*

This work was originally published alongside an article in 2018, in which we utilized Py-X Struct for collecting torsional information for ureas and thioureas. The geometric data, summarized by distribution plots in Figure 5.4, provided insight into the preferred conformations of diarylureas and diarylthioureas. When querying X=C-N-C dihedral angles (X=O for ureas, S for thioureas), we observed from crystal structures that ureas strongly prefer the anti-anti-conformation, shown by the larger cluster of datapoints in the bottom left corner of the left graph on Figure 5.4. However, diarylthioureas show a mixture of anti-anti- and anti-syn-conformers. Density functional theory calculations and noncovalent interaction analyses were also utilized to show that a preference for anti-syn-conformers in diarylthioureas was made possible to avoid steric clash between the bulkier sulfur atom and the aryl ring.

The Py-X Struct program can also return identified structures that match the SMILES query without any geometric measurements, serving as a quick method to search for a particular structure or crystal structures containing a specific structural motif. The program does require a valid CSD license, perhaps a drawback for researchers with limited resources. Another potential

drawback of this program is that at the time of release in 2018, the CSD Python API required Python version 2.7, so this program has a similar requirement. The API has since been updated to be able to account for more recent Python versions, however this program remains written for the deprecated 2.7 version.

5.5: References

1. (a) Ayala, P. Y.; Schlegel, H. B., *J. Chem. Phys.* **1998,** *108*, 2314-2325; (b) McClurg, R. B.; Flagan, R. C.; Goddard Iii, W. A., *J. Chem. Phys.* **1997,** *106*, 6675-6680; (c) Ribeiro, R. F.; Marenich, A. V.; Cramer, C. J.; Truhlar, D. G., *J. Phys. Chem. B* **2011,** *115*, 14556-14562; (d) Grimme, S., *Chemistry* **2012,** *18*, 9955-64.
2. (a) Liu, B.; Alegre-Requena, J. V.; Paton, R. S.; Miyake, G. M., *Chem. Eur. J.* **2020,** *26*, 2386-2394; (b) Koniarczyk, J. L.; Greenwood, J. W.; Alegre-Requena, J. V.; Paton, R. S.; McNally, A., *Angew. Chem.* **2019,** *131*, 15024-15028; (c) Luchini, G.; Ascough, D. M. H.; Alegre-Requena, J. V.; Gouverneur, V.; Paton, R. S., *Tetrahedron* **2019,** *75*, 697-702.
3. (a) Lewis, R. D.; Garcia-Borràs, M.; Chalkley, M. J.; Buller, A. R.; Houk, K.; Kan, S. J.; Arnold, F. H., *Proc. Natl. Acad. Sci.* **2018,** *115*, 7308-7313; (b) Svatunek, D.; Houszka, N.; Hamlin, T. A.; Bickelhaupt, F. M.; Mikula, H., *Chemistry–A European Journal* **2019,** *25*, 754-758; (c) Gomes, G. d. P.; Loginova, Y.; Vatsadze, S. Z.; Alabugin, I. V., *J. Am. Chem. Soc.* **2018,** *140*, 14272-14288; (d) Wodrich, M. D.; Busch, M.; Corminboeuf, C., *Helv. Chim. Acta* **2018,** *101*, e1800107; (e) Ye, J.; Kalvet, I.; Schoenebeck, F.; Rovis, T., *Nat. Chem.* **2018,** *10*, 1037-1041; (f) Grayson, M. N., *J. Org. Chem.* **2017,** *82*, 4396-4401; (g) Besora, M.; Vidossich, P.; Lledos, A.; Ujaque, G.; Maseras, F., *J. Phys. Chem. A* **2018,** *122*, 1392-1399.
4. Álvarez-Moreno, M.; de Graaf, C.; Lopez, N.; Maseras, F.; Poblet, J. M.; Bo, C., *J. Chem. Inf. Model.* **2015,** *55*, 95-103.
5. Alecu, I. M.; Zheng, J.; Zhao, Y.; Truhlar, D. G., *J. Chem. Theory Comput.* **2010,** *6*, 2872-2887.
6. Feller, D.; Peterson, K. A.; Grant Hill, J., *J. Chem. Phys.* **2011,** *135*, 044102.
7. Gaussian 09. Gaussian Inc.: Wallingford CT, 2016.
8. Inc., G., Gaussian 16. Wallingford CT, 2016.
9. Neese, F., *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2018,** *8*, e1327.
10. Klamt, A., *J. Phys. Chem.* **1995,** *99*, 2224-2235.
11. Hilton, M. C.; Zhang, X.; Boyle, B. T.; Alegre-Requena, J. V.; Paton, R. S.; McNally, A., *Science* **2018,** *362*, 799-804.
12. Luchini, G.; Alegre-Requena, J.; Paton, R., Data Supporting GoodVibes: Automating and applying thermodynamic corrections to harmonic frequency calculations (Version 1.0).[Data set]. Zenodo. 2020.
13. (a) Becke, A. D., *J. Chem. Phys.* **1997,** *107*, 8554-8560; (b) Jeng-Da Chai; Martin Head-Gordon, *Phys. Chem. Chem. Phys.* **2008,** *10*, 6615; (c) Hehre, W. J.; Ditchfield, R.; Pople, J. A., *J. Chem. Phys.* **1972,** *56*, 2257-2261; (d) Hariharan, P. C.; Pople, J. A., *Theor. Chim. Acta* **1973,** *28*, 213-222; (e) Krishnan, R.; Binkley, J. S.; Seeger, R.; Pople, J. A., *J. Chem. Phys.* **1980,** *72*, 650-654; (f) McLean, A.; Chandler, G., *J. Chem. Phys.* **1980,** *72*, 5639-5648; (g) Francl, M. M.; Pietro, W. J.; Hehre, W. J.; Binkley, J. S.; Gordon, M. S.; DeFrees, D. J.; Pople, J. A., *J. Chem. Phys.* **1982,** *77*, 3654-3665.
14. Marenich, A. V.; Cramer, C. J.; Truhlar, D. G., *J. Phys. Chem. B* **2009,** *113*, 6378-6396.
15. Harvey, J. N.; Himo, F.; Maseras, F.; Perrin, L., *ACS Catal.* **2019,** *9*, 6803-6813.
16. (a) Sure, R.; Grimme, S., *J. Chem. Theory Comput.* **2015,** *11*, 3785-3801; (b) Liu, Z.; Patel, C.; Harvey, J. N.; Sunoj, R. B., *Phys. Chem. Chem. Phys.* **2017,** *19*, 30647-30657.
17. (a) Purvis III, G. D.; Bartlett, R. J., *J. Chem. Phys.* **1982,** *76*, 1910-1918; (b) Pople, J. A.; Head-Gordon, M.; Raghavachari, K., *J. Chem. Phys.* **1987,** *87*, 5968-5975; (c) Riplinger, C.; Neese, F., *J. Chem. Phys.* **2013,** *138*, 034106; (d) Riplinger, C.; Sandhoefer, B.;

Hansen, A.; Neese, F., *J. Chem. Phys.* **2013,** *139*, 134101; (e) Riplinger, C.; Pinski, P.; Becker, U.; Valeev, E. F.; Neese, F., *J. Chem. Phys.* **2016,** *144*, 024109.
18. Plata, R. E.; Singleton, D. A., *J. Am. Chem. Soc.* **2015,** *137*, 3811-3826.
19. Luchini, G.; Patterson, T.; Paton, R. S., DBSTEP: DFT-Based Steric Parameters. 2021.
20. Verloop, A., Drug Design. Ariens, E. J., Ed. Academic Press: New York,, 1976; Vol. III.
21. Hillier, A. C.; Sommer, W. J.; Yong, B. S.; Petersen, J. L.; Cavallo, L.; Nolan, S. P., *Organometallics* **2003,** *22*, 4322-4326.
22. (a) Poater, A.; Cosenza, B.; Correa, A.; Giudice, S.; Ragone, F.; Scarano, V.; Cavallo, L., *Eur. J. Inorg. Chem.* **2009,** *2009*, 1759-1766; (b) Falivene, L.; Credendino, R.; Poater, A.; Petta, A.; Serra, L.; Oliva, R.; Scarano, V.; Cavallo, L., *Organometallics* **2016,** *35*, 2286-2293.
23. Jorner, K., Morfeus: Molecular Features for Machine Learning. 2021.
24. Schaefer, A. J.; Ingman, V. M.; Wheeler, S. E., *J. Comput. Chem.* **2021,** *42*, 1750-1754.
25. Brethomé, A. V.; Fletcher, S. P.; Paton, R. S., *ACS Catal.* **2019,** *9*, 2313-2323.
26. The PyMOL Molecular Graphics System. Version 2.4 ed.; Schrödinger, LLC.
27. Mantina, M.; Chamberlin, A. C.; Valero, R.; Cramer, C. J.; Truhlar, D. G., *J. Phys. Chem. A* **2009,** *113*, 5806-5812.
28. Dotson, J. J.; Anslyn, E. V.; Sigman, M. S., *J. Am. Chem. Soc.* **2021,** *143*, 19187-19198.
29. Virtanen, P.; Gommers, R.; Oliphant, T. E.; Haberland, M.; Reddy, T.; Cournapeau, D.; Burovski, E.; Peterson, P.; Weckesser, W.; Bright, J.; van der Walt, S. J.; Brett, M.; Wilson, J.; Millman, K. J.; Mayorov, N.; Nelson, A. R. J.; Jones, E.; Kern, R.; Larson, E.; Carey, C. J.; Polat, İ.; Feng, Y.; Moore, E. W.; VanderPlas, J.; Laxalde, D.; Perktold, J.; Cimrman, R.; Henriksen, I.; Quintero, E. A.; Harris, C. R.; Archibald, A. M.; Ribeiro, A. H.; Pedregosa, F.; van Mulbregt, P.; Vijaykumar, A.; Bardelli, A. P.; Rothberg, A.; Hilboll, A.; Kloeckner, A.; Scopatz, A.; Lee, A.; Rokem, A.; Woods, C. N.; Fulton, C.; Masson, C.; Häggström, C.; Fitzgerald, C.; Nicholson, D. A.; Hagen, D. R.; Pasechnik, D. V.; Olivetti, E.; Martin, E.; Wieser, E.; Silva, F.; Lenders, F.; Wilhelm, F.; Young, G.; Price, G. A.; Ingold, G.-L.; Allen, G. E.; Lee, G. R.; Audren, H.; Probst, I.; Dietrich, J. P.; Silterra, J.; Webber, J. T.; Slavič, J.; Nothman, J.; Buchner, J.; Kulick, J.; Schönberger, J. L.; de Miranda Cardoso, J. V.; Reimer, J.; Harrington, J.; Rodríguez, J. L. C.; Nunez-Iglesias, J.; Kuczynski, J.; Tritz, K.; Thoma, M.; Newville, M.; Kümmerer, M.; Bolingbroke, M.; Tartre, M.; Pak, M.; Smith, N. J.; Nowaczyk, N.; Shebanov, N.; Pavlyk, O.; Brodtkorb, P. A.; Lee, P.; McGibbon, R. T.; Feldbauer, R.; Lewis, S.; Tygier, S.; Sievert, S.; Vigna, S.; Peterson, S.; More, S.; Pudlik, T.; Oshima, T.; Pingel, T. J.; Robitaille, T. P.; Spura, T.; Jones, T. R.; Cera, T.; Leslie, T.; Zito, T.; Krauss, T.; Upadhyay, U.; Halchenko, Y. O.; Vázquez-Baeza, Y.; SciPy, C., *Nature Methods* **2020,** *17*, 261-272.
30. Groom, C. R.; Bruno, I. J.; Lightfoot, M. P.; Ward, S. C., *Acta Crystallographica Section B* **2016,** *72*, 171-179.
31. Weininger, D., *J. Chem. Inf. Comput. Sci.* **1988,** *28*, 31-36.

CHAPTER 6: CONCLUSIONS AND FUTURE DIRECTIONS

Computational molecular representations and the properties that can be measured from them are important in reaction optimization and prediction. When strong relationships are found, scientists can further understand the chemical processes occurring when a reaction is run. Properly describing the chemical system can provide insight as to which properties influence reactivity, such as reaction yield, selectivity, or rate. It may be challenging to describe a chemical system with to tools and descriptors available, which can drive the development of new chemical descriptors. This work summarizes the current field of molecular descriptors (Chapter 2), while discussing use of specific steric (Chapter 3) and electronic (Chapter 4) molecular properties. The development of open-source tools produced in these works is discussed in Chapter 5.

Chapter 2 provides an overview of methods for chemists to computationally represent molecules and their properties. The first half of the chapter details molecular representations, from “one” to “four” dimensions, discussing the different molecular properties that can be obtained from each representation. Simple 1D and 2D representations provide simple features, such as count-descriptors or connectivity information, while more complex 3D representations and 4D conformational ensembles can provide multi-variable parameter distributions for molecules and their conformational ensembles. With quantum mechanical methods, accurate geometries and energies can be computed for molecules, and further property calculations can proceed.

The second half of Chapter 2 discusses specific computational molecular descriptors designed for use in catalysis studies. Mainly obtained from DFT-optimized structures, steric and electronic properties can be collected and related to appropriate chemical systems. In catalysis, the transition structure can often provide powerful insight for how a reaction proceeds. Features collected from transition structures have been shown to be useful in predictive models.

Additionally, considering a molecule's conformational space can be important. For a flexible molecule, multiple low-energy structures may be found in a reaction flask. To account for multiple conformations, Boltzmann weights can be computed from optimized structures, and Boltzmann averaged properties can be useful in describing the individual contributions of conformers in the ensemble. It has also been shown that key conformations of molecules, not necessarily the lowest energy structure, can contribute to reactivity. Collecting properties from the conformational ensemble allows a molecule's properties to be described as a distribution, with a minimum and maximum value.

In Chapter 3, steric parameters are discussed. An identified gap in the literature, the proximity of steric bulk in relation to the chemical center of interest, is addressed with two new parameter sets, Sterimol2vec and vol2vec. The Sterimol2vec parameter set is an extension of the existing Sterimol parameter set, taking new minimum (B_{min}) and maximum (B_{max}) width measurements in set intervals along the length (L) axis for a molecule. This provides information for how the shape of the molecule changes as it ranges down the bond axis chosen as reference. An example showcasing the use of this parameter in the prediction of atropisomer rotational barriers shows how pinpointing proximal information obtained by this parameter can better describe our chemical system than Sterimol parameters alone. The vol2vec parameter set is an extension of the $\%V_{bur}$ descriptor, which measures how much a molecule occupies a sphere, typically with a radius of 3.5Å, placed at an atomic center. The vol2vec set allows this radius to expand in set intervals, measuring the percent occupancy of each "shell" as the sphere expands radially. Two examples are shown, one comparing a vol2vec analysis to historical Taft parameters, resulting in proximal sterics having an important influence on reaction rate. Another visual analysis is shown comparing vol2vec measurements for phosphine ligands for a reaction where distal steric influence from the phosphine ligand is related to reactivity. A software package resulting from this work, DBSTEP, is further detailed in Chapter 5.

Chapter 4 discusses a popular computed electronic property, the partial atomic charge. Often used to relate to experimental Hammett parameters, these properties can be computed for atoms in molecules using semi-empirical or quantum mechanical methods. Many methods exist for computing the partial atomic charge, and this study compares how well different methods compare with the experimental Hammett parameters through univariate correlations. NMR shift also captures aspects of electronic substituent effects and were also computed and compared. The charge and NMR shift values computed at carbon and hydrogen in the para positions on aryl rings are compared with experimental σ_p values, and charge and NMR shift values for carbon and hydrogen at both meta positions were averaged and compared with experimental σ_m values. We observe that the computed hydrogen values are more consistent in their correlation for para and meta values across a range of methods, while the carbon values are less consistent between para and meta correlations. We show that even in more challenging systems, for conjugated nitroarenes with multiple substituents, this result holds, using atoms outside the aryl ring has more consistent correlations with experimental rate values. It would be interesting to study this system further, perhaps with larger datasets of molecules to test with or across a wider range of methods to compute partial atomic charges available outside the Gaussian 16 software and see if the trends observed in this study hold.

Chapter 5 provides an overview of three Python packages I have contributed significantly to. To make the works published here and elsewhere accessible and reproducible, we have several open-source packages. The development process involves coding, testing, bug-fixing, interaction with users, and documentation of program features and use. This iterative process was done for GoodVibes, DBSTEP, and Py-X Struct. The GoodVibes program provides computational chemists with tools to apply commonly overlooked corrections to thermochemical data obtained from quantum mechanical software. For example, applying frequency scaling factors for known overestimations of vibrational frequencies for specific level of theory

combinations, as well as quasi-harmonic entropic corrections, to correct for large contributions in entropy due to assumptions made quantum mechanical thermochemistry calculations. GoodVibes automates tasks like computing and graphing reaction thermochemistry and can do so at user-specified temperatures and concentrations. GoodVibes is still in active development, as a larger number of quantum mechanical software gains traction, it will be useful to account for and apply corrections to a wider range of software. Features like being able to import GoodVibes into a script so that thermochemistry could be computed in a more high-throughput way would also be desirable.

The DBSTEP program is designed for the collection of steric parameters, measured from atomic coordinates. This program can compute four main steric parameters from optimized structures, including Sterimol parameters, $\%V_{bur}$, Sterimol2vec and vol2vec parameters. This package was designed to either be used on the command line or in a script for users to obtain parameters with high-throughput methods. This program also creates output scripts for visualizations in PyMOL. In the future, this program could add additional steric parameters, as the field grows, and unique parameters become prevalent. The vectorization of existing parameters has proven to be a useful method, with greater access to computing and processing power, more data can be collected from systems which could be useful in describing which aspects of size and shape are relevant in chemical reactivity.

The final program discussed in Chapter 5 is Py-X Struct, which was made to query the Cambridge Structural Database for structure matches and to obtain geometric data, including bond distances, angles, and dihedrals. This program was shown useful in studying conformational behaviors of diarylureas and thioureas, revealing differences in their preferred conformations, later explored with quantum mechanical methods. This program allows for fast searching of molecules and molecular substructures from the large database and could be modernized to study more aspects of crystallization conditions. Properties about crystallographic conditions and

additional molecules present in the crystal structure, present and available in the Cambridge Structural Database, may reveal more information about conformational behavior.

APPENDIX A: SUPPLEMENTAL INFORMATION FOR CHAPTER 3

Comparison of Sterimol Measurements with Sterimol2Vec Measurements

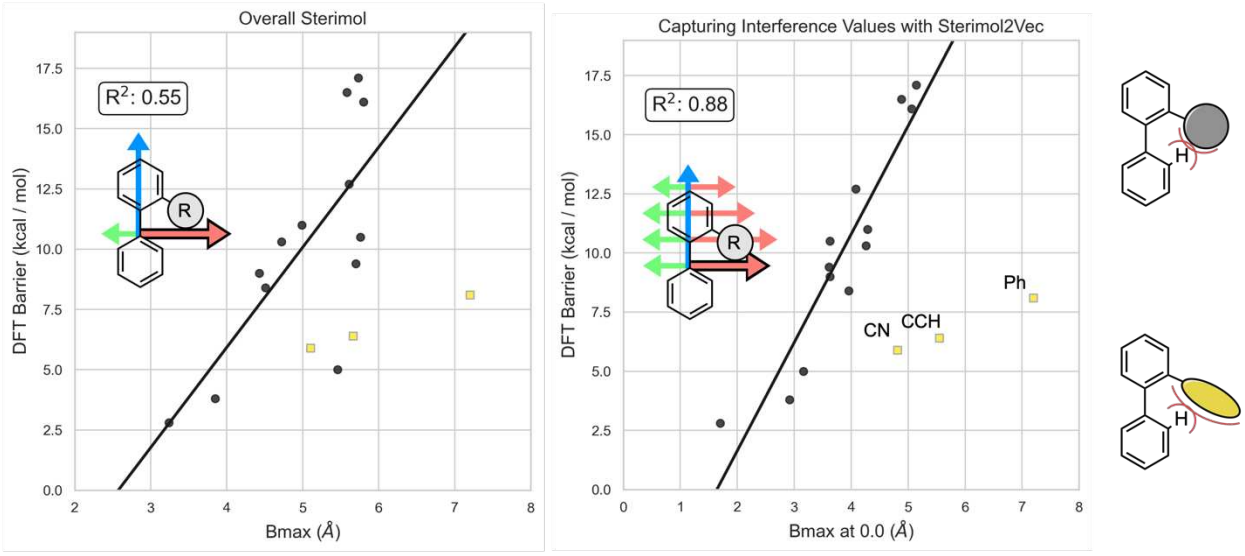


Figure A.1. *A univariate comparison between DFT computed rotational barriers with measured Sterimol parameters (left) and the B_{Max} Sterimol2vec measurement at 0.0Å, omitting the yellow datapoints. The Sterimol values do not have as strong of a correlation as the Sterimol2vec values.*

Comparing DBSTEP Sterimol Calculation with Original Fortran

From molecular structures files found on our GitHub page at:
https://github.com/patonlab/DBSTEP/tree/master/dbstep/examples
Including:

H, Me, Et, iPr, nBu, CH2iPr, cHex, nPr, Ad, tBu, CH2tBu, CHEt2, CHiPr2, CHPr2, CEt3,

Ph, Bn, 4ClPh, 4MePh, 4MeOPh, 35diMePh, 1Nap

The following plots assess the error in DBSTEP grid based Sterimol calculations when compared to the original Fortran code values. The grid spacing option in DBSTEP can be set to lower numbers to reduce the error, however with smaller grid spacing, the calculation time increases rapidly. The default grid spacing in DBSTEP is 0.05Å.

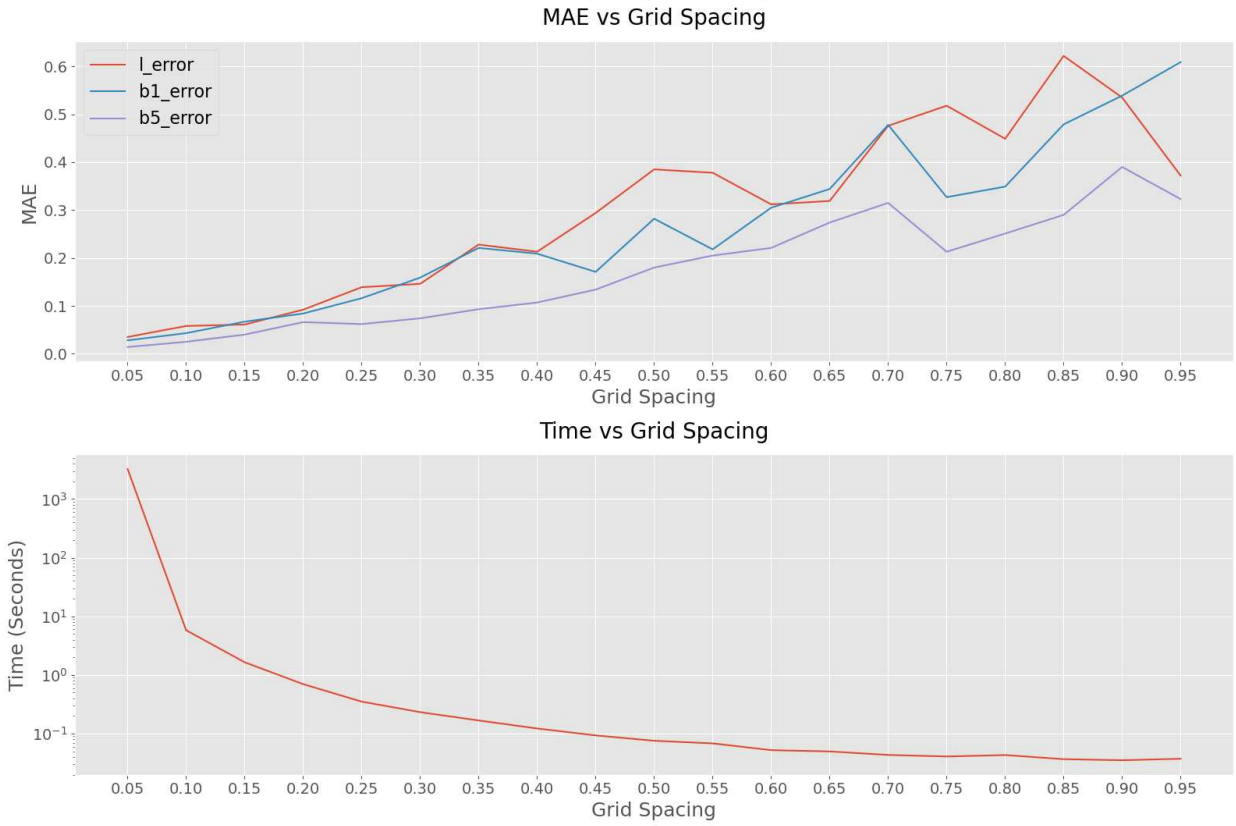


Figure A.2. *(Top) The error in Sterimol values L, B1, and B5 when computed with DBSTEP or the original Sterimol Fortran code. Error increases as grid spacing increases. (Bottom) Timing data for ranging grid spacing. As grid spacing increases, timing decreases.*

Molecular Volumes from Isodensity Surfaces

Compare DBSTEP-computed molecular volumes from isodensity surfaces to experimentally measured molecular volumes. From Weinberg and coworkers (dx.doi.org/10.1021/jp209088u). There is good agreement between the volumes.

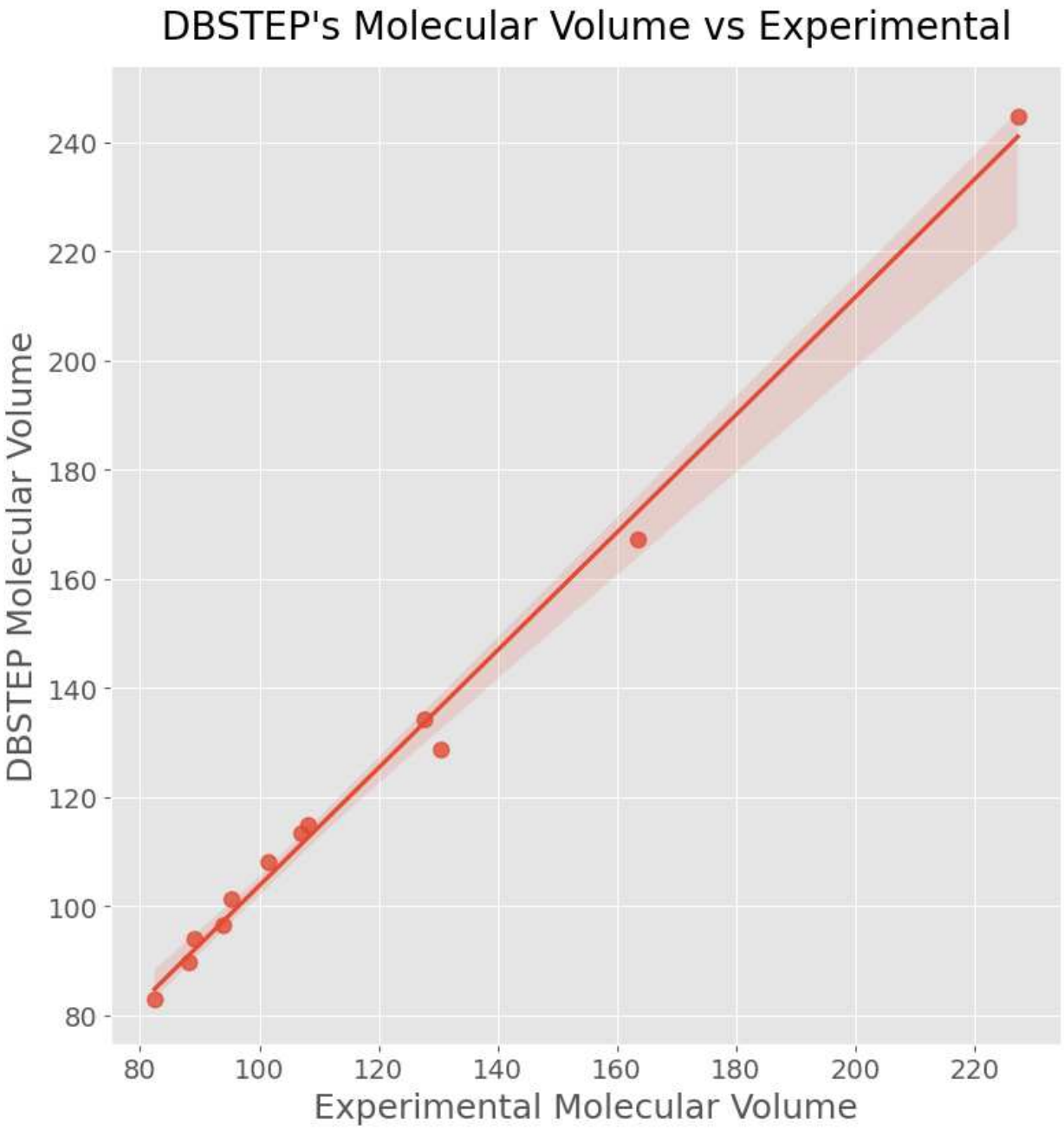


Figure A.3. *DBSTEP computed molecular volumes directly compared with experimentally measured volumes.*

Effect of hydrogens and radii scaling on volume calculations

Comparisons between no Hydrogens and all atom volume calculations, and no hydrogen and scaled all atom calculations.

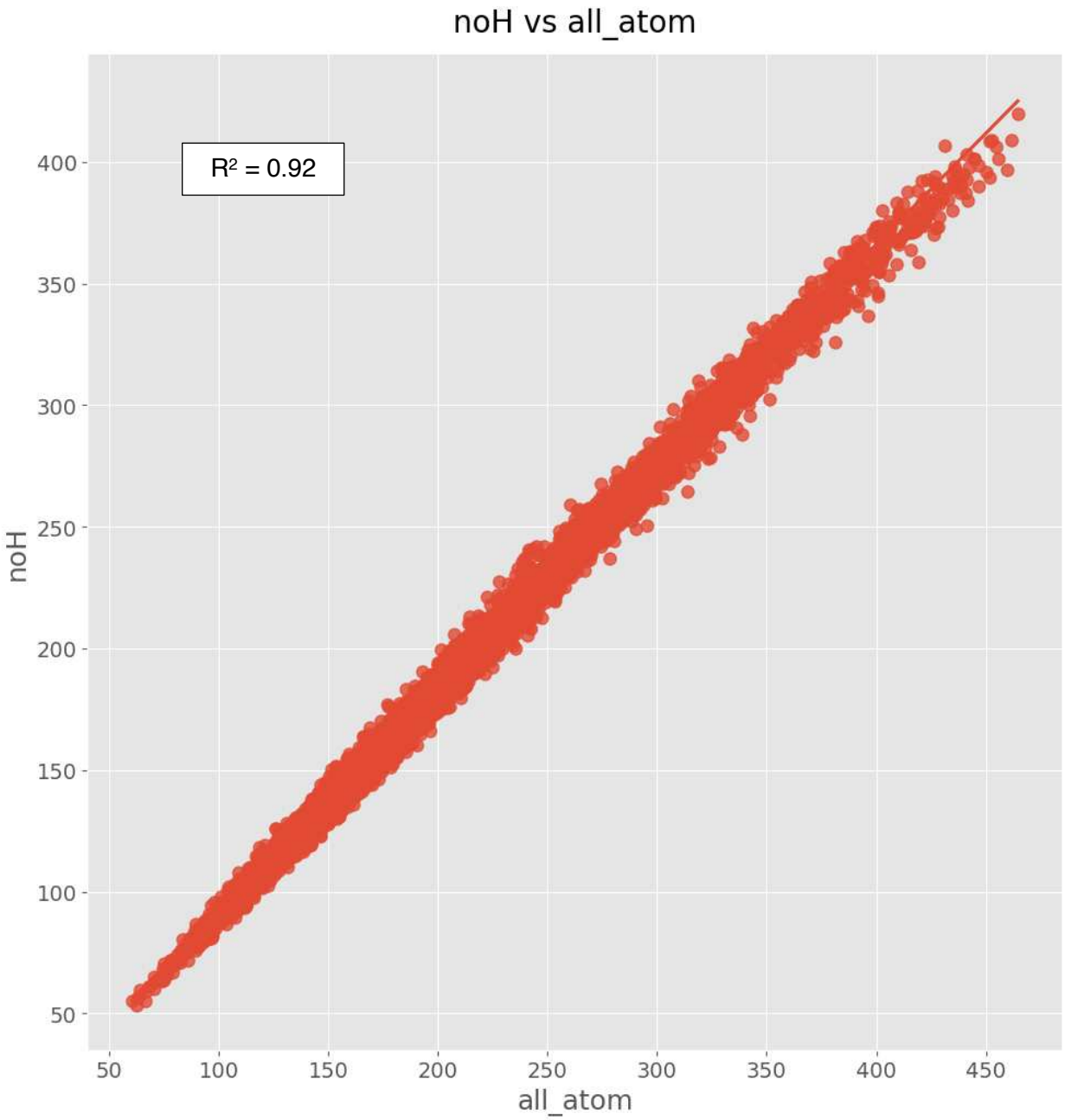


Figure A.4. *Comparisons between molecules with all atoms (x-axis) and molecules with hydrogens removed (y-axis).*

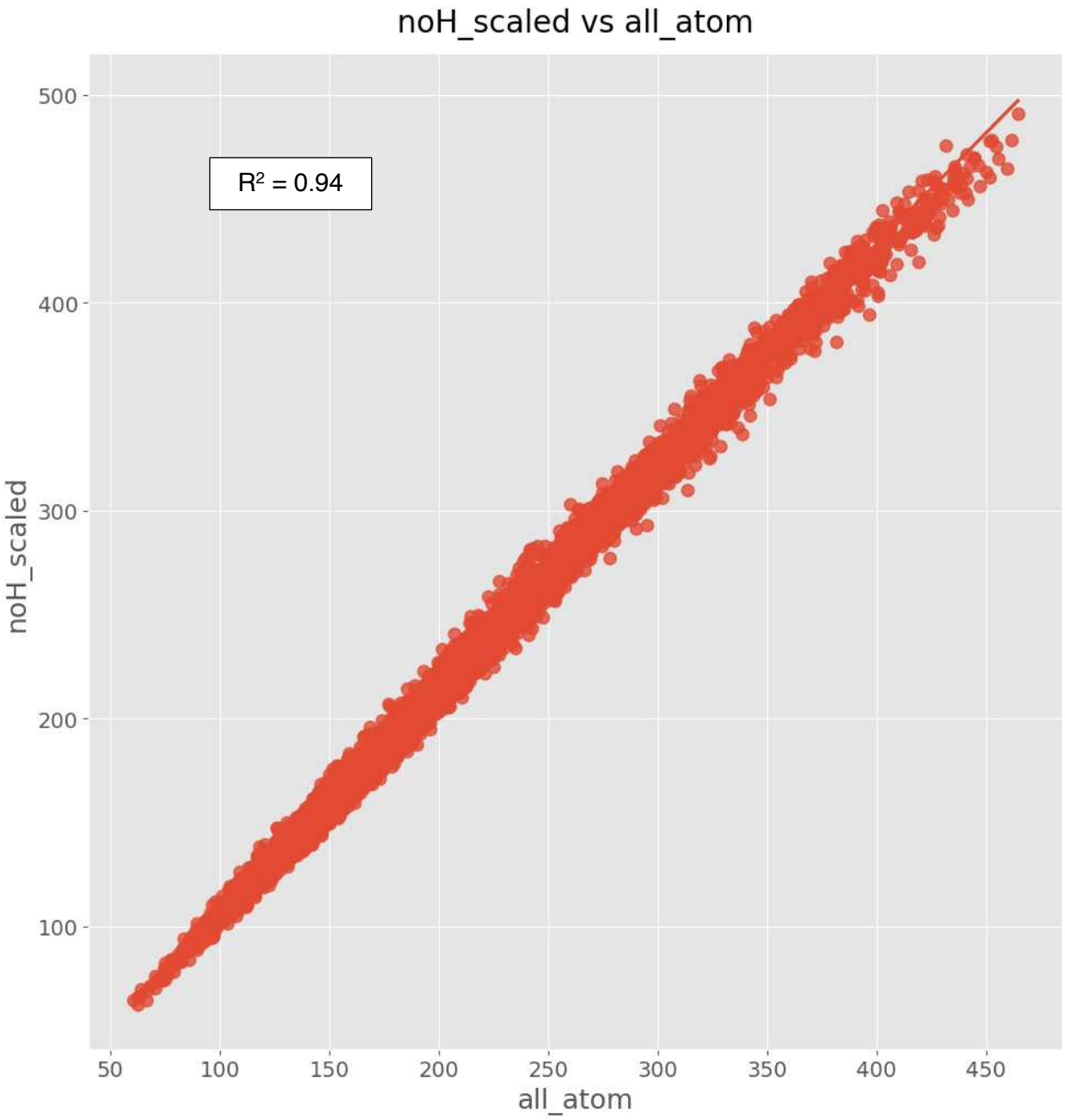


Figure A.5. *Comparisons between molecules with all atoms (x-axis) and molecules with hydrogens removed and all other atom radii scaled by (y-axis).*

APPENDIX B: SUPPLEMENTAL INFORMATION FOR CHAPTER 4

Program Versions and Computational Methods

RDKit	2022.09.1
CREST	2.12
xTB	6.5.0
Gaussian16	Revision C.01
NBO	7.0.5
qmdesc	1.0.6
CASCADE	1.0

Structures were generated from SMILES strings using RDKit functions:

```
Mol = Chem.MolFromSmiles(smiles)
hmol = AllChem.AddHs(mol)
Chem.rdDistGeom.EmbedMolecule(hmol)
Chem.rdmolfiles.MolToXYZFile(filename)
```

Coordinates from the generated structures were then submit to CREST to perform a conformational search for each molecule:

```
crest filename.xyz
```

The CREGEN workflow from CREST was used to remove duplicate structures and filter structures outside of an energy window of 4.0 kcal/mol from the lowest energy structure. Structures were also filtered as duplicates based on energy and RMSD thresholds of 0.31 kcal/mol and 0.2 Å, respectively using the crest_best.xyz and crest_conformer.xyz output files from the initial crest run with the line:

```
crest crest_best.xyz –cregen crest_conformers.xyz –ewin 4.0 –ethr 0.31 –rthr
0.2
```

Resulting structures were submitted to Gaussian16 at the B3LYP/def2TZVP level of theory with SMD implicit solvation using DMF and a Becke-Johnson damped Grimme D3-dispersion correction. Several single point calculations were then run to compute charges with resulting optimized coordinates with the following Gaussian16 keywords shown in Table S1.

Table B.1*. Charge methods and the keyword used in Gaussian 16 to calculate the values.*

Charge/NMR Method	Gaussian 16 Keyword
NBO Charge	pop=(nbo6)
Mulliken Charge	pop=(Regular)
Hirshfeld and CM5 Charge	pop=(hirshfeld)
HLYGAt Charge	pop=(hlygat)

CHELPG Charge	pop=(chelpg)
MKUFF Charge	pop=(mkuff)
Minimum Basis Mulliken Charge	pop=(MBS)
Iterative Hirshfeld and CM5 Charge	IOp(6/79=11)
GIAO NMR Shift	nmr=(GIAO)
CSGT NMR Shift	nmr=(CSGT)
IGAIM NMR Shift	nmr=(IGAIM)

Optimized coordinates were also submitted to xTB using the GFN2-xTB method for computing Mulliken and CM5 charge values with the following input line:

```
xtb filename.xyz --pop --acc 0.2 --gfn 2 --chrg 0 --uhf 1 --etemp 300
```

SMILES strings were used to obtain qmdesc charge and nmr values with the following code snippet:

```
handler = qmdesc.ReactivityDescriptorHandler()
results = handler.predict(smiles)
results = results[['partial_charge','NMR']]
```

The CASCADE web app was used for C13 and H1 NMR prediction using SMILES strings.

```
https://nova.chem.colostate.edu/cascade/predict/
```

Charge and NMR Intercorrelation Plots

This section contains Pearson R^2 values representing parameter intercorrelations for charge and NMR values taken from the C and H atoms produced by the initial dataset of 89 monosubstituted aryl groups.

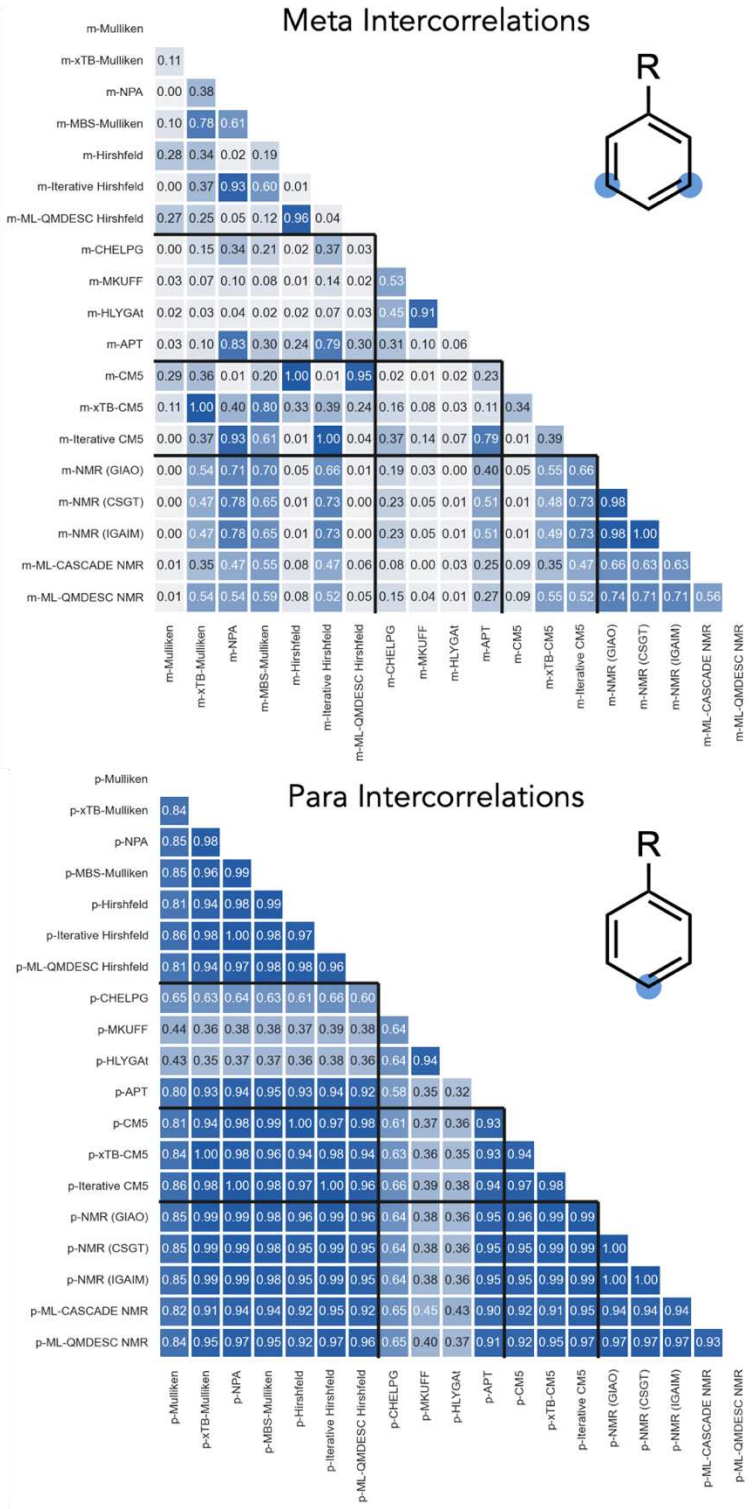


Figure B.1. *Parameter intercorrelations for values taken at the aryl carbons at the meta and para position.*

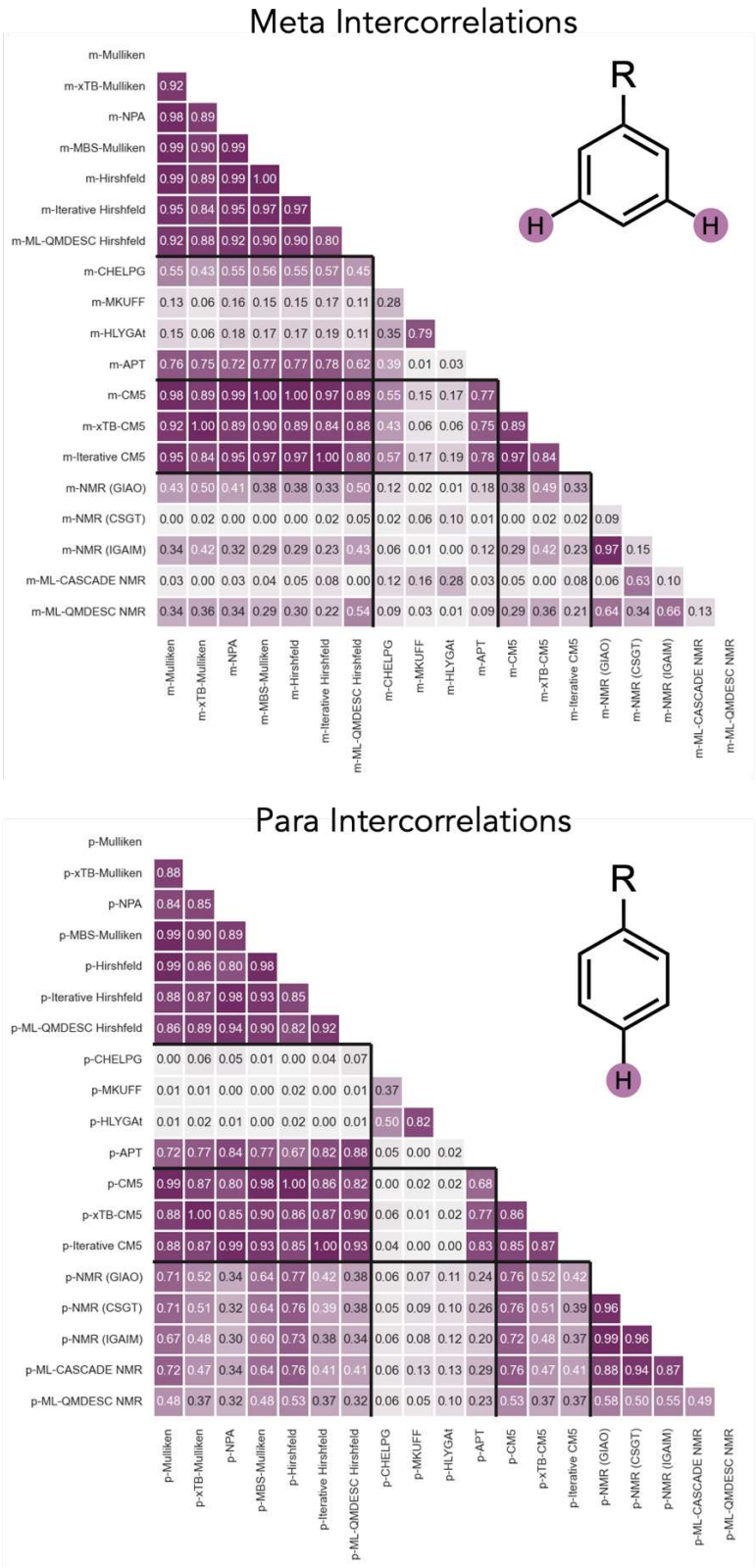


Figure B.2. *Parameter intercorrelations for values taken at the hydrogens attached to aryl carbons at meta and para positions.*

Singlet Ground State and Triplet Excited State Nitroarene Comparisons

This section contains comparisons of correlations between the singlet ground state and triplet excited state optimized structures. Visualized are the correlations between the charge and NMR values compared against the relative rate values evaluated for nitroarene photocycloaddition.

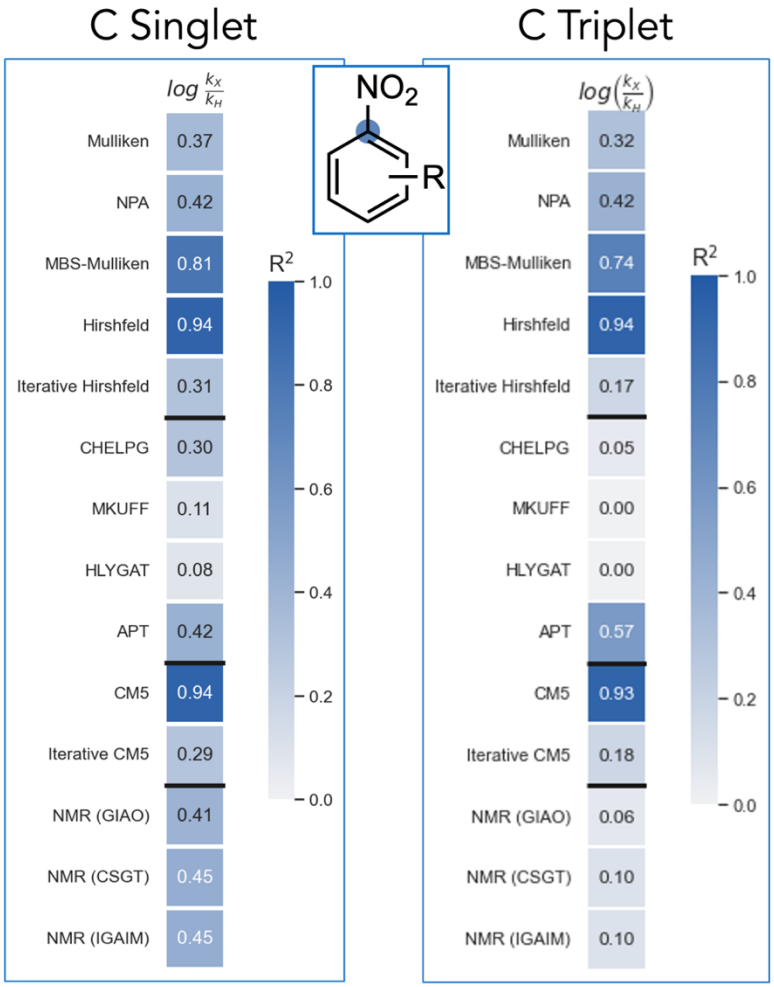


Figure B.3. *Parameter intercorrelations comparing charge and NMR values computed at the aryl C bound to the nitro group in the nitroarene optimized in the singlet ground state (left) and the triplet excited state (right) with the experimentally determined relative rates of the photocycloaddition reaction.*

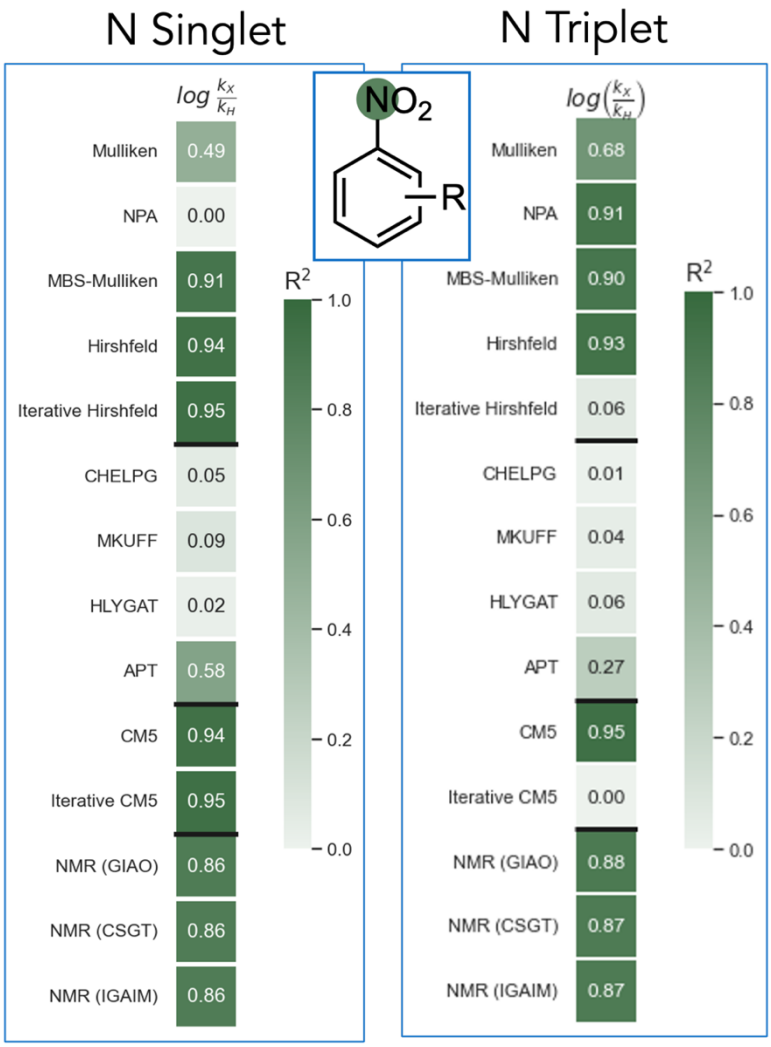


Figure B.4. *Parameter intercorrelations comparing charge and NMR values computed at the nitro group nitrogen for nitroarenes optimized in the singlet ground state (left) and the triplet excited state (right) with the experimentally determined relative rates of the photocycloaddition reaction*