Title

Geometric Learning for Quantum-Informed, Machine Learning and Analysis of Electrostatic Preorganization

Author

Vargas, Santiago

Publication Date

2024

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

Geometric Learning for Quantum-Informed, Machine Learning and Analysis of Electrostatic Preorganization

A dissertation submitted in partial satisfaction

of the requirements for the degree

Doctor of Philosophy in Chemistry

by

Santiago Vargas

2024

ABSTRACT OF THE DISSERTATION

Geometric Learning for Quantum-Informed, Machine Learning and Analysis of Electrostatic Preorganization

by

Santiago Vargas

Doctor of Philosophy in Chemistry

University of California, Los Angeles, 2024

Anastassia N. Alexandrova, Chair

This thesis is organized in a slightly unconventional fashion: algorithms lead and applications fill out the content. I think this emphasizes my interests during graduate school - I built algorithms and tools to address issues that were otherwise inaccessible to different areas of computational chemistry (including applied machine learning) and enzymology. Two sets of scientific thrusts underscore the bulk of my work: algorithms to analyze dynamic, heterogeneous fields in the context of enzymology and flexible machine learning algorithms, including those that leverage quantum descriptors, for rigorous molecular and reaction-level properties. Each section will include grounding on applications and broader impacts for the reader as well. Now we pivot to discussing the main thrusts and outlining each chapter briefly.

General ML and Quantum Theory of Atoms-in-Molecules (QTAIM): QTAIM serves as a mathematical decomposition algorithm for electronic basins within a molecule. The algorithm intakes molecular densities, as computed (typically) by density functional theory (DFT), and uses the flux of density to partition the scalar field into 3-dimensional atomic basins of density [14, 16]. These objects are known as atomic basins and represent the quantum atom within a molecule. By constructing these structures, we compute a rich set of mathematical descriptors that map to many features including energies, bonding,

and electron delocalization. These features have been correlated, in the past, to activation energies, reactivity, and overall system energies, but these uses largely relied on human intervention and small datasets [44, 62, 65, 111, 142, 287]. By developing software centered around high-throughput QTAIM calculations and machine learning, I was able to bring these descriptors to larger datasets and a wide host of applications.

In **Chapter 2**, I discuss an algorithm I implemented to predict Diels-Alder reaction barriers from QTAIM signatures alone. In this study, we showed that QTAIM features, can be used to surmise reaction barriers while also using machine learning techniques to understand what signatures were most informative to our models. Here QTAIM electrostatic potentials and delocalization indices alone were able to yield great performance on withheld datasets. In addition, we demonstrated that QTAIM features can allow a machine learning model to generalize, to an extent, to much larger Diels-Alder reactions. This chapter was adapted from the following: *Machine Learning to Predict Diels–Alder Reaction Barriers from the Reactant State Electron Density.* S. Vargas*, M. Hannefarth, Z. Liu, A.N. Alexandrova. *Journal of Chemical Theory and Computation* **2021** 17 (10), 6203-6213. 10.1021/acs.jctc.1c00623.

In **Chapter 3**, I discuss a package developed to perform high-throughput QTAIM calculations on datasets of molecules and reactions. This package is currently adapted to work with open-source packages such as ORCA and Multiwfn. These softwares, respectively, compute DFT densities at a user-specified level of theory and subsequently compute QTAIM descriptors. The package is built with high-performance compute (HPC) in mind as it can operate on a single dataset with an arbitrary number of concurrent jobs. Here I also used the package to compute QTAIM values for a diverse set of important and difficult datasets and developed graph neural networks to predict molecular and reaction properties leveraging QTAIM as inputs. This chapter was adapted from the following: This was adapted from *High-throughput quantum theory of atoms in molecules (QTAIM) for geometric deep learning of molecular and reaction properties* Santiago Vargas, Winston Gee, and Anastassia N. Alexandrova. *Digital Discovery* **2024** 3, 987-998.

Advancing Analysis of Electric Fields in Proteins: The later chapters follow our work in developing algorithms to ingest, interpret, and predict on electric fields in protein active sites. This work builds on the notion of electrostatic preorganization, a theory that posits that protein scaffolds arrange to electrostatically catalyse chemical reactions, and thereby, destabilizing reactants while suppressing transition state energies [299, 301].

Chapter 4 depicts exhaustive efforts to apply heterogenous electric field analysis to understanding directed evolution in the context of a protoglobin directed evolution (DE) trajectory. Previous DE efforts optimized protoglobin to efficiently catalyze carbene transfer reactions. We show that traditional explanations for increased catalytic activity across the DE lineage, substrate access and binding, cannot account for the dramatic improvements in protein activity. By tracking the 3-D electric field and using clustering algorithms, we pinpoint representative structures for QM/MM calculations and show that changes in the electric field, along DE, improve carbene transfer reactivity. These findings highlight the role electrostatic organization, notably its dynamic effect, has on determining protein function and points to its future importance in designing proteins for relevant chemical processes. This chapter is adapted from *Directed Evolution of Protoglobin Optimizes the Enzyme Electric Field.* Shobhit S. Chaturvedi, Santiago Vargas, Pujan Ajmera, and Anastassia N. Alexandrova. *Journal of the American Chemical Society* **2024** 146 (24), 16670-16680 DOI: 10.1021/jacs.4c03914.

In **Chapter 5**, I introduce a machine learning framework designed to predict enzyme functionality directly from the heterogeneous electric fields applied to protein active sites. We apply this method to a dataset of Heme-Iron Oxidoreductases. Previous studies here, focused on simple, point electric fields along the Fe-O bond, are insufficient for reasonable accuracy. On the otherhand, our 3-D, heterogenous model can accurately predict protein activity without relying on additional protein-specific information. In addition, feature selection elucidates what electric field components most inform our models and thus highlight important components to reactivity and selectivity. Finally, we apply previously-mentioned electric field clustering algorithms and QM/MM calculations to reveal how dynamic complexities in

protein structures can complicate predictions and thus provides a path forward for improved models in this space. This chapter is adapted from *Machine-learning prediction of protein function from the portrait of its intramolecular electric field.* S. Vargas*, S. Chaturvedi, A.N. Alexandrova. (*Accepted, Journal of the American Chemical Society*)

The dissertation of Santiago Vargas is approved.

Chong Liu

Daniel Neuhauser

Philippe Sautet

Anastassia N. Alexandrova, Committee Chair

University of California, Los Angeles

2024

To my mother Hilda, father Jaime, and sister Sarah - todo se puede, y a pesar de todo, ustedes han y siempre seran mi razon de luchar.

It's also for the immigrants. The narratives, characters, and joy you add to this country are the best part about it.

Contents

List of Figures

Acknowledgements

To my scientific and academic family in the trenches: Thank you to my advisor Anastassia Alexandrova for your infinite patience and massive vision - without you, my crazy ideas would've never panned out. I think we did some great science together and hopefully that continues. You were a source of positive energy, patience, and ideation that I have come to appreciate immensely. Without your flexibility and adaptability I would not have made it this far as a scientist, full stop. Many thanks to my committee members Phillipe Sautet, Chong Liu, and Daniel Neuhauser for constant and lucid feedback.

Without my past advisors I simply would not have become the scientist I am today: thanks to Alan Aspuru-Guzik, Peter Bloomingdale, John Calarco, Walfre Franco, Scott Joray, Antari Khot, and Tim Menke. Alan and Tim, thank you for seeing excitement and potential in me and pushing me towards computational sciences. John Calarco, thank you for being such a knowledgeable and patient mentor and a pretty solid libero as well. Peter and Antari, thank you for having faith and patience in a student from another domain and adapting my knowledge to new and exciting directions. Walfre, huge thanks for exposing me to medical research and showing me new avenues to apply my algorithms. Finally, Dr. Joray: your immense efforts to create exciting science education and research at AMSA were not lost on me.

To my bio-boys, Shobhit and Pujan, I simply could not have accomplished this without your expertise, brilliant minds, and consistency. At many points in graduate school I had lost the joy for science but working with y'all has been an absolute joy. Sam Blau, you were an absolute force in science and showed me how rewarding and fun collaborations could be. Thank you for being an open mentor willing to make academia less serious and intimidating. Prof. Evan Spotte-Smith, thank you for your constant energy and unmatched ability to generate difficult datasets. I would like to thank Dr. Rishabh Guha for being an amazing sounding board for new ideas and helping me gain experience and confidence in a new field. Wenbin Xu, it's been quite fun developing algorithms with you for some objectively insane ideas. Winston thank you for the faith and energy in our algorithms and for always being so kind and patient around my, at times, janky code. A shout of appreciation and thanks to my other collaborators Matthew Hennefarth, Dr. Patricia Poths, Hootan Roshandel, Amy Lai, and Dr. Daniel Bim.

To the members of the lab, past and present, I owe immense gratitude for creating a home away from home and a home within the science: Zerina Mehmedhovic, Thomas Cross, Patricia Poths, Harry Morgan, Dr. Chaturvedi, Nathaniel Johnston, William Laderer, Taras Khvorost, Pujan Ajmera, Dr. Julen Munarriz, Dr. Han Guo, Dr. Daniel Bim (and many more!). This journey is as much about science as the friends you make along the way.

To the folks at the Advanced Math and Science Academy: y'all were fundamental in my journey and I cannot thank you enough for equipping me with the tools and confidence to take up a challenge. Special thanks to Padmaja Bandaru, Joe Bengiovanni, Madhavi Marathe, Martha Tassi-Richardson, and Lyubov Schmidt for teaching me foundational subjects and being amazing mentors.

To the folks at Harvard's LS50, especially Andrew Murray, consider your experiment successful on this data point. Despite the wide-array of topics I learned through the course, I think the biggest takeaway from LS50 is the confidence to operate, hypothesize, and experiment in any area of science. My journey since has been a rotating circus of computational chemistry, pharmacokinetics, enzymology, machine learning, and graph algorithms where I have made progress and leveraged ideas from all these subjects to do interesting science that I think few others could (or more likely - would).

I would also like to thank the DOE CSGF Fellowship and everyone at the Krell Institute (thanks Lindsey and Michelle!) for giving me the opportunity to explore my interests and develop into the scientist I am today.

To my family, both chosen and not: Para la familia - muchisimas gracias Ma y Pa. Esto es de todos nosotros, todos fuimos parte de este logro y les tengo mucho cariño y admiración. Estos años fueron de los más difíciles de mi vida pero estamos saliendo adelante. Ma - tu mensaje diario "Hijoooooo" siempre me acordaba de que están pensando en mi. Gracias por enseñarme a trabajar duro y a dar todo por los demás. Pa - montar bici y hablar de todo han sido refugios para mi, gracias por siempre compartir eso conmigo. Sin ti no sabría lo que es meterle energía y creatividad a la vida. Sarah, it has been such a joy and privilege seeing you grow up, your resolve and ability to recover from hard times is something I think nobody can rival. Thank you for helping me grow up as well and being patient. Nobody else has shared so much of the difficult and incredible times from the last few years with me. You're a confidant, a fellow shit-talker, music-lover, and absolute menace. Glow on little shit. Finally, love and thanks to the two pets, Kiwi and Nugget, who can't read but nonetheless have been a source of unconditional love and constant support.

Priya, Adit, and Andrew, thank you for taking care of me during the hardest period of my life. I owe so much to you three and I will always love you immensely. Huge gratitude to Andrew for sharing these incredible and difficult years with me and for being an absolute constant in my life - I'm excited for the next years of joy we'll both have. Priya, I would not be the person I am today without your gentleness, love, and positive energy. Adit, thank you for always being an open door and sibling to me, you're an incredible friend and person.

Jesse, you constantly inspire me and are ultimately a massive part of my decision and ability to become a scientist. Paul, thank you for keeping me from becoming a total

curmudgeon these last few years - your unseriousness and 5 am chats were a source of refuge during some difficult times. Kim, you're goofy for real and I cannot imagine the vignettes of the last few years of life without you. Will, I cannot tell you how much you motivate me to be better - kinder, more curious, and less serious. Constant walks for lil treats and a shared sense of humor indecipherable to those around us are highlights of graduate school. Zerina - thank you for welcoming me into grad school and being a constant, caring friend along the way.

Bidart me da mucha felicidad tenerte cerca, fisicamente, emocionalmente, y cientificamente. Marley - you have been a bedrock friend since I met you and I cannot emphasize how much I appreciate your constant humor and effort in staying close. Vera, thank you for being a foundational friend growing up, I cannot thank you enough for your patience towards an insufferable me over the years.

I would also like to share unending gratitude to the creatives and artists who crafted wonderful pieces of culture that filled me with joy, curiosity, and perspective over the last years. Thank you to Arca, Michelle Zauner, Anthony Bourdain, Jhumpa Lahiri, Raveena, Turnstile, SOPHIE, and the illustrious Gabriel Garcia Marquez. I extend Jhumpa Lahiri's voice when she said "That's the thing about books. They let you travel without moving your feet" to encompass music, film, and television. In a Time of Cholera (COVID), the universes these creatives crafted inspired and contextualized my science - it is our responsibility to remain curious, empathetic and develop technologies for those outside the lab. There is immense beauty out in the world and science is neither isolated from it nor the only source of it.

Vita

Education

Harvard College 2015 - 2019

Bachelor of Arts, Chemistry and Physics

Awards

Darleane Hoffman Distinguished Postdoctoral Fellowship	April 2024
Charles J. Pederson Dissertation Award	April 2024
DOE Computational Science Graduate Fellow	April 2020
Ford Foundation Predoctoral Fellowship, Honorable Mention	March 2020
Fulbright Research Fellowship	April 2019

Publications

Theory of local fields in proteins and enzymes. P. Ajmera, S. Chaturvedi, S. Vargas, T. Wilson, A. N. Alexandrova, M. Eberhart. (Submitted).

Machine-learning prediction of protein function from the portrait of its intramolecular electric field (2023) S. Vargas*, S. Chaturvedi, A. N. Alexandrova (*Accepted, Journal of the American Chemical Society*).

A foundation model for atomistic materials chemistry. I. Batatia ..., S. Vargas, ... (10.48550/arXiv.2401.00096, Submitted).

High-throughput Quantum Theory of Atoms in Molecules (QTAIM) Applied to Geometric Deep Learning. S. Vargas*, W. Gee, A. N. Alexandrova. Digital Discovery, 2024,

10.1039/D4DD00057A.

Thermodynamic Equilibrium versus Kinetic Trapping: Thermalization of Cluster Catalyst Ensembles Can Extend Beyond Reaction Time Scales. P. Poths, S. Vargas*, P. Sautet, and A. N. Alexandrova. ACS Catalysis 0, 14, 10.1021/acscatal.3c06154.

Directed Evolution of Protoglobin Optimizes the Enzyme Electric Field (2023) S. Chaturvedi, S. Vargas*, P. Ajmera, A. N. Alexandrova. Journal of the American Chemical Society, 10.1021/jacs.4c03914.

HEPOM: A predictive framework for accelerated Hydrolysis Energy Predictions of Organic Molecules (2023) R. D. Guha, S. Vargas*, E. W. C. Spotte-Smith, A. R. Epstein, M. C. Venetos, M. Wen, R. S. Kingsbury, S. M. Blau, K. A. Persson (in press, accepted at NeurIPS AI4Mat, https://openreview.net/forum?id=eDlEn1PPJw).

An Artificial Intelligence Framework for Optimal Drug Design (2022) G. Ramey, S. Vargas*, Dinesh De Alwis, Anastassia N. Alexandrova, Joe Distefano III, Peter Bloomingdale bioRxiv 2022.10.29.514379. 10.1101/2022.10.29.514379.

Computational and Experimental Design of Quinones for Electrochemical CO2 Capture and Concentration A. M. Zito, D. Bím, S. Vargas, A. N. Alexandrova, and J. Y. Yang ACS Sustainable Chemistry and Engineering 2022 10 (34), 11387-11395. 10.1021/acssuschemeng.2c03463.

Machine Learning to Predict Diels–Alder Reaction Barriers from the Reactant State Electron Density. S. Vargas*, M. Hannefarth, Z. Liu, A.N. Alexandrova. Journal of Chemical Theory and Computation 2021 17 (10), 6203-6213. 10.1021/acs.jctc.1c00623.

Team-based Learning for Scientific Computing and Automated Experimentation: Visualization of Colored Reactions. (2019). S. Vargas*, S. Zamirpour, S. Menon, A. Rothman, S. Sim, T. Menke, and A. Aspuru-Guzik. Journal of Chemical Education 2020 97 (3), 689-694., 10.1021/acs.jchemed.9b00603.

Seasonal changes in diet and toxicity in the Climbing Mantella frog (Mantella laevigata). N. A. Moskowitz, ..., S. Vargas, ..., 2018. PLoS ONE 13(12): e0207940, 10.1371/journal.pone.0207940.

Chapter 1

Introduction

Much like the two branches of this thesis, the introduction will take two sections to introduce existing methods, studies, and algorithms in each area detailed in subsequent chapters. For QTAIM-informed geometric learing, I begin by briefly introducing the field of chemoinformatics, including descriptors, algorithms, and graph neural networks. From here, I pivot to provide grounding on *ab initio* descriptors in machine learning algorithms before providing some outlook on the future of this area (and naturally how this leads into my own work). The later half of the chapter overview electrostatic preorganization, including computational and experimental methods for its quantification, and how groups have used electric fields for a proxy to analyze the electrostatic contributions of the protein scaffold. Again, I briefly provide some prognosis for the field, including high-dimensional algorithms and dynamical studies and how these ideas tie into my own developments.

1.1 Quantum-Informed Geometric Learning for Chemistry

1.1.1 Descriptors for Machine Learning in Chemistry

0-D Descriptors

0-D descriptors encompass a set of single-valued descriptors that inform chemoinfomaticians on the global structure of a molecule without granularity towards atom or bond-level values. Many such descriptors (P_m), including molecular polarizability, molecular weight, molar refraction, diamagnetic sensitivity, and parachor [62, 66, 116, 116, 154, 313] can be computed as the simple global averaging or sum of atomic-level descriptors(p_i):

$$P_m = \sum_{i=1}^{N} p_i \tag{1.1}$$

Typically these descriptors are limited to only processing the overall chemical composition of a molecule, and thus, are limited to relatively crude information such as number of specific elements in a molecule, molecular weight, etc. Despite this, many 0-D descriptors are considered when constructing more complex models with 1-D, 2-D, and 3-D descriptors included [113, 122, 313].

1-D Descriptors

1-D descriptors extend to vectorized representations of molecular information, including, some baseline information on molecular bonding. These features can encompass concepts such as H-donors/acceptors, number of ring atoms, counts of different atom hybridizations, etc. [76]. This family of descriptors also encompasses some of the ubiqitious "fingerprints" used throughout pharamcology and traditional chemoinformatics. MACCS fingerprints, for example, count a set of 320 drug-like structural fragments that are used to encode structures [76]. Daylight fingerprints are another example that, instead of using a predefined

set of motifs, compute unique structural motifs across a dataset before hashing these vectors to 1028 or 2024-item vectors [76].

2-D Descriptors

2-D, or topological descriptors extend the previous mapping of molecular motifs to looking at the overall molecular topology, i.e., its graph structure. The earliest such descriptor, the Wiener index, was introduced in 1947 and describes the sum of distances between any two carbon atoms in the molecular graph [76, 315]. In other words, it is the number of "jumps" between carbon atoms ($d_{i,j}$), summed across the entire atom:

$$W(G) = \sum_{i=1}^{N} \sum_{j>i}^{N} d_{i,j} \tag{1.2}$$

Lipophilicity is another example that is used widely in biological applications of small molecules. Despite being a single-value, lipophilicity is computed as a convolved property across the molecular graph [186]. Topological Polar Surface Area (TPSA) is another descriptor, often used in drug design, and describes the processed property between molecular surfaces and partial charge calculations [61, 213]. This value is often calculated (and is thus a 2-D property) as the sum of tabulated *fragment* values across the molecular graph [78].

Two cornerstone 2-D descriptors are the Extended Connectivity Fingerprints (ECFP) and Morgan fingerprints. ECFPs and Morgan fingerprints are similarly constructed by considering each atom as a "seed". At each seed, the algorithms grow to consider all atoms 1, 2, etc. graph hops away where it indexes different potential graph motifs such as rings and hybridization. This leads to the consideration of larger graph fragments. Finally, each vector representative of a each seed is hashed to a fixed-sized vector [239]. motifs, it counts the number of graph degrees present at each atom node in the molecular graph representation. The Morgan algorithm will then sort atoms by how densely interconnected they are [248, 317].

Morgan Fingerprints are computed similarly but instead of growing and considering

specific molecular

3-D Descriptors

3-D descriptors extend from molecular graph structures to molecular geometries in 3-dimensional real space. This naturally leads to the question of how geometries are generated including the cost of such calculations. Most current schemes will leverage cheap conformer generation through CORINA [253], xTB [110] or UFF [231]. Given, hopefully cheap, geometries, a host of descriptors exist in this space. Radial Distribution Functions build on experimental techniques for determining 3-D molecular structure. This algorithm scans electron intensity at different observation radii(r) while inputting interatomic distances($r_{i,j}$). B is a smoothing parameter and p_i/p_j are properties associated with each atom and are often set to 1 or to atomic partial charges charges [125]:

$$g(r) = f \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} p_i p_j e^{-B(r-r_{ij})^2} \tag{1.3}$$

This method has been used extensively throughout traditional chemoinformatics as well as for machine-learned interatomic potentials (MLIPs) [23, 35, 294]. Other descriptors in this family take 3-D structures and compute single-valued structures on them, for example, radius of gyration, molecular electrostatic potential, etc. [76].

1.1.2 *Ab Initio* Descriptors/Properties for Machine Learning in Chemistry

As mentioned above, *ab initio* methods can also be leveraged as descriptors for machine learning. Here we will quickly introduce several such potential methods and their application to quantitative structure-activity relationship (QSAR) studies:

Net Atomic Charges: Partitioning overall molecular charge into substituent atoms is a common approach taken to describe a molecule at the atomic level. One family of approaches

depend on linear combinations of atomic orbitals (LCAOs) that assign electrons to atomic orbitals around a given atomic nucleus. These methods are somewhat reasonable given smaller basis sets but can quickly spiral when applied to larger basis sets with diffuse functions that overlap with adjacent atoms in a system. **Natural Population analysis (NPA)** is an alternative approach that reformulates charge under the paradigm of natural atomic orbitals rather than the basis set used for the calculation. This method, is thus, generally robust to the basis set used [234]. Other approaches include **"atoms in molecules"** approaches that divide electronic density into fragments in 3-D space. One such approach, by Streitweiser et. al [271]. projects density onto a 2-D plane and thereby partitions 3-D space. Here I also mention Bader's Quantum Theory of Atoms in Molecules (QTAIM) which will be covered in more detail in the following section [16].

Orbital energies are often used directly as descriptors, both in QSAR, and more recently, in machine learning approaches. This approach often relies on the chemical notion that interacting orbitals, often highest occupied molecular orbital (HOMO) of one molecule will interact with the lowest unoccupied molecular orbital of another species (LUMO) [67]. These values can be computed semi-empirically or through *ab initio* methods. **Surface descriptors** are another alternative that bridge 3-D structural information with electronic structure data. Much like orbital energies, these values can be approximated semi-empirically through methods such as charged polar surface area descriptors (CPSA) [269] or through DFT calculations and provide especially interpretable descriptors for tasks such as the prediction of nucleophilic/electrophilic reaction sites [78].

1.1.3 QTAIM

Bader et. al. introduced the quantum theory of atoms-in-molecules(QTAIM) in the 1980s [14, 16]. This methodology intakes the electronic density of a system, computed via any arbitrary method, and yields a rich, surjective mapping of bonding networks and interpretable descriptors for analyzing density in a system. Mathematically, QTAIM is derived from two,

Type	Eigenvalues	Interpretation
Maximum(3, -3)	$3\lambda < 0$	Nuclear Critical Point, Nonnuclear Attractor
Saddle Point(3, -1)	$2\lambda < 0$, $1\lambda > 0$	Bond Critical Point
Saddle Point(3, +1)	$1\lambda < 0$, $2\lambda > 0$	Ring Critical Point
Minimum(3, +3)	$3\lambda > 0$	Cage Critical Point

Table 1.1: Given critical point Hessians eigenvalues (λ), critical points can be interpreted.

simple conditions on electronic density(ρ): $\nabla\rho(r_c) = 0$ and $\nabla\rho \cdot \mathbf{N}S_\Omega$. In the first condition, we find points (r_c), known as critical points, where electronic density reaches critical values. Critical points can be interpreted as chemical concepts such as nuclei, bonds, rings, etc. by simply computing the eigenvalues of the second derivative of ρ:

Notably, bond critical points (BCPs) also map back to nuclear critical points in the form of bond paths, these yield a skeleton of molecular "interactions" [64]. Here interactions are used instead of bonds to separate from the notion of bonding as strictly defined as covalent, ionic, etc. - QTAIM interactions obfuscate away these distinctions. Though this is seemingly a negative feature, values derived from QTAIM are often used to map back to these traditional chemical bonding concepts [Tab. B.1]. On the other hand, this formulation allows for the treatment of non-standard bonding involving metal bonds, hydrogen bondings, and other weak bonding regimes not typically parsed correctly by traditional chemoinformatic methods [233, 286]. Thus, QTAIM is an attractive methodology for describing bonding, rigorously, to computers.

The second condition, $\nabla\rho \cdot \mathbf{N}S_\Omega$, describes boundaries between atoms in molecules as surfaces where the flux of electronic density is zero [Tab. B.1]. This bounding provides atomic basins that correspond to different atoms within a molecule, hence QTAIM is an atoms-in-molecules methodology. Given these atomic chunks, various mathematical operations including Laplacians, Hessians, Electrostatic Potentials, etc. can be performed on each bounded surface to yield a rich set of descriptors at each atom, bond, ring, and cage critical point. These descriptors have been leveraged to understand various properties of molecular

and solid-state systems [Tab. B.1].

1.1.4 Algorithms for Supervised Learning

Linear Models

As a baseline, linear regression and regularized versions of the linear regression algorithms were tested on the data set. Linear regression minimizes the residual sum of squares between predicted and training target variables. LASSO, appends the weighted sum of the weight vector to the cost function, Ridge Regression includes the L_2norm and Elastinet uses both weighted L_1 and L_2 norms appended to the cost function in order to regularize a linear model [185].

Decision Tree Regression

Decision trees and their derivatives were tested heavily throughout this work for both the QTAIM-ML and electric field efforts. Decision Trees separate data into nodes and leaves based on sets of binary decisions made for inputs. Decision trees are highly flexible with a large number of hyperparameters to tune and are leveraged in both classification and regression tasks. Tuning these hyperparameters allows us to construct generalizable models that are more resistant to overfitting. Multiple decision trees can also be combined to create an ensemble model known as a Random Forests. These models tend to outperform any single Decision Tree Algorithms by pooling weaker regressors into a weighted sum. Random Forests also utilize the bagging method where random subsets of the data set are sampled, with replacement, to train each decision tree. Further decision tree-based algorithms included the XGBoost algorithm that trains sequential predictors on the residual errors of previously trained models [102]. This algorithm is attractive due to its scalable GPU- implementation and cutting-edge performance in a plethora of different regression and classification tasks [49, 180].

Neural Networks

Neural networks encompass a wide-range of different implementations, but at their very core, neural networks are perceptrons:

$$y(\mathbf{x}, \mathbf{w}) = \sigma(\sum_{j=1}^{M} w_j x_j + b_j) \tag{1.4}$$

σ accounts for non-linear "activation functions", w_j are weights, b_j biases, and x_j is a variable within a vector $\mathbf{x}$. Multilayer perceptrons or "deep" neural network stack these non-linearities, each with their own weights and biases, to get progressively more complex functions. When we train neural networks, we are simply updating these weights and biases to reduce a training loss on a given optimization function such as mean square error, mean absolute error, cross entropy, etc. These algorithms have been tinkered with to allow for deeper training, regularization, dropouts, etc. [34] but this implementation suffers from a notable shortcoming: it necessitates a fixed-size input vector. Thus, feed-forward neural networks have been used in the context of chemistry [161], but typically with fixed-size input vectors. This leads to a discussion, in the next section, on graph neural networks, which do not suffer from this setback.

1.1.5 Graph Neural Networks (GNNs)

Graph neural networks(GNNs), unlike most convolutional and feed-forward neural networks, allow for predictions and training on data with varying input sizes. This is achieved through a few notable tweaks to standard neural networks. First, input graphs, which we will discuss in more detail briefly, are updated from their initial set of input features. These updated features qualitatively reflect their starting values while also incorporating information on edges and neighboring nodes in the graph structure. These updated features can be iteratively updated an arbitrary number of times, more iterations allow for information passing between nodes and edges further "hops" away in the graph structure [115, 282]. Following these

update steps, the new graph is embedded into a corresponding fixed-size vector [Fig. 1.1]. The global embedding function, as its known, can encompass a wide-range of functions, from mean pooling features across a graph to complex attention-based methods. Given a fixed-sized vector representing the updated graph, GNNs typically proceed with a traditional feed-forward neural network.

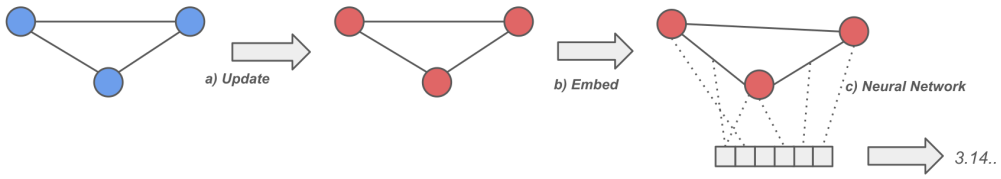


Figure 1.1: The Process of an Encoder Graph Neural Network

This begs a question on representation: how do we encode molecules as graphs? Historically, molecules as graphs have been represented with atoms as nodes and edges as bonds [115]. This translation makes sense as graphs are constructed from entirely chemical motifs. Heterographs, as opposed to homographs with bonds as edges, allow for separate relationships between each different edge type and enable the addition of a separate global node type to store important molecular-level information. Heterographs achieve this by encoding nodes and bonds, both, as nodes. This leads to the creation of difference edges bonding bonds to bonds, atoms to bonds, etc. Heterographs I have leveraged ($\mathbf{G} = (\mathbf{B}, \mathbf{A}, \mathbf{g})$) consist of $\mathbf{B}$ as bond information vectors, $\mathbf{A}$ is atom-level information, and $\mathbf{g}$ is the molecular-level feature vector (Fig. 1.2). Several works have used heterograph structures to encode molecules as graphs, in particular, for instances of molecules with varying spin and charge [24, 57, 122, 286, 313].

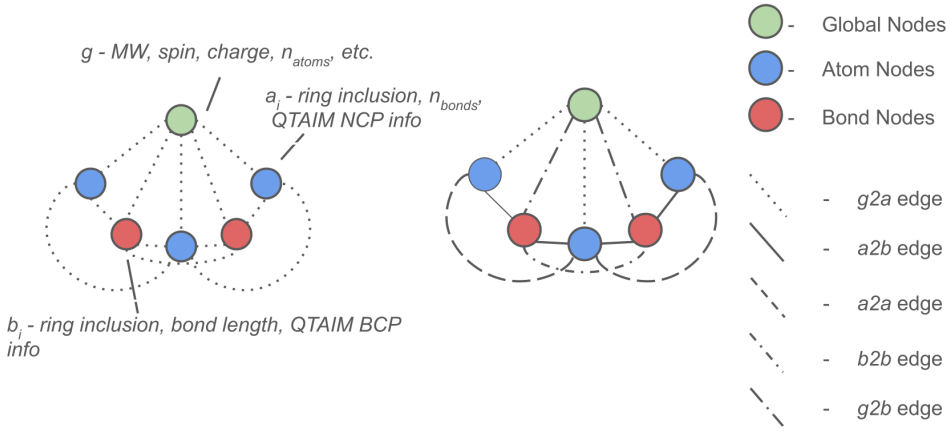


Figure 1.2: An Exemplar Heterograph Contruction for Molecules in Graph Neural Networks.

1.1.6 What Do We Need?

In 2003 Gasteiger et. al. wrote: "Why do we note have databases of quantum chemical calculations?" [76] - and while efforts such as the Opencatalyst Project [50], Materials Project [208], and Quantum Machine [252] initiatives have certainly put a dent in this assertion, there is work to be done. For one, datasets of quantum mechanical features are few and far between [68, 175, 286]. In addition, these features are not computed at concerted levels of theory that would allow for the training models across different chemical domains. The construction of unified, cross-discipline chemical datasets could see the generation of effective, general models.

Indeed, one area for future development is not just immense datasets of quantum chemical features, spanning diverse chemical domains, but using these data to train highly-general foundational models of features such as QTAIM or NBO. Similar foundational models are already making impacts in chemistry [22, 326] but a "universal", quantum feature generator could yield improved analytical techniques, and even, plug into improved MLIPs. In chapters 2 and 3 we show that QTAIM can afford improvements in model performance, including stability in out-of-domain(OOD) predictions - this could be an approach for improving existing machine learning models to unseen predictions.

1.2 Electrostatic Preorganization via Classical Electric Fields

My work focuses on improving the analysis of electrostatic preorganization via high-dimensional, dynamical algorithmic processing of classical electric fields in a protein active sites. This is, naturally, not the only method for treating electrostatistic preorganization, nor is electrostatistic preorganization the only method for analyzing protein function and activity. Here I provide a quick overview of electrostatic preorganization and build on several methods for its use in functionalizing and understanding proteins. This will build towards current methods, and finally, a cursory look at my approaches.

1.2.1 Electrostatic Preorganization

Warshel linked the notions of electrostatic environments to their effects on enzymes - a concept known today as electrostatic preorganization [300, 302, 304]. More concretely, this theory treated larger enzyme structures as a scaffold that imparts electric fields onto a protein active sites. He postulated that these aligned fields would aid chemical reactions by lowering transition barriers and/or destabilizing reactants. With time, this theory has been sustained through a slew of experimental and computational results [134,146,160,184,189,330]. Furthermore, this theory has increasingly been leveraged in *de novo* campaigns for enzyme design where focus on structure has migrated towards increasing emphasis on designing optimized charge distributions for chemical activity [39, 51, 93, 147, 155, 283, 312, 321, 323, 331]. Another advantage to this approach is that by leveraging the direct, clear relationship between charge placement in a protein structure and imparted field, electrostatic preorganization serves as a simple concept for generating mutagenic targets for enzyme enhancement [2, 30, 59, 176, 222, 274]. Naturally, no one theory can alone explain or predict protein activity but electrostatic preorganization, in conjunction with other target properties such as entropy changes, long-range interactions, and dynamics promises to add to the toolkit of approaches

for design [63, 75, 182, 203, 210, 321, 322, 324, 328].

1.2.2 Experimental Methods for Electric Field Analysis

Stark Spectroscopy

The Stark effect underpins Stark spectroscopy and thus serves as the basis for an incredibly powerful tool for probing electric field atomistically [89]. This effect is characterized by the interactions between electric fields and atomic/molecular energies. Notably, the application of electric fields was found to alter molecular spectral lines and provides a framework for analyzing electric field interactions with matter. Stark spectroscopy inverts this understanding by using changes in spectra, pertaining to specific vibrational modes, to measure electric fields at certain probes within a molecule. This interaction is calculated by measuring the shift in vibrational frequency at a given probe(ν_{obs}) in an environment field ($|F_{env}|$) and a Stark tuning rate $|\Delta\mu_{probe}|$:

$$\nu_{obs} = \nu_{probe} - |\Delta\mu_{probe}| \cdot |F_{env}| \tag{1.5}$$

This method has been used extensively to understand how proteins functionalize electric fields towards productive chemical transformations [88]. For example, Boxer et. al. [88] used the method to observe substantial intrinsic fields in proteins on the order of 100 MV/cm. In addition, VSE linked fields to catalytic activity in ketosteroid isomerase (KSI) proteins [4, 86]. Researchers have utilized VSE to explore deviations in local electric fields due to mutations and conformational changes throughout a molecular dynamics (MD) trajectory [42, 80]. Here is it vital to note the interplay between molecular dynamics (MD) and VSE as, together, they yield averaged properties over a trajectory. A shortcoming to this method is the fact it relies on probes to measure electric fields at given points - avoiding higher-dimensional analysis of the complex, heterogeneous electric fields present at protein active sites.

The popularity of VSE has coupled to computational studies that provide high-quality

information regarding fields and their effects on chemical activity. For one, molecular dynamics (MD) studies are often used to calibrate VSE measurements. Notably, MD simulations have been used to map the relationship between electric fields at specific probes and the effects of different solvents or local environments [42]. MD has explained deviations in local electric fieds from bulk electric fields as well as how mutations and ligands influence VSE measurements [38, 88, 159, 226].

Beyond pure MD studies, quantum mechanics/molecular mechanics(QM/MM) simulations have provided highly accurate benchmarks for VSE studies. For example, one study by Wang et. al. took frames from QM/MM simulations and mapped them to experimental observations [296]. This study was key in determining an interplay between electric fields at protein active sites and enzyme activity. Another study, by Hammes-Schiffer et. al., used QM/MM to *in silico* predict experimental shifts in system energies in KSI [168].

Empirical Valence Bond(EVB)

Empirical Valence Bond(EVB) theory provides a method for comparing dynamical energies of chemical reactions in the condensed phase to those in solution [306]. The method involves the construction of a pseudo-Hamiltonian mapping diabatic covalent and ionic states to their respective empirical energies (including their respective couplings). Here the interactions with the environment are treated as entirely electrostatic and thus interact with the ionic states while preserving covalent ones untouched. This method builds on valence bond theory by integrating empirical measures from quantum chemical calculations or experiment - thus providing a practical tool for studying bonding in enzyme systems [187]. The framework can thus study fields as interactions with ionic states. Similar to VSE, dynamics are treated as a convolved property across protein motions (and thus varying intrinsic fields).

EVB is ubiquitous in the study of electric fields in proteins including its initial application to study the enzymatic mechanism of lysozyme [306]. Here the decomposition of the reactants into ionic states demonstrated as strong interplay between transition states and the surrounding

ligand environment. Further studies have established the relationship between electric fields and several metrics such as pKas, reductions potentials, binding specificity, dynamics and so forth [3, 11, 179, 223]. Given this wide-breath, it is not surprising that EWB has been functionalized within the context of rational enzyme design [98]. Given the relatively-muted success of most computationally designed design efforts, EVB, and electric fields more generally, are being explored as further optimization targets for functional proteins. Several examples, including studies on Kemp eliminase demonstrated novel functionalities and improved activites via electrostatic interactions. This included a remarkable study that showed electrostatic preorganization energies contributed a 27.4 $\frac{kcal}{mol}$ stabilization in the top performing variant [92, 165].

1.2.3 Computational Field Representations Beyond Single Points

Several of the aforementioned studies relied on single-point analysis of induced electric fields. One study projected electric fields along several important bonds to design improved mutants of Kemp Eliminase [32]. Outside of proteins, Shaik et. al. [194, 297] applied uniform electric fields to study their effects on enationselectivity and activity in Diels-Alder reactions. Head-Gordon et. al [311] analyzed electric field residue contributions in KSI at critical points to illuminate the dynamical effects of fields. This study is one of the few to integrate electric field analysis with dynamic movements to understand how field fluctuations may influence protein behavior.

Induced electric fields, at a point, are represented by 3-D vectors. Extending this to a domain, say a cubic 3-D box, each infinitesimal point in real space has its respective 3-D vector. This quickly becomes a problem of sampling or convolutions over this domain as you cannot intake a continuous representation of every point in space. Sampling will require some executive decision around mesh size where input dimensions can rapidly increase to $10^5 - 10^7$ points for each electric field considered. This dimensionality complication motivates the use of single, chemically-intuited, points for analyzing induced electric fields in a protein

active site [41, 311]. Despite this, single-point analytics have demonstrated their inability to accurately describe heterogenous fields present at protein active site [54, 126, 128, 285]. Here we discuss two approaches our group has used to iterate on single-point approaches: QTAIM as a reported for electrostatics and a topological analysis of electric field slipstreams within a box.

QTAIM

QTAIM was outlined in the previous introductory section1.1.3 and thus we will avoid introducing it here. Instead we will briefly mention on how it has probed electrostatic interactions in proteins and small molecules. Our group has leveraged QTAIM as a quantum mechanical reporter of electrostatics [95, 198, 284]. Studies here have included a work linking descriptors at bond and ring critical to changes in density induced by applied, linear electric fields [95]. This work further established a relationship, via QTAIM, between applied fields and reaction barriers in KSI. Another study by Valdez et. al. simulated carboxypeptidase A (CPA) mutants, via QM/DMD, to map QTAIM parameters between activate sites with bound substrates to transition state QTAIM parameters - thereby establishing a link between electrostatics and activity for this set of proteins [284]. Finally, my own work [287] used QTAIM descriptors to train a supervised machine learning model that correctly predicted the relative activity of 3 Diels-Alderases.

Topological Representations - CPET

Our group previously developed a distance metric to measure differences between 3-D electric fields, dubbed Classical Protein Electric Field Topology (CPET) [127]. This formulation enjoys important mathematical properties such as rotational, scalar, and translational invariance - essential properties for describing dynamical structures. Our group has used this approach in a slew of different applications including the reactivity of KSI, Diels-Alderases, protoglobin, and natural hemes [54, 126, 128, 285].

The method samples points within a rectangular prism where linearizations of the electric field are computed and followed to calculate curvature. These lines are known as *streamlines, r(t),* and provide a highly parallelizable compute unit, as each streamline can be calculated independently. We compute the curvature (κ) at the beginning and end of these streamlines with:

$$\kappa = \frac{\left|\left|r^{'}\left(t\right) \mathrm{x} \, r^{''}\left(t\right)\right|\right|}{\left|\left|r^{'}\left(t\right)\right|\right|^{3}} \tag{1.6}$$

Mean curvature values of the start and end points are compiled across each individual streamline along with the Euclidian distance between the start and end points to yield a histogram distribution of curvatures and mean distances for each electric field, a form of topology. This method computes the pairwise distance between two such normalized distributions (f, g) via the χ^2 distance across N bins:

$$\chi^2 : D\left(f, g\right) = \frac{1}{2} \sum_{i=1}^{N} \frac{\left(f\left[i\right] - g\left[i\right]\right)^2}{f\left[i\right] + g\left[i\right]} \tag{1.7}$$

With a defined distance comparing electric fields we can then create a graph where the edge lengths are the distances between two electric fields. This method requires the user to specify several parameters, including box size (Å), number of streamlines, and the step size (Å) for each linearization step along a streamline.

1.2.4 What Do We Need?

Here I hope to motivate the difficulty and importance of establishing rich, informative descriptors for electrostatics in protein analysis. First, these descriptors should move beyond the notion of simply probing field magnitudes at a single point, say, at a metal center. Chemical reactions are driven by distal factors that bring reactants together, stabilize transition state complexes favorably, and finally, yield products. This is a dynamic, multi-dimensional process that cannot be parsed with such simple descriptors. Here I have introduced the CPET

method developed by our group but in chapter 4 I will discuss how we used it to create dynamical probes of electric fields in a protoglobin directed evolution study. An alternative is introduced in chapter 5 where 3-D electric fields, in a volume, are processed across a dataset or MD trajectory via principal component analysis (PCA). This yields important motifs across the analyzed electric fields as well as an intermediate dimensionality between a single point and a full 3-D electric field analysis that is more amenable to human interpretation.

An ongoing debate in the field is whether dynamics contribute meaningfully to reactivity [145]. Warshel famously stands against this notion while some have identified protein promoting vibrations linked to reactivity [13, 94]. The fact remains: non-static charged atoms will yield non-static fields, and therefore, I developed analytical techniques (Chapters 4 and 5) that both provide information on the electric field as the protein evolves with time. We use dimensional reduction and compression algorithms to wrangle and interpret the immense space of heterogenous fields at an active site along an MD trajectory. Notably, we determine that comparatively rare field configurations can yield dramatically catalytic reaction profiles. These "black swan" configurations, and their respective importance suggest that new analytic techniques should work with the varying electric field.

Many proteins, including those with metal cofactors, require more complicated electrostatic simulations including quantum mechanical treatment of metal coordination, backbone sampling, and polarization [284]. With these, approaches such as QTAIM, may be preferable to classical electric fields. The trade-off with these approaches is the computational cost of QM/MM or full QM methods to compute electronic density. Perhaps here, approaches that leverage machine learning to calculate QTAIM descriptors would be critical.

Finally, *de novo* campaigns should be built to leverage the rich set of tools available to understand protein activity - including (but not limited to) electrostatic analysis of a protein scaffold on an active site. These campaigns can look to other "handles" for tuning proteins including long-range effects, entropic effects, and dynamics. In chapters 4 and 5 we introduce methods to merge electric field analysis with dynamics - these could serve as one

such approach to merge dynamics with electrostatics but could also include more traditional energetic and distance-based approaches for tackling protein design.

Chapter 2

Machine Learning to Predict Diels–Alder Reaction Barriers from the Reactant State Electron Density

2.1 Introduction

For any reaction, we are typically interested in the transition state (TS), activation energy, and potential energy surface [131]. We often want to know how various alterations from the base reaction, or modifications of a catalyst, or reaction conditions might alter TS structures and the forward rate of reaction. Despite a wealth of different approaches, the scaling of this process with system size is poor. At the same time, it is often of interest to quickly

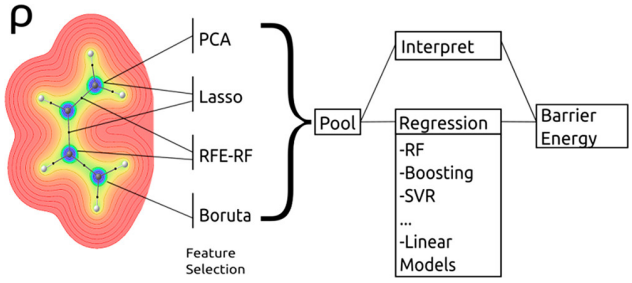


Figure 2.1: Scheme for QTAIM-Machine Learning Prediction of Diels-Alder Reaction Barriers

predict many barriers for many variations of the same reaction. Thus, being able to quickly screen reactants, reactions, and potential catalysts and accurately predict barriers without expensive TS calculations would greatly accelerate the chemical discovery process. The problem lends itself well to the realm of machine learning, particularly for extensively studied reactions. A few pioneering studies have applied machine leaning to reactivity predictions, albeit with limitations in the diversity of the data sets, quality of the fits, and/or eventual performance [90, 141, 196, 262, 292]. Here, we propose a direct prediction of the reaction barriers through quantum electronic descriptors of the reactant state: the electron density, $\rho(r)$, and its derived mathematical properties. We are building on the following previous findings: our previous work on the Ketosteroid Isomerase enzyme and its mutants [97] and the Diels–Alder reaction [128], with and without external electric field applied, have shown robust linear correlations between topological features of $\rho(\mathbf{r})$ and $\Delta G^{\ddagger}$. Furthermore, there exist works that construct linear QSAR models based on $\rho(\mathbf{r})$ to determine chemical parameters such as pKa [55, 191], binding energies [190], bond dissociation enthalpies [263], and cytotoxicity [181]. Additionally, previous studies have used topological quantities of $\rho(r)$ to predict reactivity. [144, 158, 200] This study expands on these previous works by considering a much larger host of variables and, to the best of our knowledge, is the first work in predicting reaction barriers of a family of reactions altogether. In addition, this work seeks to model more complex, nonlinear phenomena using modern machine learning algorithms. Finally, and centrally, according to the Hohenberg–Kohn theorem, [129] the total energy of the system is given as a functional of $\rho(\mathbf{r})$. We extend these ideas toward proposing that reaction barriers correlate with a set of features of the reactant state $\rho(\mathbf{r})$, which, conveniently for machine learning, are continuous and physically meaningful.

2.2 Methods

2.2.1 Density Functional Calculations

All QM calculations for the machine learning algorithm were performed in Gaussian 09 [91]. Geometries were optimized with the B3LYP functional [27, 169, 270, 293] and 6-31G* basis set. [72,121] The B3LYP functional is known to perform well for the Diels–Alder reaction; however, it has also been shown to overestimate the barrier for polar cycloadditions [79]. TS geometries were taken from the literature, and an IRC calculation with the local quadratic approximation algorithm was performed in the gas phase. We then computed the corresponding activation energy and constructed our data set from these values. QTAIM analysis of the electron density generated from Gaussian was performed using the AIMALL software [149].

2.2.2 Molecular Dynamics

A total of five replicate QM/DMD trajectories were run for each Diels–Alderase mutant, with each trajectory corresponding approximately to 15 ns. For a detailed description of the QM/DMD method, we refer the reader to [265]. CE20 QM/DMD trajectories started from the 4O5T crystal structure [224]. Mutations were performed on this structure to generate the CE11 starting structure. Residues included in the QM active site were chosen based on if they provided hydrogen bonds to the substrates or steric interactions for proper substrate alignment. All QM calculations during QM/DMD were performed with Turbomole (version 6.6) [6, 7, 279, 291, 309] with the pure meta-GGA TPSS functional [123] with D3 dispersion correction [108]. All atoms were treated with the double-ζ def2-SVP basis set [309]. The conductor-like screening model (COSMO) [156] with a constant dielectric of 4 was used to approximate the screening and solvation effects from the protein scaffold in this buried active site. [246] πDMD [225, 261] was used for DMD within QM/DMD. πDMD uses an implicit solvent along with discretized potentials.

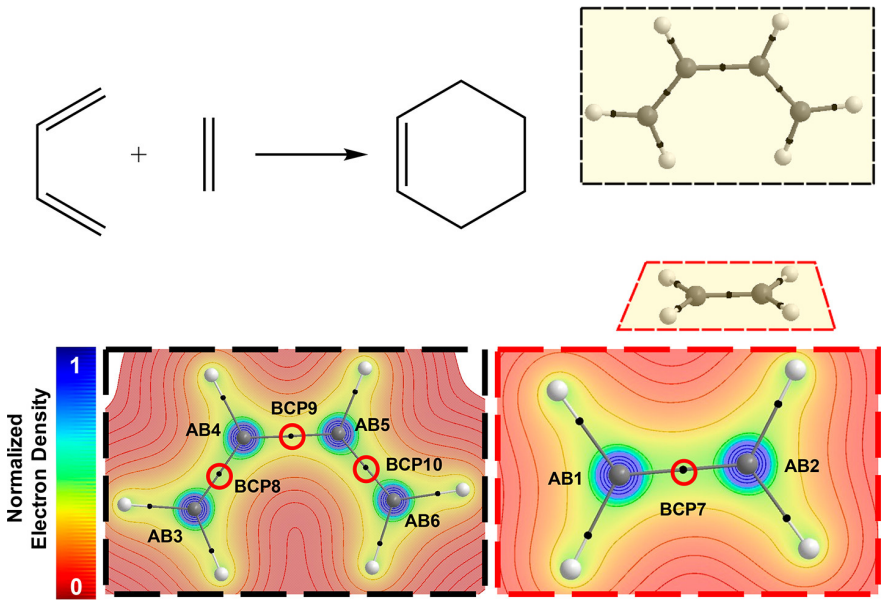


Figure 2.2: Backbone of Sampled QTAIM features

2.2.3 Quantum Theory of Atoms in Molecules

We computed $\rho(\mathbf{r})$ in the reactant state and thereafter calculated QTAIM values on these densities, a mathematically rigorous partition of the electron density into disjoint regions called atomic basins (AB), Ω. Ωs are defined by zero-flux surfaces, $S(\Omega)$, where the normal vector at any point on the surface is orthogonal to the gradient of the electron density2.1.

$$\rho(\mathbf{r}) \cdot \mathbf{n}(\mathbf{r}) \text{ for all } \mathbf{r} \in S(\Omega) \tag{2.1}$$

There are four types of critical points (CPs) of $\rho(\mathbf{r})$: nuclear (NCP), bond (BCP), ring (RCP), and cage (CCP). Each CP is defined by the curvatures of $\rho(\mathbf{r})$ at that point. A NCP is a maximum in all three spatial directions, a BCP is a maximum in two spatial directions and a minimum in one spatial directions, a RCP is a maximum in one spatial direction and a minimum in two spatial directions, and a CCP is a minimum in all three spatial directions.

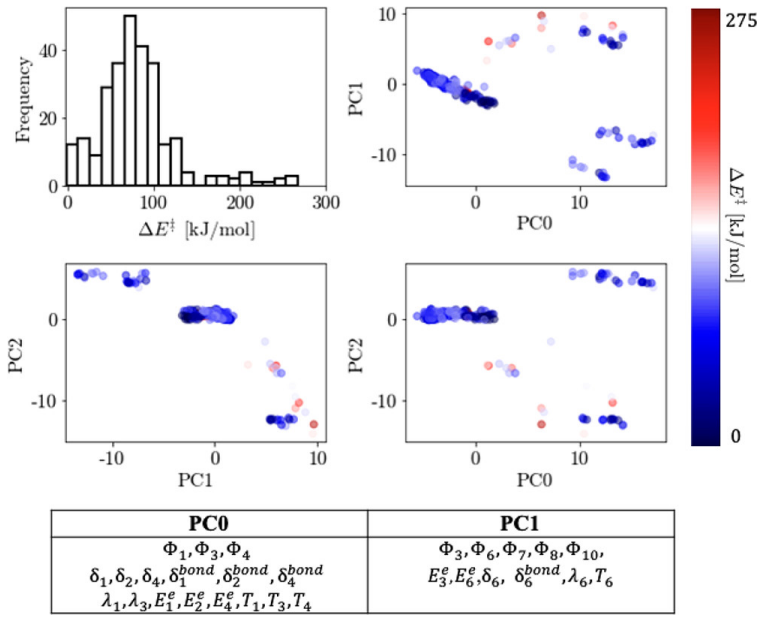
There are four types of critical points (CPs) of $\rho(\mathbf{r})$: nuclear (NCP), bond (BCP), ring (RCP), and cage (CCP). Each CP is defined by the curvatures of $\rho(\mathbf{r})$ at that point. A NCP is a maximum in all three spatial directions, a BCP is a maximum in two spatial directions

and a minimum in one spatial directions, a RCP is a maximum in one spatial direction and a minimum in two spatial directions, and a CCP is a minimum in all three spatial directions.

2.2.4 Dataset

A vast array of scientific literature detail's reaction mechanisms and barriers for important reactions, such as the Diels–Alder family of reactions. We utilize computational data on the Diels–Alder reactions collected from over a dozen articles as our case study [105, 106, 118, 138, 143, 167, 172, 173, 177, 178, 214, 215, 219, 329]. We first recompute the reaction barriers with a standardized basis set and functional to reduce artifacts generated from using a different level of theory; then, we use the quantum theory of atoms in molecules (QTAIM) [15] to generate topological parameters of $\rho(\mathbf{r})$ from our computed reactant state structures(Fig. 2.2). Jointly with more traditional descriptors, such as system mass and charge, they constitute input variables. These two sets were used to train both feature selection and regression algorithms. Feature selection was used primarily to determine a subset of factors that are essential for computing barrier energies, while also reducing dimensionality of regression algorithms and mitigating noise. This reduced space was then used to train regression algorithms that approach DFT accuracy while requiring a fraction of the compute time to find a reaction barrier. We then verify the utility of this method, including for a related but substantially more complicated system: two artificial Diels–Alderase enzymes separated by eight mutations (introduced through laboratory directed evolution) [224].

The compiled data set consists of 296 Diels–Alder reactions from over a dozen different sources, including reactions with a diverse set of functional groups, sizes, and geometries(Tab. A.3). While the canonical Diels–Alder reaction features the formation of two new C–C bonds with four new stereocenters, our data set also includes hetero Diels–Alder cycloadditions, with nitrogen and oxygen as possible heteroatoms. The reactions also encompass a large diversity of electronic barriers, with a minimum barrier of 5.6 kJ/mol (1.3 kcal/mol) and maximum of 274.5 kJ/mol (65.5 kcal/mol)(Fig. 2.3). The majority of the reactions have a barrier within



PC0	PC1
Φ_1, Φ_3, Φ_4 $\delta_1, \delta_2, \delta_4, \delta_1^{bond}, \delta_2^{bond}, \delta_4^{bond}$ $\lambda_1, \lambda_3, E_1^e, E_2^e, E_4^e, T_1, T_3, T_4$	$\Phi_3, \Phi_6, \Phi_7, \Phi_8, \Phi_{10}$, E_3^e, E_6^e, δ_6, $\delta_6^{bond}, \lambda_6, T_6$

Figure 2.3: Dataset Distribution w/ PCA components

the range of 50 to 150 kJ/mol (12 to 35.9 kcal/mol), while higher/lower reaction barriers are underrepresented within the data set. Our data set only includes Diels–Alder reactions that proceed via a concerted mechanism and does not include reactions that proceed stepwise.

2.2.5 Machine Learning and Feature Importance

Feature Selection

Feature selection reduces the dimensionality of an input space and can result in improved accuracy in regression and classification tasks by removing noise in training data. This is especially important in the low-to-medium regimes of data where noise can be severely detrimental to the accurate learning of a data set. Furthermore, decreasing the dimensions of an input space increases algorithm performance and can allow further tuning of regression/classification algorithms for a given amount of computation time. For our purposes, feature selection is also beneficial as it allows for greater interpretability in models and can inform physical understanding of how electron density properties determine reaction barriers. Wrapper, embedded, and filter feature selection algorithms were tested. Filter methods

compute statistical features of the data set including variance or correlation to a target variable. Wrapper and embedded methods both rely on training a regression or classification algorithm that reports feature importance. Wrapper methods rank features and iteratively remove them, a new model is then retrained and tested on the new subspace of input features. This process is repeated until a cutoff criterion is reached. Embedded methods generally take one training instance to select features. In our trials, wrapper feature selection algorithms and embedded schemes proved more stable in their selection despite variations in algorithm hyperparameters. Once feature selection was completed using several different algorithms, the resulting variables were compared and common features between algorithms were used to construct accurate regression algorithms.

LASSO

Least Absolute Shrinkage and Selection Operator Regression (LASSO) regression adds a regularization term to the cost function of the least squares optimization/fitting problem. The LASSO regression method consists of the following general loss function:

$$\sum_{i=1}^{n}(h(x_i) - y_i)^2 + \alpha||w||_1 \tag{2.2}$$

Equation (1) - The first term corresponds to the standard least squares' regression term with $h(x_i)$ as the predicted value of a response variable y_i is a scale parameter appended to the coefficient vector w [102].

This algorithm appends the L_1- norm of the weight vector to the cost function. This choice of a linear term specifically brings model weights towards zero for large enough, thereby eliminating non-important terms and performing feature selection. The choice of α was made through cross-validation using sklearn's LASSOCV module. This method has been used in feature selection on chemical systems prior to this study [139].

Recursive Feature Elimination

Recursive Feature elimination was also used to compile a set of statistically relevant features. The algorithm is simple as it only requires the successive training of a regression or classification algorithm that determines feature importance for a given training instance. These training instances report feature importance and the variable with the lowest rank is removed. This step is reiterated until a user-defined limit is reached. This flexibility makes this algorithm easy to tune and use, furthermore a fixed number of features can be selected for. This work used random forests of varying depths from 3 to 7 features due to computational cost and consistency. This algorithm has been used in bioinformatics [12] and agroindustrial [107] modeling with Support Vector Machine and Random Forest base regressors. Random Forest algorithms usually compute feature importance through gini importance or means decrease accuracy. Mean decrease accuracy scrambles values for a given variable across different samples and computes the loss in training accuracy. More important features result in a higher loss in accuracy when scrambled. GINI importance, which was used in this work, is computed as the loss/gain in regression variance when a variable is removed from training

Boruta

The Boruta algorithm is a wrapper feature selection algorithm that finds all relevant features within a model. The underlying mechanic removes features iteratively by fitting input variables to a random forest classification/regression algorithm and extracting feature importance from this trained fit. Some other feature selection algorithms, e.g. Sklearn's SelectFromModel, filter features based on a threshold level of feature importance on a trained model. The selection of this threshold can yield different results and is somewhat arbitrary. Boruta differs from similar methods by constructing "shadow" inputs to a regression model. These variables are constructed by shuffling values between input samples and appending these new variables to the input vector. These inputs should result in nonzero importance values only due to random noise in the input data, this serves as a baseline for the original input variables to

determine which variables were truly important in model performance. This algorithm finds all relevant features to an output variable and therefore is a viable option for our purpose in trying to elucidate important features in the large input space of our initial model [163]. The Boruta algorithm repeats this process on different shuffles in the shadow input space to ensure rejected/accepted variables are correctly sorted due to statistical importance, not just variability in data. This method has been used for feature selection in a multitude of applied scientific uses including biomechanical studies [220] and biomarker detection [84].

PCA

Finally, to determine how many independent variables might be needed to explain variance in the barrier energy, principal component analysis (PCA) was performed on a sweeping number of different components. PCA projects data into an orthogonal basis and, in the processing, groups heavily correlated variables. Specifically, PCA decomposes a data matrix X into three matrices:

$$X = U\Sigma V^t \tag{2.3}$$

The resulting matrices consist of, a diagonal matrix of singular values, U, the left singular matrix of X or the original basis of data V, and V consists of the unit vectors of the principal components. Finally, these components are sorted in descending order based on their eigenvalues which can be quantitatively compared to determine the importance of each feature. As mentioned above, lower model dimensionality can reduce noise, increase interpretability, and prevent overtraining on correlated variables that existed in the original model [76].

Permutation Importance

The final feature selection method, permutation importance, was used to establish a quantitative measure of which features were more important versus each other in the final physical feature set. Permutation is a simple, yet powerful method that trains an arbitrary regres-

sor/classifier then retrains the same algorithm but randomly permutes individual features between different samples. This algorithm will permute a single feature at a time and compute the change in a given predictive metric (mean average error, mean squared error, etc.). This sample-resample is repeated on each variable for a user-specified number of trials [36]. Features that more negatively impact the metric are deemed to be more essential and thus help establish interpretability in a quantitative sense for the construction of a physical model.

Implementation Details

A command-line interface was created to allow for rapid testing of different regression algorithms, including the choice of algorithm, number of Bayesian optimization instances, and what subspace of the features to use. Most regression and feature selection algorithms were created using the Scikit-Learn package. The notable exceptions were some of the neural network methods tested [1], Gaussian Process Regression [71], and the GPU-enable XGBoost algorithm [58] which each had their own respective methods. Tensorflow was used to test neural networks with more customizable features, namely dropout layers. Bayesian hyperparameter optimization was performed using skopt with custom dictionaries [120]. Each algorithm type was tuned for 25 training instances per tunable feature. Custom cost functions were implemented for XGBoost, Random Forest, Gradient Boost, and Extra Trees Regression algorithms. The purpose of this was to output the Mean Average Error, Mean Squared Error, and L_2 of both trained and withheld datasets for training, thus allowing for the optimization of commonly used Mean Squared Error while reporting interpretable Mean Average Error. XGBoost was used following the Python XGBoost implementation, the GPU-enabled version was used [58]. Recursive feature elimination, LASSO, and PCA methods used were the Scikit-Learn implementations and the Boruta algorithm used was from the BorutaPy implementation [163]. For Borutapy and Recursive Feature Elimination, Random Forest Algorithms of 3, 5, and 7 features were used and the features from each instance were collected. The top 20 features in Recursive Feature Elimination were selected. Visualization

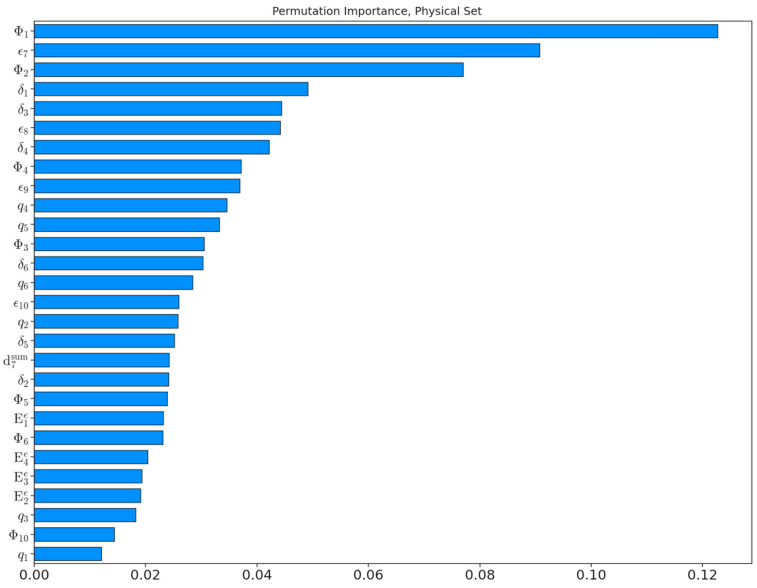


Figure 2.4: Permutation Importance of Different QTAIM Features

was performed using the seaborn package [307]. A 80-20 split of training-testing data was used to create the regression algorithms with a redefined random seed. The implementation for feature selection, regression, and plotting functions along with the dataset we created can be found here

2.3 Results and Discussion

Diels-Alder Models and Feature Importance

First, to visualize the input space of this model and understand how variables correlate within the data set, principal component analysis (PCA) analysis was performed. Along the first three principal component axis, we see that there are no apparent gradients for increasing/decreasing barrier energies(Fig. 2.3). Both high and low $\Delta E^{\ddagger}$ appear to be spread out throughout the component space implying that this data is nonlinear and that linear models might not be suitable for regression. However, the first three components only explain 50% of the variance in the data, and to account for 95% of the input space variance, 38 orthogonal components are needed. The first two eigenvectors are shown, and there is a heavy

concentration of diene variables in the primary principal component and a strong contingent of dienophile components in PC1, showing the independence between these two variable sets. We also note the almost complete set of Φ between these two components supports the notion that electrostatic potential is an important value in this quantitative structure activity relationship (QSAR) analysis.

To construct regression models, we pooled the variables (this set is labeled as "raw pooled features" in this text) selected by the three feature selection algorithms: LASSO, Boruta, Recursive Feature Elimination (Sec. 2.2.5, 2.2.5, 2.2.5) for a detailed description of each of these methods). In addition, permutation importance was used to remove multicollinear features and to gain a robust measure of feature importance relative to each other. Coupling the results from the raw pooled feature selection algorithms to the ranked list of features from the permutation ranking(Sec. 2.2.5), Φ (including both Φ_{nuc} and Φ_e) and Bader charge (e) appear to be the most physically important set of descriptors from a statistical standpoint(Tab. A.5). The permutation ranking of features from the physical data set is shown(Figs. A.1, A.3), and the permutation ranking for features in the full pooled data setA.2.

The fact that electrostatic potentials and electron density curvatures affect the Diels–Alder reaction barriers is physically meaningful. Within DFT a localized potential is used to express the potential energy in solving the one-electron Schrödinger equation, which is the sum of the external potential ($v_{ex}(\mathbf{r})$), Hartree electron–electron interaction potential, and exchange-correlation potential(Eqn. 2.4).

$$v[\rho] = v_{ex} + \int d\tau' \frac{\rho(\mathbf{r}')}{|\mathbf{r} - \mathbf{r'}|} + \frac{\delta E_{ex}[\rho]}{\delta \rho} \tag{2.4}$$

$v_{ex}(\mathbf{r})$ is the potential created by the nuclei and is exactly equivalent to Φ_{nuc}. Similarly, the middle term is exactly equivalent to Φ_e. Thus, our selection algorithms have picked out that the potential, which specifies the system's Hamiltonian in the reactant state, is also deterministic of the energy of the system at the TS. Furthermore, it seems that it is enough to know only the potential energy and contribution from the nuclei and electrons separately

feature	**type**	**raw, pooled**	**pooled, uncorrelated**	**physical**
1	AB	q, E_e, Φ, λ, T, δ, δ_{bond}	q, E_e, Φ, δ	q, E_e, Φ, δ
2	AB	q, Φ, δ_{bond}	q, Φ	q, E_e, Φ, δ
3	AB	E_e, Φ, λ, T, δ	E_e, Φ, δ	q, E^e, Φ, δ
4	AB	q, E_e, Φ, δ_{bond}, T	q, E_e, Φ	q, E^e, Φ, δ
5	AB	q, Φ, Φ_{nuc}, λ, δ	q, E^e, Φ, δ	q, Φ, Φ_{nuc}, δ
6	AB	Φ, λ, δ, δ_{bond}	Φ, δ	q, Φ, δ
7	BCP	ϵ, d_{sum}, d', d	ϵ, d_{sum}	ϵ
8	BCP	ϵ	ϵ	ϵ
9	BCP	ϵ	ϵ	ϵ
10	BCP	Φ_e, Φ	Φ_e, Φ	ϵ, Φ
total features		38	24	28

Table 2.1: **Variables Collected by Each Feature Selection Algorithm:** Features included in several algorithms that completed a set of variables were pooled to construct regression algorithms. Beyond that, features selected were used to gain physical insight and build a more general physical model. ϵ: bond ellipticity, T: electronic energy of molecule, E_e: contribution of atom to electronic energy, q: electronic charge, σ: stress, Φ: electrostatic potential, δ: delocalization index, δ_{bond}: bond delocalization index, λ: localication index, d: average number of electronic pairs formed in atom a, d': half of average number of electron pairs formed between atom A and other atoms of molecule, d_{sum} sum of d' and d.

at these nuclei and CPs, rather than the full function, to approximate the change in electronic energy at the TS. In conjunction with the electrostatic potential, the ellipticity (ϵ) at the majority of the BCPs was also selected as an important feature(Table A.4).

$$\epsilon = \frac{\lambda^{(1)}_{\mathbf{H}\rho}}{\lambda^{(2)}_{\mathbf{H}\rho}} - 1 \tag{2.5}$$

ϵ is a measure of the elliptical nature of the density within the plane orthogonal to the bond direction(Eqn. 2.5). Generally, ellipticity can be a measure of the π-character in the bond, as double bonds lack symmetry of the electron density around the bond axis, whereas axial symmetry is present for σ-bonds. Since the Diels–Alder reaction is often rationalized through the interaction between the frontier orbitals (π-orbitals), it makes physical sense that ϵ should be a strong determinant of the barrier.

Models were trained using the features selected from the selection algorithms, with an

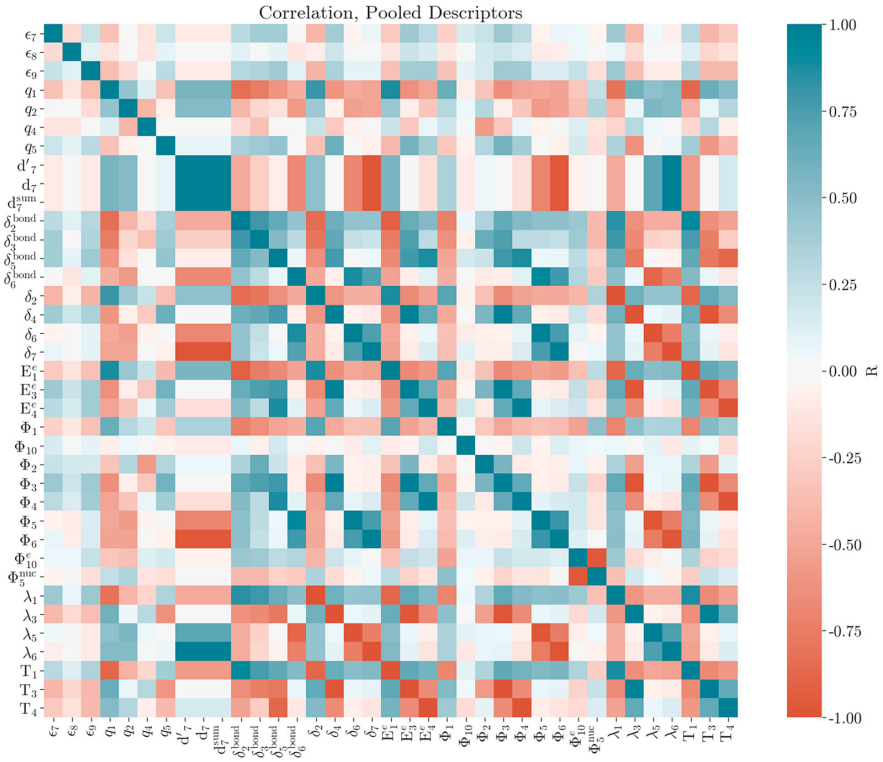


Figure 2.5: Correlation of QTAIM Features

addition of "missing features" that completed the physically meaningful set. For example, if feature selection algorithms determined that a given feature was important in all but one CP or AB, we "completed" the set by including this missed feature. The compiled data set of 38 variables still presented a large input space relative to the size of the data set; therefore, we wanted to further reduce the number of input variables. Heavily correlated features, as computed through a Pearson correlation coefficient with a magnitude above 0.8, were removed and yielded a reduced subspace of 24 variables; features with the highest permutation score were kept, while lesser important correlated features were removed(Fig. 2.5). This reduced data set (labeled "pooled, uncorrelated features") was used to train benchmark regression algorithms. The removal of heavily correlated features can be important, not just in reducing model training times (and thereby allowing the testing of more hyperparameter sets for a given computational cost) but in creating more stable, generalizable models; multicollinearity can yield models that overfit one set of highly correlated features [325]. Here we see that

physically related descriptors are often correlated with each other. For example, d_7, d'_7, $d_{7,sum}$ are all definitionally related as the latter is the sum of the former two values. In addition, some identical variables at different features also correlate heavily, as was the case with Φ at the two of the dienophile nuclei (which makes chemical sense).

The input space of uncorrelated variables was used to train a diverse array of algorithms optimized for their mean squared error to barrier energies. Performance metrics on withheld data are reported(Fig. 2.6). We see that all linear models (LASSO, Ridge) perform quite poorly, confirming the complex nature of the input space to these models. Tree based regressors (XGBoost, Gradient Boost, and Extra Trees) performed quite well, all of which achieved correlations above 0.8 on the validation set. This is not surprising as these models are quite flexible and consist of tunable parameters to prevent overfitting. Extra Trees and Gradient Boost both performed well versus other regression algorithms, withheld data, and had a baseline metric of guessing the mean barrier energy of the data set for every instance(Tab. A.2).

Beyond training the best performing model, we wished to create a more general and physically intuitive regression algorithm for predicting instances outside of our data set. To do this, we completed sets of physical features labeled as "physical feature set" by adding back some of the physically meaningful though possibly correlated variables(Tab. A.4). For example, bond ellipticity, ϵ, was originally selected in three of the four BCPs as an important feature; in the completed/physical set of variables we included of all four BCPs. In principle, reintroducing correlated variables and statistically unimportant variables would increase training loss and reduce performance metrics, but we benchmarked models trained on this data set and determined that there was almost no loss in performance(Tab. A.2). In general, these best performing algorithms were quick and accurate and could effectively be used to circumvent more expensive barrier calculations for this family of reactions.

Beyond predicting the overall barrier energy of any given Diels–Alder reaction, this model would be more practical if it were able to predict the relative energies of endo/exo reaction

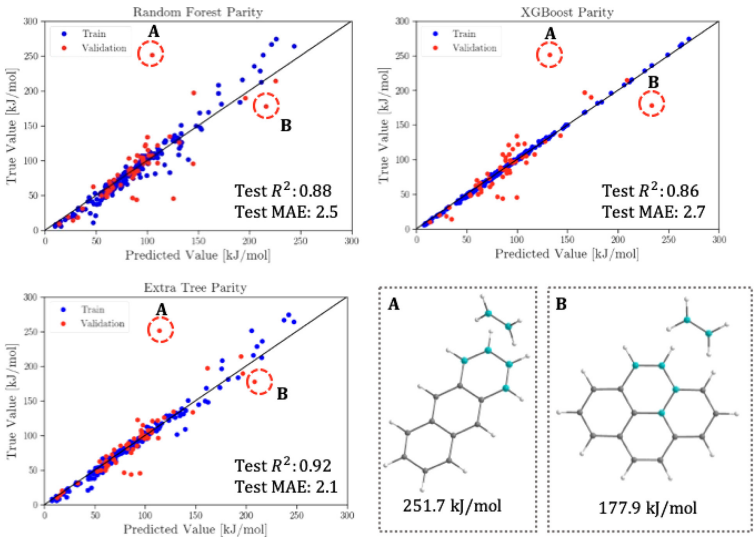


Figure 2.6: Top QTAIM-ML Model Performance

pairs and thereby predict the preferred reaction product of a Diels–Alder reaction. Our data set contained a mixture of such reaction pairs, but about half of the reactions available did not have the corresponding alternative reaction. In total, our data set contained 61 endo/exo pairs or 122 compounds. This represents less than half of the total available data set, and therefore the process of training is more difficult. To fully extend this aim, we would likely require more data, but we nonetheless retrained the best model above, Extra Trees, with a physical feature set and an 80–20 train-validation split. Our splitting scheme kept endo/exo reaction pairs in the same data set to allow for comparison after regression. We opted to avoid further hyperparameter tuning and simply reuse the model parameters from the previous models for simplicity, and therefore a test set was not used. On the validation set, the Extra Trees regression algorithm was able to correctly predict endo/exo ordering 70% of the time, although this figure could likely be improved with more data.

Next, we wish to understand the limitations of our regression models, including regimes where their predictive ability falls short. From the top four regression algorithms, we noted two data points with barrier energies of 251.8 kcal/mol (60.23 kcal/mol) and 177.9 kJ/mol (42.56 kcal/mol)(Fig. 2.6) that contributed heavily to training loss in all instances. The consistently large error for predicting these values across different families of algorithms

required further probing into the physical reasons yielding such poor performance. First, these data points fall in the underrepresented high-barrier region, where the model might have had insufficient training instances. Figure 2.6 shows the two systems responsible for these two largest testing residuals. Notably, these systems involve dienes with more delocalized -systems, and thus, the electronic density shifts during the Diels–Alder reaction within these systems extend over the entire conjugated -system of the diene. Hence, more bonds change order than in our descriptor set, and the set of mathematical features at just 10 features may prove limited. There are other conjugated systems, in both the training and test set data, but the two outliers feature the greatest extents of -delocalization. It must be noted that QTAIM properties are computed on optimized reactant geometries; therefore, our method is not agnostic to the shortcomings of the DFT methodology and basis sets, and poorly performing methods may reduce the performance of machine learning models. Our data set also includes other regioisomers for the reaction occurring in Figure 2.6, with the preferred regioisomer being the [5, 10] addition and the least preferred being the [12, 14] addition [118]. Upon testing with our best performing algorithm, Extra Trees, we can correctly predict that the [12, 14] addition is still least preferred, and the [5, 10] addition is most preferred. Hence, our algorithm, while it may not accurately predict the barrier for the [12, 14] addition, still predicts the correct regioisomer.

Diels-Alderase Study

Finally, we put the model to a stringent test and probe its expandability to a considerably more complicated regime of enzymatic catalysis, where calculating the barriers is indeed very challenging. Since the model was trained on reactions in solution, there is no guarantee that it would successfully predict the barriers for the Diels–Alder reaction catalyzed by enzymes. Artificial Diels–Alderases have been designed and undergone laboratory directed evolution to enhance the performance by several orders of magnitude. (87) These enzymes catalyze the reaction between 4-carboxylbenzyl-trans-1,3-butadiene-1-carbamate and N,N-

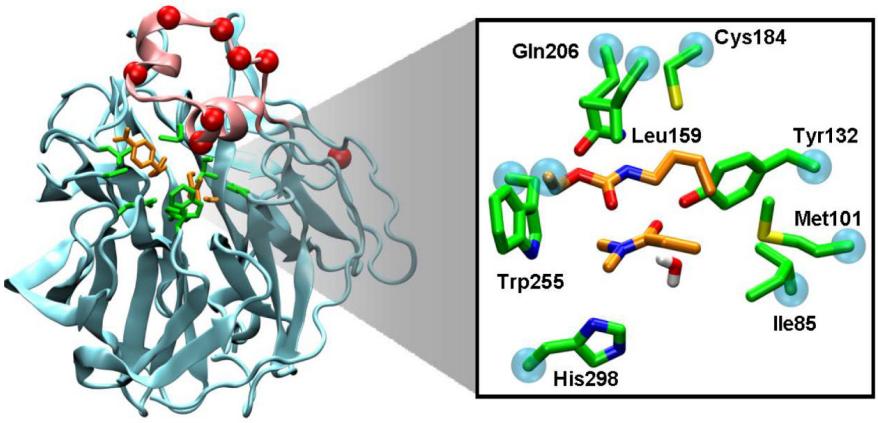


Figure 2.7: Diels Alderases Used for OOD Testing

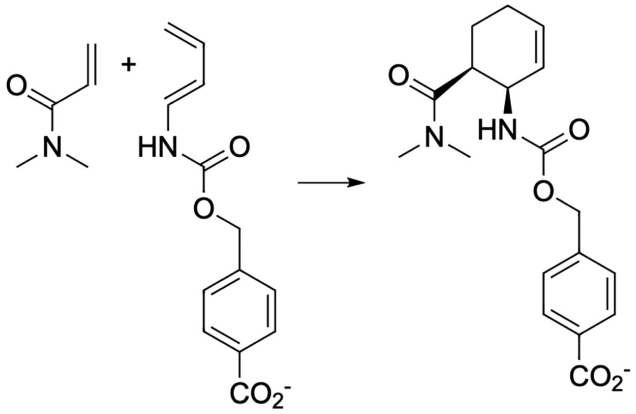


Figure 2.8: Diels–Alder Reaction between 4-Carboxylbenzyl-*trans*-1,3-butadiene-1-carbamate and *N*,*N*-Dimethylacrylamide Catalyzed by the Diels–Alderase Enzymes CE11 and CE20

dimethylacrylamide (Fig. 2.8). Using our top performing regression algorithm, we compare the barrier energies of two Diels–Alderase enzymes at the beginning and end of a directed evolution optimization (CE11 and CE20). There is a total of eight mutations between CE11 and CE20 with the majority being within the appended lid element and none within the active site(Fig. 2.7). Therefore, these mutations represent realistic, subtle changes to the active site electron density topology brought about by distant point mutations through long-range interactions.

We utilized our in-house quantum mechanical/discrete molecular dynamics (QM/DMD) engine [265] to perform sampling of the two protein variants with the bound substrates. The QM active site shown (Fig. 2.7) included Tyr132 and Gln206 which directly hydrogen bond

the dieneophile and diene respectively Additionally, in the crystal structure, a single water molecule was located near the carbonyl on the dieneophile which seemed to be a hydrogen bond donor and was included as well. Using the lowest energy QM active sites from each mutant, we performed the QTAIM analysis to generate the input vector for our machine learning algorithm.

The top-performing Extra Trees algorithm with the physical feature set was used and correctly predicted the ordering of the reaction barriers of these two Diels–Alderases: CE11 should have a higher barrier than CE20, thus being less active. We note that ranking of the artificial enzyme variants in terms of activity is often all that is needed in the protein design and optimization process. Despite the correct ordering of enzyme energies relative to each other, the barrier energy and the gap between them were considerably higher than the values estimated from experiment, 20 kJ/mol (5 kcal/mol) for the difference in electronic barriers, with a difference of 2.2 kJ/mol (0.52 kcal/mol) free energy difference at 25 °C [224]. The difference could arise from several factors including the lesser representation of low-barrier reactions in the training set and the missing entropic contributions to the free energy barrier. In this particular experiment, the choice of feature set did not change the ultimate result as we predicted the same ordering with every feature set. Note also that further investigation upon these structures is warranted to understand how the mutations alter the reaction barrier, though it is outside the scope of this present paper.

2.4 Conclusions

Here we showed that QTAIM descriptors based on the ground state electron density can be coupled to a supervised machine learning algorithm to predict reaction barrier energies. Fundamentally, QTAIM appears to be an ideal tool for feature generation in machine learning as it produces sets of physical, continuous descriptors. As a proof-of-concept, we present this study of Diels–Alder reactions. We computed reaction barriers of a diverse array of

Diels–Alder reactions from the literature and extracted a wealth of electron density and derived mathematical descriptors for their reactant states. This initially massive feature set was refined via feature selection methods to yield an interpretable set of important variables consistent with physical intuition. From there we trained and tuned several regression algorithms with excellent predictive ability based on physical descriptors. Additionally, we were able to qualitatively predict the ordering of activity for two Diels–Alderase enzymes. Thus, we were able to sidestep the necessity of finding the TS geometry to determine the TS energy with this model example. Further, since the electron density is an observable, it is possible to map the electron density experimentally and deduce the barrier directly, without computations or kinetics experiments. Therefore, this study alone could serve as a screening filter for experimental and computational studies on the Diels–Alder reaction. Beyond building a library of barrier prediction algorithms, the proposed descriptor sets could be generalized to a fixed-length descriptor compatible with any molecule, adding to the set of descriptors that might be useful in the chemoinformatics toolkit. Future studies may include building classifier algorithms to bin reactions into categories or test the ability to predict the reactivity for stepwise Diels–Alder reactions using QTAIM features. Preliminary tests with classification algorithms showed promising results with high accuracy and ROC scores, though the problem of data balance remains. We choose to avoid making classifier algorithms as regression algorithms, with a high degree of accuracy, could themselves serve as screening methods for computational chemical applications. In addition, benchmarking versus traditional fingerprinting algorithms would be a useful metric that was not possible as our diverse set of systems included a diverse length of molecular sizes and even number of molecules. Another area of interest is generalizing these descriptors to an arbitrary-size system through perhaps graph representations and corresponding graph neural networks. We do note that BCP, RCP, and CCP can disappear catastrophically (described by catastrophe theory [112]), and hence a given set of CPs may not be uniformly present across all of the systems. If this is the case, then simply supplying the null vector for the features

at that particular CP should allow fixed-length input ML algorithms to work, as well as provide incredibly important information about the system (that is, if a CP is present or not is chemically important information and includes important bonding information for that system). Hence, machine learning on QTAIM CPs can be generalized to include CPs that can disappear catastrophically. To summarize, we show that there appears to be, at minimum, a statistical relationship between the reactant state electronic density and the reaction barrier. Within DFT, the reactant state energy is a functional of the electron density; therefore, we extend this and conjecture that the TS energy is a functional of the reactant state electron density. This is of fundamental curiosity because the ground state density in principle is mostly agnostic to unoccupied states that can be important for reactivity; this could arise as a limitation in similar algorithms for some reactions. Statistical learning algorithms demonstrate a high degree of accuracy in predicting barrier energy from a small set of density descriptors, suggesting an underlying analytic relationship between these variables. This motivates further studies with different reaction families and the development of more generalizable QTAIM descriptors and algorithms.

2.5 Acknowledgements

This chapter was adopted from:

Machine Learning to Predict Diels–Alder Reaction Barriers from the Reactant State Electron Density. S. Vargas*, M. Hannefarth, Z. Liu, A.N. Alexandrova. *Journal of Chemical Theory and Computation* **2021** 17 (10), 6203-6213. 10.1021/acs.jctc.1c00623.

Here M.H. contributed DFT simulations, writing, editing. Z. L. helped with gathering data and writing. A.N.A. provided direction, writing, editing. I contributed code for machine learning, visualization, data processing, writing, and editing.

Chapter 3

High-throughput Quantum Theory of Atoms in Molecules (QTAIM) for Geometric Deep Learning of Molecular and Reaction Properties

3.1 Introduction

The Quantum Theory of Atoms in Molecules (QTAIM) is an illustrious methodology for deriving insight from the electronic density distribution of a molecule. QTAIM assigns the electronic density ρ to particular atoms and delineates bonding interactions between them. By topological analysis, QTAIM partitions ρ into atomic basins bounded by zero-flux surfaces $S(\Omega)$ that satisfy $\nabla\rho(\mathbf{r}) \cdot \mathbf{n}(\mathbf{r}) = 0$. Integrating electronic properties over each enclosed basin yields descriptors such as atomic energies and electron delocalization. Similarly, QTAIM identifies critical points (CP) at nuclei and between them where ρ is maximized according to its second derivative; these nuclear (NCP), bond (BCP), ring (RCP), and cage (CCP) critical points are differentiated by how many dimensions exhibit local maxima in ρ. Properties of the

density measured at CPs provide a compact set of descriptors that encapsulate the molecule's electronic distribution. Furthermore, a unique path of steepest ascent in ρ (i.e. the gradient path) exists from each bond CP to its two adjacent nuclear CPs, thereby linking neighboring atoms with a bonding interaction. In other words, QTAIM gives bonding networks as well as higher-order information about a molecule's electronic structure.

As a density-based theory, QTAIM builds upon either theoretical calculations or x-ray diffraction data, and is thus applicable across computational and experimental disciplines [192]. Exemplar studies utilize QTAIM to understand ligand-receptor interactions in biological systems [241], predict chemical activation barriers [288], describe toxicity [230], and estimate spectroscopic parameters in organic compounds [192]. Tab. 3.1 shows a representative set of descriptors alongside previous interpretations for properties they report on. Given QTAIM's high descriptiveness and prior use in QSAR approaches, we believe that it can be leveraged to improve machine-learned predictions of molecular, protein, and periodic system properties. QTAIM's unique bond definitions, rooted in quantum chemical information, can also serve as powerful alternatives to cheminformatic heuristics such as bond cutoffs [208] for resolving bonding in difficult chemistries involving aromaticity, multi-center bonds, and metals. Several studies have utilized QTAIM as a fine-grained analytical tool in bonding analysis, these include both covalent and non-covalent interactions. Bader previously investivated how Ti bonds to cyclopentadienyl and a substituted dienyl fragment, with QTAIM differentiating whether or not a C bonds to Ti by the presence or absence of a bonding interaction [17]. Farruga et. al. also compared the covalency of transition metal-carbon bonds based on the density and other QTAIM values at bond critical points [81]. Given these examples, we also probe whether QTAIM features could improve performance for datasets containing metals.

Value	Derived Concepts
Electron Localization Function (ELF)	electrophilic/nucleophilic sites [111, 142]
Laplacian ($\nabla^2\rho$)	electrophilic/nucleophilic sites, atomic graph [16]
Electrostatic Potential (ϕ)	interaction strengths w/ nuclei, other electrons [19, 99]
Energy Density	Valence shell polarization [44, 229]
Delocalization Index (ϵ)	π-character [65]
ETA Index	interaction type [205]
Localized Orbital Locator	e localization [249]

Table 3.1: A set of QTAIM features and how they have been interpreted in the past.

Our goal is to merge the interpretive richness and relevance of QTAIM descriptors with powerful geometric learning algorithms. Previous QTAIM/ML approaches incorporated a limited set of hand-selected features based on existing heuristics, [100, 207, 218, 288] and thus, potentially missed leveraging many useful features. With our approach, we integrate a rich set of over 20 atom and 20 bond critical point features for an exhaustive toolkit of electronic descriptors (Tab. B.1). Integrating these features into graph neural networks (GNNs) allows for greater applicability to systems with varying chemical structures and unexplored chemical motifs where heuristics have not yet been developed [115]. In addition, graphs are a flexible data structure that can readily intake spatial information such as atomic positions and/or bond lengths to further inform predictions. Given the power and ubiquity of geometric learning in chemical spaces, coupling them to electronic structure-informed features could extend their applicability and ability to generalize on smaller datasets [25, 101, 104, 320]. Notably, graph neural networks (GNNs) often perform poorly under low data regimes [282,314] — regimes where experimental and high-accuracy quantum chemical calculations may operate and electronic descriptors could offer a strategy for suitable performance. Furthermore, GNNs suffer from poor out-of-domain (OOD) extrapolation and we probe whether QTAIM features can help alleviate this shortcoming [282]. We note one other study [135] that takes a somewhat

similar approach to using QTAIM for geometric machine learning; our work differs by not having benchmarks on standard cheminformatic datasets, testing on spin/charge-varying datasets, testing out-of-domain performance, and providing tools for generating and training QTAIM-informed geometric learning models for both molecules and reactions.

We make a few important advances to the utilization of QTAIM in machine learning. First, we create a set of easy-to-use, pythonic tools for computing QTAIM descriptors at scale and using them for machine learning tasks. These tools include high-throughput job-runners for calculating QTAIM values, visualization tools for descriptive statistics, parsing utilities for compiling data into single data structures, and ready-to-use graph neural network architectures. These tools work together in an ecosystem for harnessing QTAIM in geometric learning. We also compute QTAIM values on several datasets chosen for benchmarking or developing algorithms to handle tricky chemical domains with varying charges, spins, and reactivities. In addition, we benchmark the usage of QTAIM features to demonstrate their ability to improve overall model performance, learning on smaller datasets, and out-of-domain predictions. We hope that these contributions can serve as an important foundation for further studies using hybrid QTAIM/ML approaches to tackle machine learning in difficult chemical domains with experimental or small datasets. In addition, these tools can serve as an important basis for developing more advanced QTAIM-enabled machine learning algorithms.

3.2 Methods

3.2.1 Quantum Chemical Calculations

QTAIM calculations build on top of quantum chemical density calculations. Our package can intake any format compatible with Multiwfn [183] or Critic2 [211] and thus could use a number of DFT codes such as Q-Chem [77] or Gaussian [91]. We use ORCA [204] as it's open-source, free under academic licenses, and implements a wide range of basis sets and levels of theory. For now, we have implemented options files that allow the user to write a

wide-range of custom ORCA input files, including relativistic corrections, individual atomic basis sets, and parallelization options. Generalization to other quantum chemical packages requires new methods for writing input files but otherwise can fit into our ecosystem for high-throughput QTAIM and molecular/reaction graph neural networks. We chose differing levels of theory for our dataset construction considering the relative expense of computed properties in each dataset — we wanted to ensure that the cost of DFT and subsequent QTAIM calculations did not rival the expense of computed properties. We outline the different levels of theory below for each dataset.

3.2.2 QTAIM Calculations

Our current implementation uses critic2 [211] or Multiwfn [183] to handle QTAIM calculations. All datasets here, however, leverage Multiwfn due to its richer set of QTAIM descriptors, including spin information, energies, etc. (Tab. B.1). These calculations intake any density file format supported by Multiwfn including .cube and .wfn files and yield a single text file for analysis.

3.2.3 Dataset construction

We format our datasets into standard JSONs, constructed either by standard tools from rdkit [233] and pymatgen [208] or by our built-in scripts for construction and formatting (Fig. 3.1). These scripts parse molecular charge and spin information from xyz files and produce a database. Initial guesses at bonding can optionally be handled by rdkit. The resulting json includes the following notable data structures in order to write DFT input files and perform subsequent machine learning:

Molecules (pymatgen molecules) - Pymatgen molecules, without bonding information, used to featurize the molecules for machine learning and write input files with coordinates at atomic sites.

Molecular graphs (pymatgen MoleculeGraphs) - Pymatgen molecular graphs with

added bonding information from molecules.

IDs - index of the molecule in the json, can be user specified

Names (for xyz construction) - name of the file from which a datapoint is constructed

Spin (if specified) - molecular spin state, otherwise singlet

Charge (if specified) - molecular charge, otherwise neutral

Bonds (if specified) - We include an option to specify bonding with rdkit's tools but any user-specified bonds work. These bonds can be optionally overwritten by the bond paths determined by QTAIM.

Given the dataset, our `create_files.py` script reads in several options, including information on writing DFT input files, QTAIM parser information, and reaction/molecular options. Users can also specify custom options for executables used in running DFT and/or QTAIM calculations. Folders of input files become jobs for a high-throughput job manager/runner in our package. This runner randomly selects folders and checks for pending QTAIM and DFT jobs. Incomplete tasks are performed and the implementation allows for concurrent jobs on high-performance computing resources.

Finalized directories of QTAIM properties contain either jsons (critic2) or text files (Multiwfn) with QTAIM information including bonding, energetics, and critical point types. Our `parse_data.py` script intakes these folders and merges QTAIM information into the original json. This merge process involves parsing a user-specified set of QTAIM features, and optionally, imputation. We compile all of the QTAIM values for the dataset and use these to compute mean and median values for imputation where information is missing or where QTAIM and prior bond definitions are not in alignment. The user can also select to update bond definitions using QTAIM BCPs and whether to parse the dataset as a dataset of molecules or reactions. Atom and bond mappings are computed between final bond definitions and features. This is vital for the construction of reaction-property datasets where atom/bond-mapping across different numbers of reactions/products is non-trivial. The finalized output of these processes is a new json containing pymatgen objects, bonding

information, QTAIM features, mappings, and optional features such as spin and charges. The entire pipeline allows for QTAIM calculations at scale, and as such, we include several large datasets for further experimentation and development.

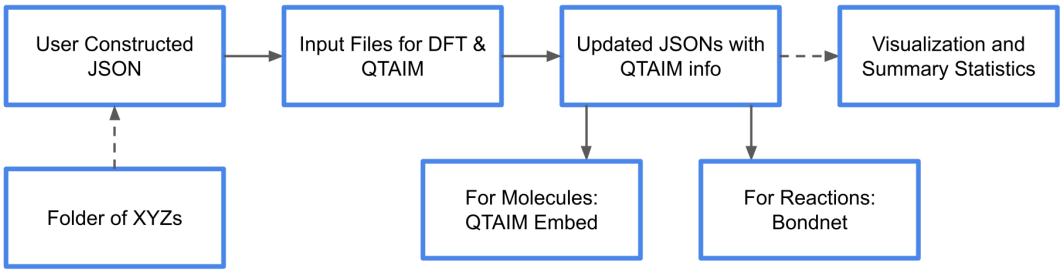


Figure 3.1: An outline of the current workflow for QTAIM property prediction. Users can either start from a JSON of data or use our helpers to parse xyz files into compatible JSON formats.

3.2.4 Dataset visualization and statistics

Included in our toolkit are also basic visualization scripts that compute summary statistics such as mean, mode, median for debugging and visualization purposes. We compute these features for each element in the dataset and output a breakdown of statistics at the elemental level as well. For visualization, we break down QTAIM descriptors at the global and element level with log scaling for highly-variable features (which is often the case for NCP energies). These tools were created to allow users to filter features with low variability and heavy outliers.

3.3 Datasets

We selected key datasets across varying levels of computational complexity and computed properties to highlight the flexibility of our package. Key considerations for these datasets and

the level of theory for subsequent QTAIM calculations were the following: first, we wanted to highlight important features of our package such as support for reactions and spin/charge. Second, we informed the level of theory in our QTAIM calculations with the relative expense of computed properties. In other words, inexpensive orbital energies of organic molecules only justified a modest level of theory in our dataset construction. Conversely, more expensive vertical excitation or vibrationally-corrected free energies on metal-containing complexes justified more expensive calculations. We wanted to reflect real use cases where computing descriptors should be considerably less expensive than the properties they are used to predict. Finally, we sought to integrate datasets that are either already in use by the community or could be adopted readily to test the limits of new models on domains such as molecules with varying spins and charges, transition metals, and reactions. Towards understanding the relationship of individual QTAIM features to individual target variables, we conducted a simple correlatory study. Here we mean-pooled each QTAIM across individual molecules and correlated these values with labels in molecular property datasets. We briefly describe the datasets we based our QTAIM datasets on as well as the labels we used to test and validate the use case for QTAIM descriptors in machine learning:

3.3.1 QM9

Perhaps the most widely-adopted dataset for structure-to-property benchmarking, QM9 is a dataset of optimized, small organic compounds consisting of 134,000 structures [227, 242]. These structures are limited to up to 9 heavy (CONF) atoms and up to 29 atoms including H. We constructed a train-test split of 90/10 and the validation set was constructed from the training set with a split of 80/20 for model selection and hyperparameter tuning. We benchmarked on 3 of the reported properties in the dataset, namely the energy of highest occupied molecular orbital (ϵ_{HOMO}), energy of lowest unoccupied molecular orbital (ϵ_{LUMO}), and the HOMO-LUMO gap (Δ_ϵ). We used this limited set as it included only size-intensive properties. Algorithms were trained in a multi-task fashion to predict all three properties.

QTAIM properties for this dataset were computed at TPSS [275]/def2-SVP [124] with D3BJ [109] dispersion. Here we aimed to study the efficacy of QTAIM features at lower levels of theory, given the comparatively low level of theory and cost of computed target values.

3.3.2 QM8

QM8 encompasses a set of time-dependent density functional theory (TD-DFT) calculations of electronic excited states [228, 242]. The dataset contains 22,000 molecules, which are a subset of QM9 with up to 8 CONF atoms, and further refinement for strained geometries. Computed properties include the vertical excitation energies for the two lowest-lying excited states and corresponding oscillator strengths. For benchmarking, we only trained/tested on the excitation energies at second-order approximate coupled-cluster (CC2) [60]/def2-TVZP [310] level of theory, yielding two target variables. We constructed a random train-test split of 90/10 and the validation set was constructed from the training set with a random split of 80/20 for model selection and hyperparameter tuning. Algorithms were trained in a multi-task fashion to predict both properties. QTAIM properties for this dataset were computed at PBE0 [216]/def2-TZVP [124] level of theory. Here we aimed to study the efficacy of QTAIM features at higher levels of theory (hybrid functionals via PBE0) given the expense of vertical excitation properties (labels for machine learning).

3.3.3 Tox21

The Toxicology in the 21st Century (Tox21) dataset measures the toxicity of 8,000 compounds across 12 different toxicity targets including nuclear receptors and stress response pathways [133, 193]. Structures in this dataset are provided as SMILES structures with RDKit [233] embedding their geometries prior to optimization. GFN2-xTB [20] further optimized these structures prior to DFT and QTAIM. We constructed a random train-test split of 90/10 and the validation set was constructed from the training set with a split of 80/20 for model selection and hyperparameter tuning. Algorithms were trained in a multi-task fashion to

predict all 12 properties (toxicity toward 12 targets). QTAIM properties for this dataset were computed at TPSS [275]/def2-SVP [124] with D3BJ [109] dispersion following geometry optimization. The dataset consists of various missing values across the 12 labels so we imputed mode values for training but at testing no imputation was performed. Here we aimed to study the efficacy of QTAIM features at high levels of theory given the experimental nature of this dataset. We did, however, use a relatively cheap method for geometry optimizations to probe how robust QTAIM is to the quality of the geometry.

3.3.4 LIBE

Lithium-ion Battery Electrolyte (LIBE) is a dataset composed of a diverse set of lithium-ion battery solid electrolyte interface (SEI) species. These structures were generated via fragmentation and combination operations on the principal molecules known to be present in the Li-ion battery SEIs. The dataset contains 17,000 structures of varying spin and charge states "labeled" with both raw and corrected enthalpies, entropies, and free energies [268]. We used the rigid-rotor harmonic oscillator (RRHO) approximated free energies [236] as a training target, units are reported in eV as in the original publication(Tab. B.2, Fig. B.1). To approximate molecular formation energies, we performed an energy correction calculation via linear regression to approximate individual atomic energies at infinite separation (Fig. B.1, Tab. B.2). We constructed a random train-test split of 90/10 and the validation set was constructed from the training set with a random split of 80/20 for model selection and hyperparameter tuning. The inclusion of the LIBE dataset was also of note as there is currently no benchmark predicting molecular properties on this dataset and it would allow us to test the ability of QTAIM descriptors to generalize across different charge and spin states. LIBE also contains metals with nonstandard bonding interactions - an instance where QTAIM's rigorous bonding definitions should fare well. QTAIM properties for this dataset were computed at TPSS [275]/def2-SVP [124] with D3BJ [109] dispersion.

3.3.5 Grambow 2022

To test QTAIM performance on predicting reaction-level properties we benchmarked a dataset recently published by Green et. al. [266]. This dataset consists of 12, 000 reactions with barrier heights and reaction enthalpies computed at three levels of theory. Reactions in the dataset involve only C, N, O, and/or H atoms with up to 7 heavy atoms. We benchmarked predicting activation energies at the highest level of theory they were computed (CCSD(T)-F12a [5]/def2-TZVP [124]). We constructed a random train-test split of 90/10 and the validation set was constructed from the training set with a random split of 80/20 for model selection and hyperparameter tuning. QTAIM here was computed at TPSS [275]/def2-SVP [124] with D3BJ [109] dispersion level of theory given the large number of individual molecules in the entire dataset.

3.4 Models

A host of geometric learning algorithms were developed or adapted to interoperate with our QTAIM generation framework: molecular graph neural networks spanning graph convolutional networks (GCNs), residual convolutions, heterograph graph attention (GAT) layers, Chemprop (albeit only for molecular property predictions with atomic QTAIM features), and a variant of the BondNet architecture for reaction-level property predictions. Further details on each architecture implementation follow.

3.4.1 Molecular Representation

Molecules, and molecules within reactions, are represented similarly as heterographs with atom, bond, and global feature nodes (Fig. 3.2). Heterographs, as opposed to homographs with bonds as edges, allow for separate relationships between each different edge type and enable the addition of a separate global node type to store important molecular-level information. Graphs ($\mathbf{G} = (\mathbf{B}, \mathbf{A}, \mathbf{g})$) consist of $\mathbf{B}$ as bond information vectors, $\mathbf{A}$ is atom-

level information, and **g** is the molecular-level feature vector. This followed prior work that also featurized molecules as complex knowledge graphs [24, 57, 122, 313]. Notably, we also intake user information on molecular charge and spin information, and transform it into one-hot encoded vectors in the global feature vector g. Features from the original graph encoding are transformed via iterative message-passing steps to yield an updated molecular graph $\mathbf{G}' = (\mathbf{B}', \mathbf{A}', \mathbf{g}')$ with updated node features $\mathbf{B}'$, $\mathbf{A}'$, $\mathbf{g}'$. These features are embedded into a fixed-size vector prior to a dense, feedforward network for property prediction similar to other molecular property graph neural networks [282].

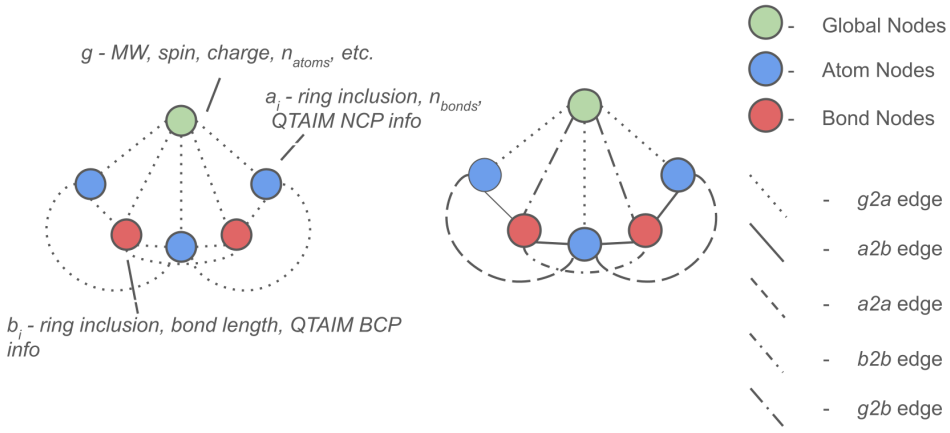


Figure 3.2: The heterograph construction of our molecular property prediction algorithm.

3.4.2 Molecular-Property Graph Neural Network

Our graph neural network models rely on complex encoder architectures where raw features are embedded to a fixed-size vector at each node prior to neural message passing. This amounts to a rectifying step that allows for greater parameterization in our models at the node level [313]. Message passing is then used to update a rich set of features in a graph. The final, updated graph is passed through a global graph pooling operation to readout the graph into a meaningful, learned vector representation (Fig. 3.1). Under the message passing paradigm, these updates are computed as a function of differentiable update and aggregate functions on neighbor features. These functions can take an arbitrary number of forms and herein lies much

of the rich diversity of graph neural networks developed [252, 255]. Typically, these functions are applied in various successive rounds to propagate information further across the initial graph. A pitfall lies with the potential of over smoothing where features become uniform across the graph. This updated graph is then embedded into vectors using one of a number of different methods we implemented to make it amenable to traditional neural networks for supervised learning tasks. These embedding schemes have also been an active area of research with schemes such as set2set [290], setTransformers [170], and self-attention graph (SAG) pooling [171] created to balance computational complexity with expressiveness. In particular, we implmented MeanPooling, WeightedMeanPooling, Self-attention pooling [171], and set2set pooling [170] as a diverse set of pooling approaches.

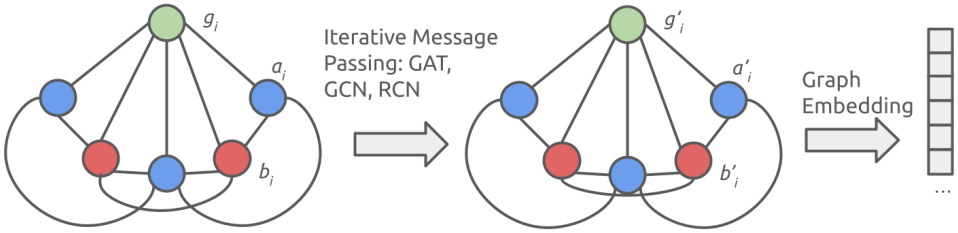


Figure 3.3: The full framework of our molecular property algorithms. Several different message passing and global pooling operations are implemented for intensive and extensive molecular properties.

We implemented several graph neural network architectures in our approach to ensure a wide-range of algorithms were benchmarked with/without QTAIM descriptors. These architectures included different update and pooling functions to ensure that relatively up-to-date models were compared. For update functions, we used traditional graph convolutions [153], graph attention mechanisms [289], and residual convolutions [119]. These layers were selected for their diversity and ability to learn at different model depths with attention and residual connections typically being more resistant to oversmoothing [243]. These layers have use across the chemical structure-to-property domain with strong results in cases including

predicting aqueous solubility [62], reactivity [237], and cost [247]. Pooling functions ultimately intake raw or processed graphs and compute a fixed-sized representation for visualization or tasks via a dense neural network. These layers are highly important and vary in complexity from simple sum operations to complex setTransformer architectures incorporating attention and beyond [170, 254]. Here we integrate 4 such operations into our potential space of graph neural networks: global summing, weighted global summing, set2set, and global attention pooling. These layers were selected to span a space of expressiveness and cost for our benchmarking and provide a wide toolkit for future QTAIM-enabled machine learning experiments. In order to merge QTAIM-features with nodes in our heterographs, we parsed Multiwfn's outputs to map features at NCP/BCPs to nodes based on "attractors" that aligned with atomic positions. For BCPs, Multiwfn also outputs NCPs that terminate bond paths, which we parsed to their respective bonds. This avoided any non-nuclear "attractors" (NNAs) appearing as atom nodes in our graphs.

3.4.3 ChemProp

Chemprop is a flexible framework for computing a host of different molecule-level and reaction-level properties [122]. The algorithm incorporates a local embedding from atom/bond features, a graph-level embedding function that transforms finalized representation graphs to a fixed-size vector, and a standard feed-forward neural network for property prediction. We adopted our QTAIM generator to construct atom-level QTAIM features in a format compatible with Chemprop's featurization. Here we limit ourselves to atom-level molecular features, excluding both bonds and reactions due to the inflexibility of Chemprop for user-defined bonds and the added complexity of atom-mappings. Hyperparameter optimization was performed using their convenient bayesian optimization functionalities.

3.4.4 BondNet

BonDNet is a reaction-property graph neural network originally constructed for the prediction of reaction ΔG_{rxn} values in single bond dissociation reactions. It consists of two modules, the graph-to-graph and graph-to-property modules, each constitute the processing of the original feature graph towards final prediction. The graph-to-graph module intakes the original knowledge graph $\mathbf{G}(\mathbf{B}, \mathbf{A}, \mathbf{g})$ and transforms it, via successive message-passing steps, to the final graph $\mathbf{G}(\mathbf{B}', \mathbf{A}', \mathbf{g}')$. Updates are performed on each separate reaction molecule prior to the construction of a global reaction difference graph. The reaction graph is constructed via the mapping of atoms and bonds in reactants to corresponding atoms in the products prior to a simple subtraction. The finalized reaction graph is embedded into a fixed-size vector via a global embedding set2set layer prior to feed-forward neural network layers for property prediction. Here QTAIM descriptors offer a promising avenue for highlighting nuanced changes in electronic structure between products and reactants, even at distal locations from the reaction site. We adapted our code to work natively with newer variants of the BonDNet architecture. This architecture was recently updated to improve generalizability for custom user descriptors and arbitrary reaction molecularity - essential quality of life updates that make it a prime model for testing an integrated QTAIM/ML approach [113]. Furthermore, this updated architecture allows for custom bond definitions, thus, we integrate QTAIM bond path connectivities to define bonds within our molecular graphs.

3.4.5 Benchmarks

QTAIM-enabled algorithms were pitted against a diverse set of molecular-graph property algorithms. Our aim here was not necessarily to outperform SOTA models but to demonstrate that models with QTAIM features could approach these models in performance and thus serve as the basis for more-advanced QTAIM-enabled algorithms. Benchmarks on molecular properties were performed using Schnet, PaiNN, and Chemprop. We briefly overview Schnet and PaiNN here. The Schnet architecture introduced the concept of continuous convolution

filters. These convolution operations allow for the arbitrary position of atoms within the model representation and give SchNet improved performance over their direct legacy algorithms, DTNNs. PaiNN is an equivariant neural network architecture, it couples ideas from SchNet to new representations, enabling more data-efficient learning. Perhaps the biggest algorithmic development of PaiNN is the use of equivariant message passing functions that incorporate not only rotationally invariant distances but also rotationally equivariant neighbor directions as part of the message-passing update function. This allows the algorithm to predict tensorial properties, as well as generalize well with less data. Its efficient representations also allow for effective models with fewer parameters and shorter inference times. We note that our baseline GNN architectures are comparatively less sophisticated than many of these algorithms, and as such, we hope to bridge performance gaps with these models via the inclusion of QTAIM features alone. These models were benchmarked competitively on QM8 and QM9 as the remaining datasets required spin/charge information or covered reaction-level properties.

Other benchmarks to note are the use of our QTAIM-enabled algorithms vs. those without on the LIBE, Green, and Tox21 datasets. Here we opted to remove the above benchmark datasets to avoid added complexities in treating classification tasks, reaction-property predictions, and spin/charge-varying molecules with algorithms that cannot encode this information. For Tox21, both models sets of models perform comparably and we include both the dataset and performance in supplementary information (Tab. B.3). To gauge the effect of QTAIM on model learning, we benchmarked model test performance on LIBE, QM8, and QM9 given 10^2, 10^3, 10^4, (and 10^5 for QM9) training data points. These learning curves are often used in machine learning to measure the learning capacity of a model and extrapolate to how accuracy varies with dataset size.

3.5 Results and Discussion

In addition to the experiments that follow, we evaluated a classifier variant of our model on the Tox21 dataset with and w/out QTAIM features (Tab. B.3). Here we see marginal but comprehensive improvements in performance with QTAIM features.

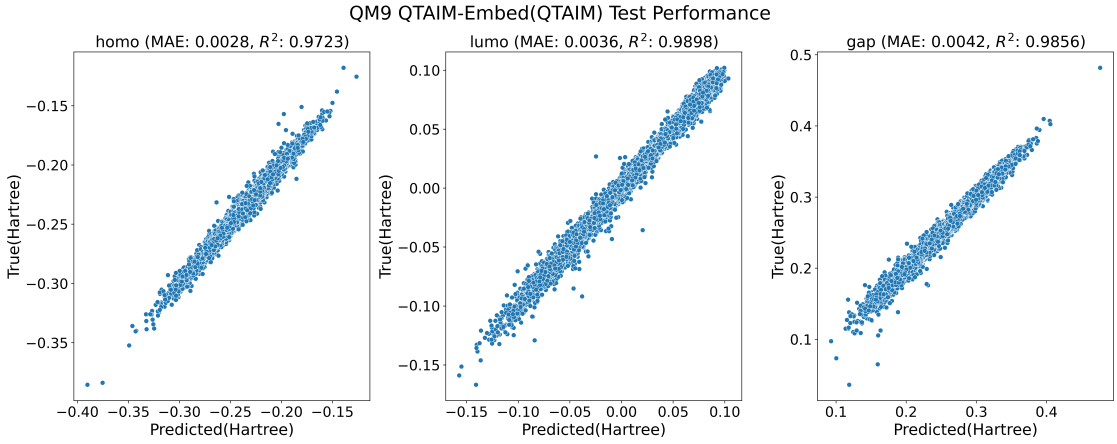


Figure 3.4: Parity plot of our model, with QTAIM, on the qm9 test set

3.5.1 QM9

Evaluating model performance on QM9, we note how our QTAIM-enhanced models are able to compete with the performance of the otherwise best-performing model, Chemprop (Tab. 3.5.1). We also augmented Chemprop with QTAIM NCP-only features but here we actually see a slight drop in testing performance. We emphasize that Chemprop does not include vital BCP QTAIM features and thus does not leverage the comprehensive set of QTAIM descriptors. Even here, the performance difference between QTAIM-enabled and non-QTAIM Chemprop models is quite small and also suggests the model is near or at capacity - not that QTAIM features are not informative. Analyzing scatterplots of QM9 test performance, we can also determine the robustness of QTAIM-informed models with few outlier points between predicted and true labels. Interestingly, our outlier points are generally some of the heaviest molecules in QM9. It is also worth noting that the QM9

dataset constitutes a comparatively-simple dataset for machine learning with the difference between the top-performing models being relatively small. Observing correlations of target variables to individual QTAIM values - QM9 exhibits the highest correlations of any of our datasets (though still quite low). Here several values emerge as important, these include electron localization values and bond Lagrangian values (Fig. B.27).

Model	**HOMO**	**LUMO**	**Gap**	**Average**
Schnet	0.0109	0.0115	0.0151	0.0125
PaiNN	0.0136	0.0148	0.0158	0.0147
ChemProp (w/out QTAIM)	**0.0028**	**0.0031**	**0.0038**	**0.0032**
Our Best (w/out QTAIM)	0.0058	0.0076	0.0090	0.0075
ChemProp (w/ QTAIM)	**0.0030**	0.0035	**0.0042**	0.0036
Our Best (w/ QTAIM)	**0.0028**	0.0036	**0.0042**	**0.0035**

Table 3.2: Test performance (MAE, Hartrees) of various geometric learning algorithms on orbital energies in QM9.

In addition, we examine the learning curves of our models with and without QTAIM features. To give each set of models even footing, we conducted hyperparameter tuning on models with and without QTAIM features separately and thus these curves (and overall test performance) correspond to the best models for each descriptor set. We see QTAIM yielding a distinctive improvement in performance in the low data regime with consistent advantages in test performance across all training set sizes (Fig. B.16). Beyond 10,000 structures, however, there is little improvement in test performance of the QTAIM-informed model suggesting the mode is at capacity to generalize or that mainly irreducible errors remain.

3.5.2 QM8

Across both tasks (first and second vertical excitation energies) QTAIM-enabled models were the top-performing algorithms (Tab. 3.3). Chemprop and our models with QTAIM yielded improved test errors over all other models with a notable gap in performance between QTAIM/ML models and all others. Again, we note that Chemprop's QTAIM featurization was limited to only QTAIM NCP features, and even then, this led to increased performance. Finally, when examining predicted versus true plots of our models, it becomes evident that QTAIM-enhanced models exhibit greater robustness, displaying fewer outlier residual errors compared to their non-QTAIM equivalent (Fig. B.2). Our correlation study (Fig. B.25) also shows remarkably low correlations between vertical excitation energies and any one QTAIM value - underlying the relative complexity of this property. We do note that some of the highest correlations are with BCP QTAIM features, suggesting Chemprop could improve with these features.

Model	**E1-CC2**	**E2-CC2**	**Average**
Schnet	0.517	0.379	0.448
PaiNN	0.0133	0.0145	0.0139
ChemProp (w/out QTAIM)	0.0373	0.0270	0.0322
Our Best (w/out QTAIM)	0.0130	0.0130	0.0130
ChemProp (w/ QTAIM)	**0.0052**	**0.0060**	**0.0056**
Our Best (w/ QTAIM)	**0.0062**	**0.0067**	**0.0064**

Table 3.3: Test performance (MAE, Hartrees) of various geometric learning algorithms on orbital energies in QM8.

The learning curves further reinforce the advantage of QTAIM-enabled models, illustrating a consistent improvement in performance across varying training set sizes (Fig. B.15). Additionally, the learning curves for both QTAIM and non-QTAIM models do not appear

to reach saturation, implying that additional training data could potentially lead to further reductions in prediction errors for both types of models.

3.5.3 LIBE

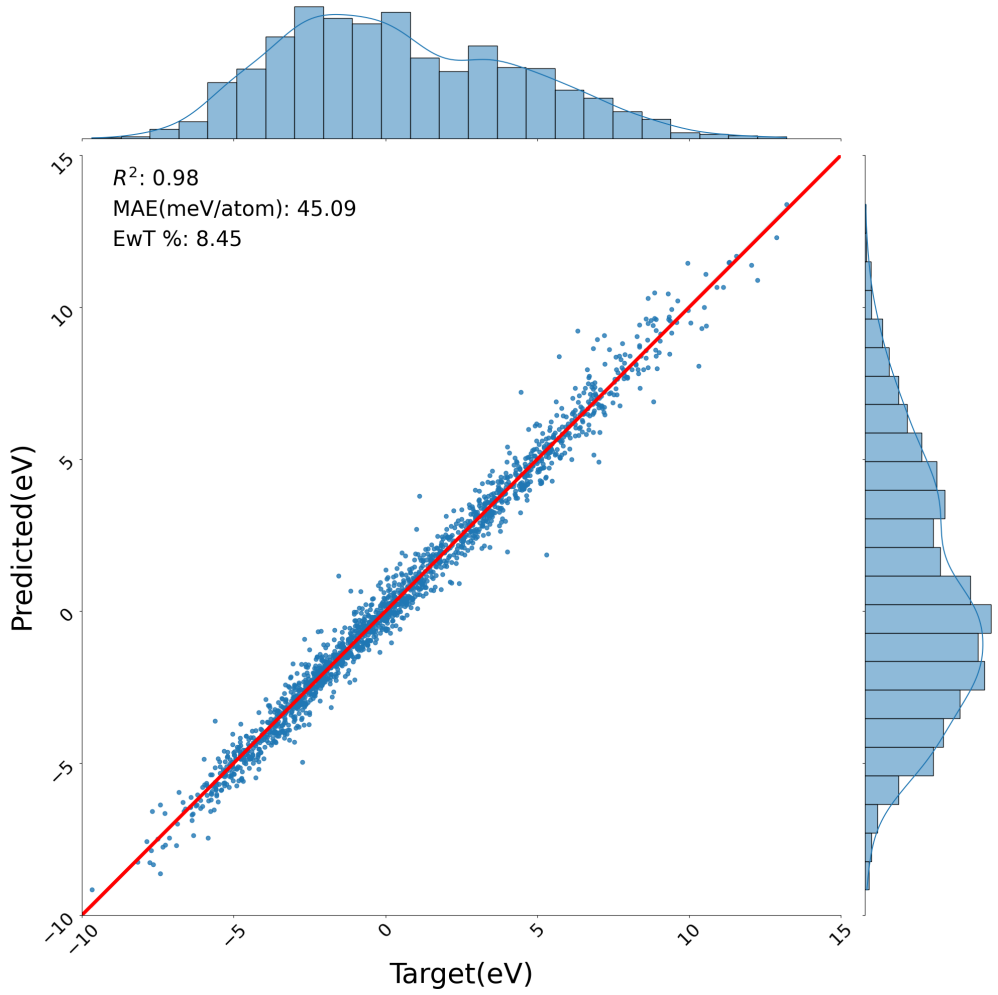


Figure 3.5: Parity plot of our model, with QTAIM, on the LIBE test set

The LIBE dataset presents a more challenging task due to its inclusion of spin-varying and charged species. Moreover, the dataset exhibits a wide range of molecular free energies which further add to the difficulty of learning energetics here. In pitting QTAIM/ML versus non-QTAIM models we note that our non-QTAIM models do directly describe spin and charge as one-hot encoded global features while the QTAIM/ML models add QTAIM features,

including α spin, β spin, and spin density at each critical point, to further inform learning. Both models perform quite well with the top QTAIM/ML model yielding a reduced error on formation energies from 76.26 meV/Atom to 45.09 meV/Atom and an increased proportion of predicted energies within chemical accuracy to true labels (8.5% vs. 5.4%) versus its non-QTAIM equivalent (Tab. 3.4). Analyzing correlation values (Fig. B.29, B.30) we see again that electron localization functions and electrostatic potentials, specifically at BCPs, emerge as the most correlated features to formation energies. This interpretation in agreement with previous studies that leveraged both electron localization and electrostatic potential values to analyzed bonding strength and orbital interactions [83, 201].

Model	**MAE (meV/Atom)**
QTAIM-Embed (Ours, No QTAIM)	76.26
QTAIM-Embed (Ours, QTAIM)	**45.09**

Table 3.4: Test performance of our geometric learning algorithms on formation energies in LIBE.

In addition, no discernible trends can be gleaned across predicted vs. true values for the QTAIM/ML models while non-QTAIM models perform slightly worse on low spin, positively-charged species (Fig. 3.5, B.7, B.6). Learning curves here present a more obfuscated picture with the non-QTAIM model outperforming the QTAIM/ML model on the smallest training set (Fig. B.14). This narrative shifts at larger dataset sizes as the QTAIM/ML model, again, outperforms the top non-QTAIM model. Here, there is no pronounced improvement in the learning curves between the two sets of models as QTAIM models have a slightly more aggressive learning curve - indicative of their ability to increase model generalizability at higher data regimes.

3.5.4 Green 2022

Model	Test MAE (kcal/mol)
Bondnet (w/out QTAIM)	4.18
Bondnet (w/ QTAIM)	2.60

Table 3.5: MAE Performance of our model with/without QTAIM on Green 2022 barriers.

The Green 2022 dataset represents a comprehensive compilation consisting of approximately 12,000 gas-phase reactions, meticulously calculated at high-level theory (CCSD(T)-F12a [5]/def2-TZVP [124]). This dataset was constructed to facilitate transfer learning approaches by incorporating two lower levels of theory (ωB97X-D3 [47]/def2-TZVP [310], and B97-D3 [29]/def2-mSVP [310]). Remarkably, our experimental results demonstrate a performance on par with the original authors' findings, achieving comparable results without necessitating a transfer learning approach at lower levels of theory [267]. Notably, the original authors employ significantly higher levels of theory for transfer learning, specifically ωB97X-D3/def2-TZVP [310] and B97-D3 [29]/def2-mSVP [310]. In contrast, our descriptors are limited to the TPSS [275]/def2-SVP [310] level, yet they enable us to attain comparable performance. It would be intriguing for the original authors to explore and compare the transfer learning process from the lowest level of theory to the highest level of theory (without the intermediate-level of theory). This would effectively simulate the relative performance of QTAIM versus transfer-learning labels at inference time. Furthermore, when evaluating their non-transfer learned models, it's observed that those roughly align (4.17 kcal/mol versus 4.07 kcal/mol) with our Bondnet training without QTAIM integration (Tab. 3.5). The incorporation of QTAIM features with Bondnet, however, elevates its performance, surpassing the non-transfer learned models with a reduced mean absolute error (MAE) of 2.6 kcal/mol (Tab. B.8, B.9). This discrepancy underscores the advantageous impact of QTAIM integration in enhancing model accuracy and predictive capabilities.

3.5.5 OOD Tests

Beyond a measure of train/test performance, we wanted to demonstrate whether QTAIM could functionalize machine learning models to make out-of-domain predictions. We conducted two sets of experiments here. First, we trained/tested models with/without QTAIM features on the LIBE dataset where the training set was trimmed to only include examples of neutral molecules and the test set was refined to only include test molecules with charges $\in \{-1, 1\}$. The baseline model included only a one-hot encoding of molecular charge in the global feature node; the QTAIM-enabled model adds QTAIM features to the model. None of the prior benchmark models include native support for spin and charge; therefore we only conducted this experiment on our own architecture. Second, we tested model performance of our GNNs with/without QTAIM features on sub-selected variants of QM9 train/test sets. Here we stratified the datasets along molecular size: molecules in the training set with fewer than 13 atoms included were included in the OOD training set and those with more than 13 atoms in the original test set included in the OOD test set.

Model	**HOMO**	**LUMO**	**Gap**	**Average**
Our Best (w/out QTAIM)	0.0177	0.0320	0.0376	0.0291
Our Best (w/ QTAIM)	**0.0155**	**0.0243**	**0.0330**	**0.0243**

Table 3.6: Test performance (MAE, Hartrees) of various geometric learning algorithms on orbital energies in QM9 OOD.

Model	**MAE (meV/Atom)**
QTAIM-Embed (Ours, No QTAIM)	191.65
QTAIM-Embed (Ours, QTAIM)	**119.13**

Table 3.7: Test performance of our geometric learning algorithms on formation energies in LIBE OOD.

For QM9 stratification, there is a significant decline in model performance between both QTAIM and non-QTAIM models (Fig. B.4, B.5, 3.6). Despite this, QTAIM-informed models demonstrate a moderate ability to generalize to much larger molecules despite being trained entirely on small molecules. We also note that the filtering of the QM9 dataset to only molecules with fewer than 13 atoms results in a training set of only 4,000 molecules. This comparatively small (2 orders of magnitude smaller than the full QM9 test set) training set also shows how QTAIM could be an effective tool for leveraging smaller datasets. We note the systematic overprediction of THE LUMO/gap energies and underprediction of THE HOMO energies in the QTAIM informed model, and couple this to the mean values for each label in the training and testing set: -0.263 Ha, -0.057 Ha, 0.206 Ha in training and -0.239 Ha, 0.0131 Ha, 0.252 Ha in the test set for HOMO, LUMO, gap respectively (Fig. B.4, B.5). Here these systematic changes can be partially attributed to the comparative difference between the two label distributions as well as to the model itself. The mean absolute error (MAE) values highlight the effectiveness of QTAIM, with an average MAE of 0.0243 Ha compared to 0.0291 Ha without QTAIM (Tab. 3.7).

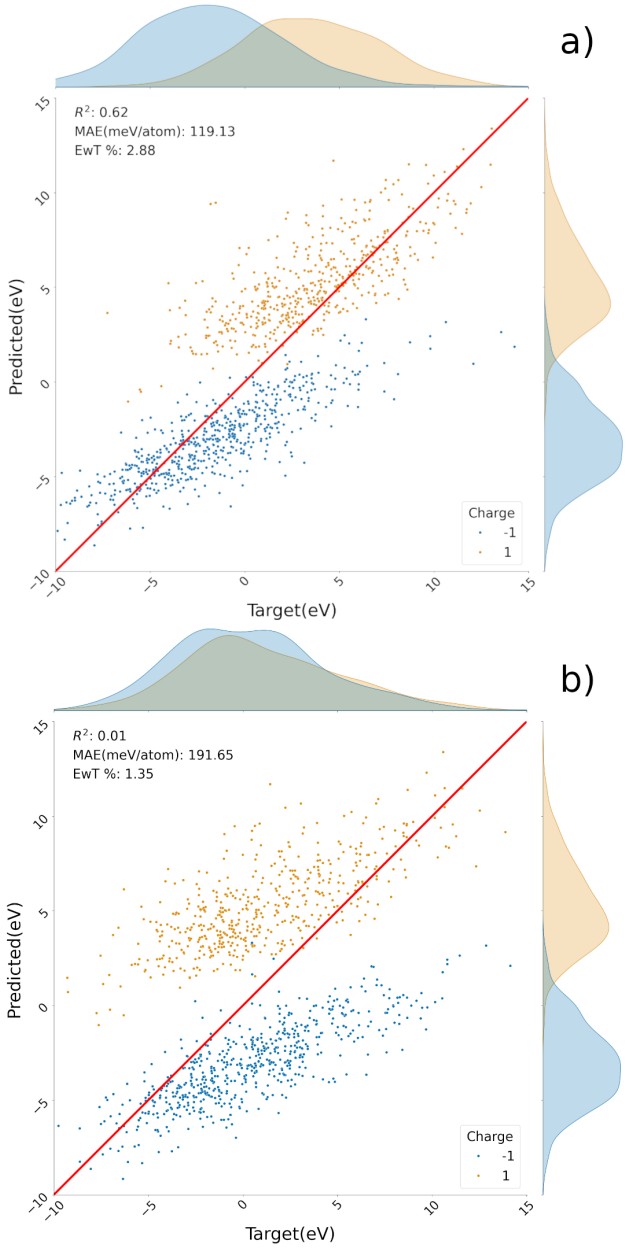


Figure 3.6: Parity plot of our model, with QTAIM (a) and without QTAIM (b)

LIBE OOD tests also show a marked drop in testing performance, though not to the extent of the QM9 OOD test (Tab. 3.7). The QTAIM model here remains quite serviceable while the model without QTAIM features is drastically worse versus in-domain testing. Changes in performance can be partially attributed to the reduction in training data (only 5,200 molecules in training). This notion is somewhat qualified by our learning tests (Fig.

B.14) where non-QTAIM models had better test errors (<125 meV/Atom) with only 1,000 training examples. Notably, both models exhibited a trend of overpredicting for positively charged molecules and underpredicting for those with a -1 charge, yet this deviation was less pronounced in QTAIM-informed models where a greater portion of test examples were within chemical accuracy (2.88% vs. 1.35%) (Fig. 3.6). These results show that QTAIM can be an effective method for improving model robustness in out-of-domain experiments, especially in the context of charged species.

3.6 Conclusions

Here we present a framework for leveraging QTAIM descriptors as general, robust features for geometric machine learning tasks. Our framework extends to both molecular and reaction-level predictive tasks and thus can be applicable in a wide-set of use cases. We created tools for both high-throughput calculation of QTAIM descriptors and a custom machine learning package for easily implementing models that use these features.

Furthermore, we performed extensive testing to demonstrate how QTAIM can functionalize machine learning models to perform better on out-of-domain tasks and smaller datasets. In the case of QM8, our test showed that QTAIM features helped both our architectures and Chemprop improve model performance given identical datasets - suggesting our featurization package could be used with outside machine learning models as well. In the future, we plan on writing more "translation" functionalities to allow users of other architectures to leverage QTAIM features for their learning tasks.

Future work in this space should see further integration beyond algorithms to include more databases and DFT codes. For example, the native dovetailing of this software into the larger Materials Project ecosystem could see QTAIM integrated into their workflows. At present, the Materials Project only natively supports Q-Chem [77] (for molecular DFT) as a DFT software - a commercial software we aimed to avoid to increase accessibility. Additional

work could also see integration of input files and execution scripts for other DFT packages such as Gaussian, NWChem, etc. We also implement reaction parsing and processing with compatibility for BondNet and Chemprop (to a lesser extent) but native dataset compatibility for more algorithms could facilitate benchmarking and development.

Also in development are graph-neural networks that could leverage QTAIM-descriptors while avoiding computationally-expensive message-passing graph neural networks. The aim here would be to rely on QTAIM descriptors to capture distal relationships between nodes (atom, bonds) rather than using iterative message-passing steps to achieve this task. From a conceptual DFT standpoint, the native integration of parsers and data structures that support ring and cage critical points would be beneficial.

3.7 Acknowledgements

This chapter was adopted from:

High-throughput quantum theory of atoms in molecules (QTAIM) for geometric deep learning of molecular and reaction properties Santiago Vargas, Winston Gee, and Anastassia N. Alexandrova. *Digital Discovery* **2024** 3, 987-998.

Here I contributed code for machine learning, clustering, visualization, data processing. I also contributed writing, high-throughput calculations, and editing. W.G. contributed to writing, editing, visualization, and data processing. A.N.A. provided direction, writing, editing.

Chapter 4

Directed Evolution of Protoglobin Optimizes the Enzyme Electric Field

4.1 Introduction

Nature has evolved enzymes as remarkably proficient biocatalysts to facilitate a vast array of chemical transformations [318]. Through billions of years of evolutionary fine-tuning, natural enzymes have unlocked extraordinary catalytic power, selectivity, and efficiency [48, 202, 209]. The drive to push beyond nature's set of catalyzed reactions, and achieve similarly efficient catalysis for other transformations has led to innovative approaches in modifying enzymes, [10, 39, 182, 281] and designing them *de novo* [132]. Indeed, enzyme design has become a frontier of innovation, with the goal of customizing enzymes for the sustainable production of a variety of chemicals, pharmaceuticals, and materials.

Creating highly active enzymes from scratch remains an unsolved task, despite the potential [18].The initially designed enzymes often need more catalytic vigor, and are subjected to subsequent rounds of directed evolution (DE) to reach appreciable activity levels [298]. DE serves as an optimization step that provides designed enzymes with properties absent from initial designs, from improving enantioselectivity in rhodium-catalyzed artificial

metalloenzymes, [235] to dramatic boosts in the activity of computationally designed retro-aldolases [9, 103],and Kemp eliminases [152, 245], to name just a few examples. DE produces stunning enhancements of $k_{\mathrm{cat}}/K_{\mathrm{M}}$ of over 4400-fold [9, 103, 152, 206, 235, 245].The need to evolve designed enzymes to attain catalytic viability underscores significant gaps in *de novo* design protocols. Understanding what DE contributes to enzyme design is crucial, as DE appears to provide essential elements that are missing in initial designs, potentially unlocking key strategies for efficient enzyme design in the future.

We study the directed evolution of *Aeropyrum pernix* Protoglobin, a Fe-heme protein, which was evolved to perform a new-to-nature selective carbene transfer to catalyze cyclopropanation of benzyl acrylate [Fig. 4.1]. [69, 221] Mutations introduced by DE are dispersed throughout the protein structure, located both close to the Fe-center (F145Q, I149L, Y60A, W59L), and as far as >15 Å away from it (F175L, C102S, V63R), and include both hydrophobic and hydrophilic residues. We use this rich evolutionary journey to gain an understanding of how DE can imbue new catalytic functions into an enzyme. We perform and analyze replica molecular dynamics (MD) simulations of wild-type (WT) Protoglobin and four evolved variants (LVRQ, LVRQL, GLVRSQL, GLAVRSQLL) that showed a progressive increase in activity, initially focusing on substrate access and binding improvements at the active site. Upon indications of changes in electrostatic preorganization at the active site along the evolutionary pathway, we develop and utilize a novel framework to study the dynamics of the heterogeneous electric field in the active site, combining electric field topological analysis, high-throughput computation, and graph compression algorithms for a comprehensive picture. Finally, we correlate changes in electrostatic preorganization with experimental yield through QM/MM reaction mechanism calculations. This workflow illuminates the critical factors DE exploits to enhance enzyme catalysis—insights crucial for refining enzyme design protocols.

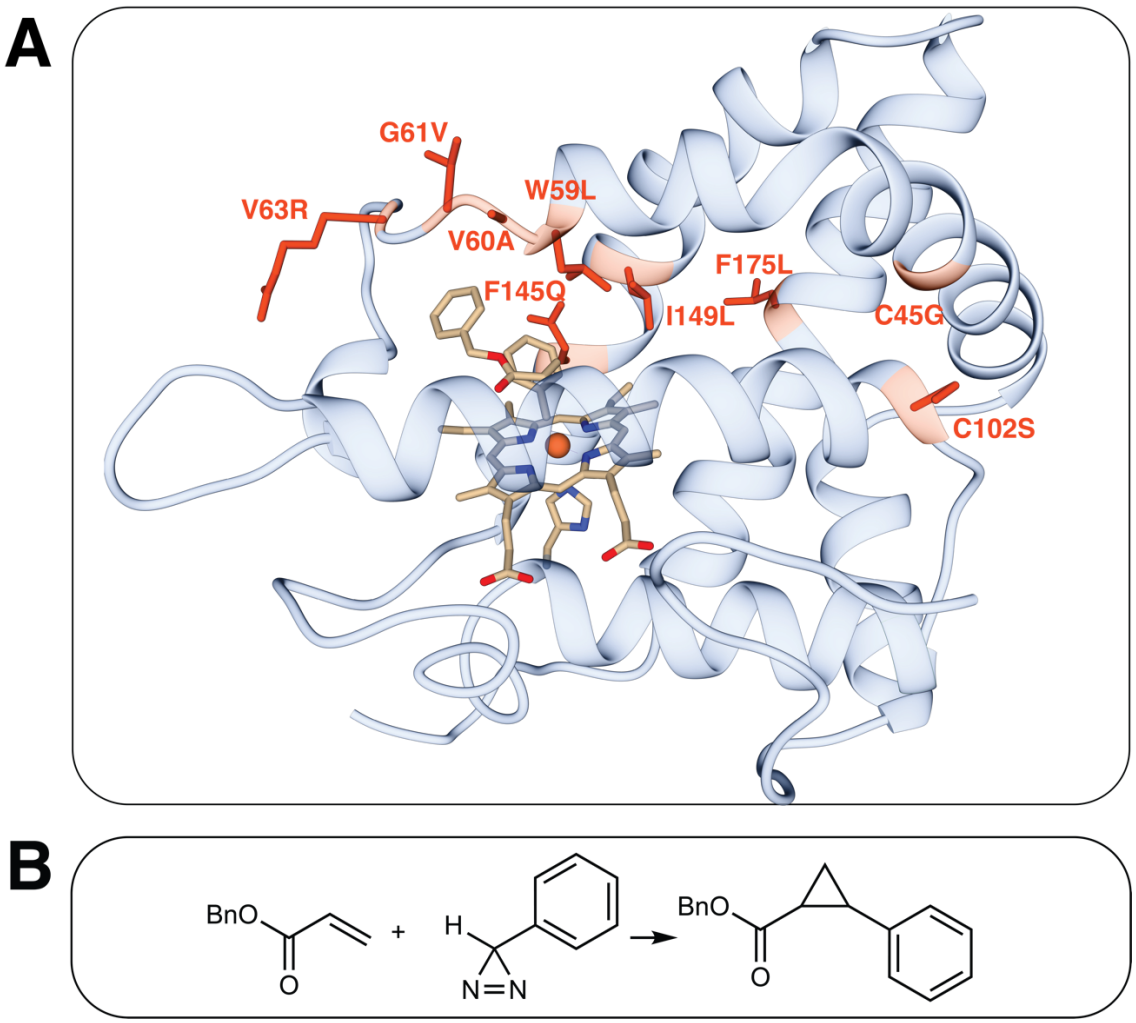


Figure 4.1: (A) Protoglobin with directed evolution mutation sites highlighted in red and labeled with the bound substrate (PDB ID: 7UTE). (B) the carbene transfer reaction being optimized along the directed evolution path.

4.2 Methods

Our developed methodology has six primary components, visualized in Fig. 4.2. We use MD simulations to sample configurations of the protein. We use methods in field analysis to calculate the point electric field and electric field topology at the active site throughout the molecular dynamics trajectories. These topologies are compared using statistical distance metrics to obtain a distance matrix for each trajectory. To analyze how these field topologies change, we then use clustering

on the distance matrices to obtain representative "snapshots" of the electric field at the active site. These snapshots are subjected to quantum mechanical/classical mechanical (QM/MM) reaction path calculations and **principal component analysis** (PCA).

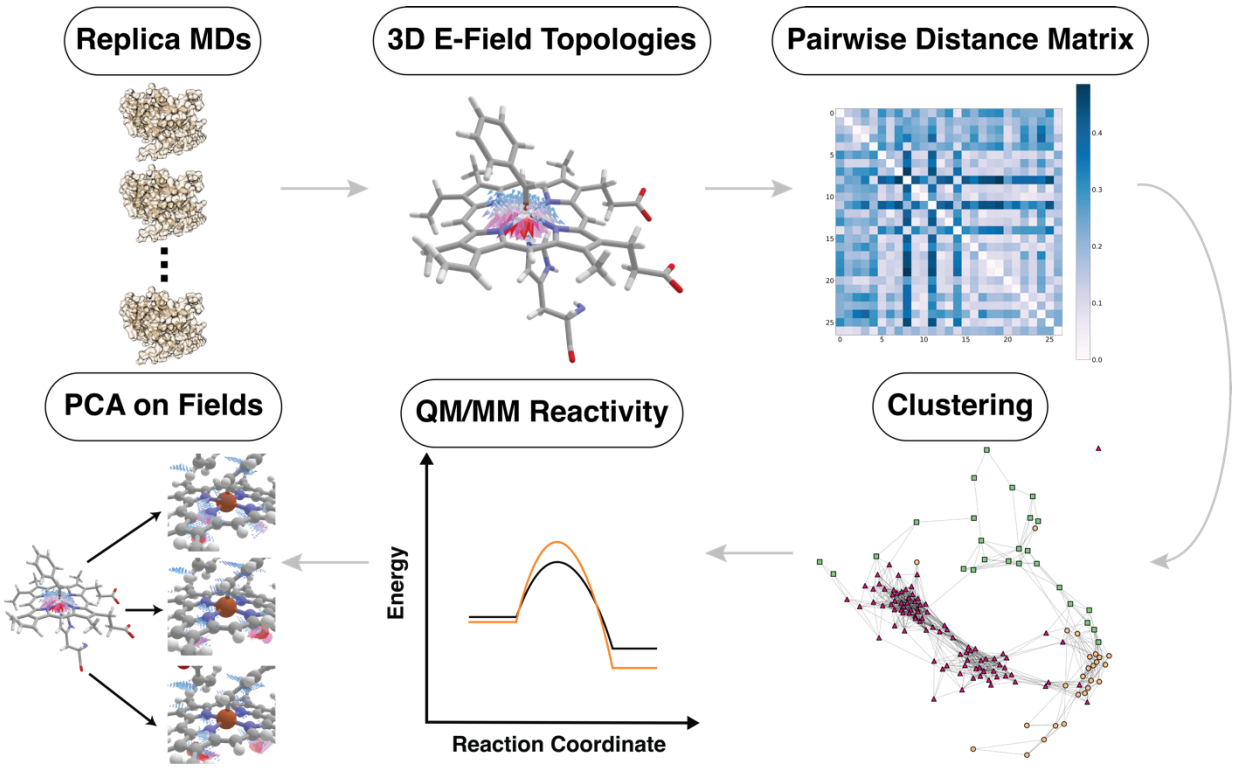


Figure 4.2: This study's approach measures electrostatic preorganization by analyzing the heterogeneous electric field topology across replica MD simulations. It further involves comparing these topologies using a pairwise distance matrix, clustering based on similarity, and then quantifying reactivity through QM/MM methods. The reactivity difference is chemically elucidated using Principal Component Analysis.

4.2.1 System Preparation and Molecular Dynamics:

We performed MD simulations on the carbene-substrate intermediate of the WT Protoglobin and the four directed evolved variants—LVRQ, LVRQL, GLVRSQL, and GLAVRSQLL. When compared to the WT, the LVRQ evolved variant presents mutations W59L, G60V, F145Q, and V63R. The LVRQL variant includes an additional I149L mutation. The GLVRSQL variant incorporates further C45G and C102S mutations. Lastly, the GLAVRSQLL evolved variant

introduces additional mutations V60A, G61V, and F175L. We used the crystal structure of the Protoglobin GLVRSQL variant as the template to model the carbene-substrate intermediate of the other evolved variants. In the absence of an experimentally determined crystal structure for WT Protoglobin ApPgb, we employed the AlphaFold2 model as a vital alternative. [45] Notably, the AlphaFold-predicted structure for WT Protoglobin ApPgb demonstrated a high average pLDDT confidence score of 96.8, indicating a robust prediction. This high level of confidence was uniformly maintained across the core structural regions, crucial for our analysis, with only a few terminal residues displaying confidence scores below 95. Furthermore, the experimentally structure of the GLAVRSQL mutant of Protoglobin [69, 221] aligns closely with an RMSD of 0.5 Å in respect to the AlphaFold2 model, reinforcing the validity of using this approach for our simulations. AlphaFold2, and its subsequent successors, have been revolutionary in yielding rapid, largely reliable structure predictions for a large swath of protein space. It excels in predicting structures of single protein chains, protein-protein complexes, and even complex hetero-multimers [21, 276]. However, AlphaFold2 is not without limitations [31, 276]. It can struggle with accurately predicting structural alterations resulting from point mutations and may misplace functionally relevant residues, [212] particularly in lower-confidence structures. Additionally, its performance can be limited when dealing with orphan proteins or proteins from less-studied families. However, these limitations did not significantly impact our study. The structure of Protoglobin used in this study has high confidence scores, ensuring the reliability of the functionally relevant residues in our analysis. Furthermore, since our study employed the crystallized GLAVRSQL structure for initial MD simulations of point mutations, the issue of predicting structural changes due to mutations was not a concern. Lastly, Protoglobin is a well-characterized protein, mitigating concerns related to lesser-known protein families [217].

The carbene was modeled taking the Micro-ED crystal structure as a reference to simulate the carbene-substrate intermediate in WT Protoglobin. The substrate benzyl acrylate was docked into the active site using AutoDock Vina [74, 280]. The setup for the MD simulation

was done via Amber 22 and AmberTools 22 modules. [46] The active site parameters for the carbene-substrate intermediates encompassing the heme, Fe, carbene, and an axial histidine were derived using AmberTool's Metal Center Parameter Builder (MCPB) v3.0 [174]. This protocol has been successfully utilized to model several metal containing systems especially hemes and non-heme iron complexes. [43, 53, 148, 327] The GAFF tool in Antechamber generated the topology for the substrate benzyl acrylate. The protonation states of the protein in the carbene-substrate intermediate were determined using Chimera routines, and the parameters for the rest of the protein were generated using the AmberFF19SB [277] force field. The LEaP module of Amber 22 neutralized the system by adding counterions. The system was then immersed into an OPC water box of at least 10 Å from the surface of the protein. Periodic boundary conditions were applied to the system, and long-range electrostatic interactions were calculated using the particle mesh Ewald method with a cut-off distance of 8 Å. The SHAKE algorithm [244] was used to constrain bonds involving a hydrogen atom. The systems were minimized in two steps: using the steepest descent (10,000 steps), and (2) the conjugate gradient (10,000 steps) methods. During this phase, the protein's heavy atoms were restrained using a harmonic potential of 100 kcal mol^{-1} $Å^2$, and the protein's hydrogen atoms, along with solvent molecules, were minimized. Subsequently, the entire system underwent a comprehensive minimization process without any restraints via steepest descent (10,000 steps) and conjugate gradient (10,000 steps) methods. The system was then heated from 0 to 300 K in 50 ps using an NVT ensemble and then remained at 300K for another 50 ps. Next, A weakly constrained MD with constant pressure was performed to achieve uniform density in the systems, followed by equilibration MD in an NPT ensemble for 10 ns with restraints on the benzyl acrylate substrate to equilibrate it in the active site and then without any restraints for 2 ns. Finally, all production runs were performed using the GPU version of the AMBER 22 package. To enhance the credibility and precision of MD analysis, five replica MD simulations, each with a 100 ns duration, were performed [157]. RMSD was performed with CPPTRAJ to validate equilibration of all runs and is provided in

Fig. D.1. Distance analysis was also done using CPPTRAJ [238]. The binding free energies were calculated with the MMPBSA/MMGBSA module implemented in AMBER 22 [197].

4.2.2 Topological Electric Field Measurements and Comparison by Distance Metric

Many previous studies incorporating electric fields as an analytical tool for protein activity use the protein structure in the context of a single structural snapshot - either crystalized variants or at a single frame within a larger MD trajectory are common [40, 127, 250, 311, 319]. These miss the effects of dynamics. Few studies, however, do incorporate dynamical information, including a study by Head-Gordon and coworkers where projections of electric field components were measured and correlated along an MD trajectory [311]. These projections amount to a low-dimensional embedding of the entire heterogeneous electric field and here we aim to increase the bandwidth to 3-D heterogenous electric fields within a sampling box with our topological distance metric. We used a distance metric to construct a matrix of pairwise distances of electric fields along an entire MD trajectory. Every 5th frame of a 100 ns trajectory was used for the description of electric fields in the active site. The atomic charges were computed for the protein in each frame using ChargeFW2 [232]. The field was calculated in a 3 Å box defined that is centered by the heme Fe – carbene carbon bond. We zero the charges on the heme, iron, and carbene moiety in all systems. This approach was taken to isolate and examine the charge effects originating from the protein scaffold alone. The pairwise distances between each electric field's topology (see eq (1) and (2)), were computed, and subsequently fed into a graph clustering algorithm. The number of streamlines used for all calculations is 10000 with a step size of 0.001 - information on testing of box size can be found in the SI **Table S1**. Raw fields were preprocessed prior to input in visualization and clustering schemes. For clustering, we determined an upper boundary CPET distance above which edges were removed [**Tab. D.1**]. This cutoff was the 10%ile distance from the collective distance matrix of all 5 WT runs. For affinity propagation, we also standardized

the remaining distances. We used max_it = 10000 and 0.5 dampening. For PCA analysis, raw fields were used.

4.2.3 Affinity Propagation

Affinity Propagation intakes the "affinity" or similitude between different data points in a distance matrix, this can include non-connected graph nodes. We refer the audience to the original implementation of Affinity propagation [85] but will provide a brief outline of the method as follows. Affinity propagation is built on the iterative message passing of responsibility and availability between nodes in a graph. If $\mathbf{X} = \{\text{x\_1}\ldots\text{x\_z}\}$ represents a set of data points and s(i, j) represents a similarity metric between points i and j. Responsibility r(i, k) describes how representative point k is for point i and availability a(i, k) measures how reasonable it is for k to pick i as a representative for itself. These values are updated using the following equations:

$$r\left(i,j\right) \leftarrow s\left(i,k\right) - \max_{k' \neq k} \left\{ a\left(i,k'\right) + s\left(i,k'\right) \right\} \tag{4.1}$$

(Update Responsibility matrix r)

$$a\left(i,k\right) \leftarrow \min\left(0,\, r\left(k,k\right) + \sum_{i' \notin \{i,k\}} \max\left(0,\, r\left(i',\, k\right)\right)\right) \tag{4.2}$$

(Update availability matrix a)

Message-passing updates are repeated until convergence of representative structures/boundaries or a maximum number of iterations are reached.

4.2.4 PCA

Principal Component Analysis (PCA) is a widely-used dimensionality reduction algorithm that intakes descriptors on a dataset and performs a basis change to orthogonal components by order

of descending variance – these new components are referred to as principal components [34]. PCA yields a few important statistical objects, namely the eigenvalues of the new principal components (PCs) and principal components themselves. Eigenvalues elucidate the variability in the dataset along the new basis and can be used to diagnose dataset dimensionality. The principal components themselves can be analyzed, along with the eigenvalues, to determine the directions of greatest variability in the dataset. We constructed PCA components from the compiled dataset of electric fields at every frame considered in the graph compression (all 5 mutants). This amounted to 25,000 electric fields where each field was centered at the heme-Fe. From here, we constructed a sampling mesh of 10 equidistant points in the six axial directions up to the boundary of 1.5 Å. This results in a 21 x 21 x 21 mesh of points spanning a 3 Å box – and thus, an input dimensionality of $>$ 27,0000 points to the PCA algorithm. Remarkably, 5 components accounted for $>$ 77% of the explained variation, 10 for $>$ 95%, and 25 for $>$ 98%. This shows that a small number of components likely can be used to understand the variability in the electric fields – though we note that variability does not signify importance and therefore we extend our analysis to include several lower-variance PCs. We selected the 10 most important components and projected the cluster centers' electric fields on those components – these 10 components account for 95% of the variance in the dataset and consist of components with $>$1% of the total variance each. This allows us to decompose the complex electric fields into simpler motifs for analysis and interpretation. Electric fields for the entire population of a mutant were analyzed along principal components to understand how the dynamic electric field evolves with mutations.

4.2.5 QM/MM Reaction Mechanism

For cluster centers obtained from affinity propagation, the reaction mechanism of carbene transfer and its energetics was elucidated with hybrid QM/MM reaction path optimizations and thermodynamics calculations. ChemShell [164, 260] was used for QM/MM calculations in combination with DL_POLY [278] for the energy of the molecular mechanics region

and TURBOMOLE [8] for the energy of the quantum mechanical region. The QM region included the Fe, carbene, reduced heme, and substrate, while the rest of the protein was in the MM region [**Fig. D.7**]. The AmberFF19SB force field generated the protein MM region parametrization. To have a well-refined reaction path, only cluster centers with the benzyl acrylate within 5 Å of the heme Fe were included in the reaction profile calculation. To determine the reaction profile, we used a collective variable that optimally combined three factors: decreasing the distance between CC and C1, increasing the distance between Fe and CC, and reducing the distance between CC and C2. For the QM reaction path optimization, the TPSS DFT functional [216, 275] was employed, with def2-TZVP and def2-SVP basis sets for the Fe atom and the remaining atoms in the QM region, respectively. The transition states and products were freely optimized. Vibrational frequency calculations were used to verify the validity of product and transition states and to compute free energies within the harmonic approximation. Single point calculations were done at the reactant, product, and transition states using the TPSSh functional, with the def2-TZVP basis set for all atoms in the QM region to provide more precise electronic energies. To ensure the robustness of our findings, the QM/MM calculations were repeated using the B3LYP functional [28]. Finally, we performed single-point QM/MM calculations at the near-gold standard DLPNO CCSD(T) [114] level using the def2-TZVP basis set to obtain accurate energy estimates for the reaction mechanism. ORCA was employed for the QM region calculations along with DL_POLY in ChemShell for these QM/MM calculations. All reported QM/MM free energies are derived from these single-point energies at the DLPNO-CCSD(T) level, incorporating thermodynamic corrections.

4.3 Results and Discussion

4.3.1 Can substrate binding explain the yield increase?

Based on the microcrystal electron diffraction structure of the GLVRSQL Protoglobin variant, it was proposed that DE facilitates the new-to-nature catalysis by enhancing substrate access to the active site [69, 221]. We analyzed the substrate access to the active site of Protoglobin, by measuring the distance between the terminal C1 atom of the benzyl acrylate substrate and the reactive CC atom of the carbene across the five replica molecular dynamics (MD) simulations of 100 ns each, for all variants [Fig. 4.3]. In agreement with experiments, the mean distance of the substrate to the active site was high in the WT enzyme, measuring 17.10 $\pm$ 9.25 Å, suggesting the benzyl acrylate substrate stays away from the active site and has a very low chance of undergoing catalysis. However, during DE, the mean distance reduced in LVRQ (7.78 $\pm$ 3.53 Å) and LVRQL (8.87 $\pm$ 8.72 Å), signifying an improvement in substrate accessibility to the active site. The large standard deviation observed for LVRQL indicates instances where the substrate approaches close to the active site but, on average, remains further away. The mean and the deviation of distances from each of the five LVRQL runs can be seen in **Fig. D.2**. The reduction in substrate distance from the active site was further pronounced in the GLVRSQL variant (4.44 $\pm$ 0.90 Å), suggesting a significant enhancement in substrate entry and stabilization within the active site. However, in the GLAVRSQLL variant, the distance of the substrate to the active site remained comparable to GLVRSQL (4.53 $\pm$ 1.09 Å), while the experimental yield dramatically increased. This observation challenges the notion that solely substrate access to the active site dictates yield enhancements. While enhanced substrate access might be a major contributor to yield improvement from WT to GLVRSQL (0-6%), it could not explain the drastic increase in yield going from GLVRSQL to GLAVRSQLL (6-28%) [221].

Further, we sought to explore if enhanced benzyl acrylate substrate binding is the reason for the observed yield increase [Fig. 4.2]. MMGBSA binding free energy calculations

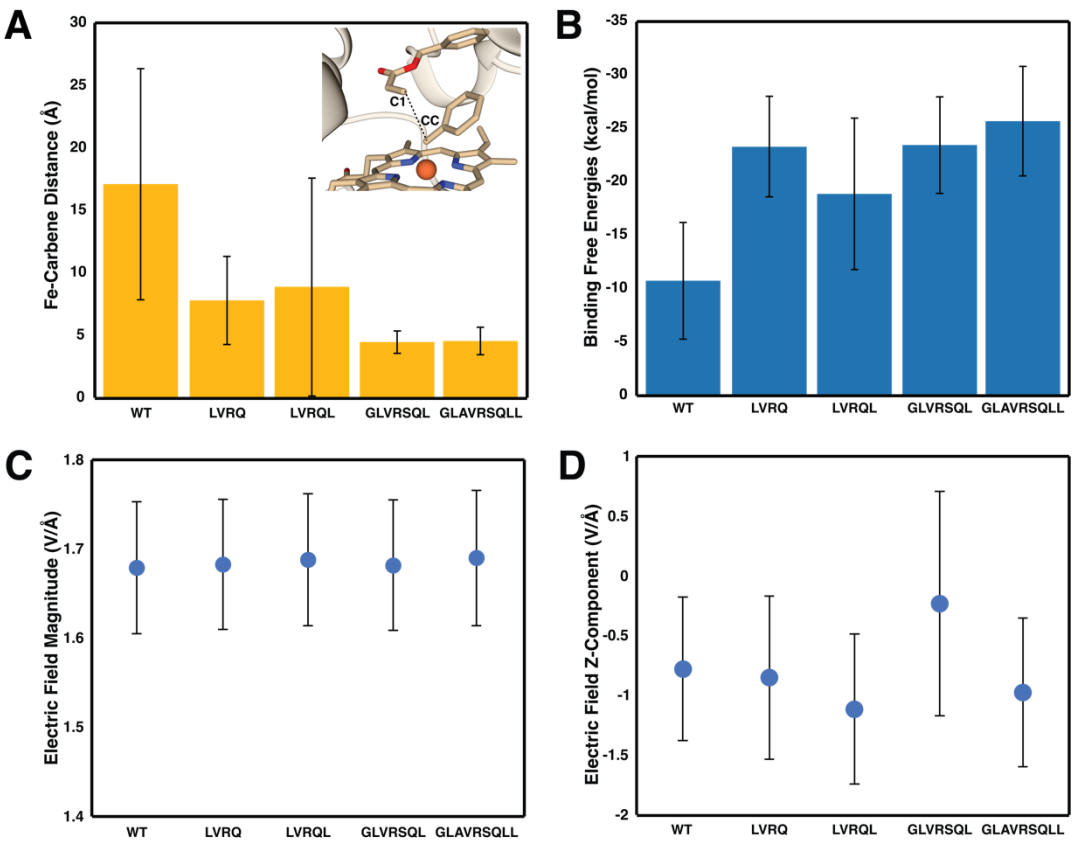


Figure 4.3: Initial parameters investigated as the cause of higher reactivity along DE path. (A) The mean and standard deviation of Fe-Carbene distance for all MD trajectories across all variants. (B) The mean and standard deviation of substrate-protein binding free energies ($G_{binding}$). (C) The total electric field magnitude computed on the Fe-Carbene bond of IPC for all systems across replica molecular dynamics. (D) The z-component of the electric field computed at the center of the Fe-Carbene bond of IPC for all systems across replica molecular dynamics.

were performed, across the five replica MDs for each variant. The MMGBSA method has demonstrated efficacy in reproducing and rationalizing experimental outcomes, accurately reflecting the trends in experimental binding free energies across a variety of ligands and protein systems. [130, 273] Notably, the substrate binding in the evolved variants (LVRQ: -23 ± 4, LVRQL: -18 ± 7, GLVRSQL: -23 ± 4, GLAVRSQLL: -25 ± 5 kcal/mol) was consistently stronger than in the WT (-10 ± 5 kcal/mol). However, the free energy of substrate binding did not reveal a discernible trend across the DE path. Specifically, LVRQ exhibited a higher binding energy, albeit at a larger distance from the active site, indicating strong substrate

binding at non-active site regions and potentially contributing to the lower yield of carbene transfer [Fig. D.3]. The mean binding free energies for GLVRSQL (-23 ± 4 kcal/mol) and GLAVRSQLL (-25 ± 5 kcal/mol) were within their respective standard deviations, and thus again, failing to provide a definitive explanation for the substantial yield increase from GLVRSQL to GLAVRSQLL. These findings suggest that while substrate binding energy is an important factor, it also does not adequately justify the enhancement in yield during DE of Protoglobin.

4.3.2 Electric field evolution during directed evolution

Now, we pivot towards analyzing if the electric fields generated by the studied enzyme changes, and its link to Protoglobin reactivity. Enzyme catalysis is often attributed to electrostatic preorganization and dynamics, [51,89,117,256,311] occasionally put in contradiction with each other [305]. We have previously observed that the reactivity of Fe-heme oxidoreductases is strongly regulated by the electric field form the protein scaffold, in addition to the regulation by the axial ligand to Fe [41]. Electrostatic preorganization has also been cited previously as a compass of directed evolution of Kemp eliminases [165]. To comprehensively address both electric fields and dynamics, we performed an electric field analysis over several replica MD trajectories to sample and compare the electrostatic behavior of the enzyme in a dynamic fashion.

We performed point electric fields calculations at the center of the Fe-carbene bond, for the carbene-substrate intermediate replica MDs of all systems. The mean electric field magnitude is evidently seen to not change meaningfully along DE, with only a very small decrease in the GLVRSQL variant [Fig. 4.3]. A more noteworthy observation emerged when examining the z-component of the electric field measured at the Fe-carbene bond center (the z-component being normal to the heme plane). The projection shows larger variation across the mutants, especially in the field directionality [Fig. 4.3]. This suggests that a point electric field-based analysis is not enough to capture the changes in the heterogeneous

3-dimensional (3D) electric field of the enzyme, therefore requiring a more comprehensive approach (introduced in Fig. 4.2).

We previously developed a method to quantify the heterogeneous 3D electric field topology in volumes within an enzyme active site, and map it to a single metric [127]. Using this method, we correlated 3D electric fields in ketosteroid isomerases to their reactivity. This is notable as electric field magnitudes at a point or along a particular bond showed no relationship between these two variables [127]. This approach involves defining a volume of interest for electric field topology calculations, which, in this study, is a cubic box centered on the Fe-carbene bond [Fig. 4.4]. The heterogeneous electric field was calculated for a total of 5,000 frames derived from 5 x 100 ns replica MD runs for each variant. To analyze this vast dataset, we employed an affinity propagation algorithm to cluster similar electric field topologies within a dynamical trajectory. Affinity propagation provides a distinct advantage by eliminating *a priori* knowledge of the number of clusters. This flexibility allows us to track the changes in the distribution and the number of clusters along the DE of Protoglobin - signaling how diverse or, inversely, tightly controlled the electric field is within the protein's active site. Additionally, this clustering algorithm yields a single best representative frame for each cluster, aiding visualization and further analysis of the 3D electric field that the active site samples. The predominant clusters (those representing >5% of the MDs) from each system were considered and subsequently compared using a distance matrix [Fig. 4.4]. A distance closer to 0 indicates high similarity in the 3D heterogeneous electric field topologies, while a score of 1 indicates high dissimilarity, for example, between WT and evolved variants.

WT Protoglobin features two highly distinct 3D electric fields, with clusters WT-EF1 (visited by the system 64.53% of the time) and WT-EF2 (34.58%). The LVRQ variant introduces four electric field clusters, LVRQ-EF1 to LVRQ-EF4, with visitations of 58.32%, 19.36%, 8.60%, and 5.68%, respectively. LVRQ-EF1 and LVRQ-EF3 closely resemble each other and WT-EF1, while LVRQ-EF2 and LVRQ-EF4 diverge significantly, marking the introduction of two novel electric fields. LVRQL further evolves this pattern, showing two main

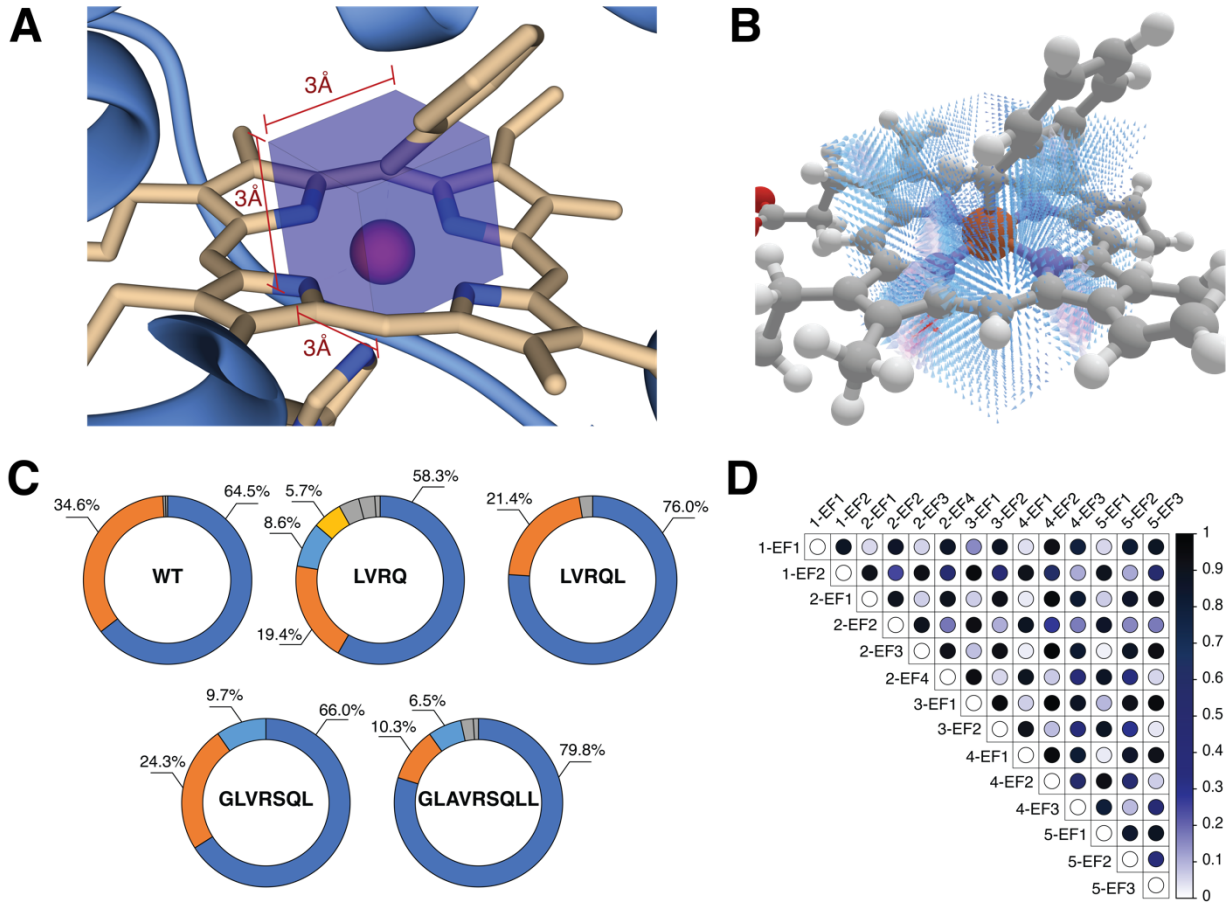


Figure 4.4: (A) Illustration of a 3Å box centered on the Fe-carbene bond for calculating the 3D heterogeneous electric field topology. (B) Example of a 3D heterogeneous electric field topology calculation. (C) Affinity Propagation clustering of electric field topologies for each variant, with blue indicating the most prevalent, orange the second, and green the third; clusters under 5% are in grey.(D) A pairwise distance matrix comparing the similarity (0) or difference (1) of electric field topology clusters across all systems. The first number in the labels indicate the stage of directed evolution (1=WT, 5=GLAVRSQLL), and the second number indicates how often the field topology is visited along the trajectory (1=the most frequently visited).

fields: LVRQL-EF1 (76.04%) and LVRQL-EF2 (21.36%), which are derivatives of LVRQ's clusters, illustrating an ongoing modification from WT through DE. GLVRSQL presents three clusters: GLVRSQL-EF1 (65.98%), GLVRSQL-EF2 (24.34%), and GLVRSQL-EF3 (9.68%), with EF1 and EF3 showing regressive similarity to WT-EF1 and EF2, respectively, while GLVRSQL-EF2 (0.92 and 0.62 from WT-EFs) remains distinct, reflecting influences from LVRQ-EF4 and LVRQL-EF2. The final GLAVRSQLL variant has three clusters:

GLAVRSQLL-EF1 (79.82%), GLAVRSQLL-EF2 (10.34%), and GLAVRSQLL-EF3 (6.46%). GLAVRSQLL-EF1 demonstrate nuanced similarities to WT-EF1 (0.05) and GLAVRSQLL-EF2 is closest related to GLVRSQL-EF3 and WT-EF2 (0.08 and 0.11, respectively), indicating evolutionary modifications. In contrast, GLAVRSQLL-EF3 introduces a distinct electric field, diverging from WT, which evolved throughout the directed evolution process. It now remains to be seen how these field variations impact the reactivity.

4.3.3 Link between evolving electric fields and reactivity changes

The carbene transfer reaction in the engineered Protoglobin proceeds through the iron porphyrin carbene (IPC) intermediate with the substrate bound nearby [69]. The IPC intermediate contains a highly reactive carbene carbon, which reacts with the double bond in the benzyl acrylate substrate, leading to the formation of two new carbon-carbon bonds and culminating in the formation of a cyclopropane ring embedded within the substrate. The IPC intermediate is capable of adopting three spin states, each potentially influencing the cyclopropanation pathway differently. However, most experimental evidence points to the existence of a closed-shell singlet spin state [150,151]. Unpaired electron states (triplet or open shell singlet) may lead to a stepwise process, while a closed shell singlet state favors a direct, concerted mechanism either synchronous or asynchronous, without intermediates [70]. To explore the cyclopropanation reactivity, we performed hybrid quantum mechanics/molecular mechanics (QM/MM) calculations on the cluster centers for WT Protoglobin and the evolved variants. The calculations indicate the preferred spin state for the IPC complex is the closed shell singlet, favored over the triplet by 14.7 kcal/mol, with several attempts to converge the open shell singlet IPC leading to the closed shell singlet structure. The spin preference for closed shell singlet IPC is also supported by similarities in Fe-CC bond lengths between the QM/MM optimized closed-shell singlet state (1.79 Å) and the crystallized Protoglobin IPC intermediate (1.74 Å), contrasting with the longer bond length (1.93 Å) in the triplet state [69]. Additional calculations across the complete reaction profile further confirm that

the triplet spin state is energetically higher than the closed-shell singlet spin state. Moreover, the only successfully optimized open-shell singlet structure also exhibits an energy level comparable to that of the corresponding triplet state [Fig. D.4]. Therefore, all further calculations were performed at the open-shell singlet spin state.

The reactivity calculations using a hybrid QM/MM method were conducted on all electric field clusters for variants containing the substrate within a reactive proximity (<5 Å) to the CC. The results showed that for the WT-RCs and EF1, EF2, and EF4 clusters for the LVRQ variant, the substrate was positioned at distances greater than the reactive range from the CC atom, classifying these states as unreactive [Tab. D.2]. Consistent with other carbene transfer studies, all reactive clusters for the LVRQ, LVRQL, GLVRSQL, and GLAVRSQLL variants, with closed shell singlet spin state, demonstrated a concerted reaction mechanism, lacking stable intermediates and characterized by the asynchronous formation and breaking of bonds [70, 240, 308]. Initially, the CC and substrate C1 atom bond formation and elongation of the Fe-CC bond is favored, followed by complete breaking of the Fe-CC bond, culminating in the bond formation between CC and C2. For the LVRQ variant with EF3, the Gibbs free energy barrier was identified as 28.8 kcal/mol, coupled with a product stabilization energy of -36.2 kcal/mol. The LVRQL variant exhibited a free energy barrier ranging between 23.8 and 32.0 kcal/mol and product stabilization energies between -38.7 and -41.9 kcal/mol. The GLVRSQL variant showed a barrier range of 27.0 to 35.9 kcal/mol and product stabilization energies between -30.8 and -33.4 kcal/mol. Lastly, the GLAVRSQLL variant displayed a barrier range from 13.4 to 34.6 kcal/mol with product stabilization energies between -34.6 and -53.5 kcal/mol.

The results indicate that LVRQ do not produce very high free energy barriers and in LVRQL the barriers are even lower with a significant reaction exothermicity, aligning with some experimental activities observed [221]. This suggests the observed low experimental reactivity in the LVRQ and LVRQL variants likely not due to a complete lack of intrinsic reactivity but rather from rare visiting of reactive configurations, as suggested by the mean

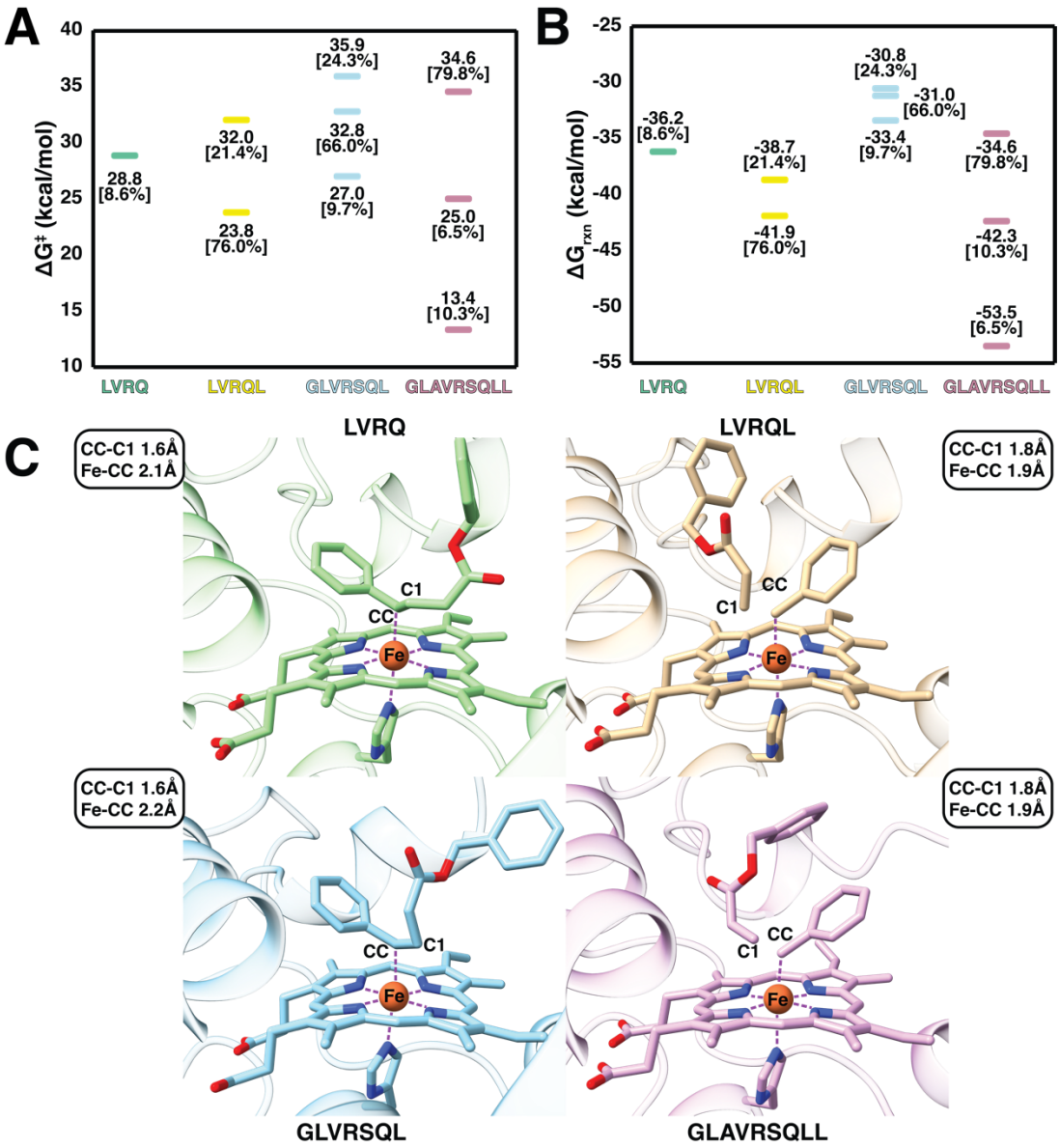


Figure 4.5: (A) Transition state free energy barriers for reactive clusters from each variant; (B) Product stabilization energies for reactive clusters from each variant. (C) Observed transition states from the best performing cluster centers of each variant. Transition state and product stabilization energies/structures were obtained from reaction path scans.

distances from molecular dynamics (MD) simulations (7.78 ± 3.53 Å for LVRQ and 8.87 ± 8.72 Å for LVRQL). Conversely, the primary reason for the low experimental yields in GLVRSQL, despite the close proximity of the benzyl acrylate substrate (MD mean distance of 4.44 ± 0.90 Å), appears to be the absence of effective electric fields necessary for effectively lowering the barrier of the cyclopropanation reaction and improving the overall reaction

energy. This underscores the principle that mere access of the substrate to the active site is insufficient for high yield; the presence of a conducive electric field is critical for enhancing reactivity. The GLAVRSQLL variant, alongside the close binding of the benzyl acrylate substrate (MD mean distances of 4.53 ± 1.09 Å), has an effective electric field leading to both low energy barriers and favorable reaction energy, indicating its proficiency in catalyzing the cyclopropanation reaction. This agrees with and rationalizes the yield increase from approximately 8% in GLVRSQLL to about 28% in GLAVRSQLL. These trends in reactivity are also in line with additional QM/MM calculations with TPSSh and B3LYP functionals [Tab. D.3].

Intriguingly, QM/MM calculations also reveal that the nature of the reaction TS within different enzyme variants is significantly influenced by the electric fields present. We identified two distinct types of TSs. The first type, observed in the most efficient EF clusters of the variants LVRQL and GLAVRSQLL, is characterized by the formation of the CC-C1 bond accompanied by a slight elongation of the Fe-CC bond. In contrast, the second type of TS, found in LVRQ and GLVRSQL variants for the same cyclopropanation reaction, showcases a fully formed CC-C1 bond and a complete dissociation of the Fe-CC bond. Thus, the distinct 3D electric fields can facilitate a mechanism change of the cyclopropanation reaction. Moreover, within the GLVRSQL variant, EF3 and EF1 both exhibit TS of the second type, whereas EF2 presents a TS of the first type. Hence, enzyme's dynamically visiting diverse electric fields has the potential for diverse mechanistic pathways to be active within the same enzyme.

4.3.4 Principal Component Analysis of the Fields

To link the 3D heterogeneous electric fields to reactivity in a chemically meaningful manner, we employed Principal Component Analysis (PCA). We mapped cluster centers to PCA components constructed from the compiled set of electric fields across all trajectories for each variant. This yields a single basis for electric field variability within

the protein active site. The population density of each mutant, as illustrated in [Fig. D.5], is mapped across PC0-9 components. This mapping reveals that every mutant, including GLVRSQL and GLAVRSQLL, exhibits significant variance from the WT Protoglobin along several PC components, confirming that DE influences the electric field and its dynamics within the active site considerably. The most dramatic shift between variants GLVRSQL and GLAVRSQLL, the mutations that incur the greatest change in activity, is observed along component PC9 [Fig. 4.6]. The population density of GLAVRSQLL shifts positively along PC9, suggesting a robust alignment of its electric field with this PC. The findings that the most pronounced changes occur in higher-order components point to a multifaceted impact of mutations on the electric field's characteristics, re-emphasizing that the full spectrum of electric field components, rather solely the dominant one, must be analyzed to elucidate the role of fields in the catalytic process.

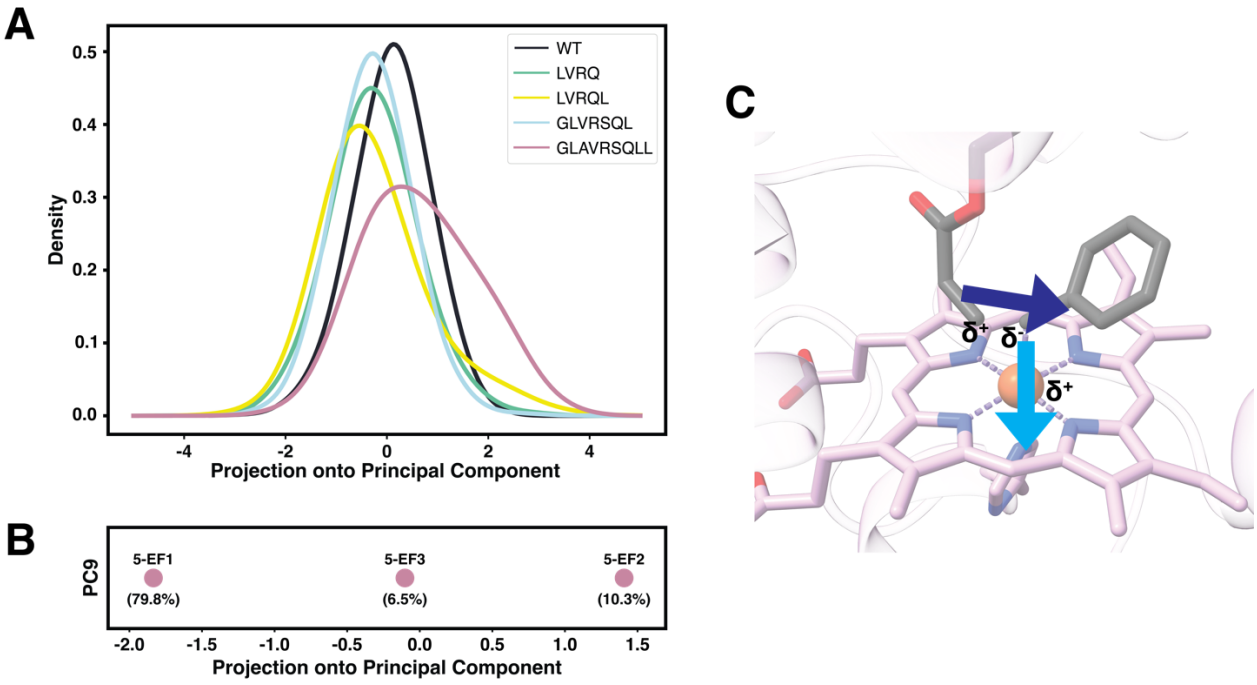


Figure 4.6: (A) Distribution of structures from replica molecular dynamics of all systems across the Principal Component 9. (B) Projections of GLAVRSQLL electric field cluster centers on PC9. (C) Schematic of the PC9 direction plotted on **TS-GLAVRSQLL-EF2** with the relative partial charges polarization marked on the atoms involved in bond rearrangements.

The isolated PC9 can be analyzed visually. It is curvy, and defined by two main directions: one tracing the path from the CC atom to the Fe and the other – from the C1 atom of the benzyl acrylate to the CC [Fig. 4.6, D.6]. This field is straightforwardly linked to chemistry: in the TSs, the C1-CC bond is formed, and the electron density shifts from CC to C1 – a shift aided by the field of opposing direction. Similarly, the field pointing from CC to Fe aids the Fe-CC bond breaking in the TS. This implies that the intrinsic electric field alignment with PC9 in GLAVRSQLL facilitates the barrier crossing. This relationship becomes even clearer when we plot the electric field clusters along PC9. We observe that the degree of alignment with PC9 in GLAVRSQLL directly corresponds with the free energy barrier [Fig. 4.5, 4.6]. Hence, the efficient catalysis observed in GLAVRSQLL is largely driven by a shift in its intrinsic electric field toward the positive direction of PC9, which plays a key role in stabilizing its TS. The development of a PC9-type electric field appears to be a key achievement of DE of Protoglobin.

4.4 Conclusions

While computational design often struggles to enable enzymes to catalyze new chemical reactions, DE has emerged as a potent method for imparting novel catalytic abilities to enzymes. This contrast poses a crucial dichotomy: despite being effective, DE is a black box method where mechanisms for improved activity are obfuscated by the enzyme complexity. We shed new light on one possible mechanism by studying Protoglobin, a protein that, through DE, has developed the ability to catalyze carbene transfer reactions, leading to the cyclopropanation of benzyl acrylate. Initial analysis of multiple MD simulations of wild-type Protoglobin and its four evolved variants indicated that merely enhancing substrate access and binding to the active site does not fully explain the improved cyclopropanation yield. Therefore, we turned our attention to the enzyme's electrostatic preorganization. We have developed a detailed and broadly applicable protocol to measure the 3D electric field topology

and dynamics, and analyzed and compared these dynamic fields along the DE path using an affinity propagation clustering algorithm. We discovered significant alterations in the active site electric field as Protoglobin evolved. Through PCA, we identified a chemically meaningful field component that emerges and takes the lead during DE and facilitates crossing the barrier to carbene transfer. The catalytic role of the evolved electric field was confirmed by QM/MM mechanistic calculations. These calculations revealed that the nature of the reaction TS (concerted Fe-CC bond breaking and C1-CC bond formation, or asynchronous and led by the Fe-CC bond breaking) can be altered by the field geometry. In summary, fine-tuning the global electric field in the active site appears to be the key achievement of DE and is, therefore, an aspirational goal for *de novo* enzyme design.

4.5 Acknowledgements

This chapter was adopted from:

Directed Evolution of Protoglobin Optimizes the Enzyme Electric Field. Shobhit S. Chaturvedi, Santiago Vargas, Pujan Ajmera, and Anastassia N. Alexandrova. *Journal of the American Chemical Society* **2024** 146 (24), 16670-16680 DOI: 10.1021/jacs.4c03914

Here I contributed code for machine learning, clustering, visualization, data processing, writing, high-throughput calculations, and editing. S.S.C. computed MDs, wrote, helped create visualizations, and edited. P.A. contributed to writing, visualization, and data processing. A.N.A. provided direction, writing, editing.

Chapter 5

Machine-Learning Prediction of Protein Function from the Portrait of its Intramolecular Electric Field

5.1 Introduction

The notion that electric fields can act as catalytic components deviates from the framework that catalysts must be purely chemical. Numerous studies have demonstrated that electric fields significantly influence both chemical reactivity and selectivity across a wide range of proteins, both metal-free and metal-dependent [3, 40, 52, 73, 87, 127, 134, 137, 140, 188, 199, 257–259, 295] Ketosteroid isomerases (KSI) became a key example demonstrating this effect through several *in silico* studies [82, 303]. Following these numerous computational demonstrations, Boxer provided experimental validation by showing that in ketosteroid isomerases, the electric field acting on the charged enolate intermediate correlated with the reaction's free energy barrier [87]. Subsequently, the quantum theory of atoms in molecules (QTAIM) was employed to examine how electric fields impact the reactivity of KSI, where it revealed that fields manipulate electron density throughout the substrate [96, 127, 316].

In the realm of metal-containing enzymes, significant attention has been devoted to exploring electric fields within Fe-heme containing enzymes and their model systems [41, 54, 166, 272]. Even with identical Fe-heme coordination, mere variation in axial ligands such as cysteine, histidine, and tyrosine, heme enzymes exhibit diverse reactivity. Our prior research unearthed a pivotal revelation: beyond the axial ligand, the electric field from the surrounding protein (excluding the heme and the axial ligand) strongly influences reactivity [41]. This underscores the heme scaffold's role as a molecular capacitor, where specific configurations of charged amino acids generate a characteristic electric field along the Fe(IV)=O bond in Compound I (F_z). Notably, we predicted that a heme equipped with the suitable axial ligand for its intended function yet situated within a protein environment typical of a different class of oxidoreductases may acquire an unintended function, such as off-pathway oxidation [41]. In a recent study of laboratory evolved protoglobin for the catalysis of carbene transfer reactions, we furthermore showed that it is the catalytic component of the electric field in the active site that the evolution develops in its course [54]. We infer the impact of fields within protein active sites on chemical reactivity, and thus offer another avenue in the pursuit of effective protein design [51]. Such insights could bridge the existing gap between computationally designed proteins and genuinely effective enzymes, whether naturally occurring or laboratory evolved.

Since electric fields are so prominent in governing enzyme reactivity, here we flip the problem and explore whether machine learning can predict enzyme reactivity solely based on the electric field of the protein scaffold. For this purpose, we use the previously reported dataset of *c.* 200 Hemoglobin proteins [41] and, with the electric field as a sole input, classify these proteins as catalases, peroxidases, or monooxygenases [Fig. 5.1]. In other words, we train a ML model that would predict the heme Fe reactivity strictly from the heterogeneous field that the protein produces. Indeed, the task is analogous to a classic image recognition problem where spatial field components act as pixel components for machine learning algorithms.

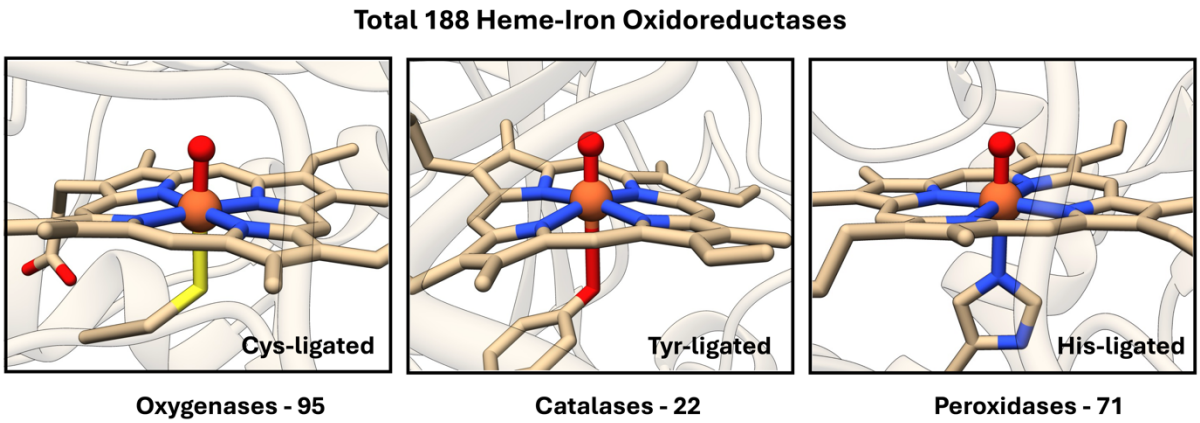


Figure 5.1: The dataset includes three classes of hemes: oxygenases, catalases, and peroxidases, each with distinct axial ligands. The total number of examples for each class is indicated on the figure, highlighting the representation of each class within the dataset.

5.2 Methods

Despite the broad success of theoretical analyses of electrostatic preorganization, they often have two shortcomings: firstly, they lack dynamic information, in the sense of the dynamics of the field itself. Naturally, the structural dynamics of the protein is included via molecular dynamics (MD) simulations and subsequent averaging of computed properties, such as reaction barriers and electric fields. Some exceptions exist; for example, the effects of KSI conformational changes on the electric field have been tracked to explain transition state stability [311]. Secondly, analysis is generally reduced to a field at a single point in an enzyme. The reason that the single point analysis is incomplete is that, for many systems, the reaction mechanism is not localized to a single bond. For example, the ubiquitous Diels-Alder reaction as an example where reactivity is delocalized across a number of atoms and bonds. Recently, a second dimension was added to field analysis, mitigating the problem to an extent [137, 295]. Here, we analyze the field in the active site in its entirety, considering also field dynamics, and then use the fundamental components of the field from dimensional reduction and machine

learning, for protein function recognition.

The issue of ingesting raw heterogeneous electric fields is dimensionally daunting - a coarse sampling of electric field values can lead to tens of thousands of input dimensions as each spatial point is associated with three directional components. This scaling leads to an intractable problem for manual analysis where we simply cannot separate signal from noise in such a high dimensional space. In addition, even statistical/machine-learned (ML) methods can struggle to find meaningful descriptors without a large enough dataset for either supervised or unsupervised machine learning tasks. We address this by using dimensionality reduction, via principal component analysis (PCA), to create a more manageable, data-informed set of input dimensions. PCA is often used as a preprocessing step before supervised machine learning tasks to reduce noise in data and simplify learning tasks. For our use case, PCA was highly attractive as it is a data-first representation scheme where prior knowledge of a system is not necessary. This establishes our framework as a universal scheme that could be used to study and explain families of proteins where domain knowledge is lacking or where representative fields are simply too complex to construct *a priori*. We envision using this methodology to recognize the functions of active sites of newly discovered proteins, distinguish active sites from areas in proteins that look like active sites but are not, and attributing selectivity to an enzyme without lengthy mechanistic investigations.

5.2.1 System set up and field calculations and analysis

To represent each protein, we take crystal structures from the Protein Data Bank, remove co-crystalized water molecules and ions, and zero the charges on the axial ligands, and the hemoglobin itself. First, we develop and ML algorithms that operate on the point field at the Fe, then – the 3-D field in a volume around the Fe without dynamics, and then extend this study to include the dynamics and clustering of the field. The fields are computed classically using the point charges of the protein, and thus excluding the Fe(IV)=O moiety, the heme, and the axial ligand. The 3-D fields were constructed on the grid over a cubic box centered

at the Fe atom in the CpdI intermediate [Fig. 5.2], the box (dimensions: 3 Å x 3 Å x 3 Å) is visualized in Fig. 5.2. The grid spacing was 21 sampling points along each dimension for a total of *c.* 9,200 points. In the context of molecular dynamics, the fields are compared to each other using the global distribution of streamlines method.

In detail, our group previously adapted a distance metric from fluid dynamics to study the differences between complex, heterogenous electric fields [127]. This method constructs a global distribution of slipstreams within a vector field, yielding histograms that describe an electric field. The formulation enjoys important mathematical properties such as rotational, scalar, and translational invariance. Here, within the 3 x 3 x 3 Å cube, random points are sampled to create linearizations, known as slipstreams. Random points along a given slipstream are selected to compute mean curvature and distance of a line where curvature is defined as

$$\kappa = \frac{\left|\left|r^{'}\left(t\right) \mathrm{x} \, r^{''}\left(t\right)\right|\right|}{\left|\left|r^{'}\left(t\right)\right|\right|^{3}} \tag{5.1}$$

A histogram of L2 distance to curvature can thereby be compiled and the distance between two discrete distributions can be computed via the χ^2 distance:

$$\chi^2 : D\left(f,g\right) = \frac{1}{2}\sum_{i=1}^{N}\frac{\left(f\left[i\right]-g\left[i\right]\right)^2}{f\left[i\right]+g\left[i\right]} \tag{5.2}$$

With a defined distance comparing electric fields, we can then create a graph where the edge weights are the distances between two electric fields. This graph encoding is ripe for graph compression algorithms, notably affinity propagation, to aid in the selection of a few representative frames entirely on the basis of the 3-D heterogeneous electric field. Our group has previously used this protocol to interpret the dynamic heterogeneous field differences along a directed evolution pathway of catalytic protoglobin complexes [54]. We used these compressed representations of the electric field along the entire MD trajectory to further study the effects of the 3-D electric field on electronic populations within the active site.

With this, we demonstrate the relationship between induced fields at the active site and the overall protein activity.

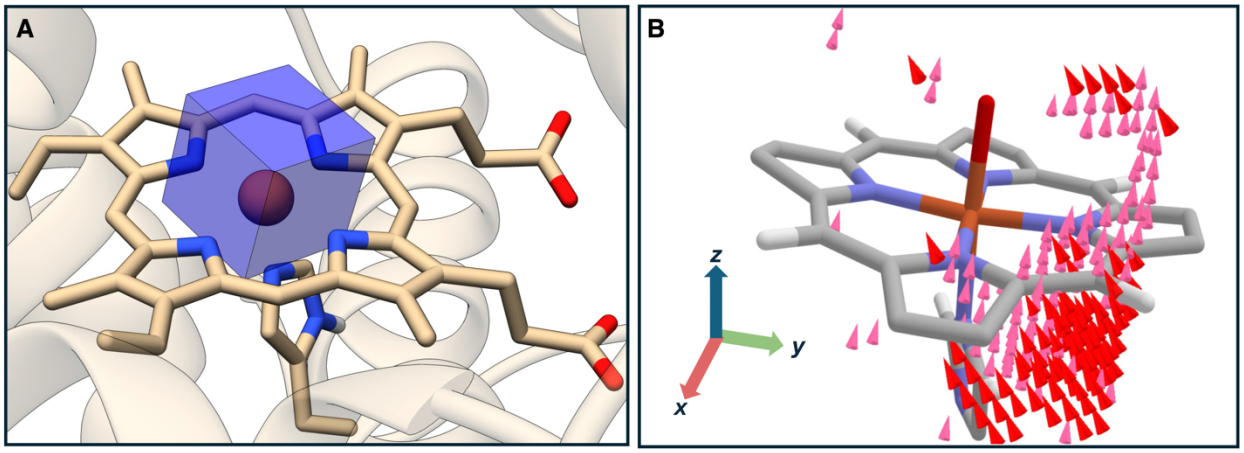


Figure 5.2: (a) The cubic box centered on Fe, used for computing the electric field on the grid. (b) An example of typical principal component computed on the dataset, plotted on the exponential scale for clarity.

5.2.2 PCA

To determine the proper number of PCA components, we swept the number of PCA components of the electric field in the model from 5 to 25 PCA components and used cross-fold validation to select the optimal number of components. We found that 9 components were optimal for performance on validation data. For validation and testing, we split our dataset into an 80-20 train-test set and used k-folds (k=5) training-validation splits to tune model parameters, PCA components were constructed entirely from the training split to avoid data leakage into the test set.

5.2.3 Molecular Dynamics

We parametrized the Fe-containing heme active site for MD simulation with the Metal Centre Parameter Builder (MCPB.py [174]. We modeled the remainder of the protein using the Amber FF19SB force field [277]. The leap module in AMBER 22 was utilized to introduce Na^+ counterions to neutralize protein systems [46]. These systems were then placed in a rectangular box, surrounded by OPC water molecules [136] extending at least 10 Å beyond the outermost boundary of the protein. We applied periodic boundary conditions throughout the simulations. The particle mesh Ewald method was used to calculate long-range electrostatic interactions, with both the direct space and the van der Waals interactions capped at a 10 Å cutoff. The protein systems was minimizated, initially with 5,000 steps of steepest descent followed by another 5,000 steps using the conjugate gradient method, all under a 100 kcal mol^{-1} $Å^2$ restraint on the solute molecules. This was succeeded by another round of full system minimization employing the same descent and gradient steps. Subsequently, the systems were gradually heated from 0 to 300 K in an NVT ensemble, controlled by a Langevin thermostat with a collision frequency of 1 ps-1 over 250 ps, while the solute molecules were held under a 50 kcal mol-1 Å2 harmonic restraint. Bonds involving hydrogen were constrained by the SHAKE algorithm [244]. Following this, a 1 ns lightly restrained MD simulation was conducted to stabilize the density under periodic boundary conditions. All systems were equilibrated at 300 K for 3 ns in an NPT ensemble, using the Berendsen barostat to maintain pressure at 1 bar, without restraints. A 100 ns productive MD simulation was carried out for each system in an NPT ensemble, maintaining a constant pressure of 1 bar with a 2 ps pressure coupling, using the GPU-accelerated version of AMBER 22 [46]. **The trajectories are subjected to field topology calculation (using the CPET code)** [127] **via embedding the active site in the point charges. The fields were then compared to each other along the trajectory and clustered by the topology similarity.** [54].

5.2.4 Quantum Mechanics/Molecular Mechanics (QM/MM) calculations

Quantum mechanics/molecular mechanics (QM/MM) calculations were conducted using the ChemShell [195] software suite, integrating Turbomole [8] for quantum mechanics calculations and DL_POLY [264] for molecular mechanics. For these calculations, water molecules beyond a 10 Å solvation layer surrounding the protein were removed using CPPTRAJ [238], leaving the protein optimally hydrated. The QM region encompassed the heme iron center, the intermediate oxo or hydroxo groups, and the axial ligand located at the active site, similar to our earlier study [41]. The unrestricted B3LYP functional [26], as previously shown to be reasonable for these systems [41], was employed for the QM calculations. The molecular mechanics region was defined as the protein area within 8 Å of the QM zone, while the remaining system components were held static. The Amber FF19SB force field was applied to the molecular mechanics region. Hydrogen link atoms capped the QM/MM boundaries, and a charge shift model was utilized. Electrostatic embedding accounted for the polarizing effects of the protein environment on the QM region. Geometry optimization and frequency analyses utilized the def2-TZVP basis set, with the exception that hydrogens were treated using the def2-SVP basis set. The CpdI Fe(IV)=O ($Por^{+\bullet}$) complex was modelled as a doublet while the CpdII Fe(IV)-OH was modelled as a triplet for all systems.

5.3 Results and Discussion

Single point fields

We used a host of traditional machine learning models due to the relatively middling amount of data, including, XGBoost, Random Forests, Ridge Regression, and K-nearest Neighbors [Fig. 5.2]. To tackle imbalanced data, present by the underrepresentation of catalases (21 proteins in training vs. roughly triple the number of **monooxygenases, peroxidases**) — we

trained Balanced Random Forests algorithms [56]. For hyperparameter tuning, we employed a 5-fold cross-validation method combined with an 80-20 train-test split for both single point and complete, heterogeneous training. To optimize parameter selection further, we used Bayesian optimization techniques in WanDB [33]. The detailed model parameter dictionaries can be found in the supplementary information.

Model	**F1 Score**	**Accuracy**
XGBoost (Single Point, 3-Comp)	0.42	0.44
Balanced Random Forest (3-D Fields, PCA)	0.75	0.82
XGBoost (3-D Fields, PCA)	0.84	0.84

Table 5.1: Performance of the two top performing ML models benchmarked against the top model to predict on a single point (x,y,z components at the Fe in Heme). This is a proxy for the previous mapping of F_z at the Fe to axial ligand. Note the dramatic improvement in performance with a richer set of electric field features.

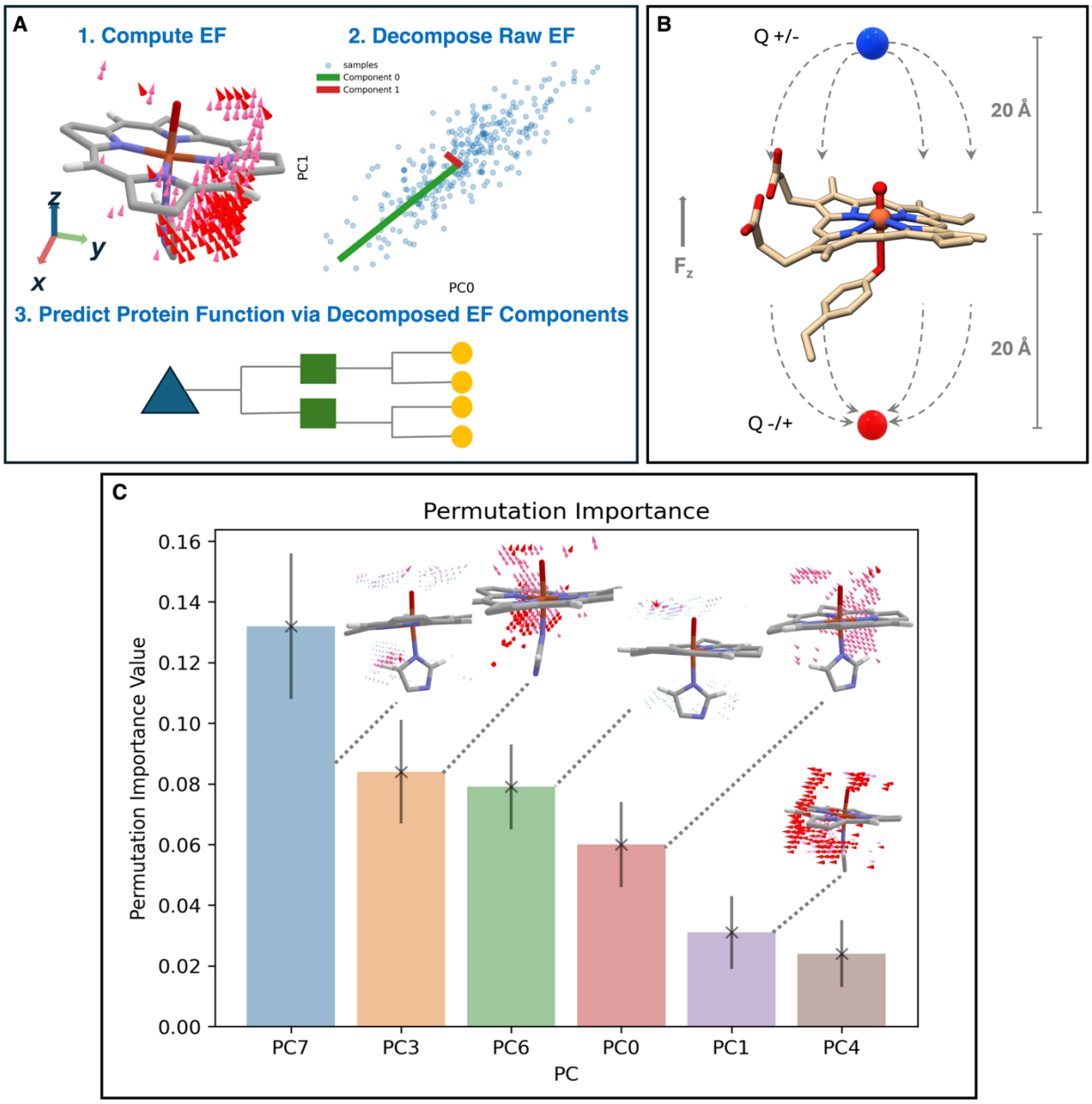


Figure 5.3: (A) Workflow for predicting protein function using Machine Learning models (B) Surrogate model to test ML machinery with applied fields. (C) Principal components selected by permutation importance and Boruta. Visualized structures (PC7, PC3, PC6, and PC4) were also flagged by Boruta as important.

Performance evaluations were conducted using accuracy and F1-scores, providing a holistic

view of model effectiveness [Tab. 5.1]. Considering our dataset's label ratio of roughly 4:3:1, we prioritized the F1-score as a fairer performance metric. All the above-mentioned models were applied to single-point electric field data, with XGBoost emerging as the best performer among them. Focusing on the three components: F_x, F_y, and F_z at the Fe atom, XGBoost achieved an underwhelming F1-score of 0.42 and an accuracy of 0.44, illustrating the limitations when relying solely on point electric fields for predicting protein functions. The results indicate that while point electric fields offer a straightforward interpretation, they are insufficient for capturing the comprehensive detail required for accurate model predictions.

5.3.1 3-D fields

In stark contrast, incorporating a full 3-D heterogeneous electric field representation, through PCA, significantly enhances model performance, achieving accuracy and F1 scores of up to 84% and 0.84 respectively. This contrasting difference underlines the inadequacy of point electric fields as simplistic, whereas 3-D heterogeneous electric fields offer more representative depictions of the enzymatic environment [Fig. 5.3]. Moreover, the ability of a machine learning model to predict functions from electric field data of a protein scaffold suggests that the scaffold is evolutionarily optimized to provide the specific fields necessary for efficient catalysis.

5.3.2 Applied Uniform Fields.

Given a machine-learning model trained on compressed electric field representations, we aim to identify which components from the heterogeneous electric field are critical for the model predictions. For this, we utilized the trained, heterogeneous electric field models to predict changes in predicted heme activity with externally applied fields. We aimed to test a crucial hypothesis: whether the magnitude of the applied F_z electric field is decisive in determining their catalytic function. Specifically, we sought to understand if changes along the F_z direction alone could flip the predicted activity of the enzyme. To explore this, we positioned positive

and negative charges 20 Å away from the Fe center of the active site, aligned along the F_z axis on each side of the heme plane [Fig. 5.3]. Here we selected a Tyr-ligated complex (PDB code 2j2m) as a test subject, allowing us to determine if the model could be biased to predict Cys-ligated/oxygenases for positive fields of large magnitudes and His-ligated/peroxidases for significant negative electric fields. This choice of protein, an unseen test example, also belongs to the category of Tyr-ligated proteins that exhibit intermediate, near-zero F_z values. We tested four distinct electric field strengths: +50, +10, -10, and -50 MV/cm along the iron-oxy bond, with the direction of the field indicated by the black arrow in Fig. 5.3. These field intensities were informed by our prior research [41], which categorized Cys, Tyr, and His-ligated heme Fe proteins, under average vertical fields of 28.5 MV/cm, 3 MV/cm, and -8.7 MV/cm, respectively.

Applied Field (MV/cm)	**Predicted Ligand/Activity**
+50	Cys-ligated Oxygenases
+10	Tyr-ligated Catalases
0 (Original)	Tyr-ligated Catalases
-10	His-ligated Peroxidases
-50	Cys-ligated Oxygenases

Table 5.2: Inducing an electric field along the oxy-iron bond modifies the predicted activity of the protein. Notably, large negative fields along the bond led to categorization as C/oxygenases—an outcome that seems unlikely based on our previous studies and thus suggesting a limitation of the low-dimensional, uniform electric field applied.

Our most effective model seemingly shows mixed success in predicting enzyme activity

with applied electric fields, as illustrated by the results presented in [Tab. 5.2]. The model correctly altered its predictions for most cases: a large (+50 MV/cm) positive field switched the accurate prediction from a Tyr-ligated catalase to a Cys-ligated oxygenase, while a moderate (-10 MV/cm) negative field led to a prediction of a His-ligated peroxidase. However, the model's limitations became apparent in certain cases; notably, a strong negative field along the Fe(IV)=O bond incorrectly predicted a Cys-ligated oxygenase—an outcome that seems unlikely considering the typically moderate F_z component magnitudes observed in this family of proteins. These discrepancies suggest that the model might be utilizing more than just the F_z electric field component from the heterogeneous 3-D electric field of the protein in making its predictions.

5.3.3 Feature Importance

Therefore, we conducted a feature importance analysis to identify all the crucial features (i.e., the key principal components) involved in the model's accurate decision-making process. A naive approach would be to consider the % explained variance of each PCA component and assert that the most variable components impact activity more. This is imperfect for several reasons. First, correlation is not causation and this signifies that components with a large variance determine ligand specificity. Looking at the correlation or variance in a single component also ignores the effects that multiple vector field components might have in conjunction. Finally, PCA does not intake labels in a supervised manner, thus these components have no mapping to function directly. To address this, we utilized Boruta [162] and permutation importance [37] feature selection. Boruta is built on top of permutation feature importance, where individual variables are shuffled between examples and the resulting change in performance gives a quantitative measure of how important that feature was to a model's prediction. Boruta extends this idea by constructing "shadow features" that are Gaussian noise with the same mean and variance as true variables in the input of a model. These features, which by construction are random, serve as a benchmark of importance for

other variables; if a variable is more important in permutation importance than a shadow feature it is more likely to be of importance in predicting a target label. This process is repeated a fixed number of times and these trials, in conjunction, creating a binomial distribution where features eventually fall into the tails of the distribution - important or not important. The resulting components from this feature selection step were studied by backtracking PCA components to their original electric field motifs.

Between Boruta and permutation importance, PC0, PC3, PC4, PC6, and PC7 were the most informative to the model. Visualizing these features [Fig. 5.3], we can summarize that a rich host of electric field features inform model predictions. Important components such as the field along the iron-oxy bond emerge, corroborating previous findings and supporting the notion that fields will shift electron distribution along this bond to promote the activation of substrates and control the selectivity. Combined, PC0 and PC3 have strong components along the Fe(IV)=O bonds, but opposite lateral components - suggesting that they together could explain the strong "vertical" component also previously proposed. PC4 is an entirely lateral field component - not previously established as an important motif in Heme selectivity. PC4 might contribute to the placement and delocalization of the radical on the porphyrin (versus on the nearby Trp residue), particularly in the His-ligated proteins. Components PC6 and PC7 are harder to decipher visually - they have strong compressive/expansive features that shift electric fields into or out of the heme center and might control the access to the active site. These components are undoubtedly complex and underscore the difficulty of fully interpreting the effect of electric field processes *a priori* without a statistical, high-throughput approach. It is also noteworthy that the most variable field components, as indicated by percentage explained variability, were not necessarily the most informative for the models. Thus, our findings demonstrate that features of the 3-D electric field, extending beyond just the F_z component, are crucial for enhancing the accuracy of model predictions related to enzyme activity. This underscores that enzymes utilize these diverse directionalities within the 3-D field at the active site to drive their catalytic functions.

5.3.4 Dynamic 3-D fields.

To build upon our static, single-frame analysis, we expanded our approach to incorporate temporal information via molecular dynamics (MD) trajectories of known proteins from each class. The premise here is that the field, as much as the protein producing it, is not static and that particularly functional fields may emerge dynamically. We selected a training set consisting of the proteins 1dgh and 1gwf (Tyr-ligated), 1ebe and 1hch (His-ligated), and 4g3j(Cys-ligated), and designated one protein from each class for the test set: 1u5u (Tyr-ligated), 3xvi (His-ligated), and 1jio (Cys-ligated) – again, ligation being linked to catalase, peroxidases, and monoxygenase activity, respectively. Employing the same suite of models, we optimized parameters using subsets of the electric fields from the training set and implemented a simple majority voting system to determine the protein class/activity. The results reveal that while the models performed well in the static single-frame analysis with a high F1-score of 0.84 using 3-D fields, their performance declined in the dynamic setting, as evidenced by the XGBoost model achieving an F1-score of 0.35 and an accuracy of 0.43 [Tab. 5.3], signifying a drop in the ability of the models to generalize to the dynamic regime. We do note that taking a majority-vote approach to predicting activity from MD trajectories, we are able to predict the activity of 2 out of 3 protein classes correctly.

MD Trials	**Test F1**	**Test Acc**
XGBoost, MD	**0.35**	**0.43**
XGBoost, MD (Combined PCAs)	**0.59**	**0.59**

Table 5.3: The data illustrates the performance outcomes for XGBoost models tailored to molecular dynamics simulations.

Protein	Ground Value	Majority Prediction	Majority Prediction (Combined PCAs)
1u5u	Y/catalase	Y/catalase	Y/catalase
3abb	C/oxygenase	Y/catalase	C/oxygenases
1apx	H/peroxidase	H/peroxidase	H/peroxidase

Table 5.4: Predicted activities for proteins in our test set, utilizing two distinct approaches: predictions made with PCA components just from the training set and those using combined PCA components constructed from the training and testing set. The comparative analysis highlights that employing combined PCA components leads to improved prediction accuracy. This improvement suggests that the previously observed poor performance was likely due to the dynamics introducing a broader variety of components.

To better understand why models extended on 3-D electric fields from MD simulations did not perform as expected, we examined the differences between electric fields derived from crystal structures and those obtained from MD simulations [Fig. 5.4]. presents the PCA explained variability, which measures the amount of variance each principal component captures from the dataset. This metric is commonly used to assess dataset dimensionality and complexity. Our analysis revealed significant differences in the cumulative variance between PCA results from crystal structures and those from dynamic simulations. This suggests that dynamic electric fields encapsulate more complex patterns and interactions, which are not as prevalent in the static fields derived from crystal structures. The increased complexity in dynamic fields likely reflects the continual conformational changes and interactions within the protein environment.

Further complicating our model training, there was a noticeable difference in the explained variance between the PCA components derived from our training set (MD simulations) and our test set. Specifically, the training set demonstrated a higher explained variance, with fewer PCA components, compared to the test set. This indicates that the PCA components from the training set may be over fit to a small set of dynamical degrees-of-freedom. Consequently, when these PCA components used to reduce dimensionality in the test set, they may not adequately capture the essential features needed for accurate predictions, leading to a mismatch in the model's ability to generalize. The model trained on less variable and comparatively simpler data from the MD training set struggles to accurately interpret and predict the behavior of complex test data. This issue highlights the need for developing strategies that can better account for and adapt to the variations in electric field complexity between different sets of molecular dynamics data.

To enhance the interpretability of our MD based 3-D electric field analysis and reduce its complexity, we have recently developed a protocol that captures dynamic information regarding the electric fields experienced by the active site of a protein [54], as illustrated in Fig. 5.4, followed by mapping these clusters onto the principal components identified as critical. By capturing the complex dynamic fluctuations within the enzyme's active site, we aim to elucidate how these variations complicate the model's ability to accurately predict enzyme activity.

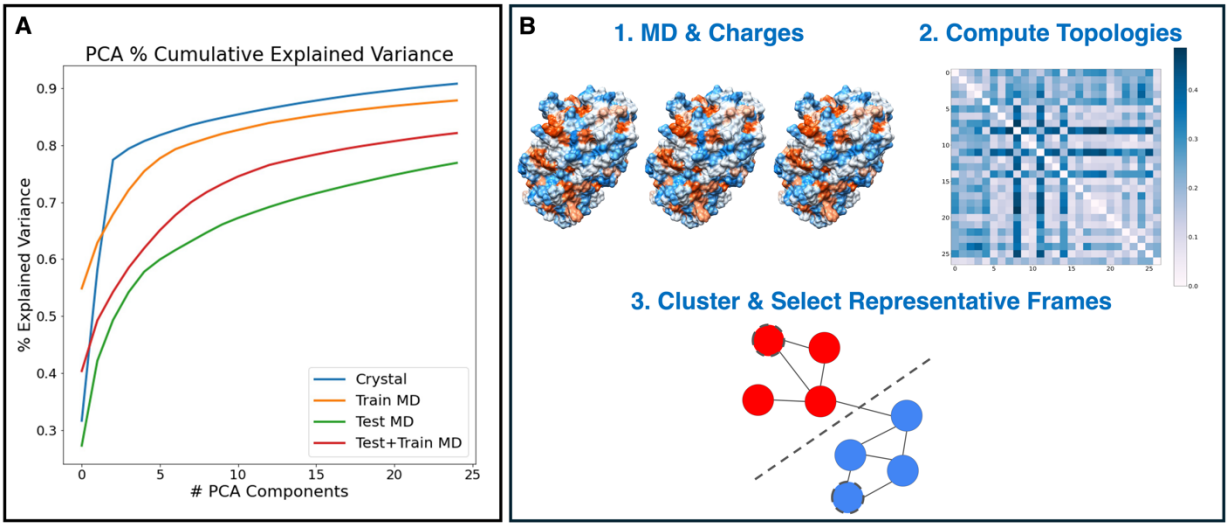


Figure 5.4: (A) Cumulative explained variance between PCAs constructed from crystal structure fields show these fields require fewer components to explain dataset variability. (B) An outline of our method for selecting representative frames based on electric field topologies.

Here we focused on the components that Boruta and permutation importance determined to be critical: PC0, PC3, PC4, PC6, and PC7. PC0, recognized as the most vertical component along the Fe-O bond, exhibited clustering trends that align with our previous studies. The ordering of His $<$ Tyr $<$ Cys within these clusters suggests that cysteine-binding proteins tend to exhibit the most positive electric field components along this direction [Fig. C.1]. However, the presence of both Tyr and His-ligated complexes in the most positive clusters of this component might affect model's accuracy. PC4 exhibits a strong vertical orientation. Notably, clusters representing 1jio are among the most positively positioned on PC4 [Fig. C.6]. While the overall trend of His $<$ Tyr $<$ Cys is maintained, there is significant overlap among the data points of the three protein classes in the projection onto this principal component. In the case of PC7, which introduces a vertical component with some compressive characteristics toward the active site, 1jio is distinctly the most positive, suggesting preorganization of the electric field to enhance activity at the active site [Fig.

C.6]. Contrarily, 1u5u and 3vxi show mixed projections on this component, aligning with prior observations of comparable F_z components between these protein categories. Our analysis on PC6 revealed a lack of clear separation between protein types, indicating that this component is less interpretable compared to others [Fig. C.8]. PC3, characterized by its predominantly horizontal orientation orthogonal to many other significant components, uniquely identified the most positive cluster associated with 1jio(C) [Fig. C.5]. This specificity did not extend to 1u5u and 3vxi, which did not separate distinctly along this component. This structured approach of clustering and principal component mapping has revealed that among the most important principal components for the model, certain components, such as PC0, can distinctly separate the three protein systems within dynamic data, while other components like PC6 and PC3 complicate the clarity of these separations.

System Description	E_H^o (kcal/mol)
1jio (Cys/Oxygenase – 2 Clusters)	68.5 – 92.3
1u5u (Tyr/Catalase – 3 Clusters)	64.7 – 68.5
3vxi (His/Peroxidase – 4 Clusters)	27.6 – 83.7

Table 5.5: Proton-coupled electron transfer potential (E_H^o) ranges for enzyme systems analyzed using QM/MM methods, highlighting variations across different clusters within each enzyme category.

Finally, we aim to explore whether the dynamic complexity, identified through PCA, truly influences factors critical to enzyme activity. To this end, we decipher how the dynamic 3-D heterogeneous electric field affects the electronic structure of the CpdI Fe(IV)=O ($Por^{+\bullet}$) and CpdII Fe(IV)-OH complex by employing quantum mechanics/molecular mechanics (QM/MM) calculations. For these calculations, we have chosen specific model systems that are representative of the enzyme classes under study: 1jio for monoxygenases, 3vxi for

peroxidases, and 1u5u for catalases. The selection of structures such that they represent unique electric field configurations, is vital. Random or field-agnostic selection methods may fail to capture variations caused by heterogeneous electric fields, potentially overlooking critical dynamic interactions that influence enzyme activity. Therefore, we used above identified cluster centers, via electric field clustering, for these calculations. For each major cluster (>10% representation), we computed the free energies of the CpdI Fe(IV)=O and CpdII Fe(IV)-OH variants to assess the relative activity of each cluster along the putative reaction pathway. The computed proton-coupled electron transfer potential (E_H^o) ranges for these clusters are as follows: 68.5 – 92.3 kcal/mol for the Cys-ligated oxygenase system 1jio, 64.7 – 68.5 kcal/mol for 1u5u, and 27.6 – 83.7 kcal/mol for 3vxi [Tab. 5.5]. These values align with the expected trend where Cys-ligated oxygenases exhibit higher reactivity compared to Tyr-ligated catalases and His-ligated peroxidases. Thus, the results indicate that the range of E_H^o values becomes less distinct between the three systems, suggesting that the introduction of dynamics extends the ranges of catalytically relevant properties and diminishes the clear segregation between them. This blurring effect might help explain why dynamics affects the machine learning model's ability to accurately classify the different systems using 3-D electric fields.

In response to these findings, we hypothesized that a model constructed with combined principal components from both the test and training datasets, providing a broader spectrum of variability for the model to learn from, might enhance classification accuracy. Indeed, this approach resulted in improved performance, where our F1 and test accuracy improved to 0.59 and majority vote approach correctly predicts all three test protein categories [Tab. 5.4, 5.3]. Here, we note that mixing train and test components between the sets is neither completely valid nor entirely invalid. On one hand, it introduces bias that can obscure the evaluation of the model's generalization. Therefore, our initial approach avoided this combination. On the other hand, in practical applications, combining electric fields to construct PCAs does not require prior knowledge of protein activity. Consequently, this approach remains a valuable

tool for protein analysis via electric fields.

For future improvements in handling dynamic electric field data, the implementation of highly efficient, sparse neural network architectures and advanced signal processing techniques could be beneficial. Equivariant neural networks, which have rapidly gained traction in scientific fields, are particularly promising due to their efficiency in learning with less data. When integrated with robust data augmentation schemes, these networks can directly process raw electric fields, minimizing data demands while ensuring that key physical symmetries are preserved. Additionally, embracing methods that intrinsically manage structured, temporal data will be essential for extending the analysis to include dynamics natively. Architectures borrowed from natural language processing, such as Long Short-Term Memory (LSTM) networks, or those that incorporate geometric learning, like message-passing graph neural networks, are well-suited for this purpose. These techniques can effectively interpret the temporal variations observed in MD trajectories, potentially enhancing the ability to predict protein behavior based on dynamic electric fields.

5.4 Conclusions

In this study, we have developed a machine-learning pipeline that ingests electric fields, reduces dimensionality via PCA, and applies these fields in a supervised learning task. Our tests on a well-studied family of Fe heme enzymes demonstrated that traditional lower-dimensional analyses of electric fields along the Fe(IV)=O bond are insufficient for accurate activity prediction. This underscores the necessity for analytical techniques capable of parsing the more complex, heterogeneous fields that are actually present at protein active sites. Our findings reveal that point electric field calculations, despite their simplicity and ease of interpretation, do not accurately reflect the true nature of electric fields within these sites. Additionally, when we applied a uniform electric field using our trained model, it failed to induce the predicted changes in a test protein, highlighting the importance of

multidirectional fields in enzyme function. Importantly, our trained machine learning model demonstrated that the enzyme's 3-D heterogeneous electric field alone can predict its function without any other protein-specific information. Through feature selection techniques such as Boruta and permutation importance, we identified key electric field components that not only corroborated previous studies but also emphasized the critical influence of several components alongside the F_z value along the Fe-O bond. Expanding our analysis to include MD trajectories and employing PCA, clustering, and QM/MM calculations, we observed that the inherent complexity in protein dynamics can complicate model predictions. However, we show that if the model is exposed to sufficient dynamic variability, its performance can improve significantly.

This research marks a significant advancement in our understanding of electrostatics in proteins. We have shown that natural enzyme scaffolds have evolved to optimize the electric field at the active site, tailored to their function. This insight offers a powerful tool for predicting potential enzyme functions based solely on their electric fields. Although our analysis focused on heme Fe proteins, the methodology is broadly applicable to any study involving electric fields at largely conserved active sites, even where there is no prior knowledge of crucial field components. Overall, the approach presented here provides a robust framework for not only understanding but also predicting enzyme functions across diverse biological systems based solely on electric field analysis.

5.5 Acknowledgements

This chapter was adopted from recently accepted work in the *Journal of The American Chemical Society*:

Machine-learning prediction of protein function from the portrait of its intramolecular electric field. S. Vargas*, S. Chaturvedi, A.N. Alexandrova.

S.C. helped write, compute molecular dynamics calculations, and comptued QM/MM

structures. A.N.A. provided direction, writing, editing. I contributed code for machine learning, visualization, and data processing. I also wrote and edited.

Appendices

Appendix A

Supporting Information for *Machine Learning to Predict Diels–Alder Reaction Barriers from the Reactant State Electron Density*

A.1 Dataset Statistics and References

Article	$\bar{N}$	$\Delta E^{\ddagger} \pm \sigma$ Barrier [kJ/mol]	Range
Transition State Distortion Energies Correlate with Activation Energies of 1,4-Dihydrogenations and Diels-Alder Cycloadditions of Aromatic Molecules [118].	36	172.3 ± 54	60.0−274.1
Computational Investigation of the Competition between the Concerted Diels-Alder Reaction and Formation of Diradicals in Reactions of Acrylonitrile with Nonpolar Dienes [138].	16	86.0 ± 10.7	74.9−106.8
Diels-Alder Reactivities of Strained and Unstrained Cycloalkenes with Normal and Inverse-Electron-Demand Dienes: Activation Barriers and Distortion/Interaction Analysis [178].	30	94.5 ± 30.9	24.5−139.3
Theoretical Elucidation of the Origins of Substituent and Strain Effects on the Rates of Diels-Alder Reactions of 1,2,4,5-Tetrazines [177].	28	78.4 ± 30.5	28.8−128.8
Origins of Stereoselectivity in Diels-Alder Cycloadditions Catalyzed by Chiral Imidazolidinones [106].	18	63.7 ± 33.2	6.3 − 107.8

Experimental Diels-Alder Reactivities of Cycloalkenones and Cyclic Dienes Explained through Transition-State Distortion Energies [214].	10	92.0 ± 15.2	$65.9-107.9$
Hydrogen Bonding Catalysis Operates by Charge Stabilization in Highly Polar Diels-Alder Reactions [105].	9	37.1 ± 20.9	$11.2 - 81.4$
Diels-Alder Reactions of Cyclopentadiene and 9,10-Dimethylanthracene with Cyanoalkenes: The Performance of Density Functional Theory and Hartree-Fock Calculations for the Prediction of Substituent Effects [143].	6	69.9 ± 14.2	$47.4 - 91.8$
Origins of Stereoselectivity in the trans Diels-Alder Paradigm [215] .	12	97.7 ± 8.8	$76.8-115.0$
Diels-Alder Exo Selectivity in Terminal-Substituted Dienes and Dienophiles: Experimental Discoveries and Computational Explanations [167].	18	29.1 ± 24.2	$5.6 - 89.2$
Hyperconjugative, Secondary Orbital, Electrostatic, and Steric Effects on the Reactivities and Endo and Exo Stereoselectivities of Cyclopropene Diels-Alder Reactions [172].	46	63.6 ± 10.9	$38.0 - 83.9$
Hyperconjugative, Secondary Orbital, Electrostatic, and Steric Effects on the Reactivities and Endo and Exo Stereoselectivities of Cyclopropene Diels-Alder Reactions [173]	17	87.0 ± 11.0	$63.0-118.4$

Lewis Acid Catalysis Alters the Shapes and Produces of Bis-Pericyclic Diels-Alder Transition States [329].	2	63.4 ± 10.3	$56.1 - 70.7$
The Origin of the Halogen Effect on Reactivity and Reversibility of Diels-Alder Cycloadditions Involving Furan [219].	27	95.3 ± 10.6	$80.5 - 114.2$
All Papers	296	87.5 ± 45.0	5.6 - 274.1

Table A.1: Table Containing Different Datasets Used.

A.2 Model Performance

Model	Test MAE (kJ/mol)	Test R^2	Test MAE (kJ/mol) w/o Outlier	Test R^2 w/o Outliers
Baseline (Mean)	7.7	0	-	-
Pooled, Feature Set				
LASSO	5.7	0.3	5.1	0.44
Bayes	5.3	0.36	4.7	0.59
Kernel Ridge	5.9	0.29	5.6	0.275
Extra Trees	2.9	0.74	2.0	0.93
Gradient Boost	3.0	0.74	2.0	0.93
Random Forest	3.3	0.69	2.4	0.91
XGBoost	3.3	0.73	2.4	0.92
Pooled, Uncorrelated Feature Set(Top Algorithms)				
Gradient Boost	2.9	0.73	2.0	0.92
Extra Trees	2.8	0.75	1.9	0.93
Random Forest	3.0	0.72	2.1	0.92
XGBoost	3.1	0.78	2.4	0.90
Physical Feature Set (Top Algorithms)				
Gradient Boost	2.9	0.76	2.1	0.92
Extra Trees	3.0	0.74	2.1	0.92
Random Forest	3.4	0.69	2.5	0.88
XGBoost	3.4	0.77	2.7	0.86

Table A.2: Performance for Different Model Types

A.3 Variable Definitions

Φ^{nuc} – Nuclear Electrostatic Potential Energy	q - Charge	λ - Localization Index
Φ - Electrostatic Potential	ϵ - Bond Ellipticity	δ - Delocalization Index
Φ^e - Electronic Electrostatic Potential	δ^{bond}– Bond Delocalization Index	E^e– Contribution to electronic energy
$x/y/z$ - atomic positions	T - Kinetic energy	

Table A.3: table of feature definitions

A.4 Feature Sets

Algorithm	Selected Features
Boruta	q_2
	$d_7,\ d_7,\ d_7^{sum}$
	$\delta_1, \delta_2, \delta_6, \delta_4^{bond}$
	$\lambda_3, \lambda_5, \lambda_6$
	Φ_1, Φ_2, Φ_4
	$\Phi_5, \Phi_6, \Phi_{10}, \Phi_5^{nuc}$
LASSO	q_4, q_5
	$d_7,\ d_7', d_7^{sum}$
	λ_1, δ_1
	Φ_1, Φ_4
	$\Phi_5^{nuc}, \Phi_{10}^{e}$
Recursive Feature Elimination	q_2
	$d_7,\ d'_7$
	$\delta_1, \delta_2, \delta_6$
	$\Phi_1, \Phi_2, \Phi_4, \Phi_5, \Phi_6, \Phi_9$
	$\Phi_{10}^{e}, \Phi_5^{nuc}$
PCA0	$q_1,$
	$\delta_1^{bond}, \delta_2^{bond}, \delta_4^{bond}$
	$\delta_1, \delta_2, \delta_4$
	Φ_1, Φ_3, Φ_4
	$T_1,\ T_3,\ T_4$
	E_1^e, E_3^e, E_4^e
	λ_1, λ_3

Table A.4: Features Selected by Different Feature Selection Algorithms.

Feature Set	Selected Features
Pooled (37 Features)	$q_1, q_2, q_4, q_5,\ d_7,\ d_7', d_7^{sum}$, $\delta_1, \delta_3, \delta_5, \delta_6, \delta_1^{bond}, \delta_2^{bond}, \delta_4^{bond}, \delta_6^{bond}$, $\lambda_1, \lambda_3, \lambda_5, \lambda_6$, E_1^e, E_3^e, E_4^e $\Phi_1, \Phi_2, \Phi_3, \Phi_4, \Phi_5, \Phi_6, \Phi_{10}, \Phi_5^{nuc}, \Phi_{10}^e$
Uncorrelated (24 Features)	$T_1,\ T_3,\ T_4$, $\epsilon_7, \epsilon_8, \epsilon_9$ $\Phi_1, \Phi_2, \Phi_3, \Phi_4, \Phi_5, \Phi_6, \Phi_{10}, \Phi_5^{nuc}, \Phi_{10}^e$ $\epsilon_7, \epsilon_8, \epsilon_9$ $E_1, q_2, q_4, q_5,\ d_7^{sum}$, $\delta_1, \delta_3, \delta_5, \delta_6$,
Physical (28 Features)	$q_1, q_2, q_3, q_4, q_5, q_6$, $\delta_1, \delta_2, \delta_3, \delta_4, \delta_5, \delta_6$, $E_1^e, E_2^e, E_3^e, E_4^e$ $\Phi_1, \Phi_2, \Phi_3, \Phi_4, \Phi_5, \Phi_6, \Phi_{10}, \Phi_5^{nuc}$, $\epsilon_7, \epsilon_8, \epsilon_9, \epsilon_{10}$

Table A.5: Different feature sets used.

A.5 Top Model Parameter Sets

Models were tuned using Sklearn's built-in Bayesian parameter optimization package. For each algorithm we performed a 4-fold cross validation for each algorithm trial and 5 times the number of parameters number of algorithms were tested in to tune the algorithms. The best performing models in cross validation performance are shown below. Dictionaries of these values and all corresponding code is available at the project github repository.

1. Extra Trees: maxdepth=49, minsamplesleaf=2, nestimators=477

2. GradientBoosting: learningrate=0.005, minsamplessplit=2, minsamplesleaf=1, maxdepth=8, nestima tors=1500, subsample=0.5

3. XGBRegressor: alpha=0.2, colsamplebynode=1, colsamplebytree=0.5, eta=0.0, gamma=0.0, lambda=0.0, learningrate=0.055, maxdepth=25, nestimators=777, regalpha=0.2, reglambda=0, scaleposweight=1, subsample= 1

4. Bayesian Ridge: alpha1=10.0, lambda1=10.0, lambda2=10.0, niter=9023, tol=0.0006

5. Lasso: alpha=0.001

6. HuberRegressor: alpha=1.8687e-06, epsilon=1.0589, maxiter=1000, tol=0.0094174

7. SGD: epsilon=0.001, eta0=0.011343, l1=0.3, tol=0.1

8. Ridge: alpha=0.010883, tol=0.1

A.6 Permutation Importance

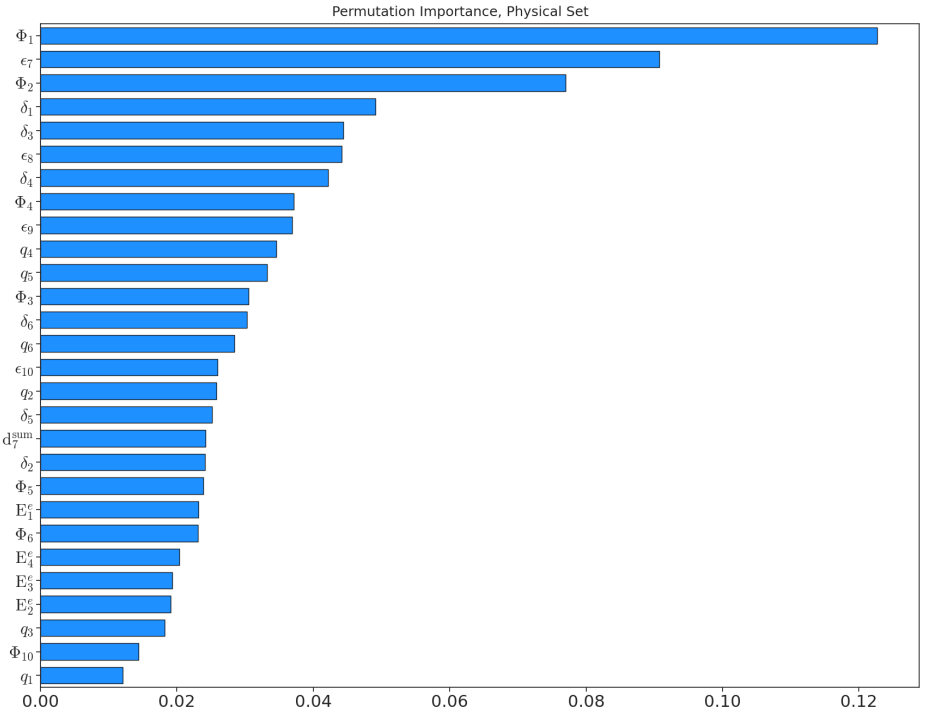


Figure A.1: Permuation Importance for the Physical Feature Set.

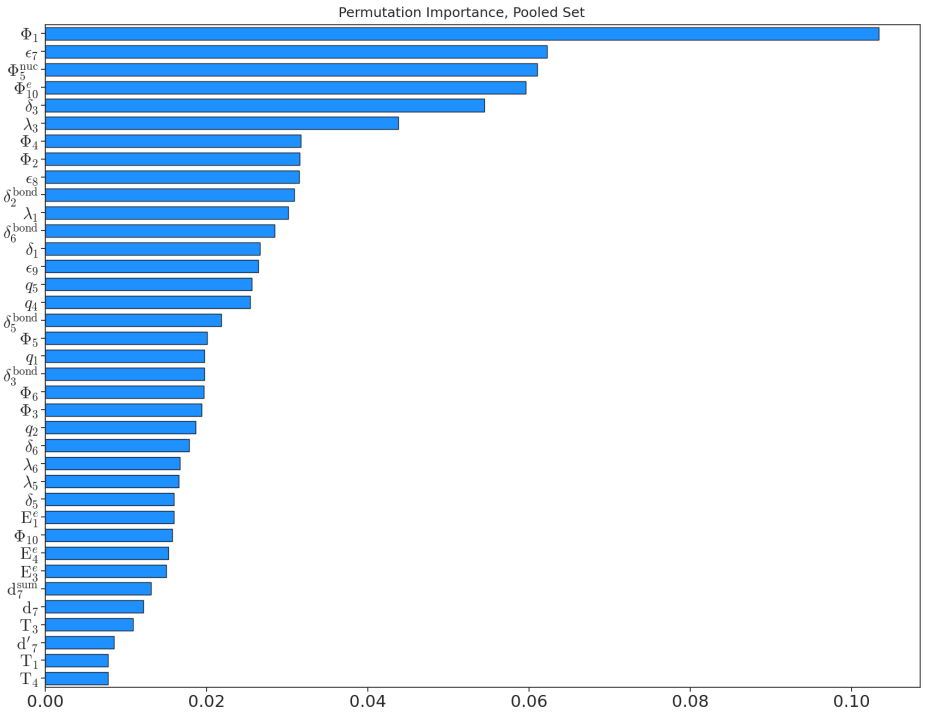


Figure A.2: Permuation Importance for the Pooled Feature Set.

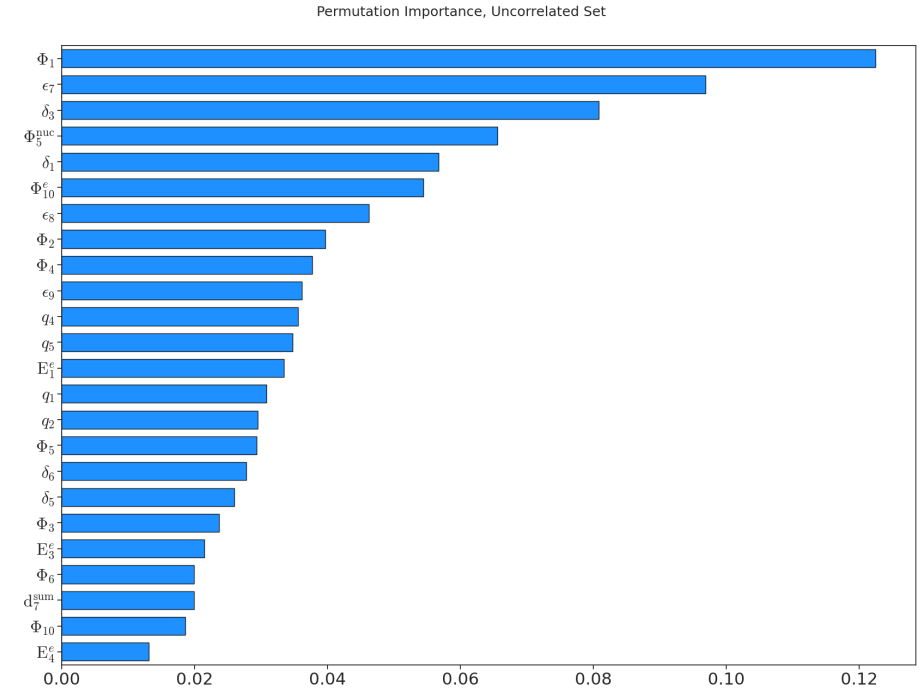


Figure A.3: Permuation Importance for the Filtered, Uncorrelated Feature Set.

A.7 Parity Plots

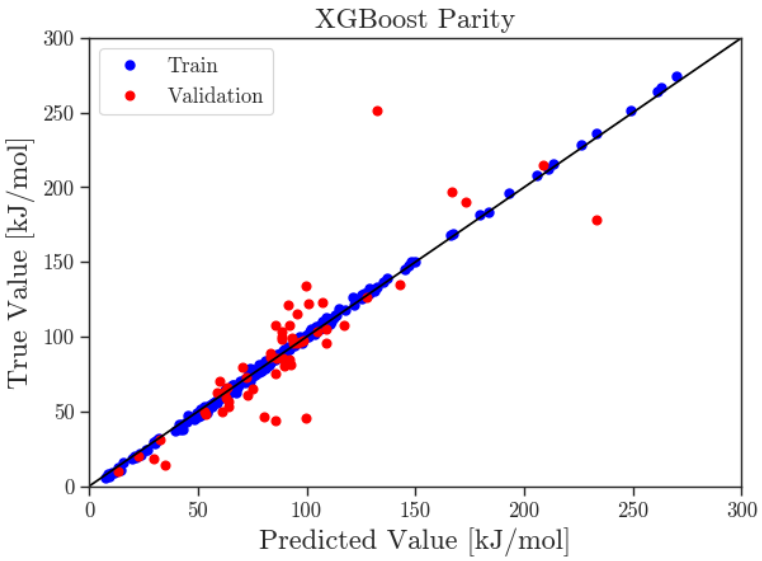


Figure A.4: Parity, XGB w/ Physical Feature Set.

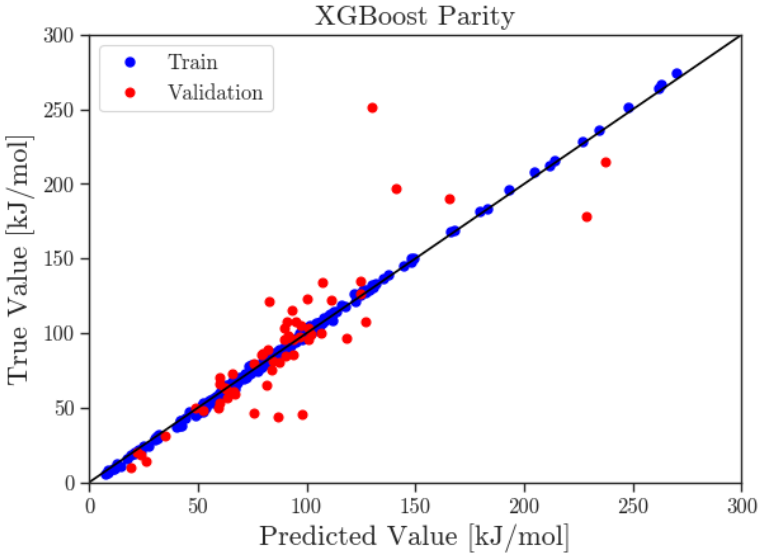


Figure A.5: Parity, XGB w/ Pooled Feature Set.

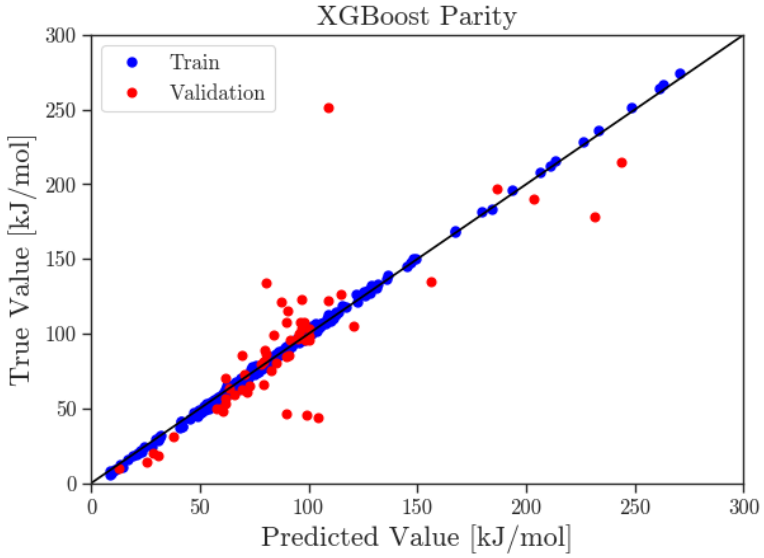


Figure A.6: Parity, XGB w/ Filtered, Uncorrelated Feature Set.

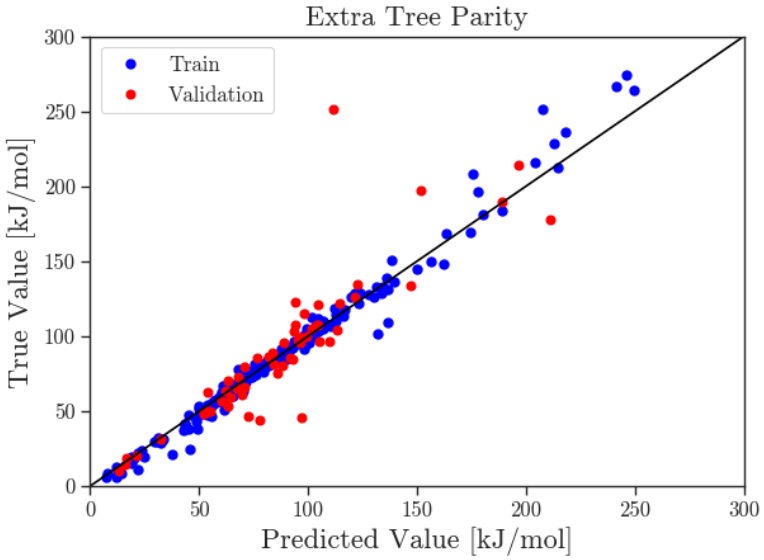


Figure A.7: Parity, Extra Trees w/ Pooled Feature Set.

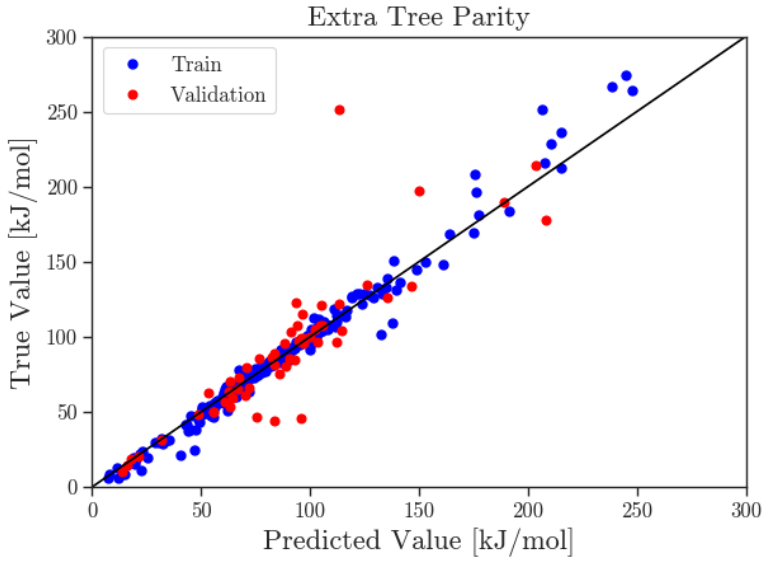


Figure A.8: Parity, Extra Trees w/ Filtered, Uncorrelated Feature Set.

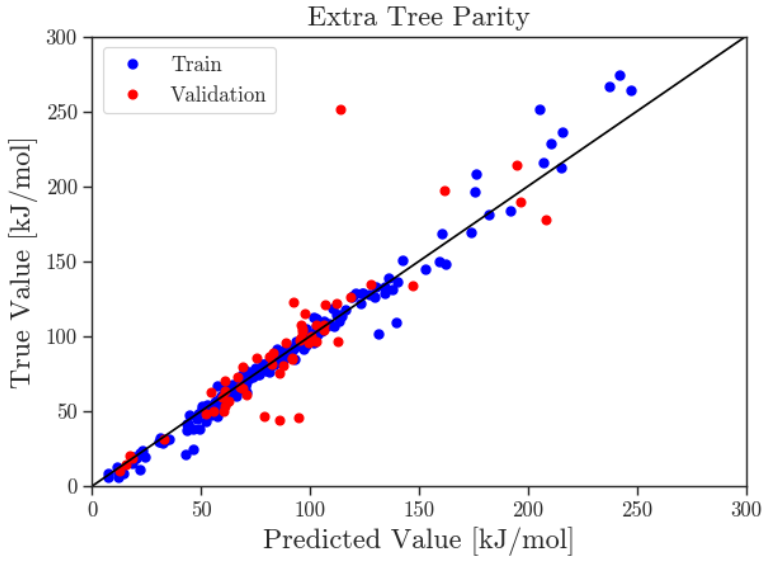


Figure A.9: Parity, Extra Trees w/ Physical Feature Set.

A.8 Variable Correlation Matrices

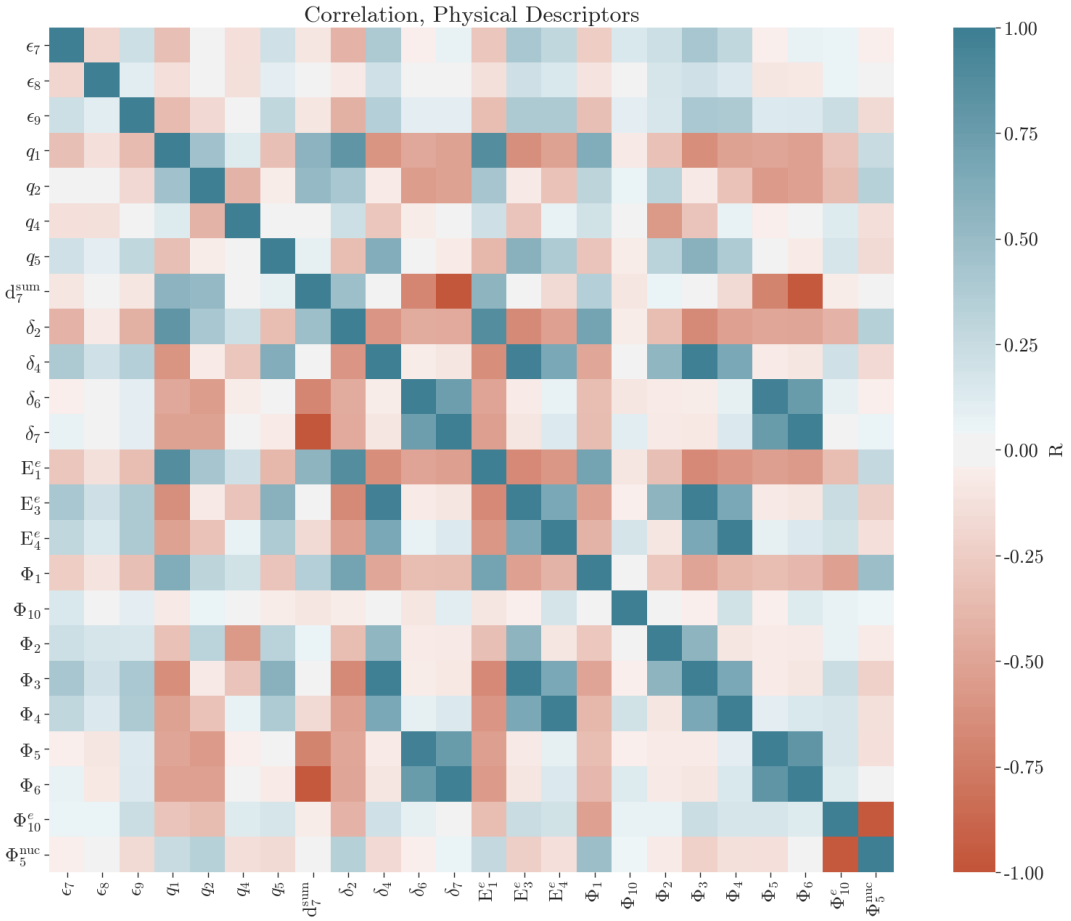


Figure A.10: Physical Feature Set Correlation With Barriers

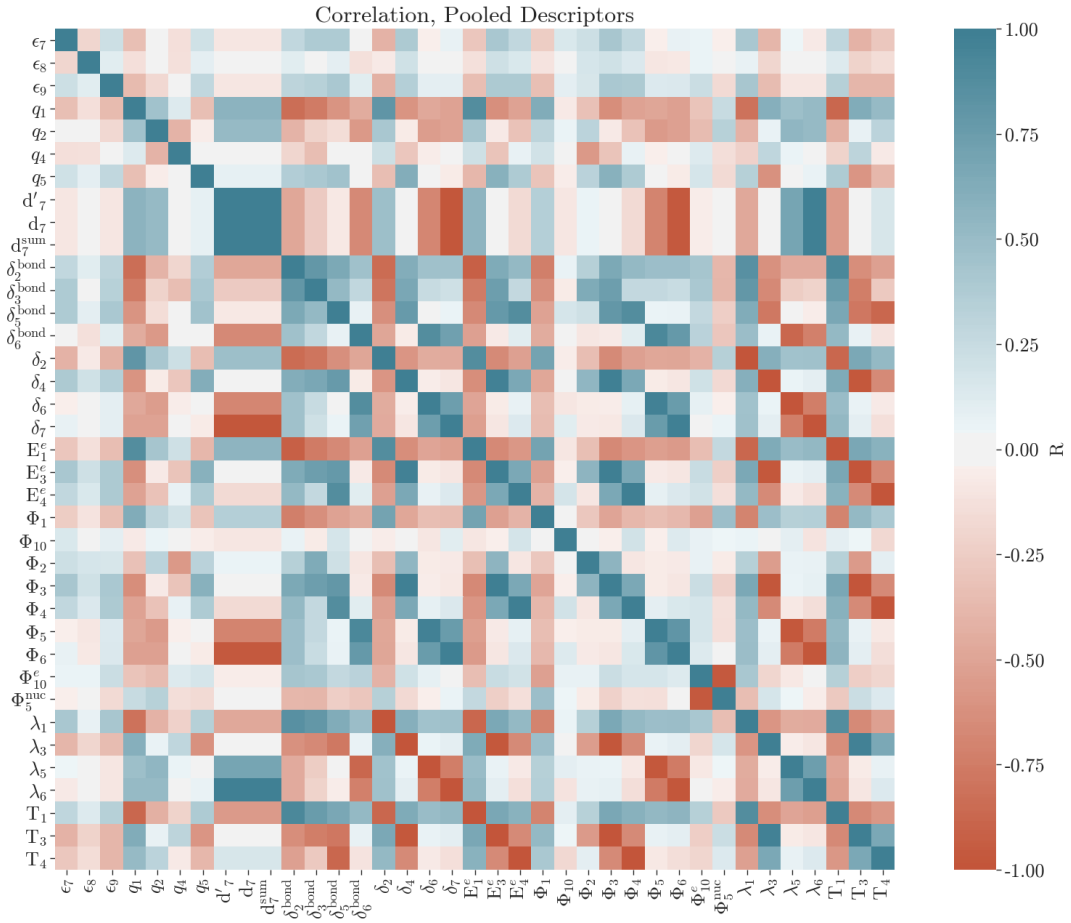


Figure A.11: Pooled Feature Set Correlation With Barriers

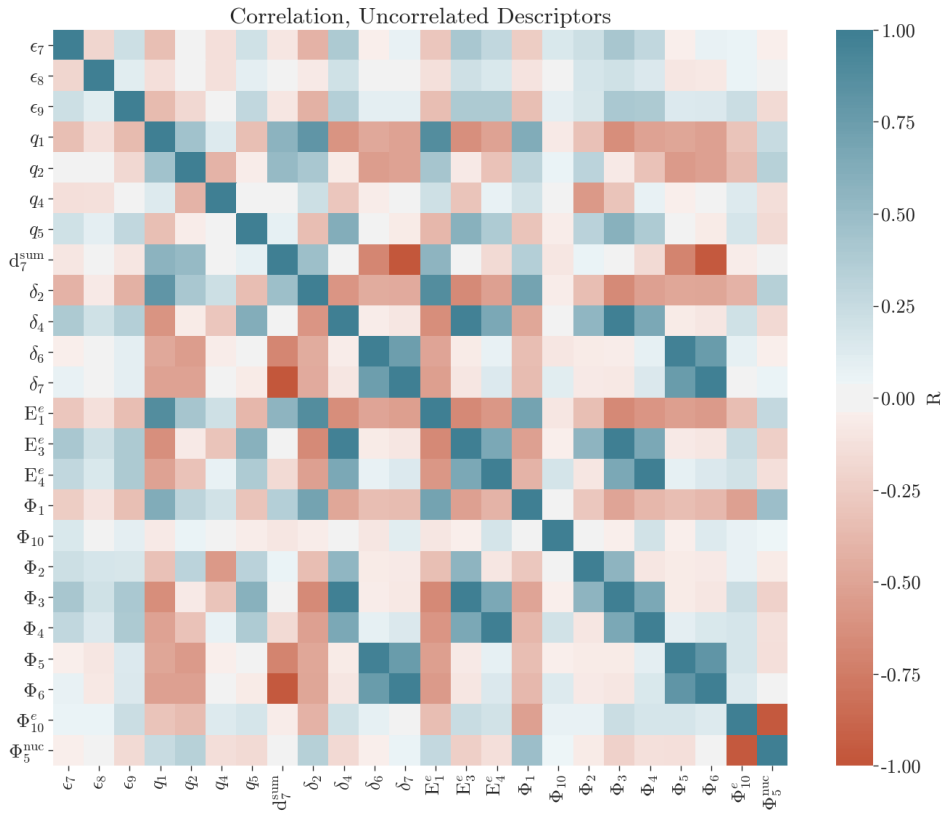


Figure A.12: Filtered, Uncorrelated Feature Set Correlation With Barriers

A.9 Barrier Correlations

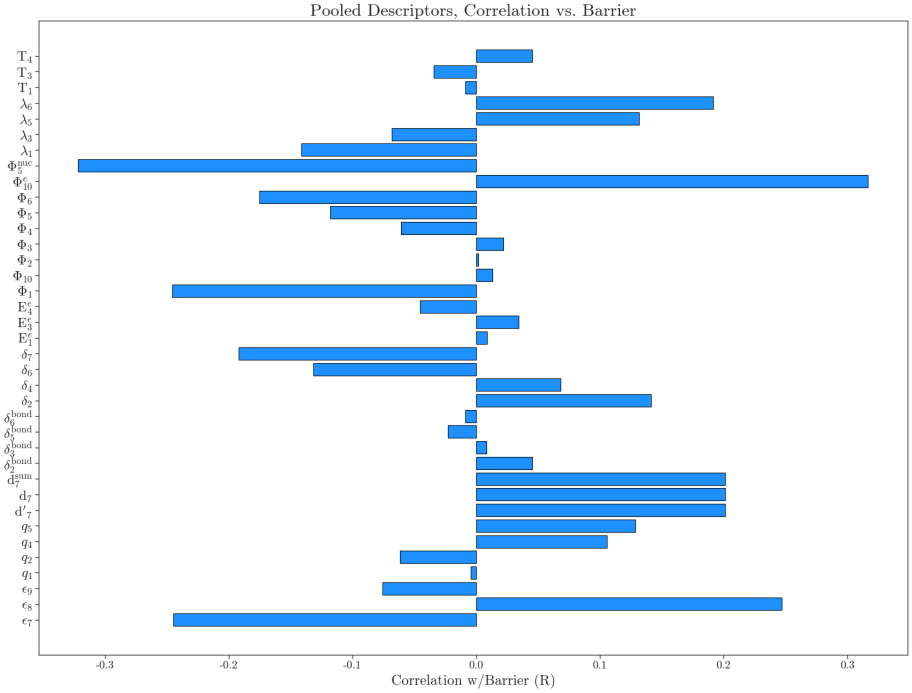


Figure A.13: Pooled Feature Set Correlation With Barriers

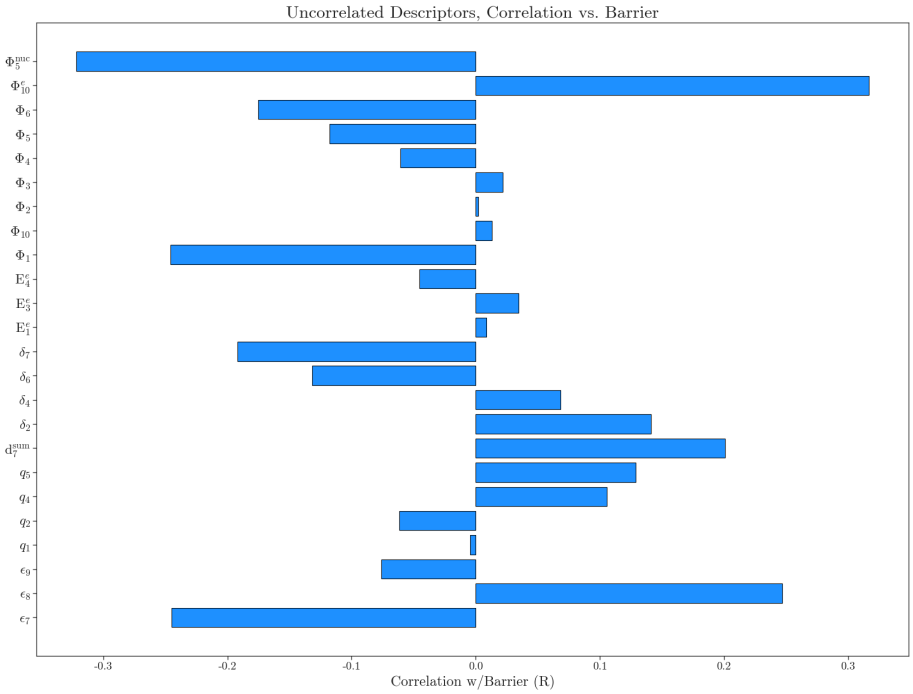


Figure A.14: Filtered, Uncorrelated Feature Set Correlation With Barriers

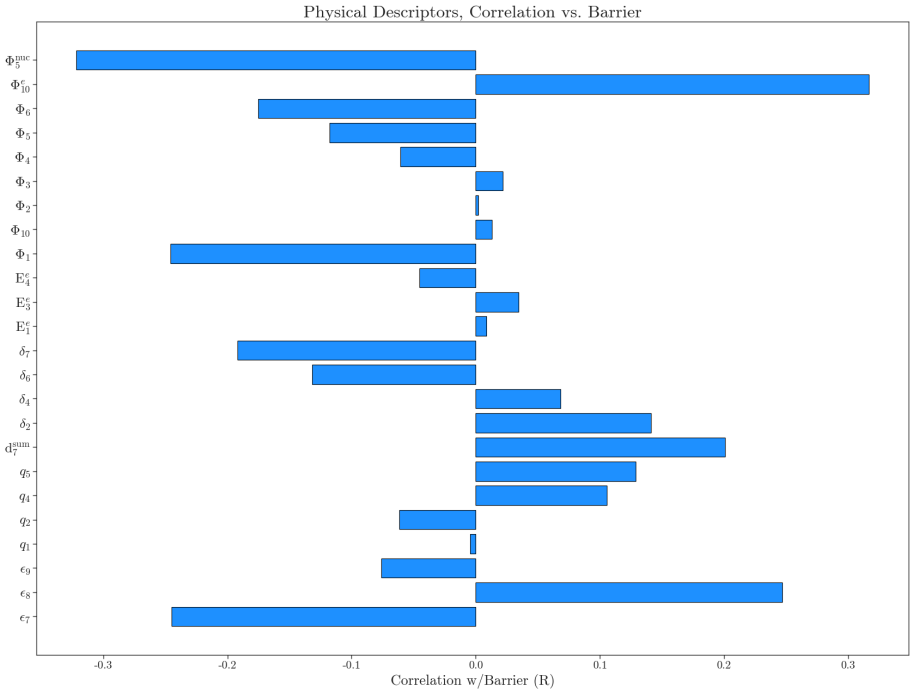


Figure A.15: Physical Feature Set Correlation With Barriers

Appendix B

Supporting Information for *High-throughput Quantum Theory of Atoms in Molecules (QTAIM) for Geometric Deep Learning of Molecular and Reaction Properties*

B.1 Full set of QTAIM descriptors

Atom Descriptors	**Bond Descriptors**
Total Electrostatic Potential Φ_{tot}	Total Electrostatic Potential Φ_{tot}
Nuclear Electrostatic Potential Φ_{nuc}	Nuclear Electrostatic Potential Φ_{nuc}
Electronic Electrostatic Potential Φ_e	Electronic Electrostatic Potential Φ_e
Lagrangrian ($\nabla^2\rho$)	Lagrangrian ($\nabla^2\rho$)
Kinetic Energy Hamiltonian	Kinetic Energy Hamiltonian

Atom Descriptors	**Bond Descriptors**
Gradient Norm	Gradient Norm
Δ_g promolecular	Δ_g promolecular
Δ_g Hirshfield	Δ_g Hirshfield
Electron Density	Electron Density$)
Laplacian Electron Density	Laplacian Electron Density
Hessian Determinant	Hessian Determinant
Electron Localization Function(ELF)	Electron Localization Function(ELF)
Laplacian Norm	Laplacian Norm
Hessian eigenvalue (1st)	Hessian eigenvalue (1st)
Electronc Ellipticity	Electronic Ellipticity
Average Location Ion E	Average Location Ion E
Eta	Eta
Localized Orbital Locator	Localized Orbital Locator
energy density	energy density
α spin	α spin
β spin	β spin
spin density	spin density

Table B.1: Full set of QTAIM Descriptors

B.2 Dataset Visualizations

B.2.1 Corrected Energies - LIBE

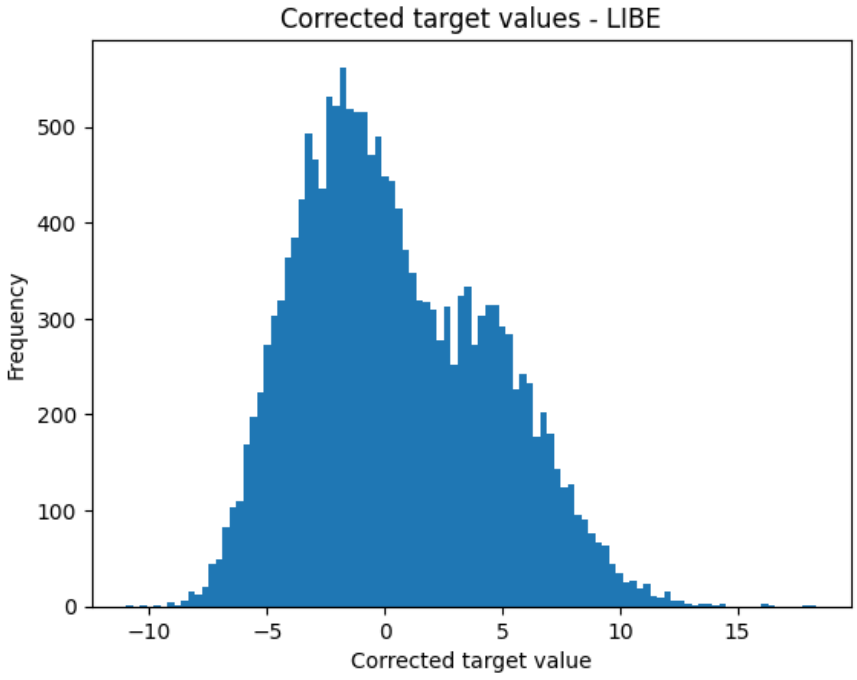


Figure B.1: LIBE corrected energies

B.2.2 Correction Values

Atomic Number	Correction Value
1	-16.77537562
3	-206.45292515
6	-1034.69861041
7	-1488.80081496
8	-2048.19270236
9	-2717.83725543
15	-9286.36995521
16	-10831.57826394

Table B.2: Correction values used from raw LIBE energies

B.3 Parity Plots

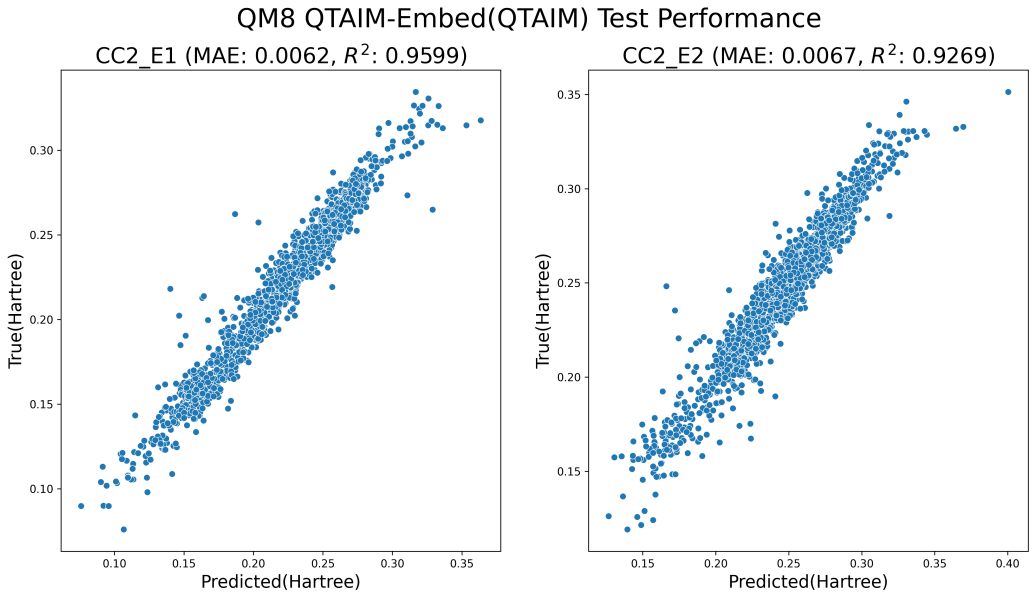


Figure B.2: QM8 QTAIM test Partity.

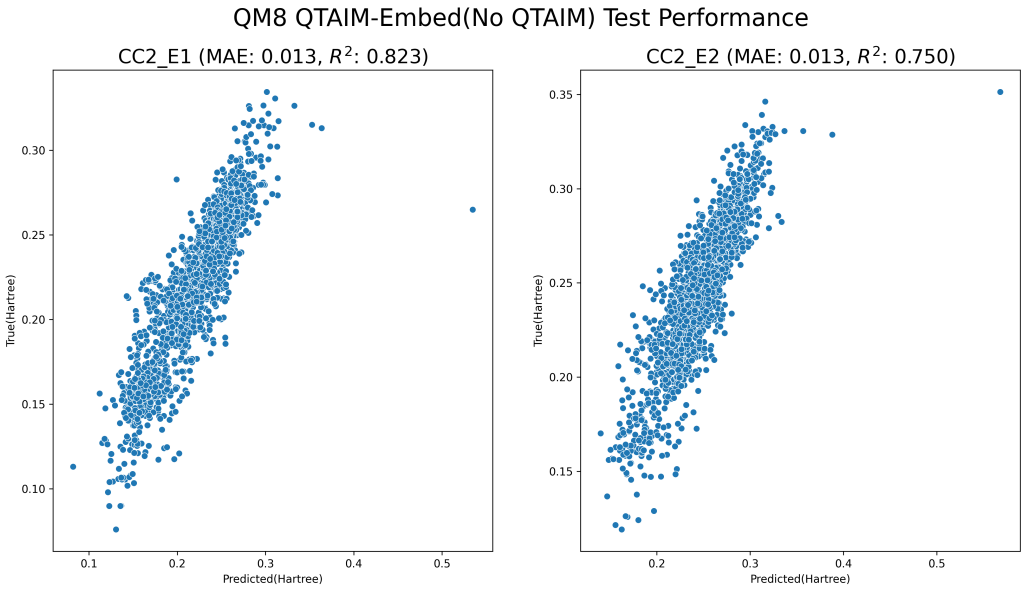


Figure B.3: QM8 non-QTAIM test Partity.

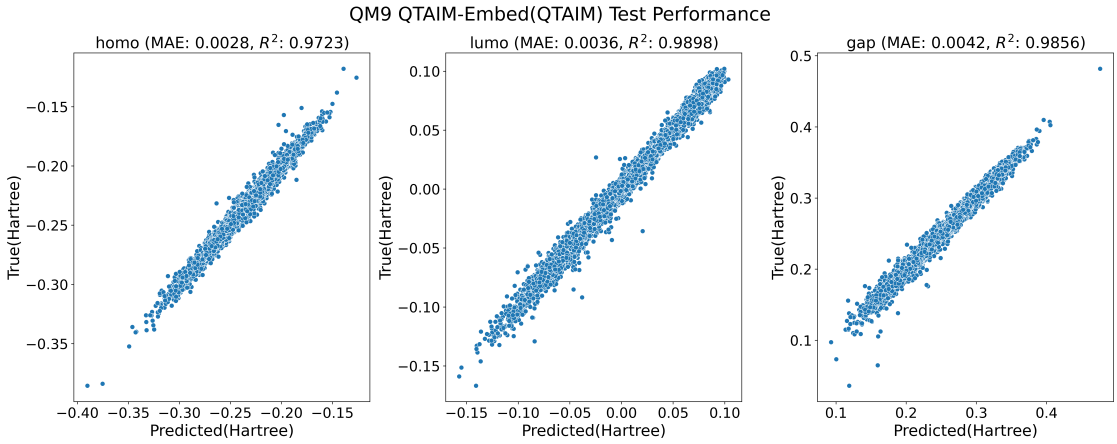


Figure B.4: QM9 QTAIM test Partity.

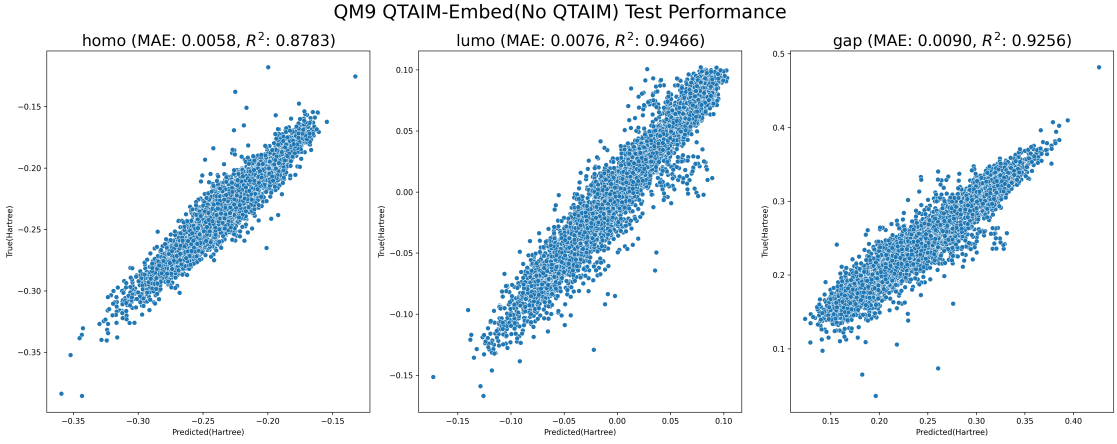


Figure B.5: QM9 non-QTAIM test Partity.

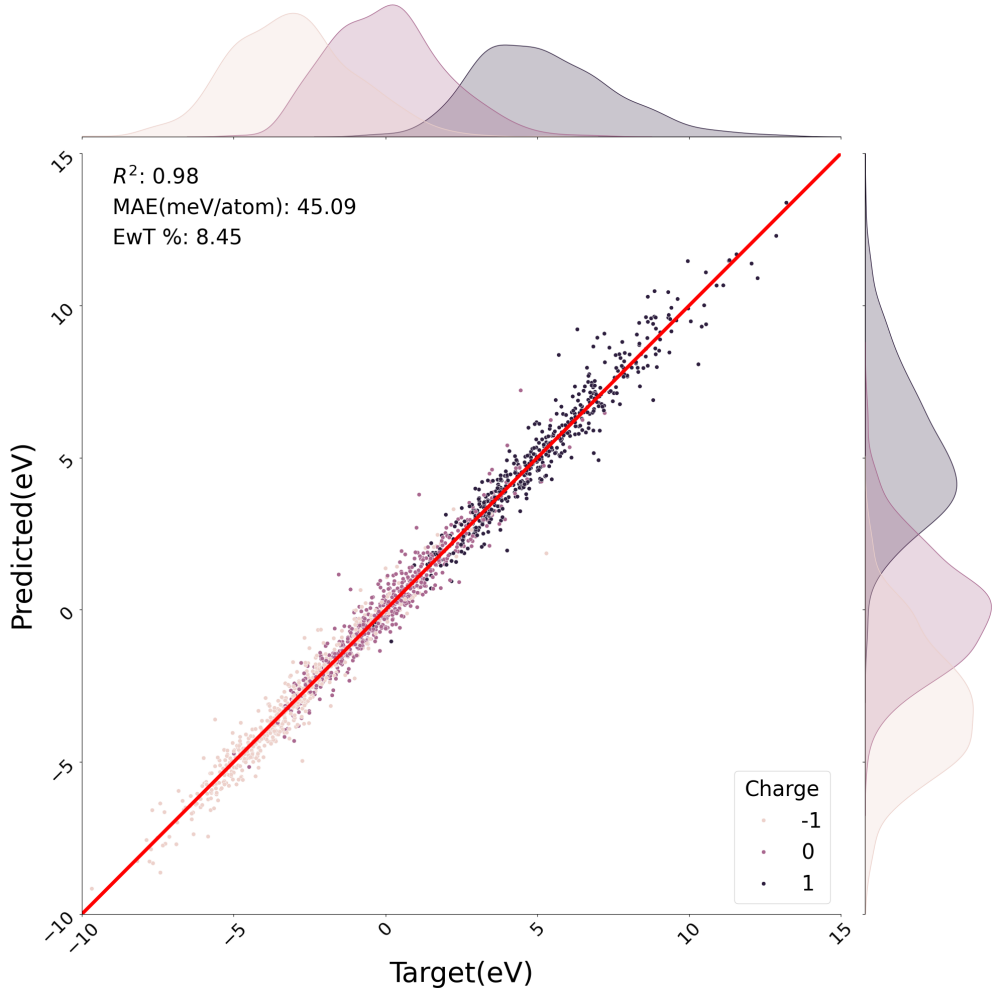


Figure B.6: LIBE QTAIM test Partity, charge-partitioned.

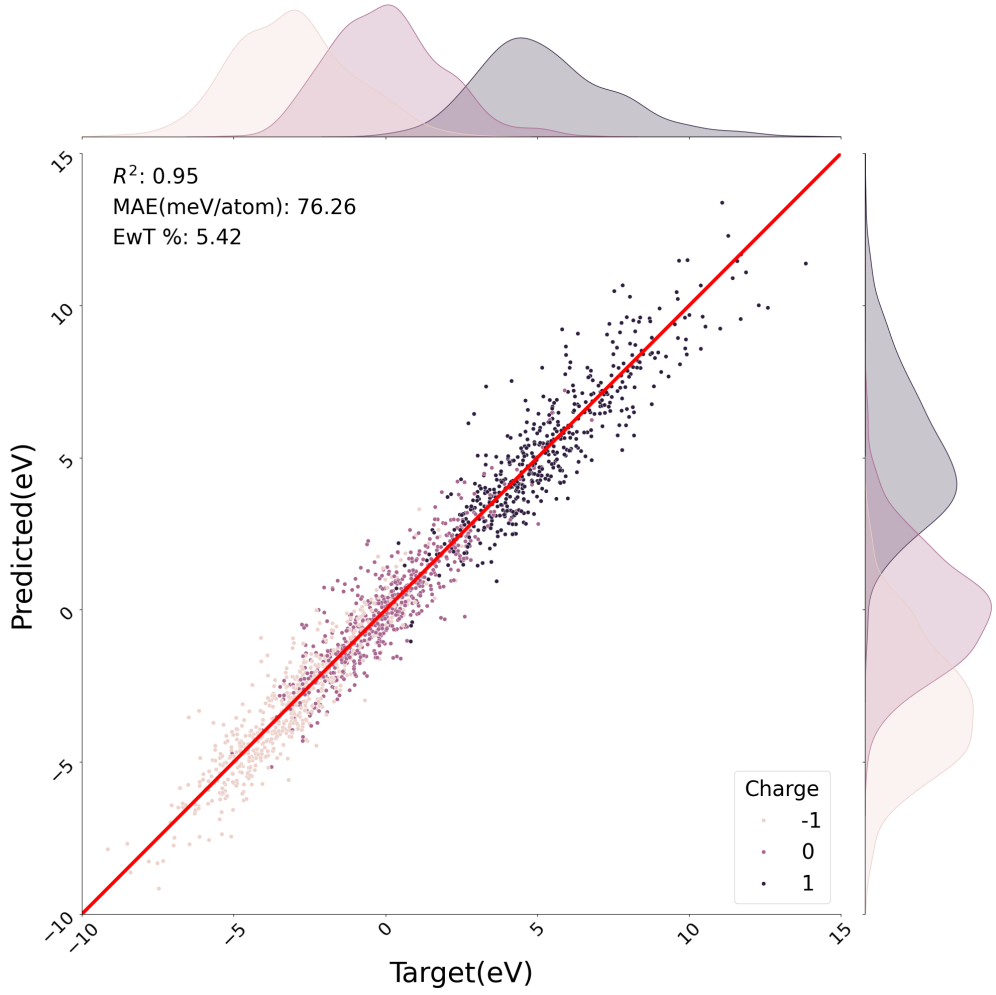


Figure B.7: LIBE non-QTAIM test Partity, charge-partitioned.

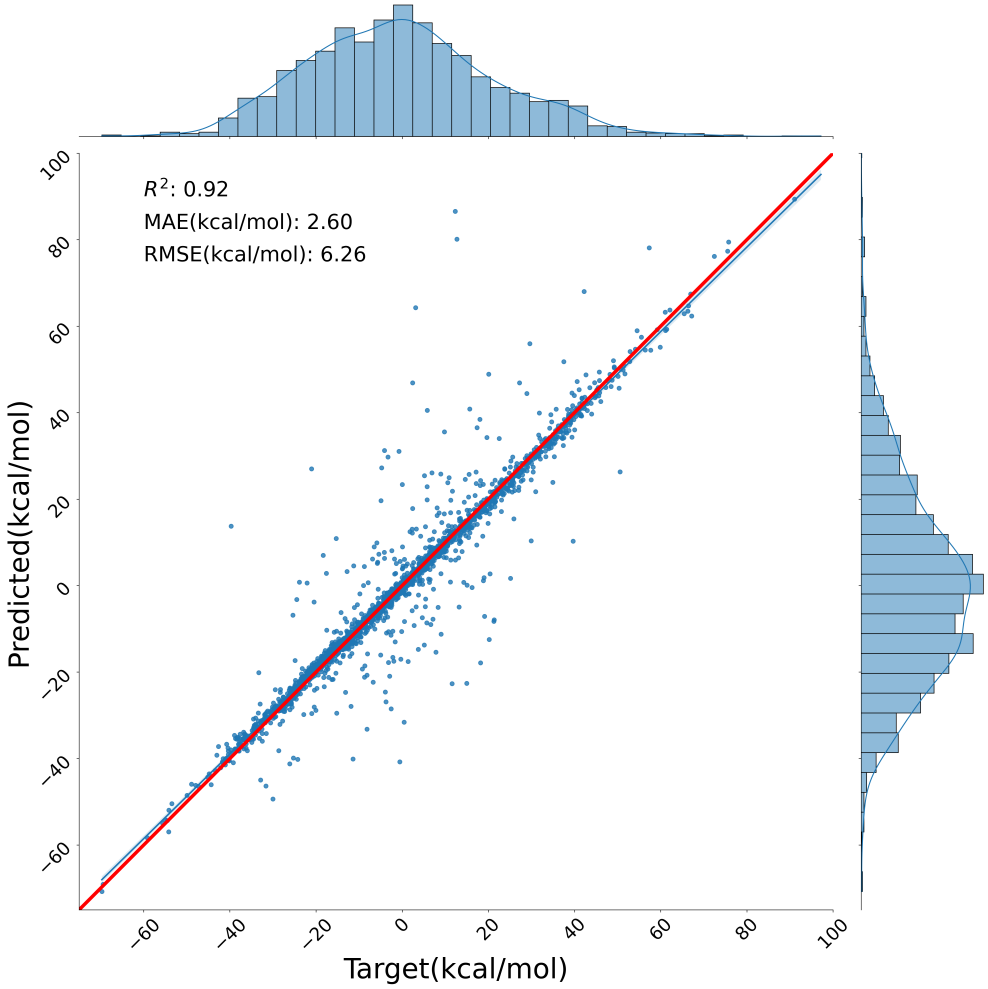


Figure B.8: Green QTAIM test Partity.

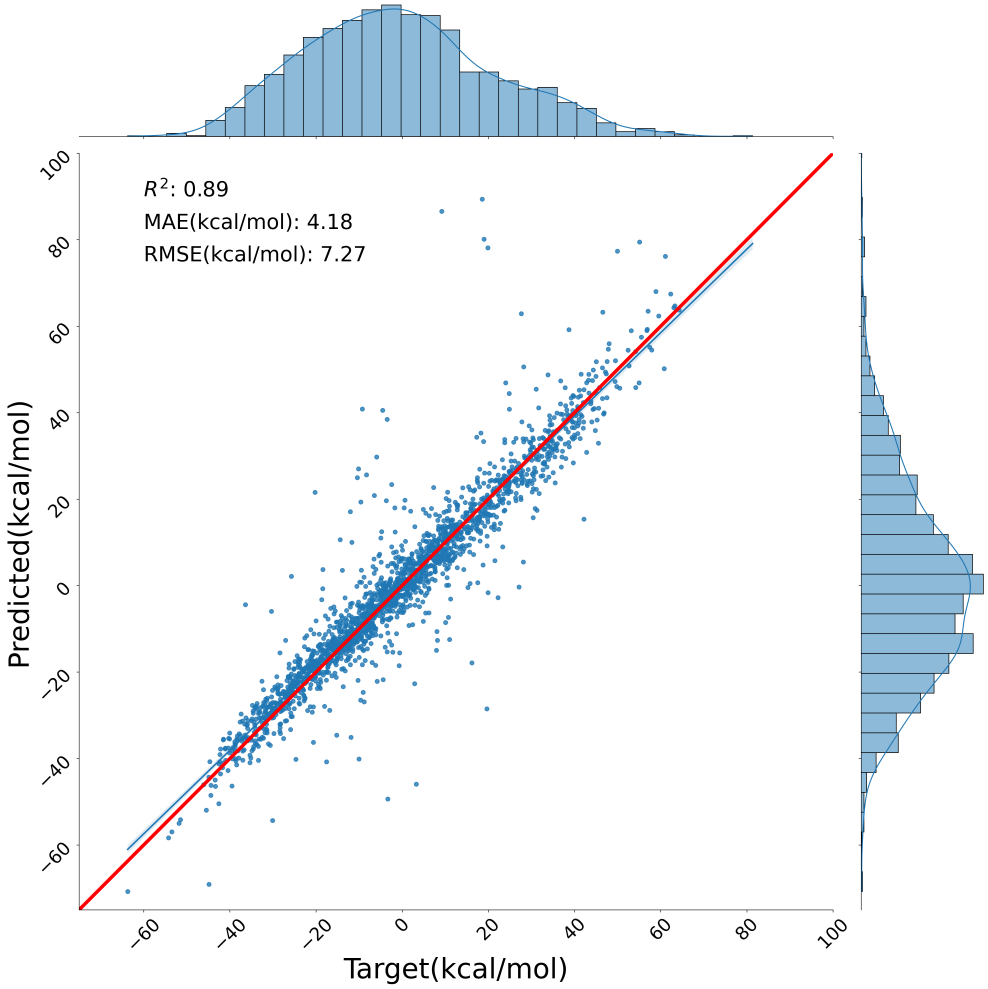


Figure B.9: Green non-QTAIM test Partity.

B.4 OOD True vs. Predicted Plots

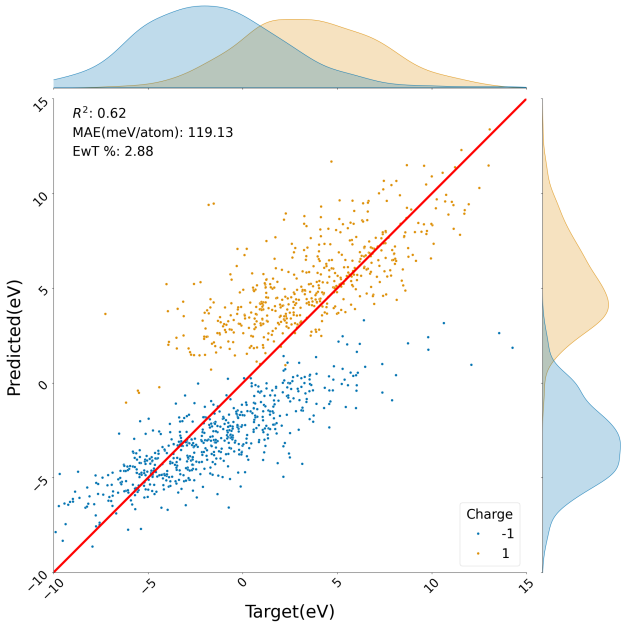


Figure B.10: LIBE OOD QTAIM charge-stratified test Partity.

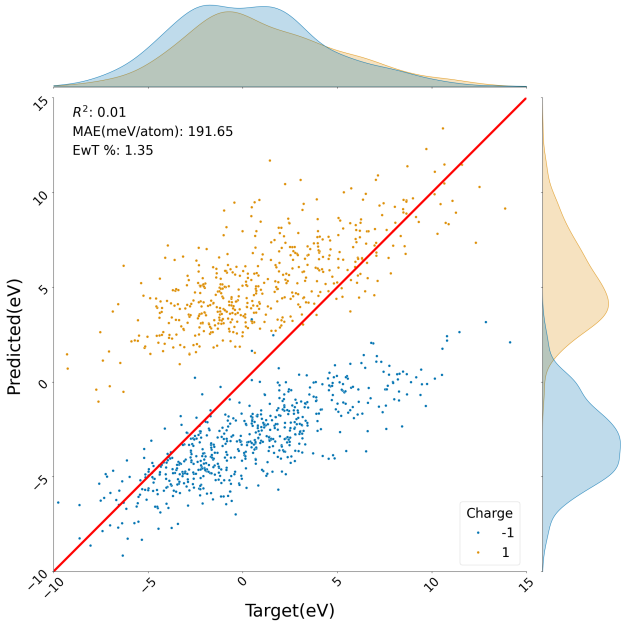


Figure B.11: LIBE OOD non-QTAIM charge-stratified test Partity.

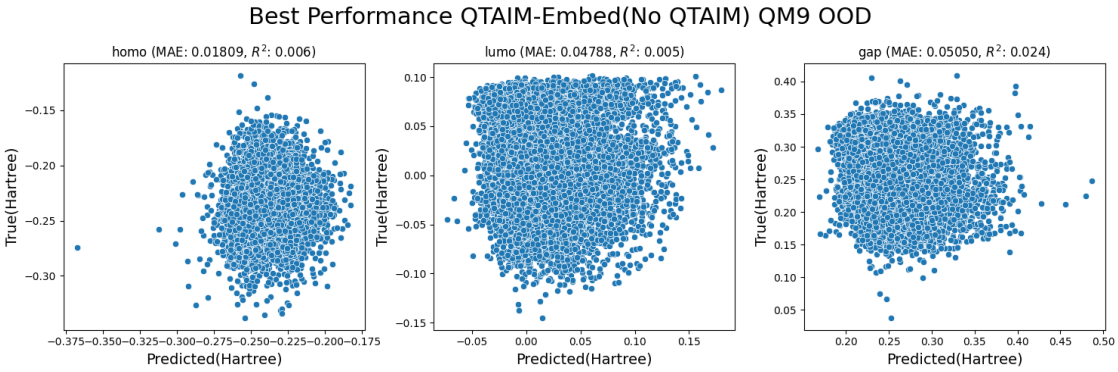


Figure B.12: QM9 OOD non-QTAIM test Partity.

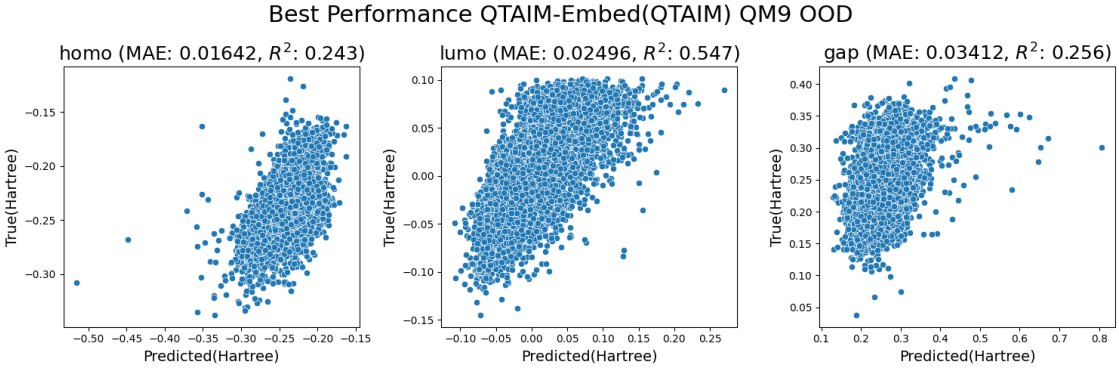


Figure B.13: QM9 OOD QTAIM test Partity.

B.5 Tox21 Results

	Our Model (QTAIM)	**Our Model (No QTAIM)**
NR-AR	**0.9722**	0.9644
NR-AR-LBD	**0.9797**	0.9734
NR-AhR	**0.8899**	0.8824
NR-Aromatase	**0.9584**	0.9502
NR-ER	**0.8988**	0.8942
NR-ER-LBD	**0.9613**	0.9567
NR-PPAR-gamma	0.9779	**0.9786**
SR-ARE	**0.8567**	0.8413
SR-ATAD5	**0.9775**	0.9748
SR-HSE	**0.9506**	0.9467
SR-MMP	**0.8552**	0.8405
SR-p53	**0.9477**	0.9458
Average AUROC	**0.9355**	0.9291

Table B.3: Tox21 Test Performance

B.6 Full Learning Curves

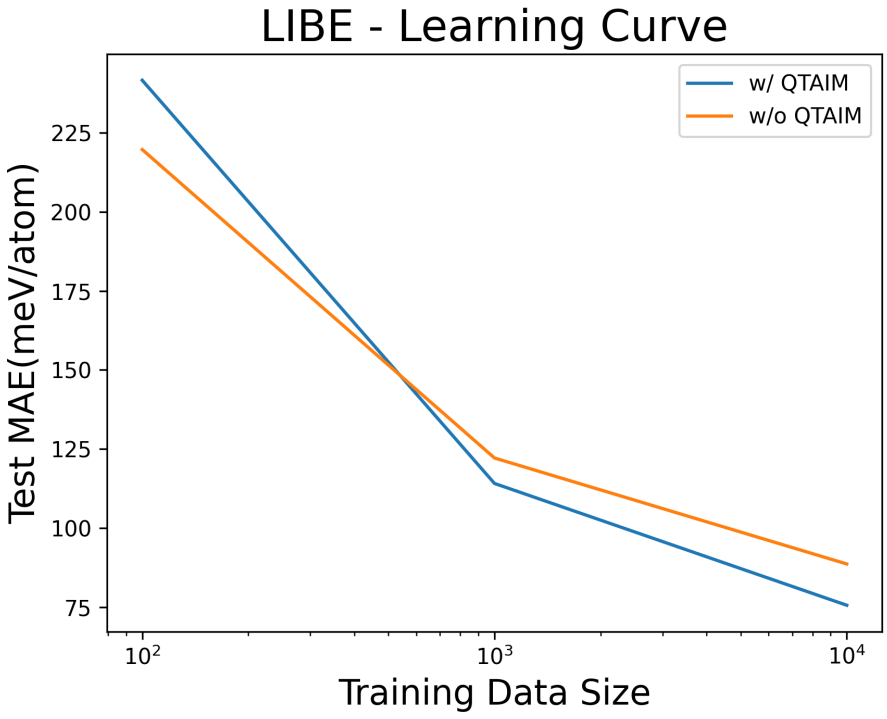


Figure B.14: LIBE Learning Curve on MAE

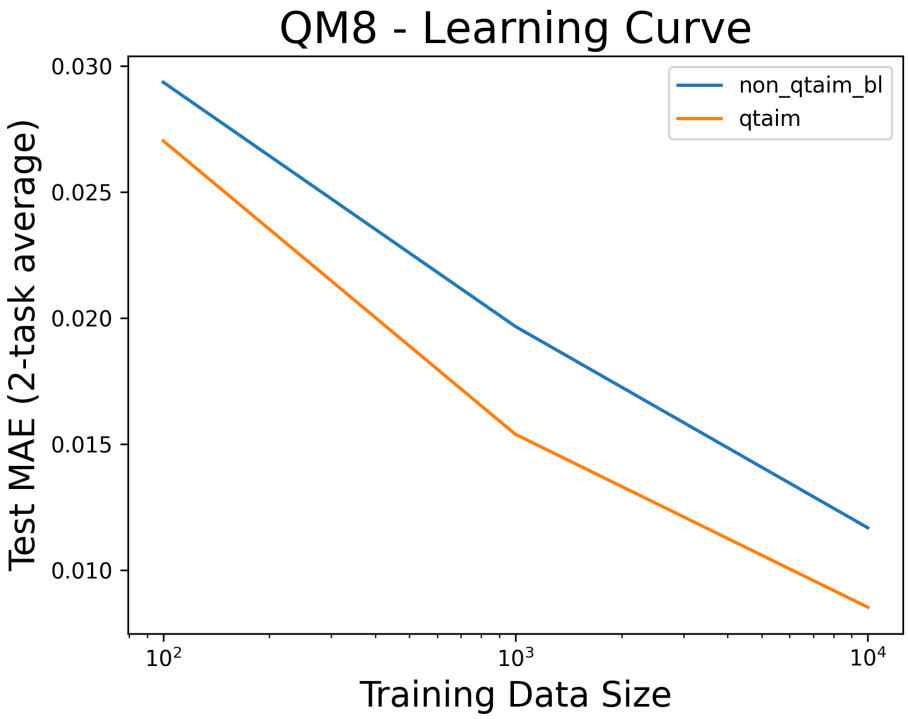


Figure B.15: QM8 Learning Curve on MAE

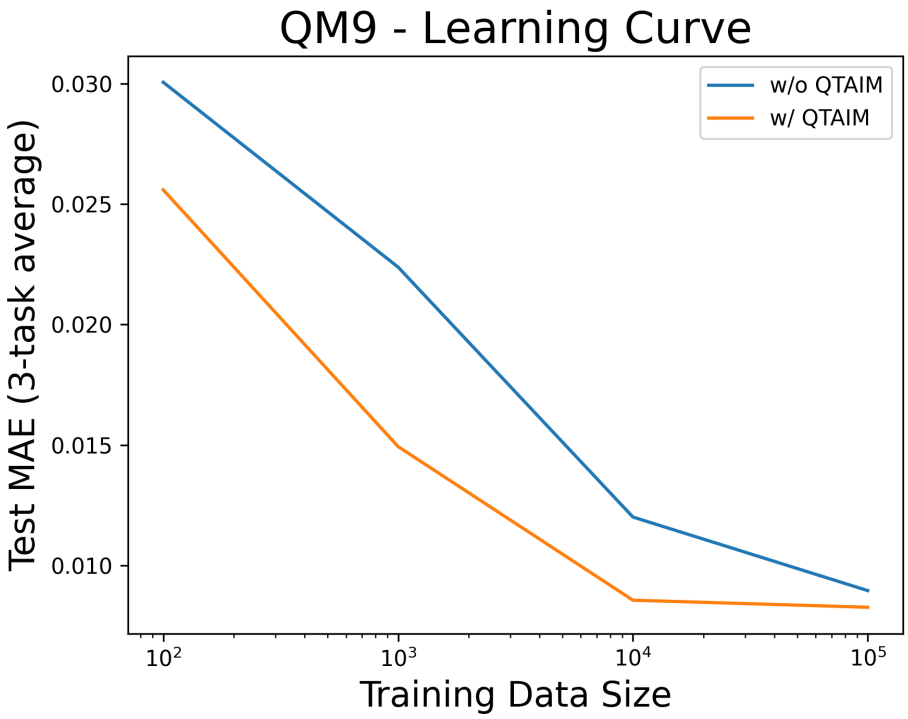


Figure B.16: QM9 Learning Curve on MAE

B.7 Hyperparameter Selection

For ChemProp hyperparameter optimizations we used their inbuilt hyperopt functionality [122]. Here we used the following set of parameters as sweep values:

Schnet/PaiNN - Here we used their default values [251, 252] a select set of values:

Hyperparameter	**Values Swept**
N_atom_basis	10, 20, 50
Shared interaction	T F
LR	0.01, 0.001, 0.0001

Table B.4: Hyperparameters Swept for SchNet and PaiNN

QTAIM-embed - for our own in-house algorithms, we leveraged Wandb's parameter

selection tool. [33] We used the same hyperparameter sweep configs for each model set. Finalized trained models for LIBE used the complete set of QTAIM descriptors above, other models removed α and β spin descriptors.

Hyperparameter	**Values Swept**
weight_decay	0.0, 0.00001
Embedding_size	16, 20, 24
Gated_dropout	0.0, 0.1, 0.2
Gated_hidden_size	64, 128
Gated_batch_norm	T, f
Gated_graph_norm	T, f
Num_lstm_iters	9, 11, 13, 15
Num_lstm_layers	1, 2
Fc_dropout	0.1, 0.2
Fc_hidden_size_1	256, 128
Fc_hidden_shape	flat, cone
Precision	bf16, 32
Gradient_clip_val	10, 100
Accumulated_grad_batches	1, 3, 5

Table B.5: QTAIM-embed (our) model hyperparameter sweeps

B.8 Scatterplots of competing models

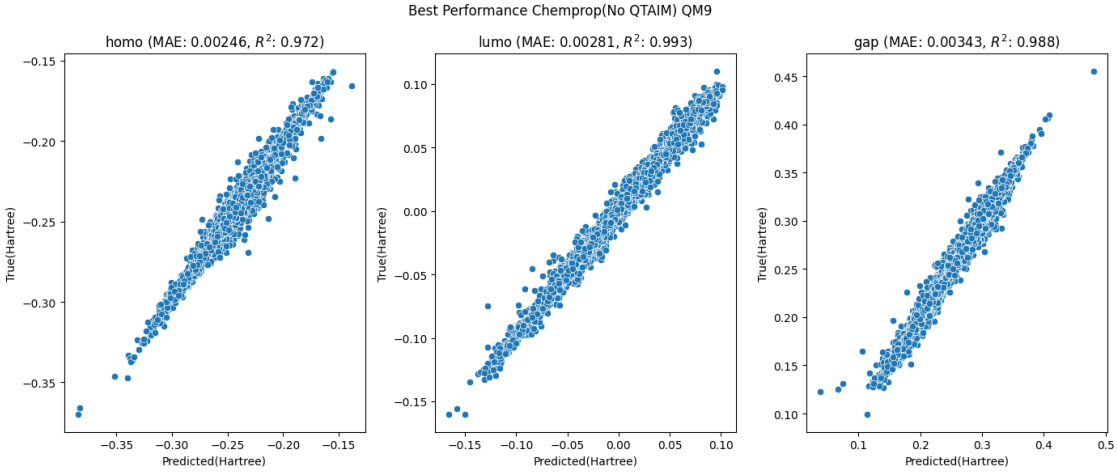


Figure B.17: Parity Plot QM9 chemprop no QTAIM

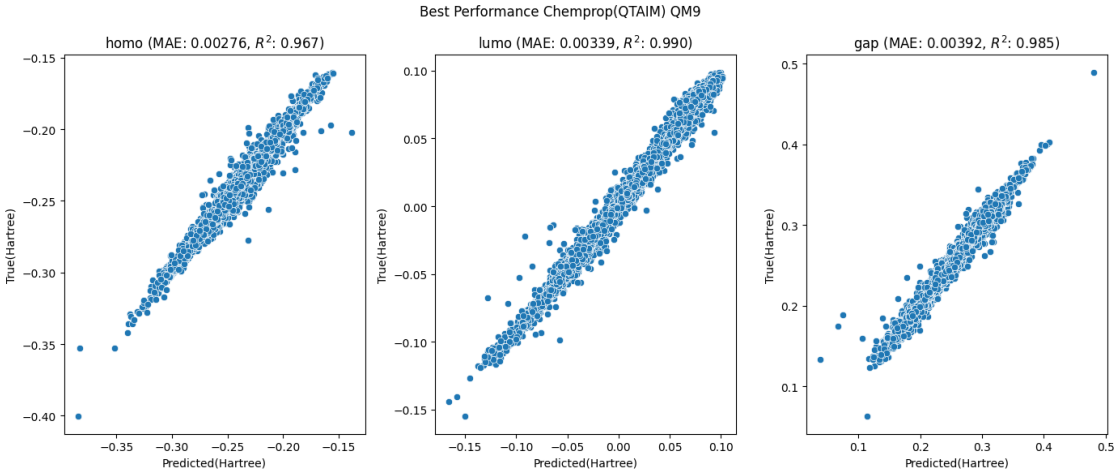


Figure B.18: Parity Plot QM9 chemprop QTAIM

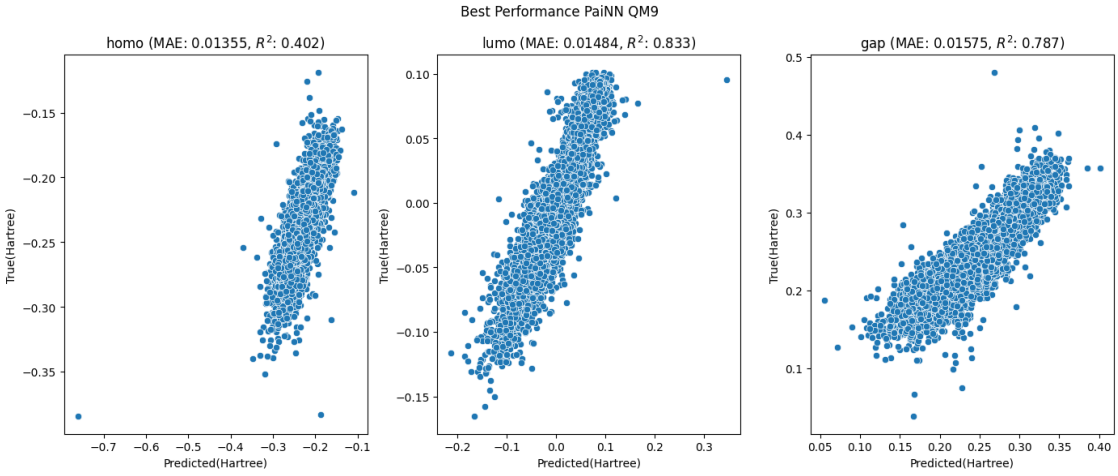


Figure B.19: Parity Plot QM9 PaiNN

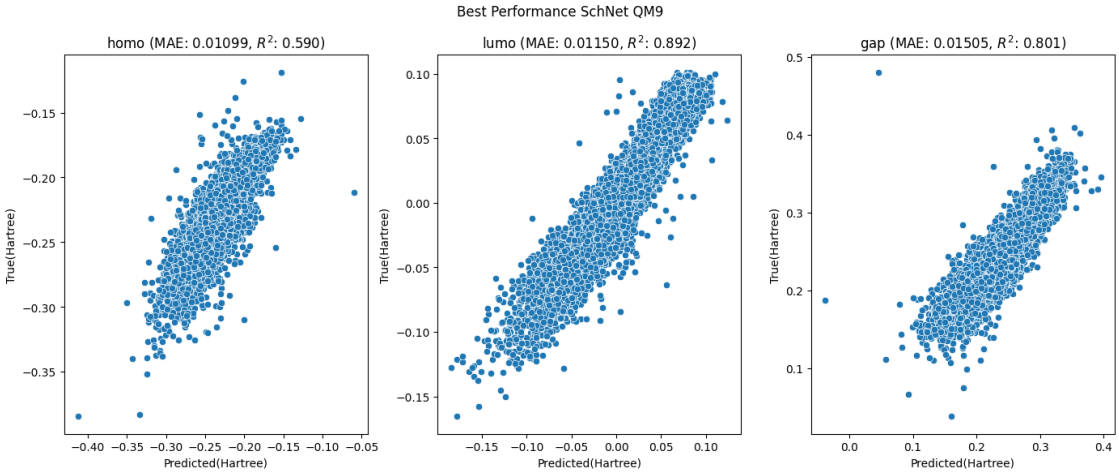


Figure B.20: Parity Plot QM9 Schnet

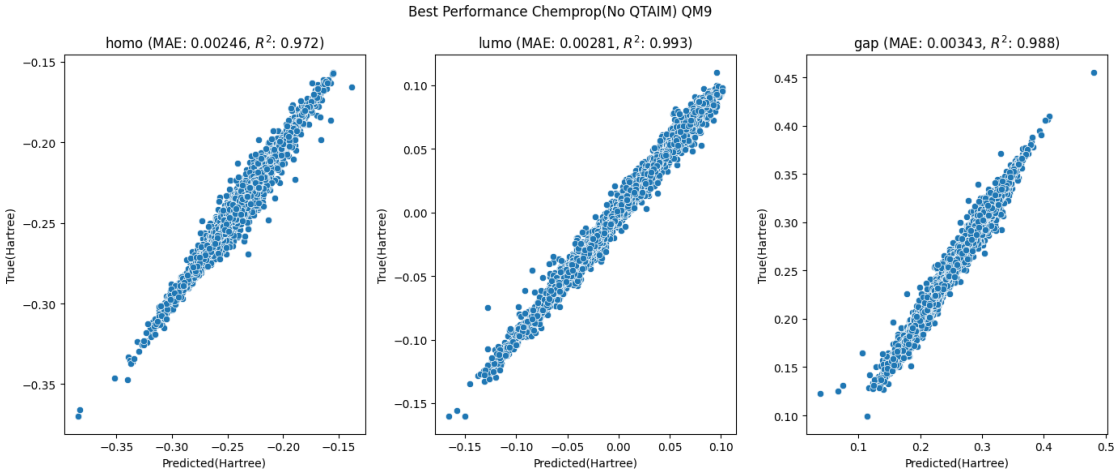


Figure B.21: Parity Plot QM8 chemprop, no QTAIM

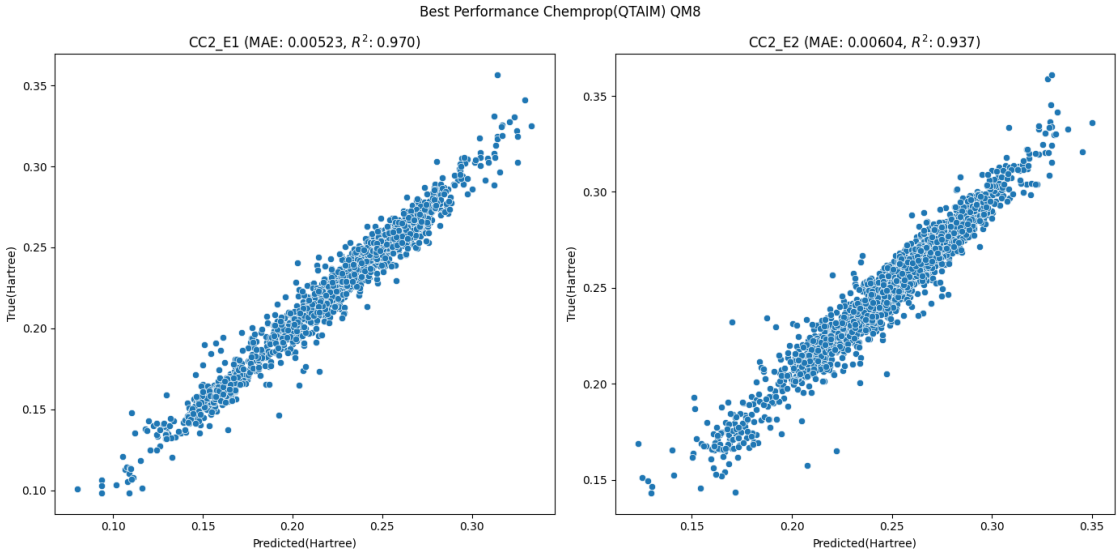


Figure B.22: Parity Plot QM8 chemprop, QTAIM

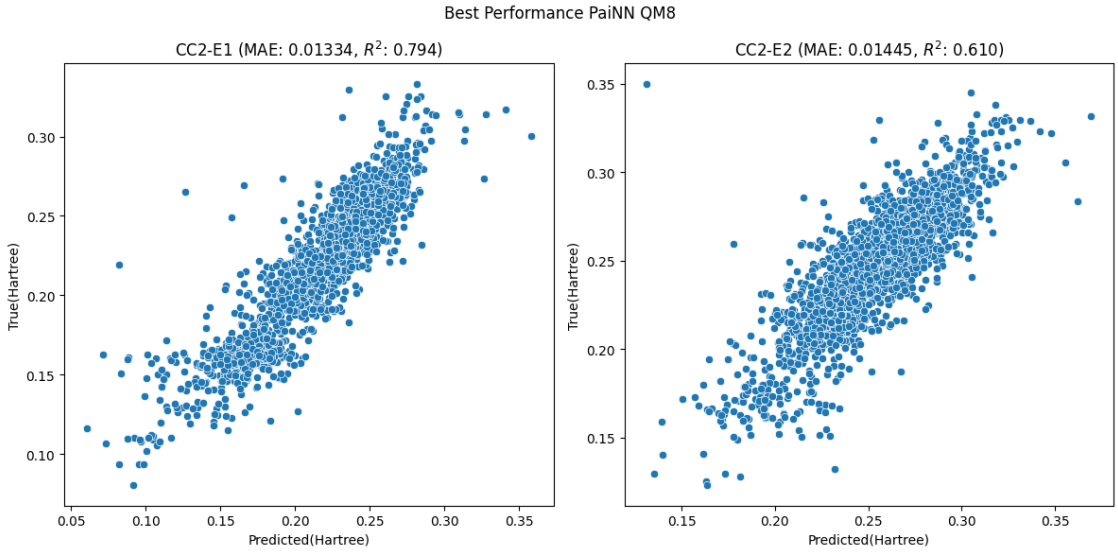


Figure B.23: Parity Plot QM8 PaiNN

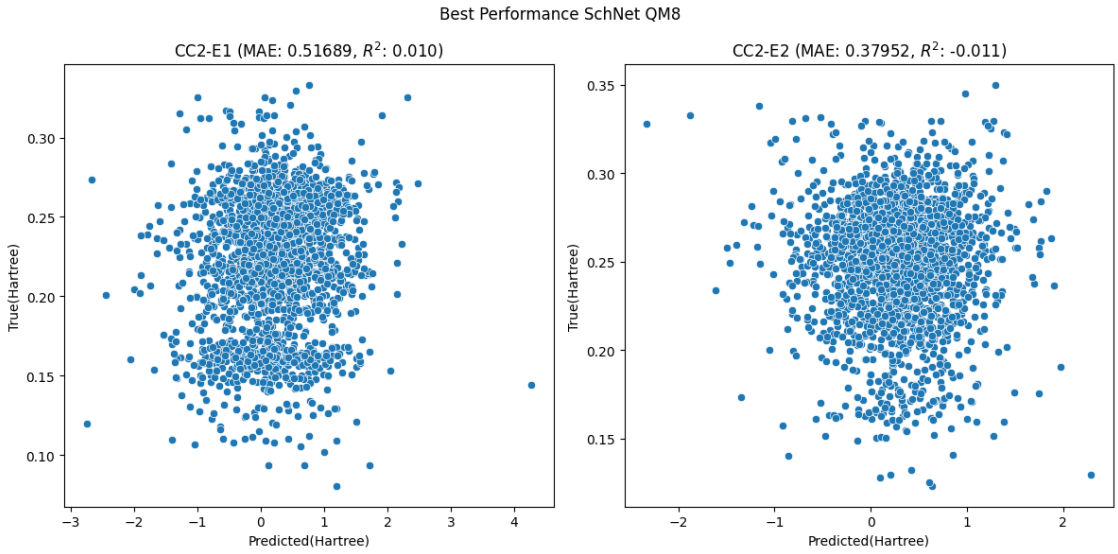


Figure B.24: Parity Plot QM8 Schnet

B.9 Correlation of QTAIM Values to Targets

B.9.1 QM8

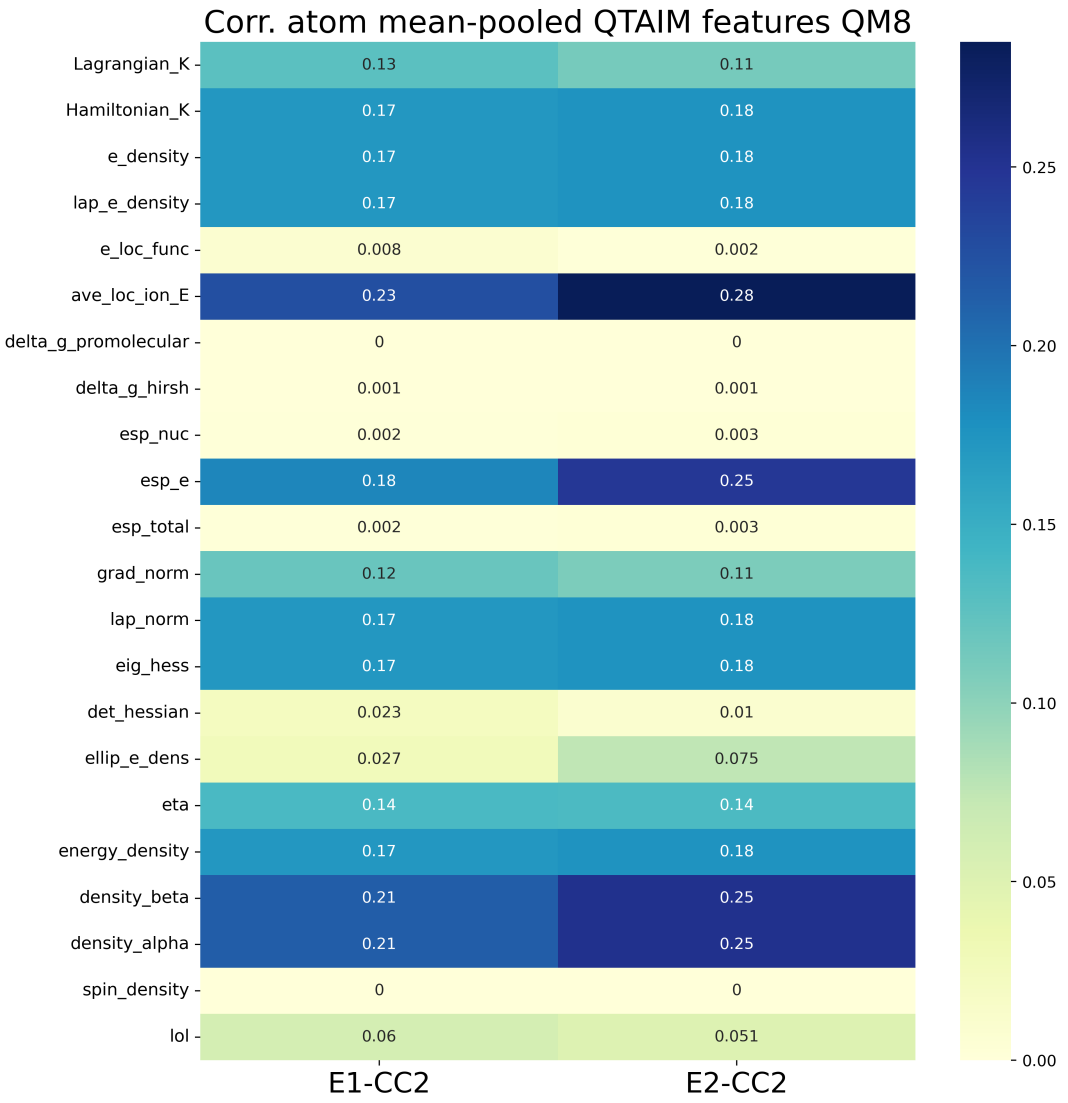


Figure B.25: Correlation of NCP values with QM8 target values

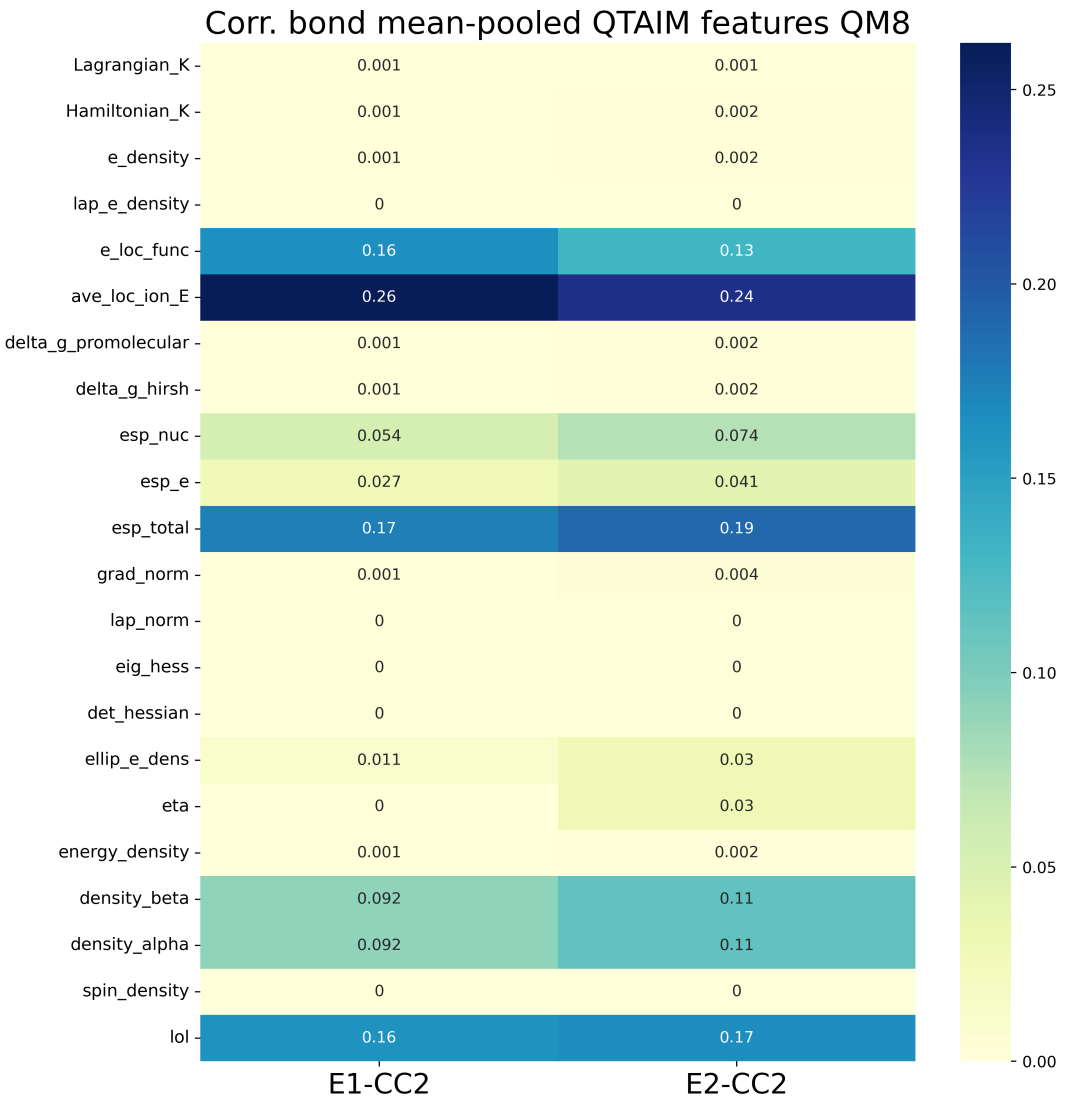


Figure B.26: Correlation of BCP values with QM8 target values

B.9.2 QM9

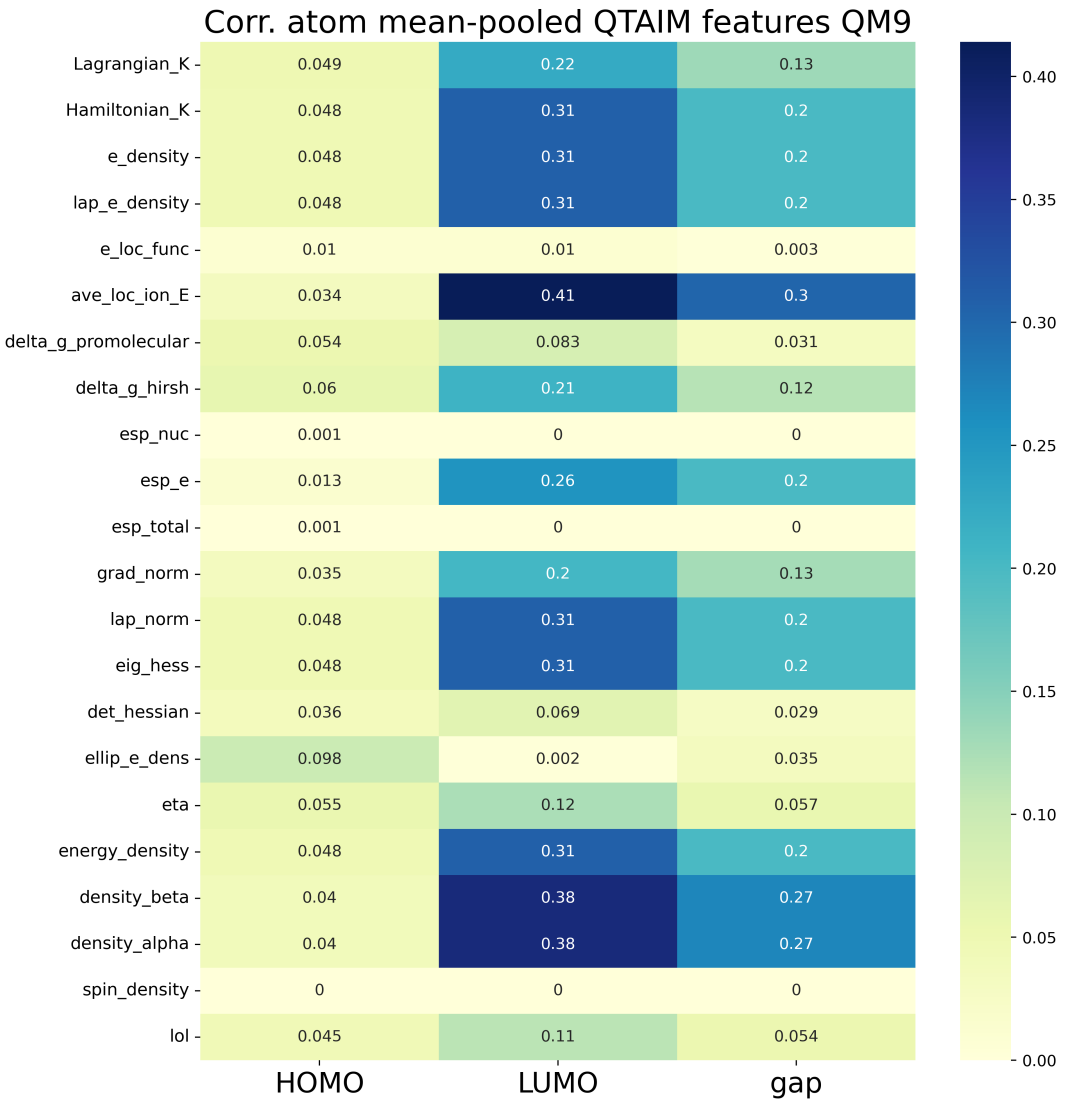


Figure B.27: Correlation of NCP values with QM9 target values

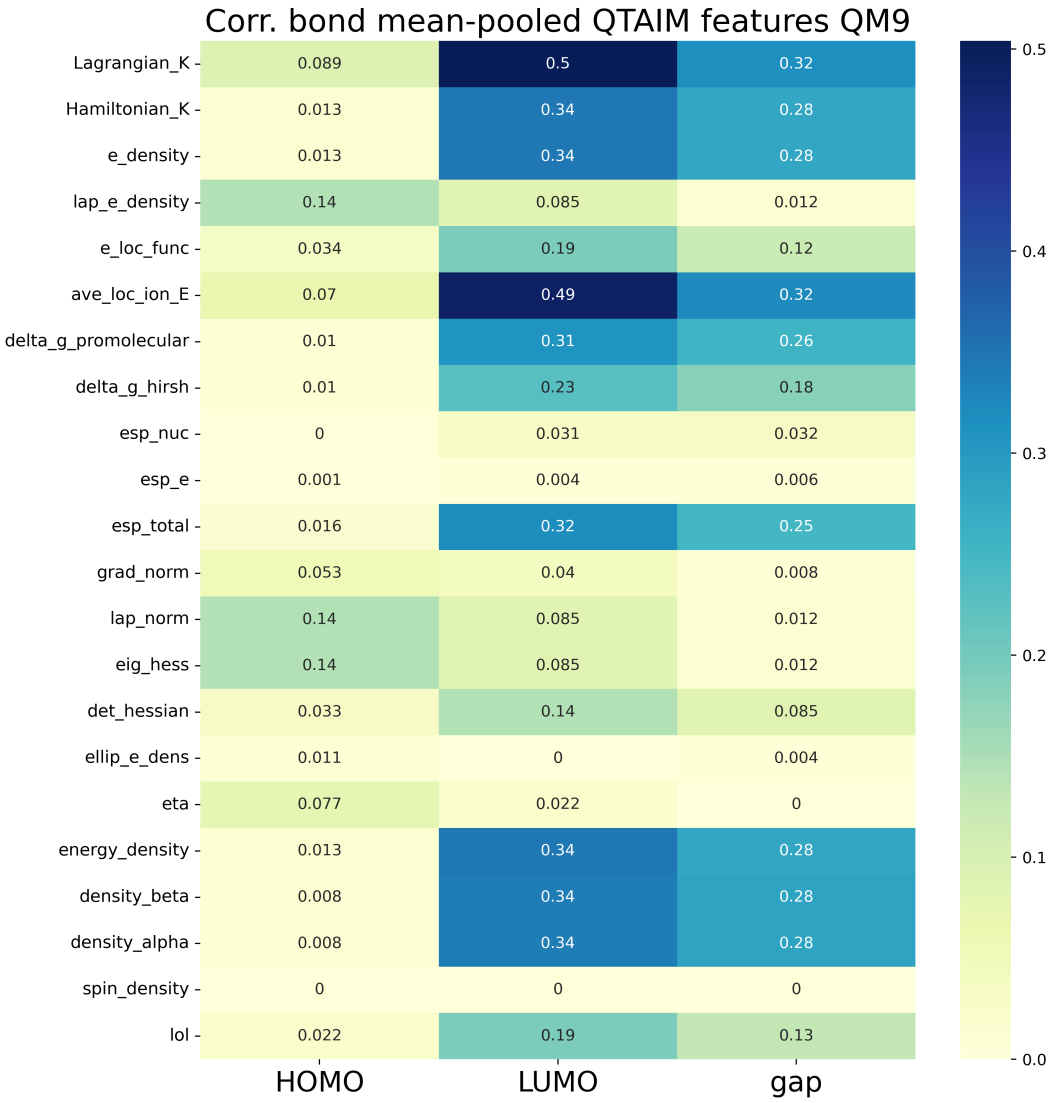


Figure B.28: Correlation of BCP values with QM9 target values

B.9.3 LIBE

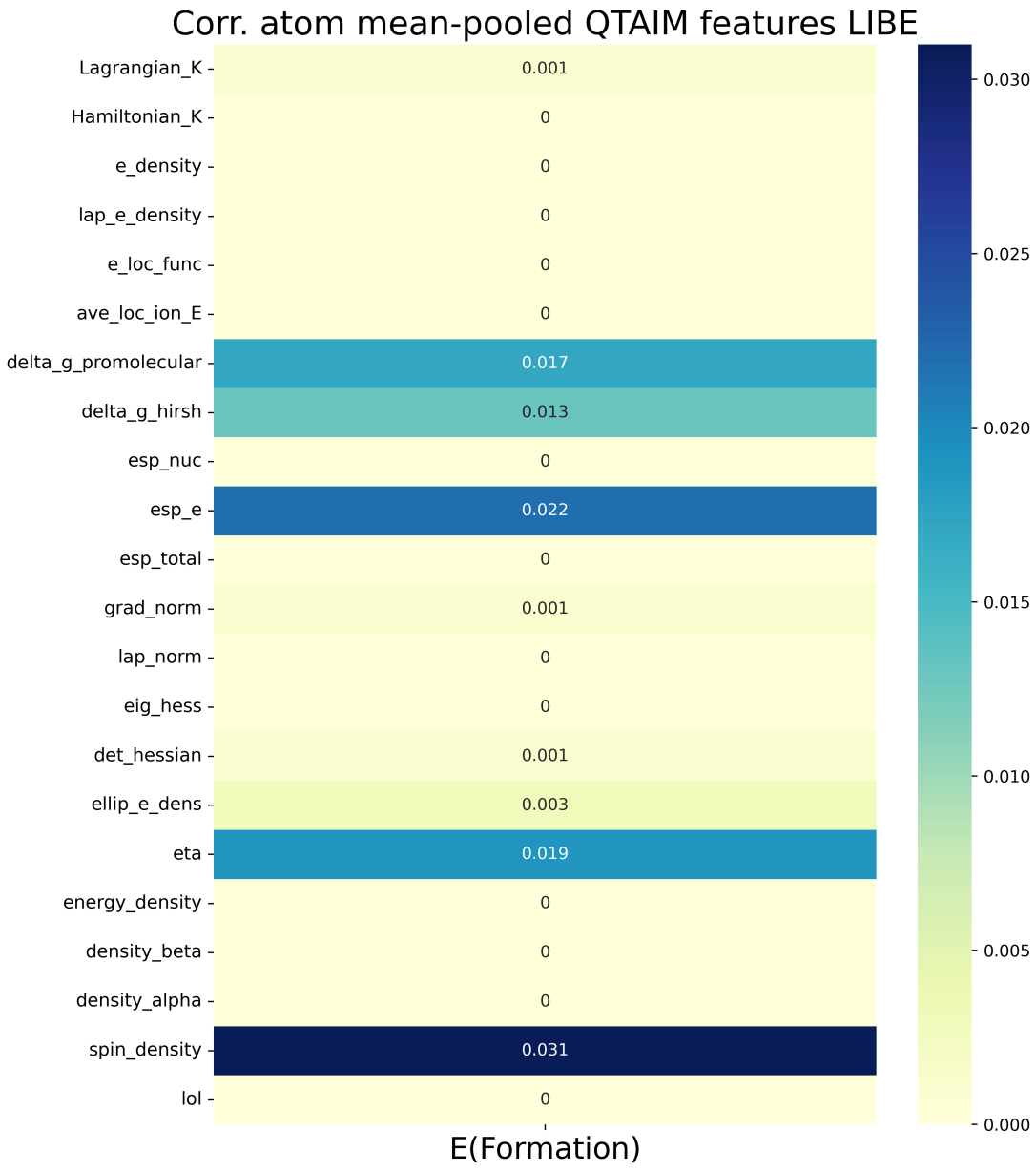


Figure B.29: Correlation of NCP values with LIBE target values

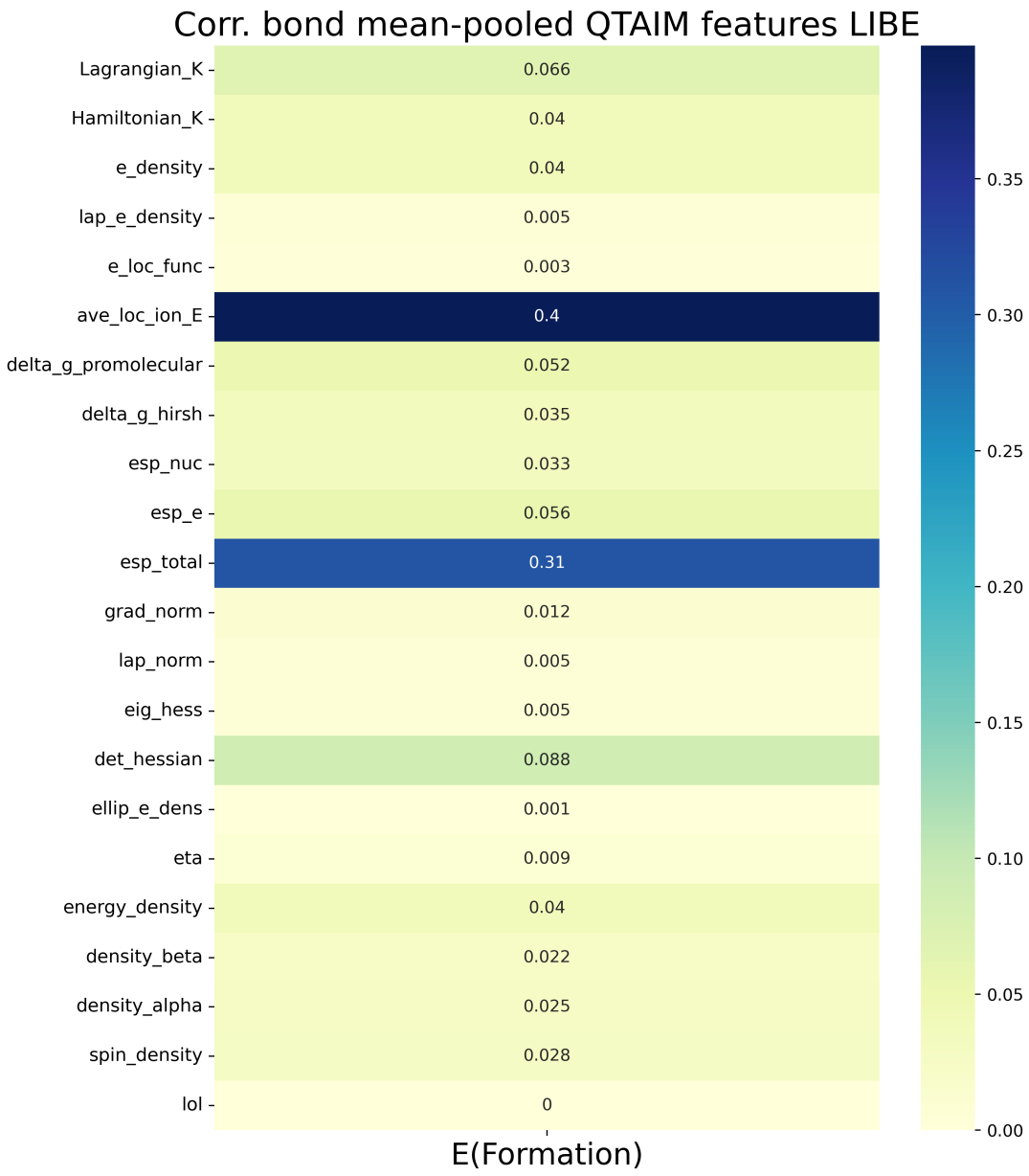


Figure B.30: Correlation of BCP values with LIBE target values

Appendix C

Supporting Information for *Machine-Learning Prediction of Protein Function from the Portrait of its Intramolecular Electric Field*

C.1 Dataset Description

Component	F_x	F_y	F_z
Overall	0.098	-0.074	0.144
Center (Fe)	0.160	-0.022	0.216

Table C.1: Average Electric Field Components (V/Ang).

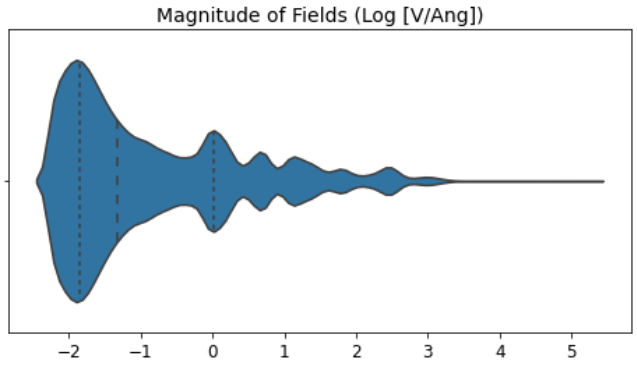


Figure C.1: Magnitude of Fields at Sampled Points Along the dataset. Average: 1.14, StD: 2.94.

C.2 Hyperparameter Tuning Information on Crystal Structure Prediction

Hyperparameter	**Values**
Model	Xgboost, RF, BalancedRandomForest
N_estimators	Min: 50, Max: 600
max_depth	Min: 2, Max: 8
Min_samples_leaf	2, 4
Bootstrap	T/F
Augmentation	T/F
Scaling	Log(Magnitude + 1), standard scaling, None

Table C.2: Hyperparameters used in bayesian optimization.

C.3 Crystal Structure PCAs Visualized

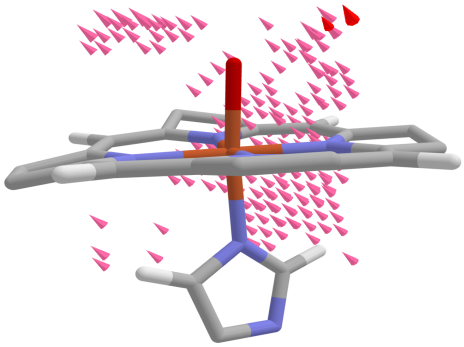


Figure C.2: Crystal Structure Training Set PC0.

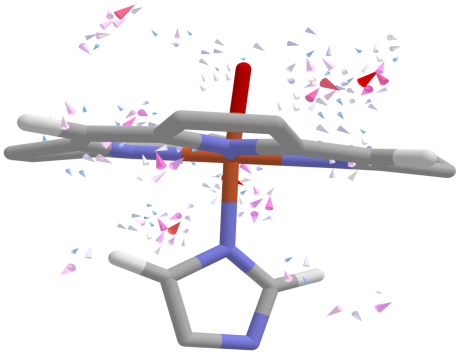


Figure C.3: Crystal Structure Training Set PC1.

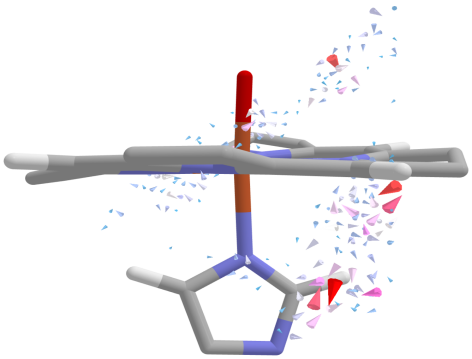


Figure C.4: Crystal Structure Training Set PC2.

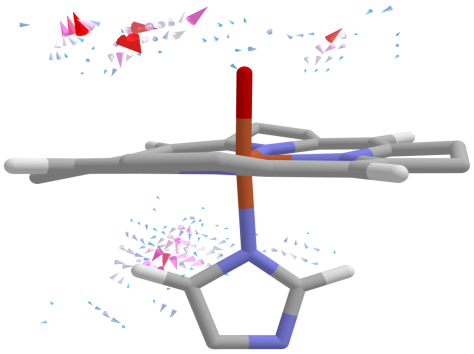


Figure C.5: Crystal Structure Training Set PC3.

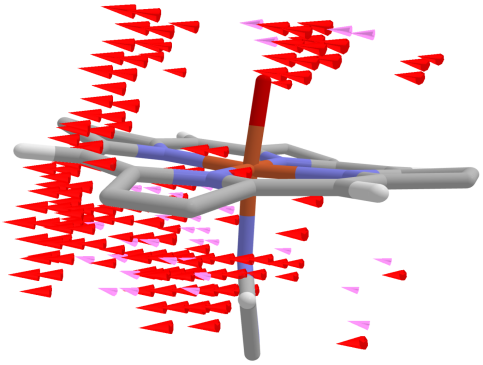


Figure C.6: Crystal Structure Training Set PC4.

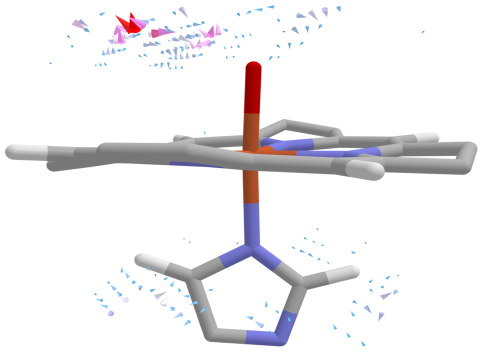


Figure C.7: Crystal Structure Training Set PC5.

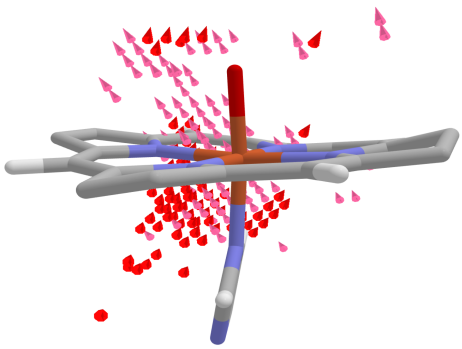


Figure C.8: Crystal Structure Training Set PC6.

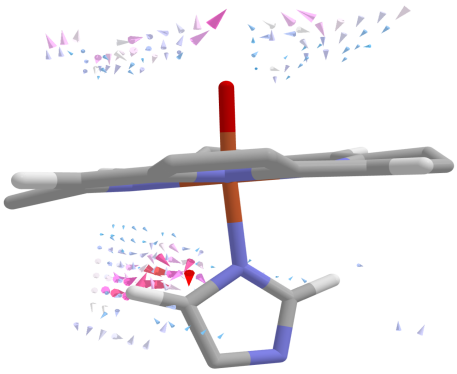


Figure C.9: Crystal Structure Training Set PC7.

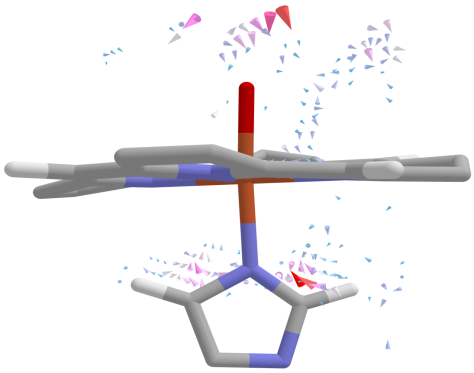


Figure C.10: Crystal Structure Training Set PC8.

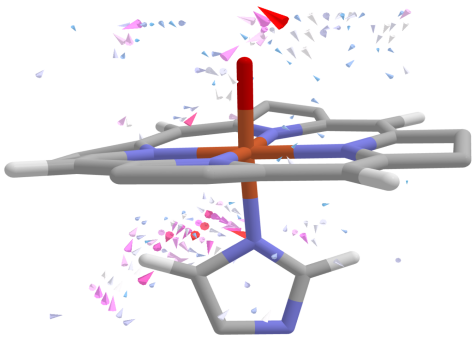


Figure C.11: Crystal Structure Training Set PC9.

C.4 MD Prediction Distribution

Train PCA	
Protein	Correct %
1U5U	47.9% (Plurality)
1APX	97.4%
3ABB	5.5%
Train-Test Combined PCA	
Protein	Correct %
1U5U	60.7%
1APX	38.3% (Plurality)
3ABB	70%

Table C.3: Table of pooled test predictions, by % of frames correctly labels, for model trained on fields along MD trajectories.

C.5 Cluster Center Breakdown

Cluster Ind	%	# Frame
1U5U		
0	10.9%	404
1	10%	473
2	79.1%	983
1JIO		

0	52.7%	122
1	47.3%	840
3VXI		
0	80.5%	004
2	12.2%	489

Table C.4: Distribution of Major Clusters used in QM/MM simulations.

C.6 Compressed Frames along PCA components

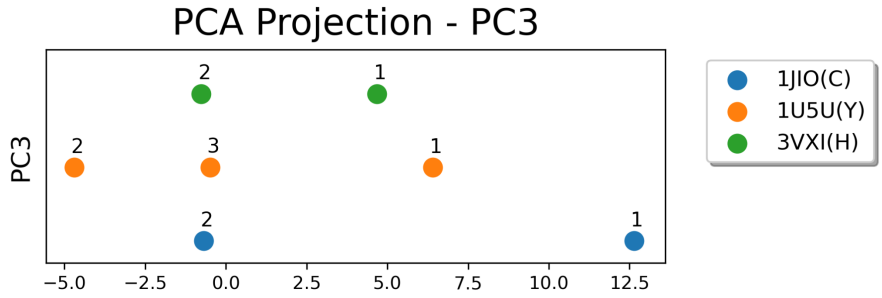


Figure C.12: Cluster Centers Projected Along PC3.

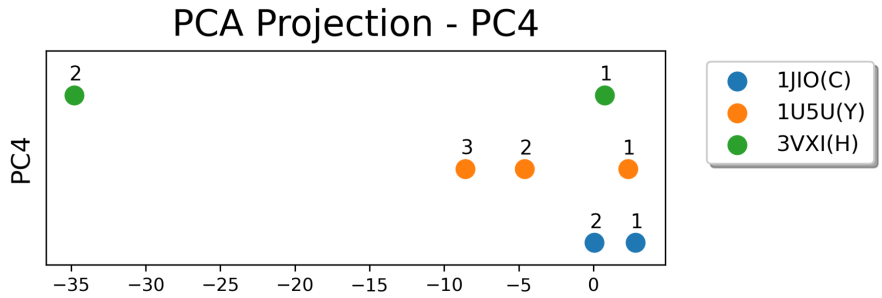


Figure C.13: Cluster Centers Projected Along PC4.

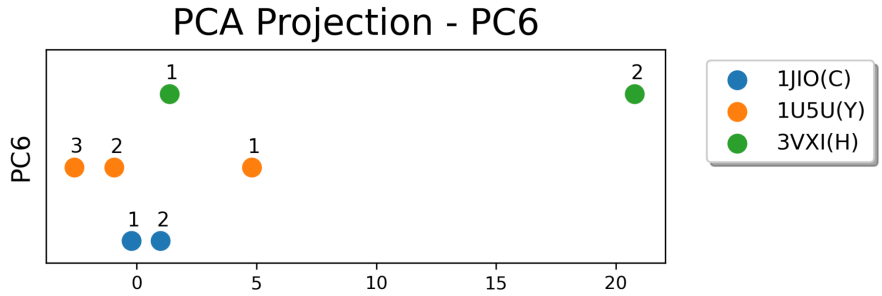


Figure C.14: Cluster Centers Projected Along PC6.

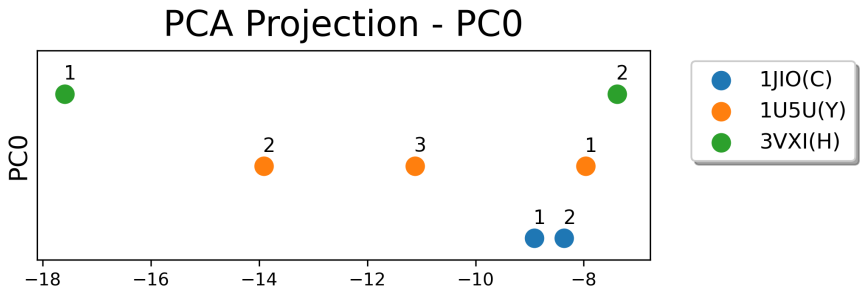


Figure C.15: Cluster Centers Projected Along PC0.

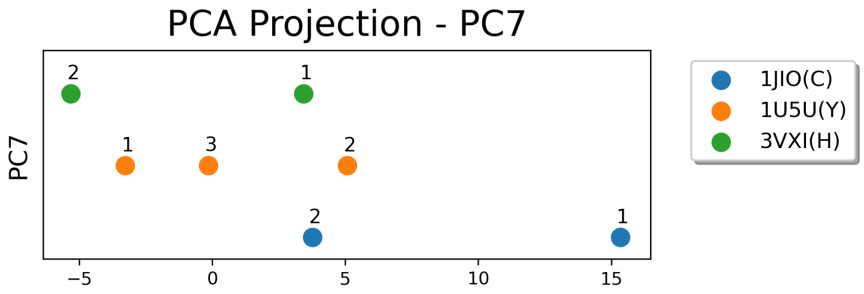


Figure C.16: Cluster Centers Projected Along PC7.

C.7 MD combined PCAs

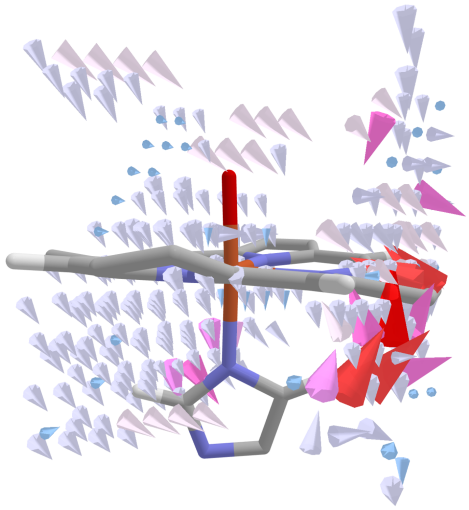


Figure C.17: Combined Train/Test PC0.

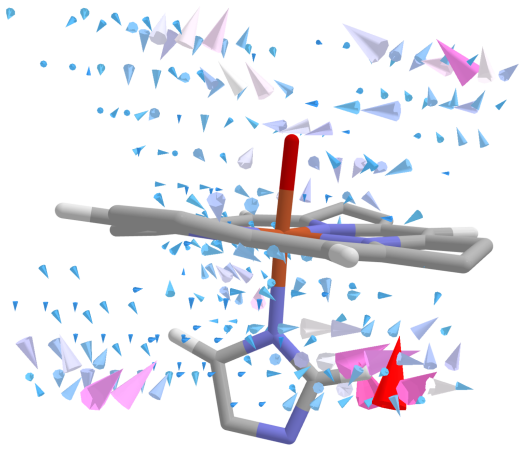


Figure C.18: Combined Train/Test PC1.

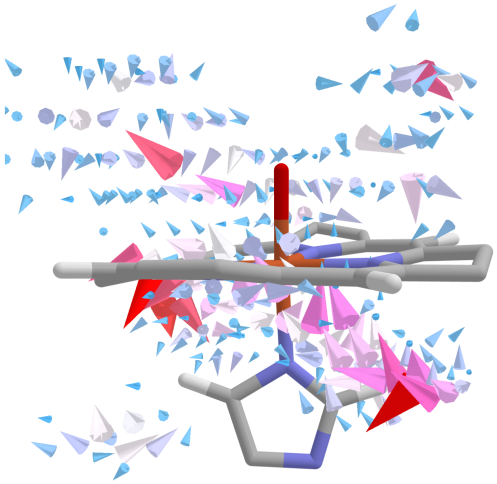


Figure C.19: Combined Train/Test PC2.

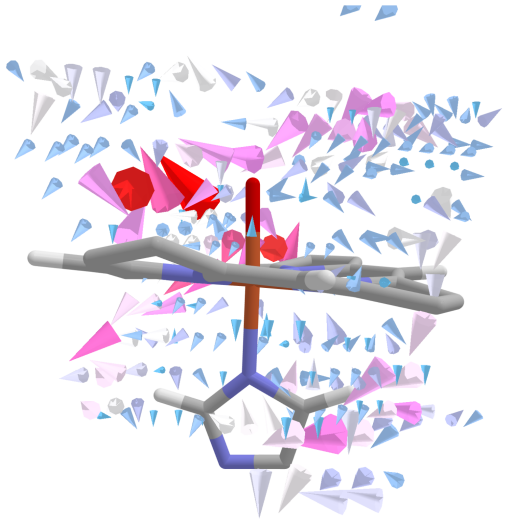


Figure C.20: Combined Train/Test PC3.

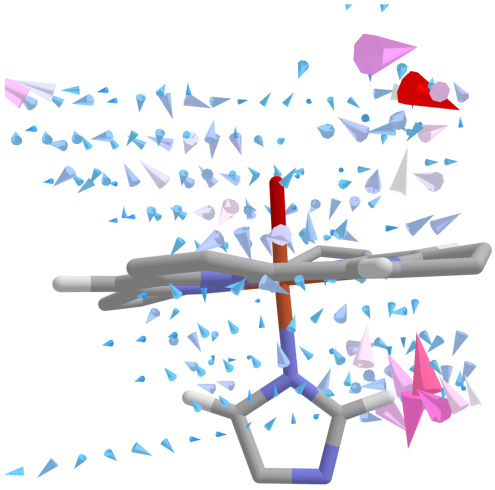


Figure C.21: Combined Train/Test PC4.

C.8 MD train only PCAs

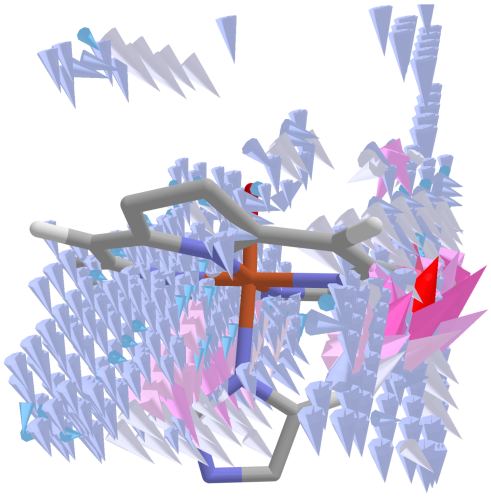


Figure C.22: Train PC0.

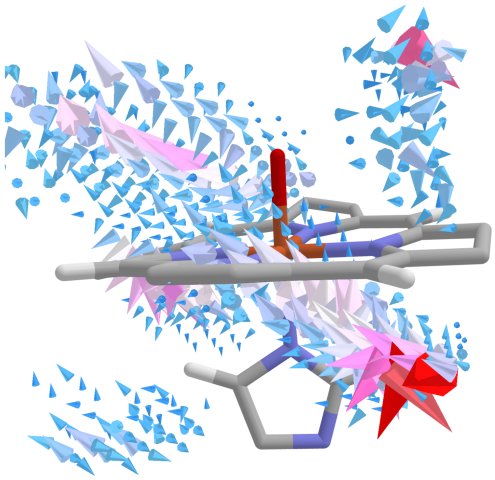


Figure C.23: Train PC1.

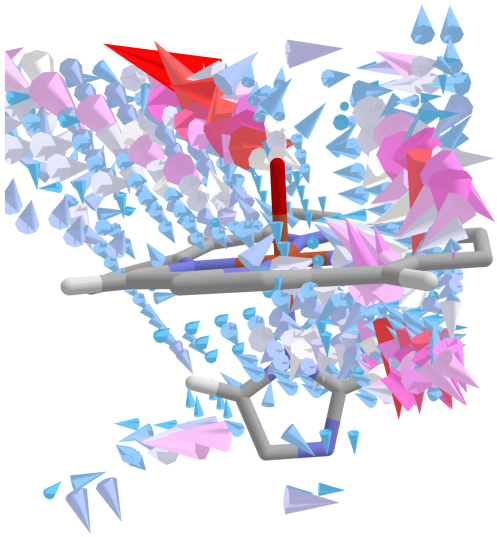


Figure C.24: Train PC2.

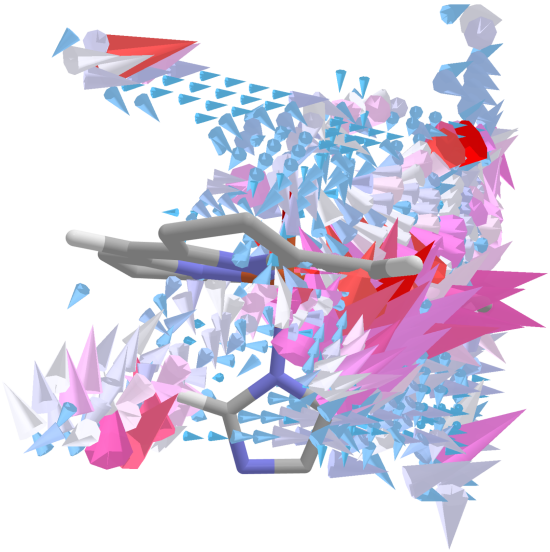


Figure C.25: Train PC3.

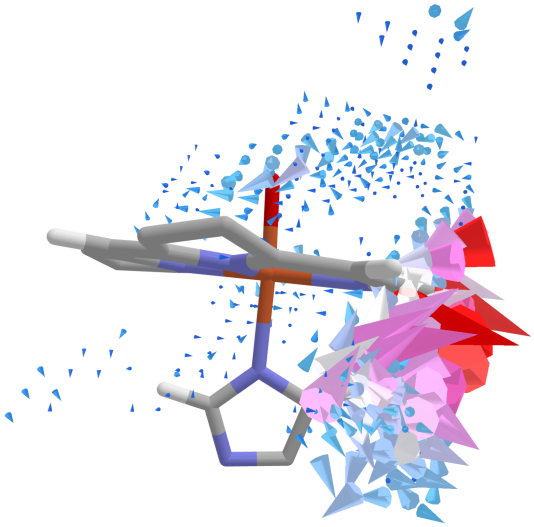


Figure C.26: Train PC4.

Appendix D

Supporting Information for *Directed Evolution of Protoglobin Optimizes the Enzyme Electric Field*

D.1 MD RMSD

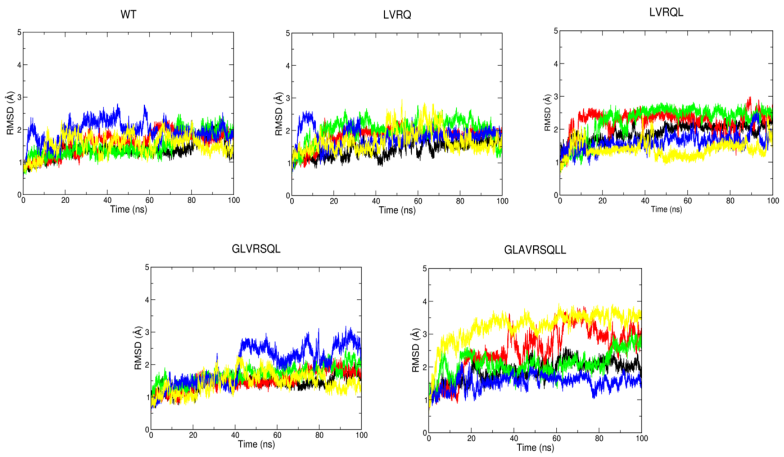


Figure D.1: RMSD Analysis of the Alpha Carbon Atoms of the Wild-Type Protoglobin and the Four Directed Evolved Variants.

D.2 Traditional Analysis

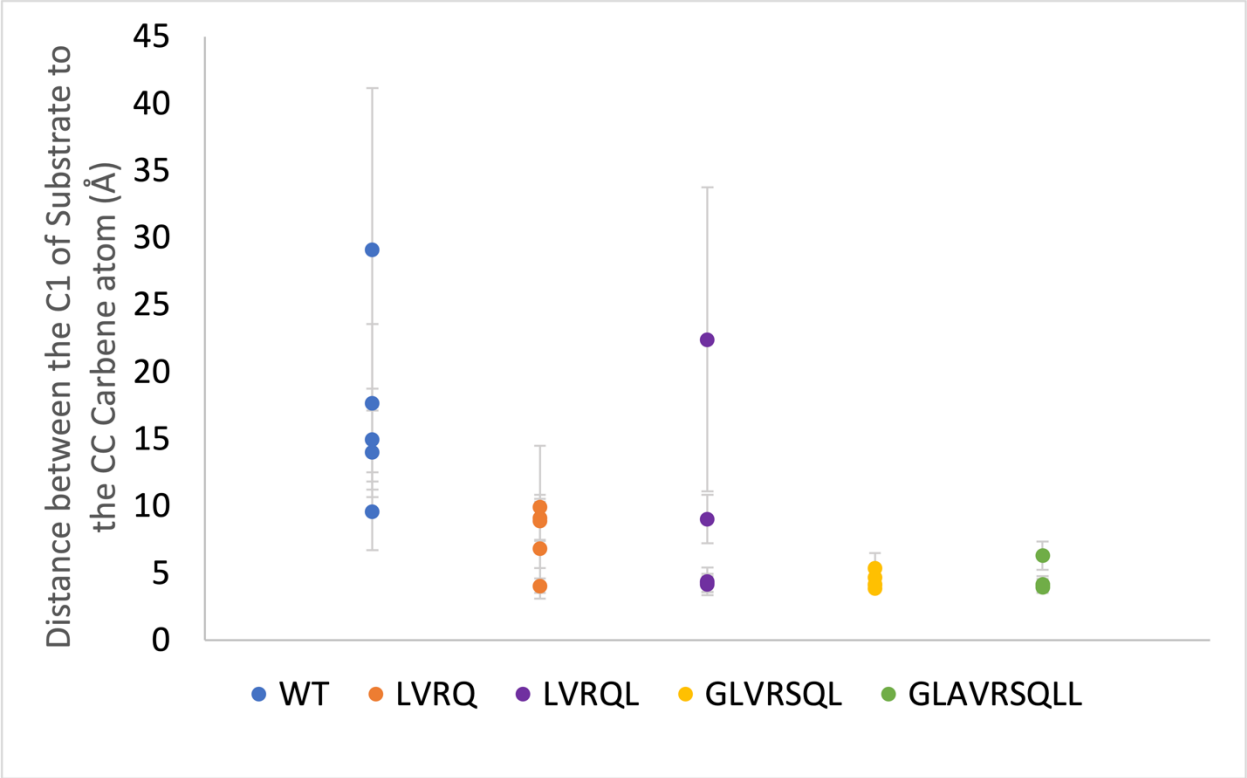


Figure D.2: Mean distances and standard deviations between the benzyl acrylate substrate and the carbene across each replica run for all analyzed systems.

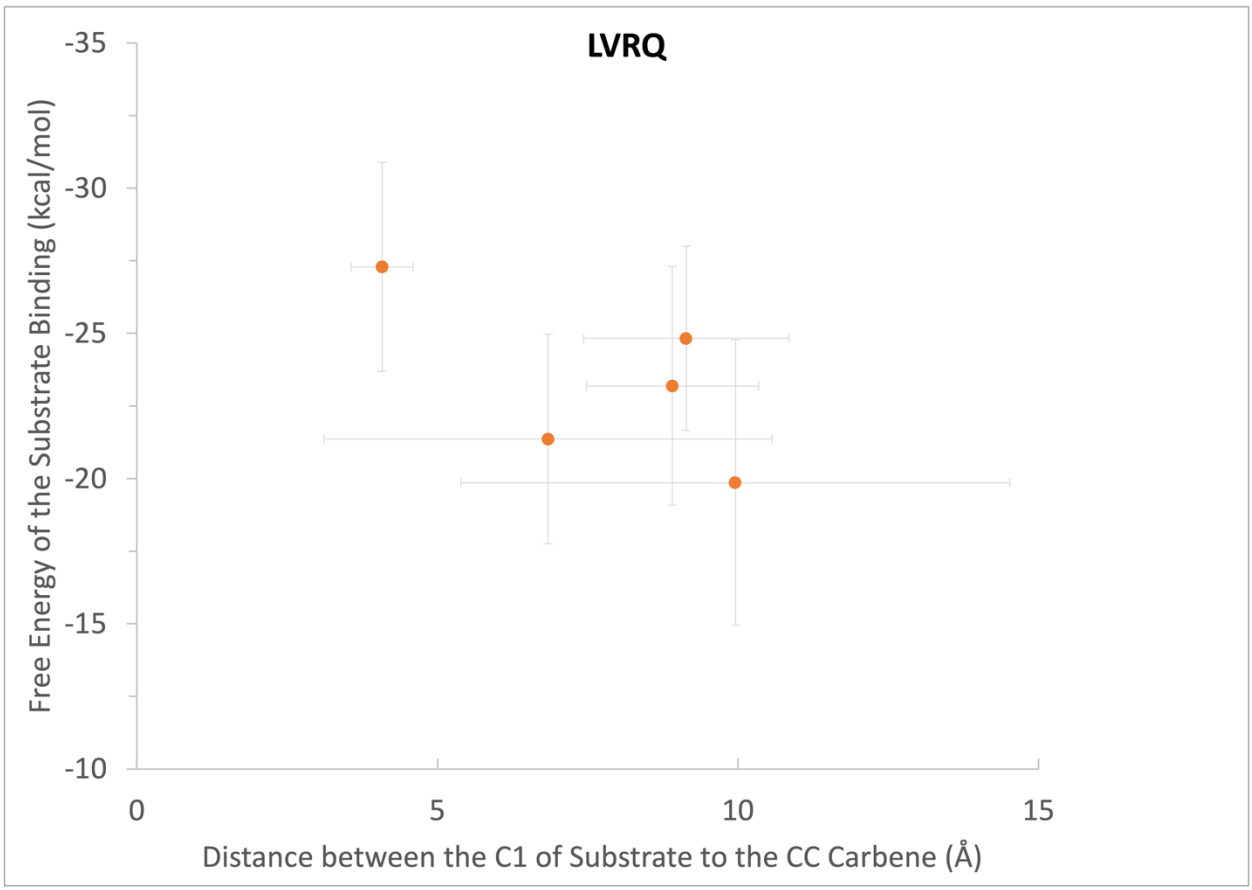


Figure D.3: Correlation between the mean distance from the benzyl acrylate substrate to carbene and the binding free energy of the benzyl acrylate substrate in LVRQ.

D.3 Spin State Benchmarking

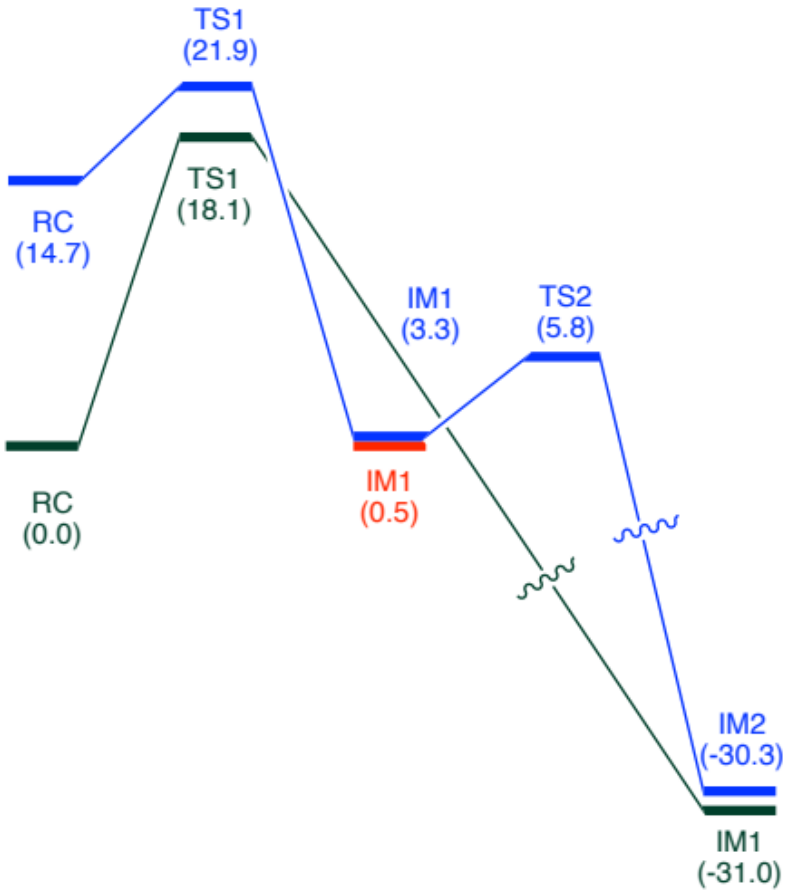


Figure D.4: Comparison of the free energy of the cyclopropanation reaction at the triplet (blue), open-shell singlet (red) and closed shell-singlet (green) spin state at the most reactive GLAVRSQLL cluster [10.3%]. Note several attempts to optimize the missing open-shell structures were not successful. The QM/MM calculations are at TPSSh functional with def2-TZVP basis set for all atoms.

D.4 PCA Data

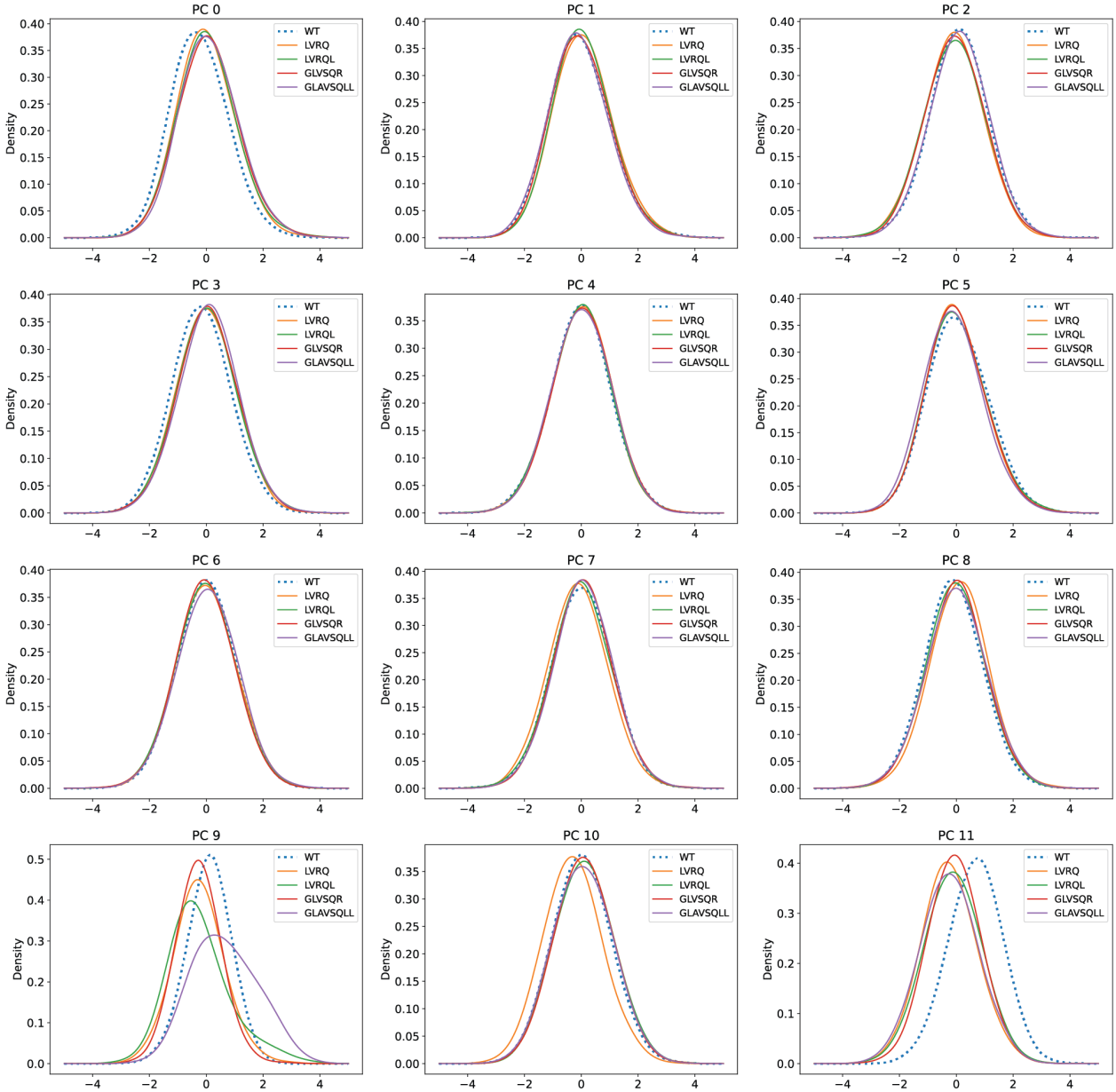


Figure D.5: Distribution of structures from replica molecular dynamics of all systems across the top Principal Components.

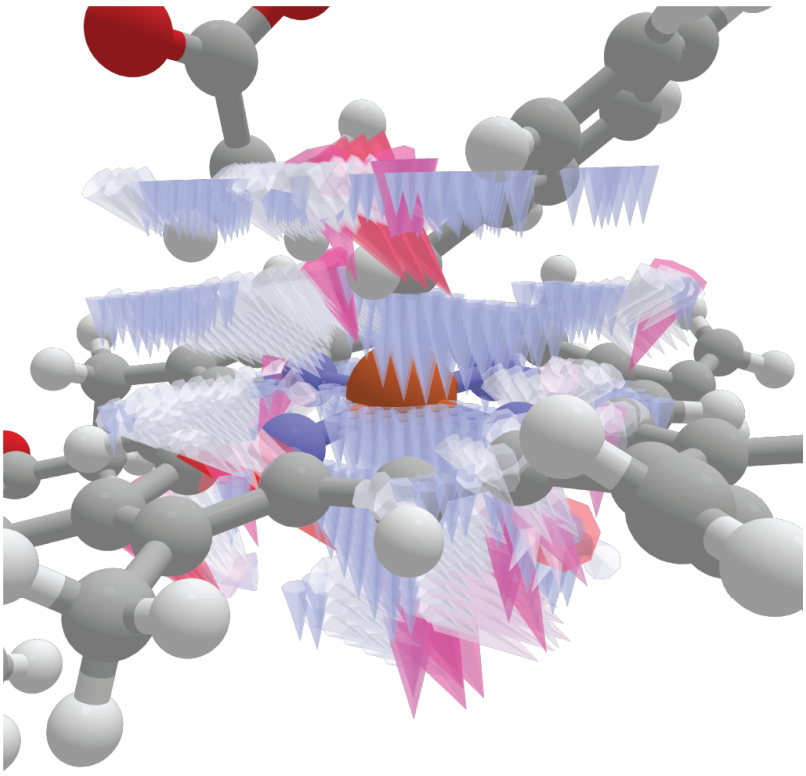


Figure D.6: Visualization of the Principal Component 9 directions plotted on the TS-GLAVRSQLL-EF2 structure.

D.5 QM Region

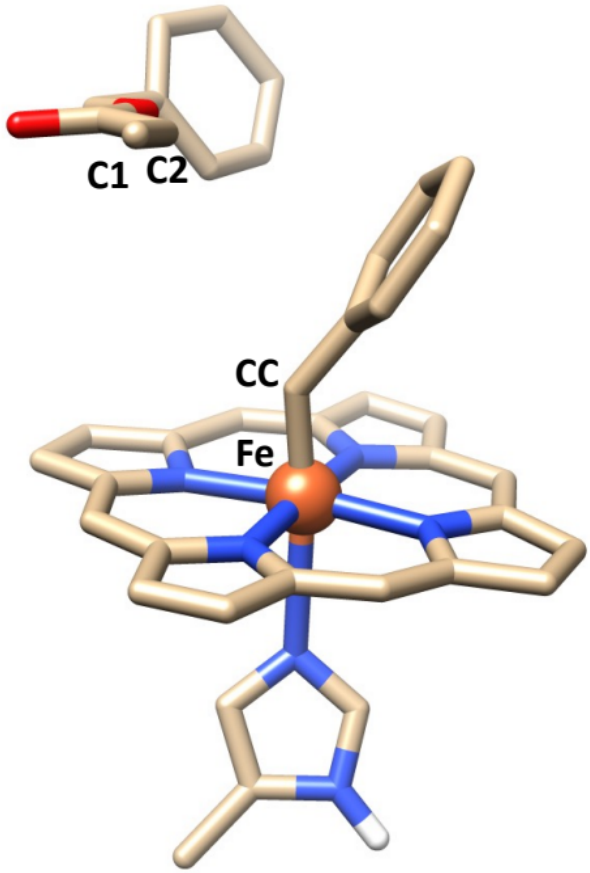


Figure D.7: QM region selected for all the QM/MM calculations.

D.6 Topology Data

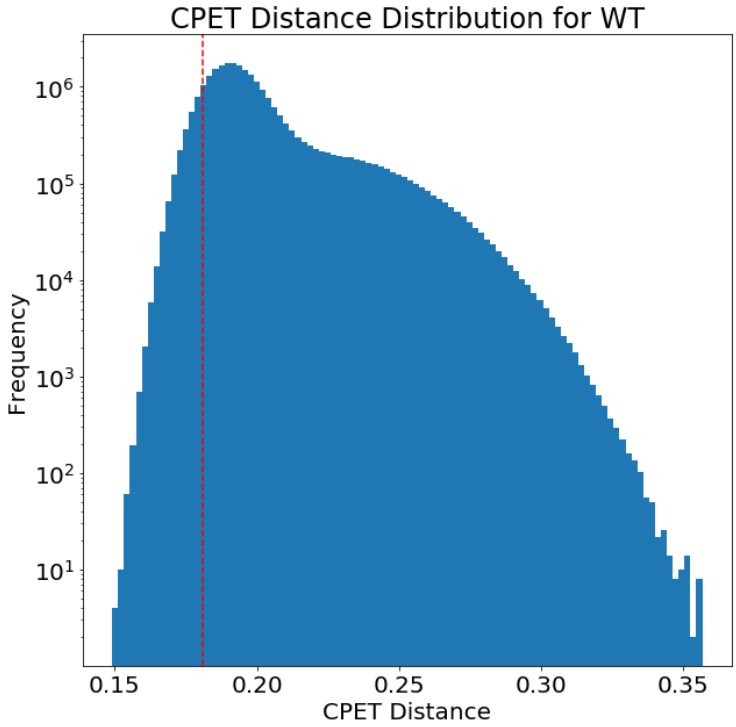


Figure D.8: Distribution of CPET distances for WT trajectories. The vertical denotes the cutoff distance we used prior to compression.

D.6.1 Box Distance Benchmarking

Index	Name	Count	%
	Box size: 1.5 Å		
	WT		
1	WT-run4-892.top	2879	57.58

Index	Name	Count	%
2	WT-run1-539.top	985	19.70
3	WT-run3-147.top	449	8.98
4	WT-run2-266.top	423	8.46
5	WT-run5-221.top	155	3.10
6	WT-run3-079.top	69	1.38
7	WT-run3-192.top	40	0.80
	LVRQ		
1	LVRQ-run4-589.top	3526	70.52
2	LVRQ-run2-191.top	1474	29.48
	LVRQL		
1	LVRQL-run4-334.top	3928	78.56
2	LVRQL-run3-017.top	462	9.24
3	LVRQL-run4-154.top	403	8.06
4	LVRQL-run5-886.top	207	4.14
	GLVRSQL		
1	GLVRSQL-run2-048.top	2796	55.92
2	GLVRSQL-run5-573.top	1924	38.48
3	GLVRSQL-run5-337.top	230	4.60
4	GLVRSQL-run3-716.top	50	1.00

Index	Name	Count	%
	GLAVRSQLL		
1	GLAVRSQLL-run5-846.top	1693	33.86
2	GLAVRSQLL-run5-238.top	1296	25.92
3	GLAVRSQLL-run5-464.top	824	16.48
4	GLAVRSQLL-run1-040.top	709	14.18
5	GLAVRSQLL-run1-196.top	478	9.56
	Box size: 2.0 Å		
	WT		
1	WT-run4-892.top	2254	45.07
2	WT-run1-539.top	1599	31.97
3	WT-run5-221.top	391	7.82
4	WT-run3-147.top	369	7.38
5	WT-run5-065.top	272	5.44
6	WT-run2-266.top	92	1.84
7	WT-run1-120.top	24	0.48
	LVRQ		
1	LVRQ-run4-589.top	3416	68.32
2	LVRQ-run1-716.top	556	11.12

Index	Name	Count	%
3	LVRQ-run2-239.top	408	8.16
4	LVRQ-run5-619.top	350	7.00
5	LVRQ-run2-191.top	270	5.40

LVRQL

1	LVRQL-run4-334.top	4078	81.56
2	LVRQL-run4-998.top	465	9.30
3	LVRQL-run3-017.top	457	9.14

GLVRSQL

1	GLVRSQL-run2-048.top	2123	42.46
2	GLVRSQL-run3-931.top	1397	27.94
3	GLVRSQL-run1-845.top	1018	20.36
4	GLVRSQL-run3-453.top	249	4.98
5	GLVRSQL-run4-268.top	213	4.26

GLAVRSQLL

1	GLAVRSQLL-run5-464.top	2287	45.74
2	GLAVRSQLL-run5-846.top	1290	25.80
3	GLAVRSQLL-run5-238.top	670	13.40
4	GLAVRSQLL-run1-040.top	532	10.64
5	GLAVRSQLL-run5-605.top	148	2.96

Index	Name	Count	%
6	GLAVRSQLL-run5-009.top	73	1.46

Box size: 2.5 Å

WT

1	WT-run1-539.top	2455	49.10
2	WT-run5-065.top	1346	26.92
3	WT-run4-892.top	1117	22.34
4	WT-run1-127.top	82	1.64

LVRQ

1	LVRQ-run5-972.top	1441	28.82
2	LVRQ-run5-701.top	956	19.12
3	LVRQ-run5-914.top	624	12.48
4	LVRQ-run4-170.top	541	10.82
5	LVRQ-run2-695.top	536	10.72
6	LVRQ-run1-590.top	322	6.44
7	LVRQ-run5-496.top	273	5.46
8	LVRQ-run5-915.top	212	4.24
9	LVRQ-run1-716.top	95	1.90

LVRQL

Index	Name	Count	%
1	LVRQL-run4-334.top	4019	80.38
2	LVRQL-run4-688.top	737	14.74
3	LVRQL-run3-017.top	244	4.88

GLVRSQL

1	GLVRSQL-run1-845.top	3503	70.06
2	GLVRSQL-run3-931.top	715	14.30
3	GLVRSQL-run1-361.top	399	7.98
4	GLVRSQL-run3-936.top	333	6.66
5	GLVRSQL-run4-268.top	50	1.00

GLAVRSQLL

1	GLAVRSQLL-run5-464.top	3497	69.94
2	GLAVRSQLL-run5-605.top	747	14.94
3	GLAVRSQLL-run5-846.top	385	7.70
4	GLAVRSQLL-run5-446.top	241	4.82
5	GLAVRSQLL-run5-217.top	130	2.60

Box size: 3.0 Å

WT

1	WT-run1-539.top	3226	64.52

Index	Name	Count	%
2	WT-run5-065.top	1729	34.58
3	WT-run1-127.top	24	0.48
4	WT-run2-855.top	21	0.42
	LVRQ		
1	LVRQ-run5-972.top	2916	58.32
2	LVRQ-run4-258.top	968	19.36
3	LVRQ-run4-675.top	430	8.60
4	LVRQ-run1-635.top	284	5.68
5	LVRQ-run3-699.top	196	3.92
6	LVRQ-run1-590.top	154	3.08
7	LVRQ-run5-915.top	52	1.04
	LVRQL		
1	LVRQL-run4-334.top	3802	76.04
2	LVRQL-run3-017.top	1068	21.36
3	LVRQL-run1-168.top	130	2.60
	GLVRSQL		
1	GLVRSQL-run1-845.top	3299	65.98
2	GLVRSQL-run1-361.top	1217	24.34
3	GLVRSQL-run3-936.top	484	9.68

Index	Name	Count	%
	GLAVRSQLL		
1	GLAVRSQLL-run5-464.top	3991	79.82
2	GLAVRSQLL-run5-605.top	517	10.34
3	GLAVRSQLL-run5-300.top	323	6.46
4	GLAVRSQLL-run5-446.top	119	2.38
5	GLAVRSQLL-run5-217.top	50	1.00

Table D.1: Box size parameter sweep for electric field clustering. The box with the largest size of 3.0 A provides the most space for sampling without getting too close to other nearby charge residues, as indicated by similar cluster centers as the 2.5 A box.

Distances	Fe-CC (Å)	CC-C1 (Å)	CC-C2 (Å)
WT			
1-01-WT-1RC	1.80	13.52	13.42
1-02-WT-1RC	1.78	15.40	14.63
LVRQ			

Distances	Fe-CC (Å)	CC-C1 (Å)	CC-C2 (Å)
2-01-LVRQ-1RC	1.78	8.00	9.10
2-02-LVRQ-1RC	1.78	6.01	5.58
2-03-LVRQ-1RC	1.78	3.69	4.47
2-03-LVRQ-2TS	1.97	1.71	2.68
2-03-LVRQ-3PD	4.48	1.48	1.52
2-04-LVRQ-1RC	1.78	9.42	8.21
LVRQLL			
3-01-LVRQL-1RC	1.79	3.15	4.13
3-01-LVRQL-2TS	1.95	1.87	2.68
3-01-LVRQL-3PD	4.03	1.49	1.54
3-02-LVRQL-1RC	1.78	3.39	4.41
3-02-LVRQL-2TS	1.92	1.89	2.76
3-02-LVRQL-3PD	3.17	1.48	1.54
GLVRSQL			
4-01-GLVRSQL-1RC	1.78	3.14	3.79
4-01-GLVRSQL-2TS	2.00	1.71	2.60
4-01-GLVRSQL-3PD	3.53	1.49	1.53

Distances	**Fe-CC (Å)**	**CC-C1 (Å)**	**CC-C2 (Å)**
4-02-GLVRSQL-1RC	1.78	3.56	3.92
4-02-GLVRSQL-2TS	1.97	1.80	2.57
4-02-GLVRSQL-3PD	3.87	1.49	1.55
4-03-GLVRSQL-1RC	1.97	1.80	2.57
4-03-GLVRSQL-2TS	2.22	1.60	2.37
4-03-GLVRSQL-3PD	4.68	1.49	1.54
GLAVRSQLL			
5-01-GLAVRSQLL-1RC	1.78	5.67	4.91
5-01-GLAVRSQLL-2TS	2.02	1.72	2.45
5-01-GLAVRSQLL-3PD	3.21	1.49	1.52
5-02-GLAVRSQLL-1RC	1.77	3.72	4.03
5-02-GLAVRSQLL-2TS	1.95	1.88	2.45
5-02-GLAVRSQLL-3PD	2.97	1.48	1.53
5-03-GLAVRSQLL-1RC	1.78	3.92	5.24
5-03-GLAVRSQLL-2TS	1.95	1.87	2.67
5-03-GLAVRSQLL-3PD	3.64	1.48	1.54

Table D.2: Summary of key distances in QM/MM optimized reactants, transition states, and products across all systems for the cyclopropanation reaction.

D.6.2 Cluster Center Energetics

Cluster[MD%]	$\Delta G^{\ddagger}$ (B3LYP)	$\Delta G^{\ddagger}$ (TPSSh)	$\Delta G_{\mathbf{rxn}}$ (B3LYP)	$\Delta G_{\mathbf{rxn}}$ (TPSSh)
$LVRQ_1$(8.6%)	27.07	21.18	-23.17	-13.61
$LVRQL_1$(76.0%)	22.80	19.66	-35.31	-33.51
$LVRQL_2$(21.4%)	29.09	22.62	-25.39	-20.83
$GLVRSQL_1$(66.0%)	30.67	25.99	-24.17	-15.73
$GLVRSQL_2$(24.3%)	34.92	31.83	-24.52	-19.44
$GLVRSQL_3$(9.7%)	27.11	22.25	-23.28	-16.24
$GLAVRSQLL_1$(79.8%)	35.06	35.96	-27.37	-20.67
$GLAVRSQLL_2$(10.3%)	15.42	18.18	-31.15	-31.03
$GLAVRSQLL_3$(6.5%)	24.90	19.40	-38.51	-37.40

Table D.3: Transition state and reaction free energies for reactive clusters from each variant at B3-LYP and TPSSh functionals (in kcal/mol).

System[MD%]	ZPE (kcal/mol)	S_{vib} (kcal/mol/K)	Imaginary Frequency
$LVRQ_1$(8.6%) - RC	366.43	0.13	N/A
$LVRQ_1$(8.6%) - TS	365.73	0.12	-127.3
$LVRQ_1$(8.6%) - PC	369.76	0.12	N/A

System[MD%]	ZPE (kcal/mol)	S_{vib} (kcal/mol/K)	Imaginary Frequency
LVRQL$_1$(76.0%) - RC	366.12	0.13	N/A
LVRQL$_1$(76.0%) - TS	365.44	0.12	-346.8
LVRQL$_1$(76.0%)- PC	369.77	0.12	N/A
LVRQL$_2$(21.4%) - RC	366.22	0.13	N/A
LVRQL$_2$(21.4%) - TS	365.59	0.12	-390.3
LVRQL$_2$(21.4%) - PC	369.20	0.12	N/A
GLVRSQL$_1$(66.0%) - RC	366.13	0.13	N/A
GLVRSQL$_1$(66.0%) - TS	366.37	0.12	-179.0
GLVRSQL$_1$(66.0%) - PC	369.18	0.12	N/A
GLVRSQL$_2$(24.3%) - RC	366.74	0.13	N/A
GLVRSQL$_2$(24.3%)- TS	365.05	0.12	-279.6
GLVRSQL$_2$(24.3%) - PC	369.34	0.12	N/A
GLVRSQL$_3$(9.7%) - RC	365.21	0.13	N/A
GLVRSQL$_3$(9.7%) - TS	365.30	0.13	-236.9
GLVRSQL$_3$(9.7%) - PC	369.61	0.12	N/A
GLAVRSQLL$_1$(79.8%) - RC	366.57	0.13	N/A
GLAVRSQLL$_1$(79.8%) - TS	365.42	0.12	-85.7
GLAVRSQLL$_1$(79.8%) - PC	370.18	0.12	N/A
GLAVRSQLL$_2$(10.3%)- RC	366.59	0.13	N/A

System[MD%]	ZPE (kcal/mol)	S_{vib} (kcal/mol/K)	Imaginary Frequency
$GLAVRSQLL_2$(10.3%) - TS	366.29	0.13	-418.9
$GLAVRSQLL_2$(10.3%) - PC	369.53	0.12	N/A
$GLAVRSQLL_3$(6.5%)- RC	366.65	0.13	N/A
$GLAVRSQLL_3$(6.5%) - TS	365.80	0.12	-397.5
$GLAVRSQLL_3$(6.5%) - PC	369.22	0.12	N/A

Table D.4: Thermodynamics corrections and transition state frequencies for the reactive clusters from each variant.

D.7 Mulliken Charges of Cluster Centers

Charges	Fe	CC	C1	C2
WT				
1-01-WT-1RC	-0.14164	-0.13955	0.01580	-0.09994
1-02-WT-1RC	0.00286	-0.16927	0.01008	-0.09707
LVRQ				
2-01-LVRQ-1RC	-0.02758	-0.16726	-0.00602	-0.08271
2-02-LVRQ-1RC	-0.00785	-0.14163	0.01284	-0.08780

Charges	Fe	CC	C1	C2
2-03-LVRQ-1RC	0.01003	-0.23359	-0.02879	-0.11094
2-03-LVRQ-2TS	-0.04858	-0.44006	0.34295	-0.15394
2-03-LVRQ-3PD	-0.18849	-0.45495	0.19898	0.02425
2-04-LVRQ-1RC	-0.04754	-0.16185	-0.00256	-0.08949
LVRQLL				
3-01-LVRQL-1RC	-0.12884	-0.29870	0.01422	-0.09565
3-01-LVRQL-2TS	-0.21566	-0.62815	0.35124	-0.08816
3-01-LVRQL-3PD	-0.06324	-0.58806	0.24895	0.02414
3-02-LVRQL-1RC	-0.06794	-0.26950	-0.02176	-0.10175
3-02-LVRQL-2TS	-0.11214	-0.46187	0.28051	-0.05757
3-02-LVRQL-3PD	-0.12255	-0.57562	0.27153	0.02784
GLVRSQL				
4-01-GLVRSQL-1RC	0.03082	-0.34774	0.01833	-0.12559
4-01-GLVRSQL-2TS	-0.03006	-0.47951	0.36358	-0.14595
4-01-GLVRSQL-3PD	-0.02063	-0.54774	0.28930	-0.00963
4-02-GLVRSQL-1RC	0.03090	-0.25857	-0.00520	-0.12825
4-02-GLVRSQL-2TS	-0.02515	-0.54564	0.33862	-0.13077
4-02-GLVRSQL-3PD	-0.00293	-0.45000	0.27096	-0.02807

Charges	Fe	CC	C1	C2
4-03-GLVRSQL-1RC	-0.01020	-0.22215	-0.00707	-0.10875
4-03-GLVRSQL-2TS	-0.11771	-0.56735	0.48559	-0.10915
4-03-GLVRSQL-3PD	-0.00509	-0.51902	0.25709	0.00180
GLAVRSQLL				
5-01-GLAVRSQLL-1RC	0.01745	-0.16811	0.01632	-0.10940
5-01-GLAVRSQLL-2TS	-0.05329	-0.41845	0.39818	-0.16977
5-01-GLAVRSQLL-3PD	-0.13911	-0.61906	0.27256	-0.01745
5-02-GLAVRSQLL-1RC	0.01512	-0.25447	0.03834	-0.13135
5-02-GLAVRSQLL-2TS	-0.03592	-0.50015	0.33124	-0.16915
5-02-GLAVRSQLL-3PD	-0.22741	-0.71862	0.30531	0.01956
5-03-GLAVRSQLL-1RC	-0.00191	-0.20706	-0.04438	-0.09098
5-03-GLAVRSQLL-2TS	-0.10740	-0.59598	0.29873	-0.10723
5-03-GLAVRSQLL-3PD	-0.05270	-0.51187	0.24003	-0.00137

Table D.5: Summary of Mulliken charges at key atoms in QM/MM optimized reactants, transition states, and products across all systems for the cyclopropanation reaction.

Bibliography

[1] Martín Abadi. Tensorflow: learning functions at scale. *ACM SIGPLAN Notices*, 51(9):1–1, September 2016.

[2] Carles Acosta-Silva, Joan Bertran, Vicenç Branchadell, and Antoni Oliva. Kemp elimination reaction catalyzed by electric fields. *ChemPhysChem*, 21(4):295–306, January 2020.

[3] Andrew J. Adamczyk, Jie Cao, Shina C. L. Kamerlin, and Arieh Warshel. Catalysis by dihydrofolate reductase and other enzymes arises from electrostatic preorganization, not conformational motions. *Proceedings of the National Academy of Sciences*, 108(34):14115–14120, August 2011.

[4] Aduragbemi S. Adesina, Katarzyna Świderek, Louis Y. P. Luk, Vicent Moliner, and Rudolf K. Allemann. Electric Field Measurements Reveal the Pivotal Role of Cofactor–Substrate Interaction in Dihydrofolate Reductase Catalysis. *ACS Catalysis*, 10(14):7907–7914, July 2020.

[5] Thomas B. Adler, Gerald Knizia, and Hans-Joachim Werner. A simple and efficient CCSD(T)-F12 approximation. *The Journal of Chemical Physics*, 127(22):221106, 12 2007.

[6] Reinhart Ahlrichs. Efficient evaluation of three-center two-electron integrals over gaussian functions. *Phys. Chem. Chem. Phys.*, 6:5119–5121, 2004.

[7] Reinhart Ahlrichs, Michael Bär, Marco Häser, Hans Horn, and Christoph Kölmel. Electronic structure calculations on workstation computers: The program system turbomole. *Chemical Physics Letters*, 162(3):165–169, 1989.

[8] Reinhart Ahlrichs, Michael Bär, Marco Häser, Hans Horn, and Christoph Kölmel. Electronic structure calculations on workstation computers: The program system turbomole. *Chemical Physics Letters*, 162(3):165–169, October 1989.

[9] Eric A. Althoff, Ling Wang, Lin Jiang, Lars Giger, Jonathan K. Lassila, Zhizhi Wang, Matthew Smith, Sanjay Hari, Peter Kast, Daniel Herschlag, Donald Hilvert, and David Baker. Robust design and optimization of retroaldol enzymes. *Protein Science*, 21(5):717–726, 2012. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/pro.2059.

[10] Frances H. Arnold. Directed Evolution: Bringing New Chemistry to Life. *Angewandte Chemie International Edition*, 57(16):4143–4148, April 2018.

[11] Mojgan Asadi and Arieh Warshel. Analyzing the reaction of orotidine 5-phosphate decarboxylase as a way to examine some key catalytic proposals. *Journal of the American Chemical Society*, 145(2):1334–1341, December 2022.

[12] Mohammad Aziz, Mohamed Hussein, and Musa Gabere. Filtered selection coupled with support vector machines generate a functionally relevant prediction model for colorectal cancer. *OncoTargets and Therapy*, page 3313, June 2016.

[13] Ram Prasad B, Nikolay V. Plotnikov, Jeronimo Lameira, and Arieh Warshel. Quantitative exploration of the molecular origin of the activation of gtpase. *Proceedings of the National Academy of Sciences*, 110(51):20509–20514, November 2013.

[14] R. F. W. Bader. Atoms in molecules. *Accounts of Chemical Research*, 18(1):9–15, January 1985.

[15] R. F. W. Bader. Atoms in molecules. *Accounts of Chemical Research*, 18(1):9–15, 1985.

[16] R. F. W. Bader, P. J. MacDougall, and C. D. H. Lau. Bonded and nonbonded charge concentrations and their relation to molecular geometry and reactivity. *Journal of the American Chemical Society*, 106(6):1594–1605, March 1984.

[17] Richard F. W. Bader and Chérif F. Matta. Bonding to titanium. *Inorganic Chemistry*, 40(22):5603–5611, Oct 2001.

[18] David Baker. An exciting but challenging road ahead for computational enzyme design. *Protein Science*, 19(10):1817–1819, 2010. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/pro.481.

[19] P. Balanarayan, Ritwik Kavathekar, and Shridhar R. Gadre. Electrostatic potential topography for exploring electronic reorganizations in 1,3 dipolar cycloadditions. *The Journal of Physical Chemistry A*, 111(14):2733–2738, March 2007.

[20] Christoph Bannwarth, Sebastian Ehlert, and Stefan Grimme. Gfn2-xtb—an accurate and broadly parametrized self-consistent tight-binding quantum chemical method with multipole electrostatics and density-dependent dispersion contributions. *Journal of Chemical Theory and Computation*, 15(3):1652–1671, February 2019.

[21] Tara K. Bartolec, Xabier Vázquez-Campos, Alexander Norman, Clement Luong, Marcus Johnson, Richard J. Payne, Marc R. Wilkins, Joel P. Mackay, and Jason K. K. Low. Cross-linking mass spectrometry discovers, evaluates, and corroborates structures and protein–protein interactions in the human cell. *Proceedings of the National Academy of Sciences*, 120(17):e2219418120, April 2023.

[22] Ilyes Batatia, Philipp Benner, Yuan Chiang, Alin M. Elena, Dávid P. Kovács, Janosh Riebesell, Xavier R. Advincula, Mark Asta, Matthew Avaylon, William J. Baldwin, Fabian Berger, Noam Bernstein, Arghya Bhowmik, Samuel M. Blau, Vlad Cărare, James P. Darby, Sandip De, Flaviano Della Pia, Volker L. Deringer, Rokas Elijošius, Zakariya El-Machachi, Fabio Falcioni, Edvin Fako, Andrea C. Ferrari, Annalena Genreith-

Schriever, Janine George, Rhys E. A. Goodall, Clare P. Grey, Petr Grigorev, Shuang Han, Will Handley, Hendrik H. Heenen, Kersti Hermansson, Christian Holm, Jad Jaafar, Stephan Hofmann, Konstantin S. Jakob, Hyunwook Jung, Venkat Kapil, Aaron D. Kaplan, Nima Karimitari, James R. Kermode, Namu Kroupa, Jolla Kullgren, Matthew C. Kuner, Domantas Kuryla, Guoda Liepuoniute, Johannes T. Margraf, Ioan-Bogdan Magdău, Angelos Michaelides, J. Harry Moore, Aakash A. Naik, Samuel P. Niblett, Sam Walton Norwood, Niamh O'Neill, Christoph Ortner, Kristin A. Persson, Karsten Reuter, Andrew S. Rosen, Lars L. Schaaf, Christoph Schran, Benjamin X. Shi, Eric Sivonxay, Tamás K. Stenczel, Viktor Svahn, Christopher Sutton, Thomas D. Swinburne, Jules Tilly, Cas van der Oord, Eszter Varga-Umbrich, Tejs Vegge, Martin Vondrák, Yangshuai Wang, William C. Witt, Fabian Zills, and Gábor Csányi. A foundation model for atomistic materials chemistry, 2024.

[23] Ilyes Batatia, Dávid Péter Kovács, Gregor N. C. Simm, Christoph Ortner, and Gábor Csányi. Mace: Higher order equivariant message passing neural networks for fast and accurate force fields, 2023.

[24] Peter W. Battaglia, Jessica B. Hamrick, Victor Bapst, Alvaro Sanchez-Gonzalez, Vinicius Zambaldi, Mateusz Malinowski, Andrea Tacchetti, David Raposo, Adam Santoro, Ryan Faulkner, Caglar Gulcehre, Francis Song, Andrew Ballard, Justin Gilmer, George Dahl, Ashish Vaswani, Kelsey Allen, Charles Nash, Victoria Langston, Chris Dyer, Nicolas Heess, Daan Wierstra, Pushmeet Kohli, Matt Botvinick, Oriol Vinyals, Yujia Li, and Razvan Pascanu. Relational inductive biases, deep learning, and graph networks, 2018.

[25] Simon Batzner, Albert Musaelian, Lixin Sun, Mario Geiger, Jonathan P. Mailoa, Mordechai Kornbluth, Nicola Molinari, Tess E. Smidt, and Boris Kozinsky. E(3)-equivariant graph neural networks for data-efficient and accurate interatomic potentials. *Nature Communications*, 13(1):2453, May 2022.

[26] Axel D. Becke. Density-functional thermochemistry. III. The role of exact exchange. *The Journal of Chemical Physics*, 98(7):5648–5652, April 1993.

[27] Axel D. Becke. Density-functional thermochemistry. III. The role of exact exchange. *The Journal of Chemical Physics*, 98(7):5648–5652, 04 1993.

[28] Axel D. Becke. Density-functional thermochemistry. III. The role of exact exchange. *The Journal of Chemical Physics*, 98(7):5648–5652, April 1993. Publisher: American Institute of Physics.

[29] Axel D. Becke. Density-functional thermochemistry. v. systematic optimization of exchange-correlation functionals. *The Journal of Chemical Physics*, 107(20):8554–8560, November 1997.

[30] Wiktor Beker and W. Andrzej Sokalski. Bottom-up nonempirical approach to reducing search space in enzyme design guided by catalytic fields. *Journal of Chemical Theory and Computation*, 16(5):3420–3429, April 2020.

[31] Letícia M. F. Bertoline, Angélica N. Lima, Jose E. Krieger, and Samantha K. Teixeira. Before and after AlphaFold2: An overview of protein structure prediction. *Frontiers in Bioinformatics*, 3:1120370, February 2023.

[32] Asmit Bhowmick, Sudhir C. Sharma, and Teresa Head-Gordon. The Importance of the Scaffold for *de Novo* Enzymes: A Case Study with Kemp Eliminase. *Journal of the American Chemical Society*, 139(16):5793–5800, April 2017.

[33] Lukas Biewald. Experiment tracking with weights and biases, 2020. Software available from wandb.com.

[34] Christopher M. Bishop. *Pattern recognition and machine learning*. Information science and statistics. Springer, New York, 2006.

[35] Mariana Boiani, Hugo Cerecetto, Mercedes González, and Johann Gasteiger. Modeling anti-trypanosoma cruzi activity of n-oxide containing heterocycles. *Journal of Chemical Information and Modeling*, 48(1):213–219, December 2007.

[36] Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

[37] Leo Breiman. Random Forests. *Machine Learning*, 45(1):5–32, 2001.

[38] Gerold U. Bublitz and Steven G. Boxer. STARK SPECTROSCOPY: Applications in Chemistry, Biology, and Materials Science. *Annual Review of Physical Chemistry*, 48(1):213–242, October 1997.

[39] H. Adrian Bunzel, J.L. Ross Anderson, and Adrian J. Mulholland. Designing better enzymes: Insights from directed evolution. *Current Opinion in Structural Biology*, 67:212–218, April 2021.

[40] Daniel Bím and Anastassia N. Alexandrova. Electrostatic regulation of blue copper sites. *Chemical Science*, 12(34):11406–11413, 2021.

[41] Daniel Bím and Anastassia N. Alexandrova. Local Electric Fields As a Natural Switch of Heme-Iron Protein Reactivity. *ACS Catalysis*, 11(11):6534–6546, June 2021. Publisher: American Chemical Society.

[42] Bartosz Błasiak, Andrew W. Ritchie, Lauren J. Webb, and Minhaeng Cho. Vibrational solvatochromism of nitrile infrared probes: beyond the vibrational Stark dipole approach. *Physical Chemistry Chemical Physics*, 18(27):18094–18111, July 2016. Publisher: The Royal Society of Chemistry.

[43] Yuanxin Cao, Sam Hay, and Sam P. De Visser. An Active Site Tyr Residue Guides the Regioselectivity of Lysine Hydroxylation by Nonheme Iron Lysine-4-hydroxylase Enzymes through Proton-Coupled Electron Transfer. *Journal of the American Chemical Society*, 146(17):11726–11739, May 2024.

[44] Pablo Carpio-Martínez, José E. Barquera-Lozada, Angel Martín Pendás, and Fernando Cortés-Guzmán. Laplacian of the hamiltonian kinetic energy density as an indicator of binding and weak interactions. *ChemPhysChem*, 21(3):194–203, December 2019.

[45] Guillem Casadevall, Cristina Duran, and Sílvia Osuna. AlphaFold2 and Deep Learning for Elucidating Enzyme Conformational Flexibility and Its Application for Design. *JACS Au*, 3(6):1554–1562, June 2023.

[46] D.A. Case, I.Y Ben-Shalom, S.R. Brozell, D.S. Cerutti, T.E. Cheatham, III Cruzeiro, V.W.D., T.A. Darden, R.E. Duke, D. Ghoreishi, M.K. Gilson, H. Gohlke, A.W. Goetz, D. Greene, R Harris, N. Homeyer, Y. Huang, S. Izadi, A. Kovalenko, T. Kurtzman, T.S. Lee, S. LeGrand, P. Li, C. Lin, J. Liu, T. Luchko, R. Luo, D.J. Mermelstein, K.M. Merz, Y. Miao, G. Monard, C. Nguyen, H. Nguyen, I. Omelyan, A. Onufriev, F. Pan, R. Qi, D.R. Roe, A. Roitberg, C. Sagui, S. Schott-Verdugo, J. Shen, C.L. Simmerling, J. Smith, R. Salomon-Ferrer, J. Swails, R.C. Walker, J. Wang, H. Wei, R.M. Wolf, X. Wu, L. Xiao, D.M. York, and P.A. Kollman. AMBER 2018.

[47] Jeng-Da Chai and Martin Head-Gordon. Long-range corrected hybrid density functionals with damped atom–atom dispersion corrections. *Physical Chemistry Chemical Physics*, 10(44):6615, 2008.

[48] H. C. Stephen Chan, Lu Pan, Yi Li, and Shuguang Yuan. Rationalization of stereoselectivity in enzyme reactions. *WIREs Computational Molecular Science*, 9(4):e1403, July 2019.

[49] Wenbing Chang, Yinglai Liu, Yiyong Xiao, Xinglong Yuan, Xingxing Xu, Siyue Zhang, and Shenghan Zhou. A machine-learning-based prediction method for hypertension outcomes based on medical data. *Diagnostics*, 9(4):178, November 2019.

[50] Lowik Chanussot, Abhishek Das, Siddharth Goyal, Thibaut Lavril, Muhammed Shuaibi, Morgane Riviere, Kevin Tran, Javier Heras-Domingo, Caleb Ho, Weihua Hu, Aini

Palizhati, Anuroop Sriram, Brandon Wood, Junwoong Yoon, Devi Parikh, C. Lawrence Zitnick, and Zachary Ulissi. Open catalyst 2020 (oc20) dataset and community challenges. *ACS Catalysis*, 11(10):6059–6072, May 2021.

[51] Shobhit S. Chaturvedi, Daniel Bím, Christo Z. Christov, and Anastassia N. Alexandrova. From random to rational: improving enzyme design through electric fields, second coordination sphere interactions, and conformational dynamics. *Chemical Science*, 14(40):10997–11011, 2023.

[52] Shobhit S. Chaturvedi, Simahudeen Bathir Jaber Sathik Rifayee, Rajeev Ramanan, Joel A. Rankin, Jian Hu, Robert P. Hausinger, and Christo Z. Christov. Can an external electric field switch between ethylene formation and l -arginine hydroxylation in the ethylene forming enzyme? *Physical Chemistry Chemical Physics*, 25(19):13772–13783, 2023.

[53] Shobhit S. Chaturvedi, Simahudeen Bathir Jaber Sathik Rifayee, Sodiq O. Waheed, Jon Wildey, Cait Warner, Christopher J. Schofield, Tatyana G. Karabencheva-Christova, and Christo Z. Christov. Can Second Coordination Sphere and Long-Range Interactions Modulate Hydrogen Atom Transfer in a Non-Heme Fe(II)-Dependent Histone Demethylase? *JACS Au*, 2(9):2169–2186, September 2022.

[54] Shobhit S. Chaturvedi, Santiago Vargas, Pujan Ajmera, and Anastassia N. Alexandrova. Directed Evolution of Protoglobin Optimizes the Enzyme Electric Field. *Journal of the American Chemical Society*, page jacs.4c03914, June 2024.

[55] U. A. Chaudry and P. L. A. Popelier. Estimation of pka using quantum topological molecular similarity descriptors: application to carboxylic acids, anilines and phenols. *The Journal of Organic Chemistry*, 69(2):233–241, 2004. PMID: 14725434.

[56] Chao Chen. Using Random Forest to Learn Imbalanced Data. *University of California, Berkeley*, 110:1–12, 2004.

[57] Chi Chen, Weike Ye, Yunxing Zuo, Chen Zheng, and Shyue Ping Ong. Graph networks as a universal machine learning framework for molecules and crystals. *Chemistry of Materials*, 31(9):3564–3572, April 2019.

[58] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16. ACM, August 2016.

[59] Martyna Chojnacka, Mikolaj Feliks, Wiktor Beker, and W. Andrzej Sokalski. Predicting substituent effects on activation energy changes by static catalytic fields. *Journal of Molecular Modeling*, 24(1), December 2017.

[60] Ove Christiansen, Henrik Koch, and Poul Jørgensen. The second-order approximate coupled cluster singles and doubles model cc2. *Chemical Physics Letters*, 243(5):409–418, 1995.

[61] David E. Clark. Rapid calculation of polar molecular surface area and its application to the prediction of transport phenomena. 1. prediction of intestinal absorption. *Journal of Pharmaceutical Sciences*, 88(8):807–814, August 1999.

[62] Connor W. Coley, Regina Barzilay, William H. Green, Tommi S. Jaakkola, and Klavs F. Jensen. Convolutional embedding of attributed molecular graphs for physical property prediction. *Journal of Chemical Information and Modeling*, 57(8):1757–1772, 2017. PMID: 28696688.

[63] Marina Corbella, Gaspar P. Pinto, and Shina C. L. Kamerlin. Loop dynamics and the evolution of enzyme activity. *Nature Reviews Chemistry*, pages 1–12, May 2023. Publisher: Nature Publishing Group.

[64] Fernando Cortés-Guzmán, Juan I. Rodríguez, and James S.M. Anderson. Chapter 1 - introduction to qtaim and beyond. In Juan I. Rodríguez, Fernando Cortés-Guzmán,

and James S.M. Anderson, editors, *Advances in Quantum Chemical Topology Beyond QTAIM*, pages 1–19. Elsevier, 2023.

[65] Fernando Cortés-Guzmán, Juan I. Rodríguez, and James S.M. Anderson. *Introduction to QTAIM and beyond*, page 1–19. Elsevier, 2023.

[66] Christopher J. Cramer. *Essentials of computational chemistry: theories and models*. Wiley, Chichester, West Sussex, England ; Hoboken, NJ, 2nd ed edition, 2004.

[67] Christopher J. Cramer, George R. Famini, and Alfred H. Lowrey. Use of calculated quantum chemical properties as surrogates for solvatochromic parameters in structure-activity relationships. *Accounts of Chemical Research*, 26(11):599–605, November 1993.

[68] Rebecca Crawshaw, Amy E. Crossley, Linus Johannissen, Ashleigh J. Burke, Sam Hay, Colin Levy, David Baker, Sarah L. Lovelock, and Anthony P. Green. Engineering an efficient and enantioselective enzyme for the Morita–Baylis–Hillman reaction. *Nature Chemistry*, 14(3):313–320, March 2022.

[69] Emma Danelius, Nicholas J. Porter, Johan Unge, Frances H. Arnold, and Tamir Gonen. MicroED Structure of a Protoglobin Reactive Carbene Intermediate. *Journal of the American Chemical Society*, 145(13):7159–7165, April 2023.

[70] E. De Brito Sá, A. Rimola, L. Rodríguez-Santiago, M. Sodupe, and X. Solans-Monfort. Reactivity of Metal Carbenes with Olefins: Theoretical Insights on the Carbene Electronic Structure and Cyclopropanation Reaction Mechanism. *The Journal of Physical Chemistry A*, 122(6):1702–1712, February 2018.

[71] Alexander G. de G. Matthews, Mark van der Wilk, Tom Nickson, Keisuke Fujii, Alexis Boukouvalas, Pablo León-Villagrá, Zoubin Ghahramani, and James Hensman. Gpflow: A gaussian process library using tensorflow. *Journal of Machine Learning Research*, 18(40):1–6, 2017.

[72] R. Ditchfield, W. J. Hehre, and J. A. Pople. Self-Consistent Molecular-Orbital Methods. IX. An Extended Gaussian-Type Basis for Molecular-Orbital Studies of Organic Molecules. *The Journal of Chemical Physics*, 54(2):724–728, 01 1971.

[73] Kshatresh Dutta Dubey, Thijs Stuyver, and Sason Shaik. Local Electric Fields: From Enzyme Catalysis to Synthetic Catalyst Design. *The Journal of Physical Chemistry B*, 126(49):10285–10294, December 2022.

[74] Jerome Eberhardt, Diogo Santos-Martins, Andreas F. Tillack, and Stefano Forli. Autodock vina 1.2.0: New docking methods, expanded force field, and python bindings. *Journal of Chemical Information and Modeling*, 61(8):3891–3898, July 2021.

[75] Elan Z. Eisenmesser, Oscar Millet, Wladimir Labeikovsky, Dmitry M. Korzhnev, Magnus Wolf-Watz, Daryl A. Bosco, Jack J. Skalicky, Lewis E. Kay, and Dorothee Kern. Intrinsic dynamics of an enzyme underlies catalysis. *Nature*, 438(7064):117–121, November 2005. Number: 7064 Publisher: Nature Publishing Group.

[76] Thomas Engel and Johann Gasteiger, editors. *Chemoinformatics*. Wiley-VCH Verlag, Weinheim, Germany, August 2018.

[77] Evgeny Epifanovsky, Andrew T B Gilbert, Xintian Feng, Joonho Lee, Yuezhi Mao, Narbe Mardirossian, Pavel Pokhilko, Alec F White, Marc P Coons, Adrian L Dempwolff, Zhengting Gan, Diptarka Hait, Paul R Horn, Leif D Jacobson, Ilya Kaliman, Jörg Kussmann, Adrian W Lange, Ka Un Lao, Daniel S Levine, Jie Liu, Simon C McKenzie, Adrian F Morrison, Kaushik D Nanda, Felix Plasser, Dirk R Rehn, Marta L Vidal, Zhi-Qiang You, Ying Zhu, Bushra Alam, Benjamin J Albrecht, Abdulrahman Aldossary, Ethan Alguire, Josefine H Andersen, Vishikh Athavale, Dennis Barton, Khadiza Begam, Andrew Behn, Nicole Bellonzi, Yves A Bernard, Eric J Berquist, Hugh G A Burton, Abel Carreras, Kevin Carter-Fenk, Romit Chakraborty, Alan D Chien, Kristina D Closser, Vale Cofer-Shabica, Saswata Dasgupta, Marc de Wergifosse, Jia Deng, Michael

Diedenhofen, Hainam Do, Sebastian Ehlert, Po-Tung Fang, Shervin Fatehi, Qingguo Feng, Triet Friedhoff, James Gayvert, Qinghui Ge, Gergely Gidofalvi, Matthew Goldey, Joe Gomes, Cristina E González-Espinoza, Sahil Gulania, Anastasia O Gunina, Magnus W D Hanson-Heine, Phillip H P Harbach, Andreas Hauser, Michael F Herbst, Mario Hernández Vera, Manuel Hodecker, Zachary C Holden, Shannon Houck, Xunkun Huang, Kerwin Hui, Bang C Huynh, Maxim Ivanov, Ádám Jász, Hyunjun Ji, Hanjie Jiang, Benjamin Kaduk, Sven Kähler, Kirill Khistyaev, Jaehoon Kim, Gergely Kis, Phil Klunzinger, Zsuzsanna Koczor-Benda, Joong Hoon Koh, Dimitri Kosenkov, Laura Koulias, Tim Kowalczyk, Caroline M Krauter, Karl Kue, Alexander Kunitsa, Thomas Kus, István Ladjánszki, Arie Landau, Keith V Lawler, Daniel Lefrancois, Susi Lehtola, Run R Li, Yi-Pei Li, Jiashu Liang, Marcus Liebenthal, Hung-Hsuan Lin, You-Sheng Lin, Fenglai Liu, Kuan-Yu Liu, Matthias Loipersberger, Arne Luenser, Aaditya Manjanath, Prashant Manohar, Erum Mansoor, Sam F Manzer, Shan-Ping Mao, Aleksandr V Marenich, Thomas Markovich, Stephen Mason, Simon A Maurer, Peter F McLaughlin, Maximilian F S J Menger, Jan-Michael Mewes, Stefanie A Mewes, Pierpaolo Morgante, J Wayne Mullinax, Katherine J Oosterbaan, Garrette Paran, Alexander C Paul, Suranjan K Paul, Fabijan Pavošević, Zheng Pei, Stefan Prager, Emil I Proynov, Ádám Rák, Eloy Ramos-Cordoba, Bhaskar Rana, Alan E Rask, Adam Rettig, Ryan M Richard, Fazle Rob, Elliot Rossomme, Tarek Scheele, Maximilian Scheurer, Matthias Schneider, Nickolai Sergueev, Shaama M Sharada, Wojciech Skomorowski, David W Small, Christopher J Stein, Yu-Chuan Su, Eric J Sundstrom, Zhen Tao, Jonathan Thirman, Gábor J Tornai, Takashi Tsuchimochi, Norm M Tubman, Srimukh Prasad Veccham, Oleg Vydrov, Jan Wenzel, Jon Witte, Atsushi Yamada, Kun Yao, Sina Yeganeh, Shane R Yost, Alexander Zech, Igor Ying Zhang, Xing Zhang, Yu Zhang, Dmitry Zuev, Alán Aspuru-Guzik, Alexis T Bell, Nicholas A Besley, Ksenia B Bravaya, Bernard R Brooks, David Casanova, Jeng-Da Chai, Sonia Coriani, Christopher J Cramer, György Cserey, A Eugene DePrince, 3rd, Robert A DiStasio, Jr, Andreas Dreuw, Barry D

Dunietz, Thomas R Furlani, William A Goddard, 3rd, Sharon Hammes-Schiffer, Teresa Head-Gordon, Warren J Hehre, Chao-Ping Hsu, Thomas-C Jagau, Yousung Jung, Andreas Klamt, Jing Kong, Daniel S Lambrecht, Wanzhen Liang, Nicholas J Mayhall, C William McCurdy, Jeffrey B Neaton, Christian Ochsenfeld, John A Parkhill, Roberto Peverati, Vitaly A Rassolov, Yihan Shao, Lyudmila V Slipchenko, Tim Stauch, Ryan P Steele, Joseph E Subotnik, Alex J W Thom, Alexandre Tkatchenko, Donald G Truhlar, Troy Van Voorhis, Tomasz A Wesolowski, K Birgitta Whaley, H Lee Woodcock, 3rd, Paul M Zimmerman, Shirin Faraji, Peter M W Gill, Martin Head-Gordon, John M Herbert, and Anna I Krylov. Software for the frontiers of quantum chemistry: An overview of developments in the Q-Chem 5 package. *J. Chem. Phys.*, 155(8):084801, August 2021.

[78] Peter Ertl, Bernhard Rohde, and Paul Selzer. Fast calculation of molecular polar surface area as a sum of fragment-based contributions and its application to the prediction of drug transport properties. *Journal of Medicinal Chemistry*, 43(20):3714–3717, September 2000.

[79] DanielH. Ess, GavinO. Jones, and K.N. Houk. Conceptual, qualitative, and quantitative theories of 1,3-dipolar and diels–alder cycloadditions used in synthesis. *Advanced Synthesis & Catalysis*, 348(16-17):2337–2361, 2006.

[80] Aaron T. Fafarman, Paul A. Sigala, Daniel Herschlag, and Steven G. Boxer. Decomposition of Vibrational Shifts of Nitriles into Electrostatic and Hydrogen-Bonding Effects. *Journal of the American Chemical Society*, 132(37):12811–12813, September 2010.

[81] Louis J. Farrugia, Cameron Evans, Dieter Lentz, and Max Roemer. The qtaim approach to chemical bonding between transition metals and carbocyclic rings: A combined experimental and theoretical study of (η5-c5h5)mn(co)3, (η6-c6h6)cr(co)3, and (e)-{(η5-c5h4)cfcf(η5-c5h4)}(η5-c5h5)2fe2. *Journal of the American Chemical Society*, 131(3):1251–1268, Jan 2009.

[82] Isabella Feierberg and Johan Åqvist. The Catalytic Power of Ketosteroid Isomerase Investigated by Computer Simulation. *Biochemistry*, 41(52):15728–15735, December 2002.

[83] David Ferro-Costas, Ángel Martín Pendás, Leticia González, and Ricardo A. Mosquera. Beyond the molecular orbital conception of electronically excited states through the quantum theory of atoms in molecules. *Phys. Chem. Chem. Phys.*, 16:9249–9258, 2014.

[84] Vittorio Fortino, Pia Kinaret, Nanna Fyhrquist, Harri Alenius, and Dario Greco. A robust and accurate method for feature selection and prioritization from multi-class omics data. *PLoS ONE*, 9(9):e107801, September 2014.

[85] Brendan J. Frey and Delbert Dueck. Clustering by passing messages between data points. *Science*, 315(5814):972–976, February 2007.

[86] Stephen D. Fried, Sayan Bagchi, and Steven G. Boxer. Extreme electric fields power catalysis in the active site of ketosteroid isomerase. *Science*, 346(6216):1510–1514, December 2014.

[87] Stephen D. Fried, Sayan Bagchi, and Steven G. Boxer. Extreme electric fields power catalysis in the active site of ketosteroid isomerase. *Science*, 346(6216):1510–1514, December 2014.

[88] Stephen D. Fried and Steven G. Boxer. Measuring Electric Fields and Noncovalent Interactions Using the Vibrational Stark Effect. *Accounts of Chemical Research*, 48(4):998–1006, April 2015.

[89] Stephen D Fried and Steven G Boxer. Electric Fields and Enzyme Catalysis. 2017.

[90] Pascal Friederich, Gabriel dos Passos Gomes, Riccardo De Bin, Alán Aspuru-Guzik, and David Balcells. Machine learning dihydrogen activation in the chemical space surrounding vaska's complex. *Chem. Sci.*, 11:4584–4601, 2020.

[91] M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, G. Scalmani, V. Barone, G. A. Petersson, H. Nakatsuji, X. Li, M. Caricato, A. V. Marenich, J. Bloino, B. G. Janesko, R. Gomperts, B. Mennucci, H. P. Hratchian, J. V. Ortiz, A. F. Izmaylov, J. L. Sonnenberg, D. Williams-Young, F. Ding, F. Lipparini, F. Egidi, J. Goings, B. Peng, A. Petrone, T. Henderson, D. Ranasinghe, V. G. Zakrzewski, J. Gao, N. Rega, G. Zheng, W. Liang, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, T. Vreven, K. Throssell, J. A. Montgomery, Jr., J. E. Peralta, F. Ogliaro, M. J. Bearpark, J. J. Heyd, E. N. Brothers, K. N. Kudin, V. N. Staroverov, T. A. Keith, R. Kobayashi, J. Normand, K. Raghavachari, A. P. Rendell, J. C. Burant, S. S. Iyengar, J. Tomasi, M. Cossi, J. M. Millam, M. Klene, C. Adamo, R. Cammi, J. W. Ochterski, R. L. Martin, K. Morokuma, O. Farkas, J. B. Foresman, and D. J. Fox. Gaussian˜16 Revision C.01, 2016. Gaussian Inc. Wallingford CT.

[92] Maria P. Frushicheva, Jie Cao, Zhen T. Chu, and Arieh Warshel. Exploring challenges in rational enzyme design by simulating the catalysis in artificial kemp eliminase. *Proceedings of the National Academy of Sciences*, 107(39):16869–16874, September 2010.

[93] Maria P Frushicheva, Matthew JL Mills, Patrick Schopf, Manoj K Singh, Ram B Prasad, and Arieh Warshel. Computer aided enzyme design and catalytic concepts. *Current Opinion in Chemical Biology*, 21:56–62, August 2014.

[94] Maria P. Frushicheva and Arieh Warshel. Towards quantitative computer-aided studies of enzymatic enantioselectivity: The case of candida antarctica lipase a. *ChemBioChem*, 13(2):215–223, December 2011.

[95] Jack Fuller, Tim R. Wilson, Mark E. Eberhart, and Anastassia N. Alexandrova. Charge Density in Enzyme Active Site as a Descriptor of Electrostatic Preorganization. *Journal of Chemical Information and Modeling*, 59(5):2367–2373, May 2019.

[96] Jack Fuller, Tim R. Wilson, Mark E. Eberhart, and Anastassia N. Alexandrova. Charge Density in Enzyme Active Site as a Descriptor of Electrostatic Preorganization. *Journal of Chemical Information and Modeling*, 59(5):2367–2373, May 2019.

[97] Jack III Fuller, Tim R. Wilson, Mark E. Eberhart, and Anastassia N. Alexandrova. Charge density in enzyme active site as a descriptor of electrostatic preorganization. *Journal of Chemical Information and Modeling*, 59(5):2367–2373, 2019. PMID: 30793899.

[98] Monika Fuxreiter and Letif Mones. The empirical valence bond approach as a tool for designing artificial catalysts, February 2017.

[99] Shridhar R. Gadre, Cherumuttathu H. Suresh, and Neetha Mohan. Electrostatic potential topology for probing molecular structure, bonding and reactivity. *Molecules*, 26(11):3289, May 2021.

[100] Miguel Gallegos, José Manuel Guevara-Vela, and Ángel Martín Pendás. NNAIMQ: A neural network model for predicting QTAIM charges. *The Journal of Chemical Physics*, 156(1):014112, 01 2022.

[101] Johannes Gasteiger, Janek Groß, and Stephan Günnemann. Directional message passing for molecular graphs, 2022.

[102] Aurelien Geron. *Hands-on machine learning with scikit-learn, keras, and TensorFlow*. O'Reilly Media, Sebastopol, CA, 2 edition, October 2019.

[103] Lars Giger, Sami Caner, Richard Obexer, Peter Kast, David Baker, Nenad Ban, and Donald Hilvert. Evolution of a designed retro-aldolase leads to complete active site remodeling. *Nature Chemical Biology*, 9(8):494–498, August 2013.

[104] Justin Gilmer, Samuel S. Schoenholz, Patrick F. Riley, Oriol Vinyals, and George E. Dahl. Neural message passing for quantum chemistry, 2017.

[105] Ruth Gordillo, Travis Dudding, Christopher D. Anderson, and K. N. Houk. Hydrogen bonding catalysis operates by charge stabilization in highly polar dielsalder reactions. *Organic Letters*, 9(3):501–503, 2007. PMID: 17249797.

[106] Ruth Gordillo and K. N. Houk. Origins of stereoselectivity in dielsalder cycloadditions catalyzed by chiral imidazolidinones. *Journal of the American Chemical Society*, 128(11):3543–3553, 2006. PMID: 16536527.

[107] Pablo M. Granitto, Cesare Furlanello, Franco Biasioli, and Flavia Gasperi. Recursive feature elimination with random forest for ptr-ms analysis of agroindustrial products. *Chemometrics and Intelligent Laboratory Systems*, 83:83–90, 2006.

[108] Stefan Grimme, Jens Antony, Stephan Ehrlich, and Helge Krieg. A consistent and accurate ab initio parametrization of density functional dispersion correction (DFT-D) for the 94 elements H-Pu. *The Journal of Chemical Physics*, 132(15):154104, 04 2010.

[109] Stefan Grimme, Jens Antony, Stephan Ehrlich, and Helge Krieg. A consistent and accurate ab initio parametrization of density functional dispersion correction (dft-d) for the 94 elements h-pu. *The Journal of Chemical Physics*, 132(15):154104, Apr 2010.

[110] Stefan Grimme, Christoph Bannwarth, and Philip Shushkov. A robust and accurate tight-binding quantum chemical method for structures, vibrational frequencies, and noncovalent interactions of large molecular systems parametrized for all spd-block elements (z = 1–86). *Journal of Chemical Theory and Computation*, 13(5):1989–2009, April 2017.

[111] Yuri Grin, Andreas Savin, and Bernard Silvi. The elf perspective of chemical bonding, May 2014.

[112] John Guckenheimer. Book review: Stabilité structurelle et morphogénèse, essai d'une théorie générale des modèles. *Bulletin of the American Mathematical Society*, 79(5):878–891, September 1973.

[113] Rishabh Debraj Guha, Santiago Vargas, Evan Walter Clark Spotte-Smith, Alex R Epstein, Maxwell Christopher Venetos, Mingjian Wen, Ryan Kingsbury, Samuel M Blau, and Kristin Persson. HEPOM: A predictive framework for accelerated hydrolysis energy predictions of organic molecules. In *AI for Accelerated Materials Design - NeurIPS 2023 Workshop*, 2023.

[114] Yang Guo, Christoph Riplinger, Ute Becker, Dimitrios G. Liakos, Yury Minenkov, Luigi Cavallo, and Frank Neese. Communication: An improved linear scaling perturbative triples correction for the domain based local pair-natural orbital based singles and doubles coupled cluster method [DLPNO-CCSD(T)]. *The Journal of Chemical Physics*, 148(1):011101, January 2018.

[115] William L. Hamilton. Graph representation learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, 14(3):1–159.

[116] Supa Hannongbua, Kanda Nivesanond, Luckhana Lawtrakul, Pornpan Pungpo, and Peter Wolschann. 3d-quantitative structureactivity relationships of hept derivatives as hiv-1 reverse transcriptase inhibitors, based on ab initio calculations. *Journal of Chemical Information and Computer Sciences*, 41(3):848–855, April 2001.

[117] Philip Hanoian, C. Tony Liu, Sharon Hammes-Schiffer, and Stephen Benkovic. Perspectives on Electrostatics and Conformational Motions in Enzyme Catalysis. *Accounts of Chemical Research*, 48(2):482–489, February 2015. Publisher: American Chemical Society.

[118] Amy E. Hayden and K. N. Houk. Transition state distortion energies correlate with activation energies of 1,4-dihydrogenations and dielsalder cycloadditions of aromatic molecules. *Journal of the American Chemical Society*, 131(11):4084–4089, 2009. PMID: 19256544.

[119] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, June 2016.

[120] Tim Head, Gilles Louppe MechCoder, Iaroslav Shcherbatyi, et al. scikit-optimize/scikit-optimize: v0. 5.2. *Version v0*, 5, 2018.

[121] W. J. Hehre, R. Ditchfield, and J. A. Pople. Self—Consistent Molecular Orbital Methods. XII. Further Extensions of Gaussian—Type Basis Sets for Use in Molecular Orbital Studies of Organic Molecules. *The Journal of Chemical Physics*, 56(5):2257–2261, 03 1972.

[122] Esther Heid, Kevin P. Greenman, Yunsie Chung, Shih-Cheng Li, David E. Graff, Florence H. Vermeire, Haoyang Wu, William H. Green, and Charles J. McGill. Chemprop: A machine learning package for chemical property prediction. *Journal of Chemical Information and Modeling*, 64(1):9–17, December 2023.

[123] Arnim Hellweg, Sarah A. Grün, and Christof Hättig. Benchmarking the performance of spin-component scaled cc2 in ground and electronically excited states. *Phys. Chem. Chem. Phys.*, 10:4119–4127, 2008.

[124] Arnim Hellweg and Dmitrij Rappoport. Development of new auxiliary basis functions of the karlsruhe segmented contracted basis sets including diffuse basis functions (def2-svpd, def2-tzvppd, and def2-qvppd) for ri-mp2 and ri-cc calculations. *Physical Chemistry Chemical Physics*, 17(2):1010–1017, 2015.

[125] Markus C Hemmer, Valentin Steinhauer, and Johann Gasteiger. Deriving the 3d structure of organic molecules from their infrared spectra. *Vibrational Spectroscopy*, 19(1):151–164, February 1999.

[126] Matthew R. Hennefarth and Anastassia N. Alexandrova. Direct Look at the Electric Field in Ketosteroid Isomerase and Its Variants. *ACS Catalysis*, 10(17):9915–9924, September 2020.

[127] Matthew R. Hennefarth and Anastassia N. Alexandrova. Direct Look at the Electric Field in Ketosteroid Isomerase and Its Variants. *ACS Catalysis*, 10(17):9915–9924, September 2020.

[128] Matthew R. Hennefarth and Anastassia N. Alexandrova. Heterogeneous intramolecular electric field as a descriptor of diels–alder reactivity. *The Journal of Physical Chemistry A*, 125(5):1289–1298, 2021. PMID: 33523664.

[129] P. Hohenberg and W. Kohn. Inhomogeneous electron gas. *Phys. Rev.*, 136:B864–B871, Nov 1964.

[130] Tingjun Hou, Junmei Wang, Youyong Li, and Wei Wang. Assessing the Performance of the MM/PBSA and MM/GBSA Methods. 1. The Accuracy of Binding Free Energy Calculations Based on Molecular Dynamics Simulations. *Journal of Chemical Information and Modeling*, 51(1):69–82, January 2011.

[131] Hrant P. Hratchian and H. Bernhard Schlegel. Chapter 10 - finding minima, transition states, and following reaction pathways on ab initio potential energy surfaces. In Clifford E. Dykstra, Gernot Frenking, Kwang S. Kim, and Gustavo E. Scuseria, editors, *Theory and Applications of Computational Chemistry*, pages 195–249. Elsevier, Amsterdam, 2005.

[132] Po-Ssu Huang, Scott E. Boyken, and David Baker. The coming of age of de novo protein design. *Nature*, 537(7620):320–327, September 2016. Number: 7620 Publisher: Nature Publishing Group.

[133] Ruili Huang, Menghang Xia, Dac-Trung Nguyen, Tongan Zhao, Srilatha Sakamuru, Jinghua Zhao, Sampada A. Shahane, Anna Rossoshek, and Anton Simeonov.

Tox21challenge to build predictive models of nuclear receptor and stress response pathways as mediated by exposure to environmental chemicals and drugs. *Frontiers in Environmental Science*, 3:1, 2016. Original Research.

[134] Geir Villy Isaksen, Kathrin Helen Hopmann, Johan Åqvist, and Bjørn Olav Brandsdal. Computer Simulations Reveal Substrate Specificity of Glycosidic Bond Cleavage in Native and Mutant Human Purine Nucleoside Phosphorylase. *Biochemistry*, 55(14):2153–2162, April 2016.

[135] Clemens Isert, Kenneth Atz, Sereina Riniker, and Gisbert Schneider. Exploring protein–ligand binding affinity prediction with electron density-based geometric deep learning. *RSC Adv.*, 14:4492–4502, 2024.

[136] Saeed Izadi, Ramu Anandakrishnan, and Alexey V. Onufriev. Building Water Models: A Different Approach. *The Journal of Physical Chemistry Letters*, 5(21):3863–3871, November 2014.

[137] Hira Jabeen, Michael Beer, James Spencer, Marc W. Van Der Kamp, H. Adrian Bunzel, and Adrian J. Mulholland. Electric Fields Are a Key Determinant of Carbapenemase Activity in Class A -Lactamases. *ACS Catalysis*, 14(9):7166–7172, May 2024.

[138] Natalie C. James, Joann M. Um, Anne B. Padias, H. K. Jr. Hall, and K. N. Houk. Computational investigation of the competition between the concerted diels–alder reaction and formation of diradicals in reactions of acrylonitrile with nonpolar dienes. *The Journal of Organic Chemistry*, 78(13):6582–6592, 2013. PMID: 23758325.

[139] Jon Paul Janet and Heather J. Kulik. Resolving transition metal chemical space: Feature selection for machine learning and structure–property relationships. *The Journal of Physical Chemistry A*, 121(46):8939–8954, November 2017.

[140] Zhe Ji and Steven G. Boxer. -Lactamases Evolve against Antibiotics by Acquiring Large Active-Site Electric Fields. *Journal of the American Chemical Society*, 144(48):22289–22294, December 2022.

[141] Ryosuke Jinnouchi and Ryoji Asahi. Predicting catalytic activity of nanoparticles by a dft-aided machine-learning algorithm. *The Journal of Physical Chemistry Letters*, 8(17):4279–4283, 2017. PMID: 28837771.

[142] Erin R. Johnson, Shahar Keinan, Paula Mori-Sánchez, Julia Contreras-García, Aron J. Cohen, and Weitao Yang. Revealing noncovalent interactions. *Journal of the American Chemical Society*, 132(18):6498–6506, April 2010.

[143] Gavin O. Jones, Vildan A. Guner, and K. N. Houk. Dielsalder reactions of cyclopentadiene and 9,10-dimethylanthracene with cyanoalkenes: the performance of density functional theory and hartreefock calculations for the prediction of substituent effects. *The Journal of Physical Chemistry A*, 110(4):1216–1224, 2006. PMID: 16435782.

[144] Travis E. Jones, Mark E. Eberhart, Scott Imlay, and Craig Mackey. Bond bundles and the origins of functionality. *The Journal of Physical Chemistry A*, 115(45):12582–12585, 2011. PMID: 21809887.

[145] Shina C. L. Kamerlin and Arieh Warshel. At the dawn of the 21st century: Is dynamics the missing link for understanding enzyme catalysis? *Proteins: Structure, Function, and Bioinformatics*, 78(6):1339–1375, 2010. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/prot.22654.

[146] Shina C.L. Kamerlin, Pankaz K. Sharma, Zhen T. Chu, and Arieh Warshel. Ketosteroid isomerase provides further support for the idea that enzymes work by electrostatic preorganization. *Proceedings of the National Academy of Sciences*, 107(9):4075–4080, March 2010.

[147] J. Kaplan and W. F. DeGrado. De novo design of catalytic proteins. *Proceedings of the National Academy of Sciences*, 101(32):11566–11570, August 2004.

[148] Vandana Kardam and Kshatresh Dutta Dubey. Tyr 118-Mediated Electron Transfer is Key to the Chlorite Decomposition in Heme-Dependent Chlorite Dismutase. *Inorganic Chemistry*, 62(44):18322–18330, November 2023.

[149] Todd A. Keith. Aimall (version 19.10.12), 2019. TK Gristmill Software, Overland Park KS, USA.

[150] Rahul L. Khade, Wenchao Fan, Yan Ling, Liu Yang, Eric Oldfield, and Yong Zhang. Iron Porphyrin Carbenes as Catalytic Intermediates: Structures, Mössbauer and NMR Spectroscopic Properties, and Bonding. *Angewandte Chemie International Edition*, 53(29):7574–7578, July 2014.

[151] Rahul L. Khade and Yong Zhang. CH Insertions by Iron Porphyrin Carbene: Basic Mechanism and Origin of Substrate Selectivity. *Chemistry – A European Journal*, 23(70):17654–17658, December 2017.

[152] Olga Khersonsky, Gert Kiss, Daniela Röthlisberger, Orly Dym, Shira Albeck, Kendall N. Houk, David Baker, and Dan S. Tawfik. Bridging the gaps in design methodologies by evolutionary optimization of the stability and proficiency of designed Kemp eliminase KE59. *Proceedings of the National Academy of Sciences*, 109(26):10358–10363, June 2012. Publisher: Proceedings of the National Academy of Sciences.

[153] Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks, 2017.

[154] Dmitri B. Kireev, Jacques R. Chrétien, David S. Grierson, and Claude Monneret. A 3d qsar study of a series of hept analogues: the influence of conformational mobility on hiv-1 reverse transcriptase inhibition. *Journal of Medicinal Chemistry*, 40(26):4257–4264, December 1997.

[155] Gert Kiss, Nihan Çelebi Ölçüm, Rocco Moretti, David Baker, and K. N. Houk. Computational Enzyme Design. *Angewandte Chemie International Edition*, 52(22):5700–5725, 2013. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/anie.201204077.

[156] A. Klamt and G. Schüürmann. Cosmo: a new approach to dielectric screening in solvents with explicit expressions for the screening energy and its gradient. *J. Chem. Soc., Perkin Trans. 2*, pages 799–805, 1993.

[157] Bernhard Knapp, Luis Ospina, and Charlotte M. Deane. Avoiding False Positive Conclusions in Molecular Simulation: The Importance of Replicas. *Journal of Chemical Theory and Computation*, 14(12):6127–6138, December 2018. Publisher: American Chemical Society.

[158] Erika H. Knoerr and M. E. Eberhart. Toward a density-based representation of reactivity: sn2 reaction. *The Journal of Physical Chemistry A*, 105(5):880–884, 2001.

[159] Bryan E. Kohler and Jörg C. Woehl. Measuring internal electric fields with atomic resolution. *The Journal of Chemical Physics*, 102(20):7773–7781, May 1995.

[160] Agnieszka Krzemińska, Vicent Moliner, and Katarzyna Świderek. Dynamic and electrostatic effects on the reaction catalyzed by hiv-1 protease. *Journal of the American Chemical Society*, 138(50):16283–16298, December 2016.

[161] Maksim Kulichenko, Justin S. Smith, Benjamin Nebgen, Ying Wai Li, Nikita Fedik, Alexander I. Boldyrev, Nicholas Lubbers, Kipton Barros, and Sergei Tretiak. The rise of neural networks for materials and chemical dynamics. *The Journal of Physical Chemistry Letters*, 12(26):6227–6243, July 2021.

[162] Miron B. Kursa and Witold R. Rudnicki. Feature Selection with the **Boruta** Package. *Journal of Statistical Software*, 36(11), 2010.

[163] Miron B. Kursa and Witold R. Rudnicki. Feature selection with theborutapackage. *Journal of Statistical Software*, 36(11), 2010.

[164] Johannes Kästner, Joanne M. Carr, Thomas W. Keal, Walter Thiel, Adrian Wander, and Paul Sherwood. DL-FIND: An Open-Source Geometry Optimizer for Atomistic Simulations. *The Journal of Physical Chemistry A*, 113(43):11856–11865, October 2009. Publisher: American Chemical Society.

[165] A. Labas, E. Szabo, L. Mones, and M. Fuxreiter. Optimization of reorganization energy drives evolution of the designed Kemp eliminase KE07. *Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics*, 1834(5):908–917, May 2013.

[166] Wenzhen Lai, Hui Chen, Kyung-Bin Cho, and Sason Shaik. External Electric Field Can Control the Catalytic Cycle of Cytochrome P450cam: A QM/MM Study. *The Journal of Physical Chemistry Letters*, 1(14):2082–2087, July 2010. Publisher: American Chemical Society.

[167] Yu-hong Lam, Paul Ha-Yeon Cheong, José M. Blasco Mata, Steven J. Stanway, Véronique Gouverneur, and K. N. Houk. Dielsalder exo selectivity in terminal-substituted dienes and dienophiles: Experimental discoveries and computational explanations. *Journal of the American Chemical Society*, 131(5):1947–1957, 2009. PMID: 19154113.

[168] Joshua P. Layfield and Sharon Hammes-Schiffer. Calculation of vibrational shifts of nitrile probes in the active site of ketosteroid isomerase upon ligand binding. *Journal of the American Chemical Society*, 135(2):717–725, December 2012.

[169] Chengteh Lee, Weitao Yang, and Robert G. Parr. Development of the colle-salvetti correlation-energy formula into a functional of the electron density. *Phys. Rev. B*, 37:785–789, Jan 1988.

[170] Juho Lee, Yoonho Lee, Jungtaek Kim, Adam R. Kosiorek, Seungjin Choi, and Yee Whye Teh. Set transformer: A framework for attention-based permutation-invariant neural networks, 2019.

[171] Junhyun Lee, Inyeop Lee, and Jaewoo Kang. Self-attention graph pooling. In *International conference on machine learning*, pages 3734–3743. PMLR, 2019.

[172] Brian J. Levandowski and K. N. Houk. Hyperconjugative, secondary orbital, electrostatic, and steric effects on the reactivities and endo and exo stereoselectivities of cyclopropene diels–alder reactions. *Journal of the American Chemical Society*, 138(51):16731–16736, 2016. PMID: 27977194.

[173] Brian J. Levandowski, Lufeng Zou, and K. N. Houk. Hyperconjugative aromaticity and antiaromaticity control the reactivities and -facial stereoselectivities of 5-substituted cyclopentadiene diels–alder cycloadditions. *The Journal of Organic Chemistry*, 83(23):14658–14666, 2018. PMID: 30395708.

[174] Pengfei Li and Kenneth M. Merz. MCPB.py: A Python Based Metal Center Parameter Builder. *Journal of Chemical Information and Modeling*, 56(4):599–604, April 2016.

[175] Shih-Cheng Li, Haoyang Wu, Angiras Menon, Kevin A. Spiekermann, Yi-Pei Li, and William H. Green. When do quantum mechanical descriptors help graph neural networks to predict chemical properties? *Journal of the American Chemical Society*, August 2024.

[176] C. Tony Liu, Joshua P. Layfield, Robert J. Stewart, Jarrod B. French, Philip Hanoian, John B. Asbury, Sharon Hammes-Schiffer, and Stephen J. Benkovic. Probing the electrostatics of active site microenvironments along the catalytic cycle for escherichia coli dihydrofolate reductase. *Journal of the American Chemical Society*, 136(29):10349–10360, July 2014.

[177] Fang Liu, Yong Liang, and K. N. Houk. Theoretical elucidation of the origins of substituent and strain effects on the rates of diels–alder reactions of 1,2,4,5-tetrazines. *Journal of the American Chemical Society*, 136(32):11483–11493, 2014. PMID: 25041719.

[178] Fang Liu, Robert S. Paton, Seonah Kim, Yong Liang, and K. N. Houk. Diels–alder reactivities of strained and unstrained cycloalkenes with normal and inverse-electron-demand dienes: Activation barriers and distortion/interaction analysis. *Journal of the American Chemical Society*, 135(41):15642–15649, 2013. PMID: 24044412.

[179] Hanbin Liu and Arieh Warshel. The catalytic effect of dihydrofolate reductase and its mutants is determined by reorganization energies. *Biochemistry*, 46(20):6011–6025, May 2007.

[180] Liangliang Liu, Ying Yu, Zhihui Fei, Min Li, Fang-Xiang Wu, Hong-Dong Li, Yi Pan, and Jianxin Wang. An interpretable boosting model to predict side effects of analgesics for osteoarthritis. *BMC Systems Biology*, 12(S6), November 2018.

[181] R.J. Loader, N. Singh, P.J. O'Malley, and P.L.A. Popelier. The cytotoxicity of ortho alkyl substituted 4-x-phenols: A qsar based on theoretical bond lengths and electron densities. *Bioorganic Medicinal Chemistry Letters*, 16(5):1249–1254, 2006.

[182] Sarah L. Lovelock, Rebecca Crawshaw, Sophie Basler, Colin Levy, David Baker, Donald Hilvert, and Anthony P. Green. The road to fully programmable protein catalysis. *Nature*, 606(7912):49–58, June 2022. Number: 7912 Publisher: Nature Publishing Group.

[183] Tian Lu and Feiwu Chen. Multiwfn: A multifunctional wavefunction analyzer. *Journal of Computational Chemistry*, 33(5):580–592, December 2011.

[184] Nadia G. Léonard, Rakia Dhaoui, Teera Chantarojsiri, and Jenny Y. Yang. Electric Fields in Catalysis: From Enzymes to Molecular Catalysts. *ACS Catalysis*, 11(17):10923–10932, September 2021.

[185] Christopher M. *Pattern Recognition and Machine Learning*. Information Science and Statistics. Springer, New York, NY, 1 edition, August 2006.

[186] Raimund Mannhold, Gennadiy I. Poda, Claude Ostermann, and Igor V. Tetko. Calculation of molecular lipophilicity: State-of-the-art and comparison of logp methods on more than 96, 000 compounds. *Journal of Pharmaceutical Sciences*, 98(3):861–893, March 2009.

[187] R. A. Marcus. On the Theory of Oxidation-Reduction Reactions Involving Electron Transfer. I. *The Journal of Chemical Physics*, 24(5):966–978, 05 1956.

[188] Nicholas M. Marshall, Dewain K. Garner, Tiffany D. Wilson, Yi-Gui Gao, Howard Robinson, Mark J. Nilges, and Yi Lu. Rationally tuning the reduction potential of a single cupredoxin beyond the natural range. *Nature*, 462(7269):113–116, November 2009.

[189] Sergio Martí, Maite Roca, Juan Andrés, Vicent Moliner, Estanislao Silla, Iñaki Tuñón, and Juan Bertrán. Theoretical insights in enzyme catalysis. *Chem. Soc. Rev.*, 33(2):98–107, 2004.

[190] Chérif F. Matta. Modeling biophysical and biological properties from the characteristics of the molecular electron density, electron localization and delocalization matrices, and the electrostatic potential. *Journal of Computational Chemistry*, 35(16):1165–1198, 2014.

[191] Chérif F Matta and Alya A Arabi. Electron-density descriptors as predictors in quantitative structure–activity/property relationships and drug design. *Future Medicinal Chemistry*, 3(8):969–994, June 2011.

[192] Chérif F Matta and Alya A Arabi. Electron-density descriptors as predictors in quantitative structure–activity/property relationships and drug design. *Future Medicinal Chemistry*, 3(8):969–994, June 2011.

[193] Andreas Mayr, Günter Klambauer, Thomas Unterthiner, and Sepp Hochreiter. Deeptox: Toxicity prediction using deep learning. *Frontiers in Environmental Science*, 3:1, feb 2016.

[194] Rinat Meir, Hui Chen, Wenzhen Lai, and Sason Shaik. Oriented Electric Fields Accelerate Diels–Alder Reactions and Control the *endo* / *exo* Selectivity. *ChemPhysChem*, 11(1):301–310, January 2010.

[195] Sebastian Metz, Johannes Kästner, Alexey A. Sokol, Thomas W. Keal, and Paul Sherwood. Chemshell—a modular software package for QM/MM simulations. *WIREs Computational Molecular Science*, 4(2):101–110, March 2014.

[196] Benjamin Meyer, Boodsarin Sawatlon, Stefan Heinen, O. Anatole von Lilienfeld, and Clémence Corminboeuf. Machine learning meets volcano plots: computational discovery of cross-coupling catalysts. *Chem. Sci.*, 9:7069–7077, 2018.

[197] Bill R. Miller, T. Dwight McGee, Jason M. Swails, Nadine Homeyer, Holger Gohlke, and Adrian E. Roitberg. *MMPBSA.py* : An Efficient Program for End-State Free Energy Calculations. *Journal of Chemical Theory and Computation*, 8(9):3314–3321, September 2012.

[198] Amanda Morgenstern, Matthew Jaszai, Mark E. Eberhart, and Anastassia N. Alexandrova. Quantified electrostatic preorganization in enzymes using the geometry of the electron charge density. *Chemical Science*, 8(7):5010–5018, 2017.

[199] Amanda Morgenstern, Matthew Jaszai, Mark E. Eberhart, and Anastassia N. Alexandrova. Quantified electrostatic preorganization in enzymes using the geometry of the electron charge density. *Chemical Science*, 8(7):5010–5018, 2017.

[200] Amanda Morgenstern, Charles Morgenstern, Jonathan Miorelli, Tim Wilson, and M. E. Eberhart. The influence of zero-flux surface motion on chemical reactivity. *Phys. Chem. Chem. Phys.*, 18:5638–5646, 2016.

[201] Abigail R. E. Mountain and Nikolas Kaltsoyannis. Do qtaim metrics correlate with the strength of heavy element–ligand bonds? *Dalton Trans.*, 42:13477–13486, 2013.

[202] Ruipu Mu, Zhaoshuai Wang, Max C. Wamsley, Colbee N. Duke, Payton H. Lii, Sarah E. Epley, London C. Todd, and Patty J. Roberts. Application of Enzymes in Regioselective and Stereoselective Organic Reactions. *Catalysts*, 10(8):832, July 2020.

[203] Vishal C Nashine, Sharon Hammes-Schiffer, and Stephen J Benkovic. Coupled motions in enzyme catalysis. *Current Opinion in Chemical Biology*, 14(5):644–651, October 2010.

[204] Frank Neese, Frank Wennmohs, Ute Becker, and Christoph Riplinger. The orca quantum chemistry program package. *The Journal of Chemical Physics*, 152(22):224108, Jun 2020.

[205] Benedikt Niepötter, Regine Herbst-Irmer, Daniel Kratzert, Prinson P. Samuel, Kartik Chandra Mondal, Herbert W. Roesky, Paul Jerabek, Gernot Frenking, and Dietmar Stalke. Experimental charge density study of a silylone. *Angewandte Chemie International Edition*, 53(10):2766–2770, January 2014.

[206] Richard Obexer, Alexei Godina, Xavier Garrabou, Peer R. E. Mittl, David Baker, Andrew D. Griffiths, and Donald Hilvert. Emergence of a catalytic tetrad during evolution of a highly active artificial aldolase. *Nature Chemistry*, 9(1):50–56, January 2017. Number: 1 Publisher: Nature Publishing Group.

[207] Evgeniia E. Ondar, Mikhail V. Polynski, and Valentine P. Ananikov. Predicting 195pt nmr chemical shifts in water-soluble inorganic/organometallic complexes with a fast and simple protocol combining semiempirical modeling and machine learning. *ChemPhysChem*, 24(11):e202200940, 2023.

[208] Shyue Ping Ong, William Davidson Richards, Anubhav Jain, Geoffroy Hautier, Michael Kocher, Shreyas Cholia, Dan Gunter, Vincent L. Chevrier, Kristin A. Persson, and

Gerbrand Ceder. Python materials genomics (pymatgen): A robust, open-source python library for materials analysis. *Computational Materials Science*, 68:314–319, February 2013.

[209] Paul R. Ortiz De Montellano, editor. *Cytochrome P450: Structure, Mechanism, and Biochemistry*. Springer International Publishing, Cham, 2015.

[210] Sílvia Osuna. The challenge of predicting distal active site mutations in computational enzyme design. *WIREs Computational Molecular Science*, 11(3):e1502, 2021.

[211] A. Otero-de-la Roza, Erin R. Johnson, and Víctor Luaña. Critic2: A program for real-space analysis of quantum chemical interactions in solids. *Computer Physics Communications*, 185(3):1007–1018, March 2014.

[212] Marina A. Pak, Karina A. Markhieva, Mariia S. Novikova, Dmitry S. Petrov, Ilya S. Vorobyev, Ekaterina S. Maksimova, Fyodor A. Kondrashov, and Dmitry N. Ivankov. Using AlphaFold to predict the impact of single mutations on protein stability and function. *PLOS ONE*, 18(3):e0282689, March 2023.

[213] Katrin Palm, Patric Stenberg, Kristina Luthman, and Per Artursson1. Polar molecular surface properties predict the intestinal absorption of drugs in humans. *Pharmaceutical Research*, 14(5):568–571, 1997.

[214] Robert S. Paton, Seonah Kim, Audrey G. Ross, Samuel J. Danishefsky, and K. N. Houk. Experimental diels–alder reactivities of cycloalkenones and cyclic dienes explained through transition-state distortion energies. *Angewandte Chemie*, 123(44):10550–10552, September 2011.

[215] Robert S. Paton, Joel L. Mackey, Woo Han Kim, Jun Hee Lee, Samuel J. Danishefsky, and K. N. Houk. Origins of stereoselectivity in the trans dielsalder paradigm. *Journal of the American Chemical Society*, 132(27):9335–9340, 2010.

[216] John P. Perdew, Matthias Ernzerhof, and Kieron Burke. Rationale for mixing exact exchange with density functional approximations. *The Journal of Chemical Physics*, 105(22):9982–9985, December 1996.

[217] Alessandra Pesce, Martino Bolognesi, and Marco Nardini. Protoglobin: Structure and Ligand-Binding Properties. In *Advances in Microbial Physiology*, volume 63, pages 79–96. Academic Press, 2013.

[218] Vlada V. Petrova, Anton V. Domnin, Yuri B. Porozov, Pavel O. Kuliaev, and Yaroslav V. Solovev. Implementation of machine learning protocols to predict the hydrolysis reaction properties of organophosphorus substrates using descriptors of electron density topology. *Journal of Computational Chemistry*, 45(3):170–182, September 2023.

[219] Susan N. Pieniazek and Kendall N. Houk. The origin of the halogen effect on reactivity and reversibility of diels–alder cycloadditions involving furan. *Angewandte Chemie International Edition*, 45(9):1442–1445, 2006.

[220] Nitesh K. Poona and Riyad Ismail. Using boruta-selected spectroscopic wavebands for the asymptomatic detection of fusarium circinatum stress. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 7(9):3764–3772, September 2014.

[221] Nicholas J. Porter, Emma Danelius, Tamir Gonen, and Frances H. Arnold. Biocatalytic Carbene Transfer Using Diazirines. *Journal of the American Chemical Society*, 144(20):8892–8896, May 2022.

[222] Alja Prah, Eric Frančišković, Janez Mavri, and Jernej Stare. Electrostatics as the driving force behind the catalytic function of the monoamine oxidase a enzyme confirmed by quantum computations. *ACS Catalysis*, 9(2):1231–1240, January 2019.

[223] Alja Prah, Miha Purg, Jernej Stare, Robert Vianello, and Janez Mavri. How monoamine oxidase a decomposes serotonin: An empirical valence bond simulation of the reactive step. *The Journal of Physical Chemistry B*, 124(38):8259–8265, August 2020.

[224] Nathalie Preiswerk, Tobias Beck, Jessica D. Schulz, Peter Milovník, Clemens Mayer, Justin B. Siegel, David Baker, and Donald Hilvert. Impact of scaffold rigidity on the design and evolution of an artificial diels-alderase. *Proceedings of the National Academy of Sciences*, 111(22):8013–8018, 2014.

[225] Elizabeth A. Proctor, Feng Ding, and Nikolay V. Dokholyan. Discrete molecular dynamics. *WIREs Computational Molecular Science*, 1(1):80–92, 2011.

[226] Christina M. Ragain, Robert W. Newberry, Andrew W. Ritchie, and Lauren J. Webb. Role of Electrostatics in Differential Binding of RalGDS to Rap Mutations E30D and K31E Investigated by Vibrational Spectroscopy of Thiocyanate Probes. *The Journal of Physical Chemistry B*, 116(31):9326–9336, August 2012.

[227] Raghunathan Ramakrishnan, Pavlo O. Dral, Matthias Rupp, and O. Anatole von Lilienfeld. Quantum chemistry structures and properties of 134 kilo molecules. *Scientific Data*, 1(1):140022, Aug 2014.

[228] Raghunathan Ramakrishnan, Mia Hartmann, Enrico Tapavicza, and O. Anatole von Lilienfeld. Electronic spectra from tddft and machine learning in chemical space. *The Journal of Chemical Physics*, 143(8):1, August 2015.

[229] David I. Ramírez-Palma and Fernando Cortés-Guzmán. From the linnett–gillespie model to the polarization of the spin valence shells of metals in complexes. *Physical Chemistry Chemical Physics*, 22(42):24201–24212, 2020.

[230] Uriel J. Rangel-Peña, Luis A. Zárate-Hernández, Rosa L. Camacho-Mendoza, Carlos Z. Gómez-Castro, Simplicio González-Montiel, Miriam Pescador-Rojas, Amilcar Meneses-Viveros, and Julián Cruz-Borbolla. Conceptual dft, machine learning and molecular

docking as tools for predicting ld50 toxicity of organothiophosphates. *Journal of Molecular Modeling*, 29(7):217, Jun 2023.

[231] A. K. Rappe, C. J. Casewit, K. S. Colwell, W. A. Goddard, and W. M. Skiff. Uff, a full periodic table force field for molecular mechanics and molecular dynamics simulations. *Journal of the American Chemical Society*, 114(25):10024–10035, December 1992.

[232] Tomáš Raček, Ondřej Schindler, Dominik Toušek, Vladimír Horský, Karel Berka, Jaroslav Koča, and Radka Svobodová. Atomic charge calculator ii: web-based tool for the calculation of partial atomic charges. *Nucleic Acids Research*, 48(W1):W591–W596, May 2020.

[233] RDKit: Open-source cheminformatics. http://www.rdkit.org.

[234] Alan E. Reed, Larry A. Curtiss, and Frank Weinhold. Intermolecular interactions from a natural bond orbital, donor-acceptor viewpoint. *Chemical Reviews*, 88(6):899–926, September 1988.

[235] Manfred T. Reetz. Directed Evolution of Artificial Metalloenzymes: A Universal Means to Tune the Selectivity of Transition Metal Catalysts? *Accounts of Chemical Research*, 52(2):336–344, February 2019. Publisher: American Chemical Society.

[236] Raphael F. Ribeiro, Aleksandr V. Marenich, Christopher J. Cramer, and Donald G. Truhlar. Use of solution-phase vibrational frequencies in continuum models for the free energy of solvation. *The Journal of Physical Chemistry B*, 115(49):14556–14562, November 2011.

[237] Kai Riedmiller, Patrick Reiser, Elizaveta Bobkova, Kiril Maltsev, Ganna Gryn'ova, Pascal Friederich, and Frauke Gräter. Substituting density functional theory in reaction barrier calculations for hydrogen atom transfer in proteins. *Chem. Sci.*, 15:2518–2527, 2024.

[238] Daniel R. Roe and Thomas E. Cheatham. PTRAJ and CPPTRAJ: Software for Processing and Analysis of Molecular Dynamics Trajectory Data. *Journal of Chemical Theory and Computation*, 9(7):3084–3095, July 2013.

[239] David Rogers and Mathew Hahn. Extended-connectivity fingerprints. *Journal of Chemical Information and Modeling*, 50(5):742–754, April 2010.

[240] Torben Rogge, Qingyang Zhou, Nicholas J. Porter, Frances H. Arnold, and K. N. Houk. Iron Heme Enzyme-Catalyzed Cyclopropanations with Diazirines as Carbene Precursors: Computational Explorations of Diazirine Activation and Cyclopropanation Mechanism. *Journal of the American Chemical Society*, 146(5):2959–2966, February 2024.

[241] Sebastián Rojas, Oscar Parravicini, Marcela Vettorazzi, Rodrigo Tosso, Adriana Garro, Lucas Gutiérrez, Sebastián Andújar, and Ricardo Enriz. Combined md/qtaim techniques to evaluate ligand-receptor interactions. scope and limitations. *European Journal of Medicinal Chemistry*, 208:112792, December 2020.

[242] Lars Ruddigkeit, Ruud van Deursen, Lorenz C. Blum, and Jean-Louis Reymond. Enumeration of 166 billion organic small molecules in the chemical universe database gdb-17. *Journal of Chemical Information and Modeling*, 52(11):2864–2875, November 2012.

[243] T. Konstantin Rusch, Michael M. Bronstein, and Siddhartha Mishra. A survey on oversmoothing in graph neural networks, 2023.

[244] Jean-Paul Ryckaert, Giovanni Ciccotti, and Herman J.C Berendsen. Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. *Journal of Computational Physics*, 23(3):327–341, March 1977.

[245] Daniela Röthlisberger, Olga Khersonsky, Andrew M. Wollacott, Lin Jiang, Jason DeChancie, Jamie Betker, Jasmine L. Gallaher, Eric A. Althoff, Alexandre Zanghellini,

Orly Dym, Shira Albeck, Kendall N. Houk, Dan S. Tawfik, and David Baker. Kemp elimination catalysts by computational enzyme design. *Nature*, 453(7192):190–195, May 2008.

[246] Aleksandr B. Sahakyan. Computational studies of dielectric permittivity effects on chemical shifts of alanine dipeptide. *Chemical Physics Letters*, 547:66–72, 2012.

[247] Ruben Sanchez-Garcia, Dávid Havasi, Gergely Takács, Matthew C. Robinson, Alpha Lee, Frank von Delft, and Charlotte M. Deane. Coprinet: graph neural networks provide accurate and rapid compound price prediction for molecule prioritisation. *Digital Discovery*, 2:103–111, 2023.

[248] Hiroko Satoh, Oliver Sacher, Tadashi Nakata, Lingran Chen, Johann Gasteiger, and Kimito Funatsu. Classification of organic reactions: similarity of reactions based on changes in the electronic features of oxygen atoms at the reaction sites. *Journal of Chemical Information and Computer Sciences*, 38(2):210–219, February 1998.

[249] H.L Schmider and A.D Becke. Chemical content of the kinetic energy density. *Journal of Molecular Structure: THEOCHEM*, 527(1–3):51–61, August 2000.

[250] Samuel H. Schneider and Steven G. Boxer. Vibrational Stark Effects of Carbonyl Probes Applied to Reinterpret IR and Raman Data for Enzyme Inhibitors in Terms of Electric Fields at the Active Site. *The Journal of Physical Chemistry B*, 120(36):9672–9684, September 2016.

[251] Kristof Schütt, Oliver Unke, and Michael Gastegger. Equivariant message passing for the prediction of tensorial properties and molecular spectra. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 9377–9388. PMLR, 18–24 Jul 2021.

[252] Kristof T. Schütt, Farhad Arbabzadah, Stefan Chmiela, Klaus R. Müller, and Alexandre Tkatchenko. Quantum-chemical insights from deep tensor neural networks. *Nature Communications*, 8(1):13890, Jan 2017.

[253] Christof H. Schwab. Conformations and 3d pharmacophore searching. *Drug Discovery Today: Technologies*, 7(4):e245–e253, December 2010.

[254] Artur M. Schweidtmann, Jan G. Rittig, Jana M. Weber, Martin Grohe, Manuel Dahmen, Kai Leonhard, and Alexander Mitsos. Physical pooling functions in graph neural networks for molecular property prediction. *Computers & Chemical Engineering*, 172:108202, April 2023.

[255] Kristof T. Schütt, Oliver T. Unke, and Michael Gastegger. Equivariant message passing for the prediction of tensorial properties and molecular spectra, 2021.

[256] S. Shaik, D. Danovich, K. D. Dubey, and T. Stuyver. CHAPTER 2:The Impact of Electric Fields on Chemical Structure and Reactivity. In *Effects of Electric Fields on Structure and Reactivity*, pages 12–70. March 2021.

[257] Sason Shaik, David Danovich, Jyothish Joy, Zhanfeng Wang, and Thijs Stuyver. Electric-Field Mediated Chemistry: Uncovering and Exploiting the Potential of (Oriented) Electric Fields to Exert Chemical Catalysis and Reaction Control. *Journal of the American Chemical Society*, 142(29):12551–12562, July 2020. Publisher: American Chemical Society.

[258] Sason Shaik, Debasish Mandal, and Rajeev Ramanan. Oriented electric fields as future smart reagents in chemistry. *Nature Chemistry*, 8(12):1091–1098, December 2016. Number: 12 Publisher: Nature Publishing Group.

[259] Pankaz K. Sharma, Zhen T. Chu, Mats H. M. Olsson, and Arieh Warshel. A new paradigm for electrostatic catalysis of radical reactions in vitamin B $_{12}$ enzymes. *Proceedings of the National Academy of Sciences*, 104(23):9661–9666, June 2007.

[260] Paul Sherwood, Alex H. de Vries, Martyn F. Guest, Georg Schreckenbach, C.Richard A. Catlow, Samuel A. French, Alexey A. Sokol, Stefan T. Bromley, Walter Thiel, Alex J. Turner, Salomon Billeter, Frank Terstegen, Stephan Thiel, John Kendrick, Stephen C. Rogers, John Casci, Mike Watson, Frank King, Elly Karlsen, Merethe Sjøvoll, Adil Fahmi, Ansgar Schäfer, and Christian Lennartz. QUASI: A general purpose implementation of the QM/MM approach and its application to problems in catalysis. *Journal of Molecular Structure: THEOCHEM*, 632(1-3):1–28, August 2003.

[261] David Shirvanyants, Feng Ding, Douglas Tsao, Srinivas Ramachandran, and Nikolay V. Dokholyan. Discrete molecular dynamics: An efficient and versatile simulation method for fine protein characterization. *The Journal of Physical Chemistry B*, 116(29):8375–8382, 2012. PMID: 22280505.

[262] Aayush R. Singh, Brian A. Rohr, Joseph A. Gauthier, and Jens K. Nørskov. Predicting chemical reaction barriers with a machine learning model. *Catalysis Letters*, 149(9):2347–2354, March 2019.

[263] Nakul Singh, Robert J. Loader, Patrick J. O'Malle, and Paul L. A. Popelier. Computation of relative bond dissociation enthalpies (bde) of phenolic antioxidants from quantum topological molecular similarity (qtms). *The Journal of Physical Chemistry A*, 110(20):6498–6503, 2006. PMID: 16706407.

[264] W. Smith, C.W. Yong, and P.M. Rodger. DL_poly: Application to molecular simulation. *Molecular Simulation*, 28(5):385–471, May 2002.

[265] Manuel Sparta, David Shirvanyants, Feng Ding, Nikolay V. Dokholyan, and Anastassia N. Alexandrova. Hybrid dynamics simulation engine for metalloproteins. *Biophysical Journal*, 103(4):767–776, 2012.

[266] Kevin Spiekermann, Lagnajit Pattanaik, and William H. Green. High accuracy barrier heights, enthalpies, and rate coefficients for chemical reactions. *Scientific Data*, 9(1):417, Jul 2022.

[267] Kevin A. Spiekermann, Lagnajit Pattanaik, and William H. Green. Fast predictions of reaction barrier heights: Toward coupled-cluster accuracy. *The Journal of Physical Chemistry A*, 126(25):3976–3986, June 2022.

[268] Evan Walter Clark Spotte-Smith, Samuel M. Blau, Xiaowei Xie, Hetal D. Patel, Mingjian Wen, Brandon Wood, Shyam Dwaraknath, and Kristin Aslaug Persson. Quantum chemical calculations of lithium-ion battery electrolyte and interphase species. *Scientific Data*, 8(1):203, Aug 2021.

[269] David T. Stanton and Peter C. Jurs. Development and use of charged partial surface area structural descriptors in computer-assisted quantitative structure-property relationship studies. *Analytical Chemistry*, 62(21):2323–2329, November 1990.

[270] P. J. Stephens, F. J. Devlin, C. F. Chabalowski, and M. J. Frisch. Ab initio calculation of vibrational absorption and circular dichroism spectra using density functional force fields. *The Journal of Physical Chemistry*, 98(45):11623–11627, 1994.

[271] Andrew Streitwieser, John B. Collins, John M. McKelvey, David Grier, John Sender, and A. Glenn Toczko. Integrated spatial electron populations in molecules: The electron projection function. *Proceedings of the National Academy of Sciences*, 76(6):2499–2502, June 1979.

[272] Thijs Stuyver, Rajeev Ramanan, Dibyendu Mallick, and Sason Shaik. Oriented (Local) Electric Fields Drive the Millionfold Enhancement of the H-Abstraction Catalysis Observed for Synthetic Metalloenzyme Analogues. *Angewandte Chemie International Edition*, 59(20):7915–7920, 2020. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/anie.201916592.

[273] Huiyong Sun, Youyong Li, Sheng Tian, Lei Xu, and Tingjun Hou. Assessing the performance of MM/PBSA and MM/GBSA methods. 4. Accuracies of MM/PBSA and MM/GBSA methodologies evaluated by various simulation protocols using PDBbind data set. *Phys. Chem. Chem. Phys.*, 16(31):16719–16729, 2014.

[274] Borys Szefczyk, Adrian J. Mulholland, Kara E. Ranaghan, and W. Andrzej Sokalski. Differential transition-state stabilization in enzyme catalysis: quantum chemical analysis of interactions in the chorismate mutase reaction and prediction of the optimal catalytic field. *Journal of the American Chemical Society*, 126(49):16148–16159, November 2004.

[275] Jianmin Tao, John P. Perdew, Viktor N. Staroverov, and Gustavo E. Scuseria. Climbing the density functional ladder: Nonempirical meta–generalized gradient approximation designed for molecules and solids. *Phys. Rev. Lett.*, 91:146401, Sep 2003.

[276] Thomas C. Terwilliger, Dorothee Liebschner, Tristan I. Croll, Christopher J. Williams, Airlie J. McCoy, Billy K. Poon, Pavel V. Afonine, Robert D. Oeffner, Jane S. Richardson, Randy J. Read, and Paul D. Adams. AlphaFold predictions are valuable hypotheses and accelerate but do not replace experimental structure determination. *Nature Methods*, 21(1):110–116, January 2024.

[277] Chuan Tian, Koushik Kasavajhala, Kellon A. A. Belfon, Lauren Raguette, He Huang, Angela N. Migues, John Bickel, Yuzhang Wang, Jorge Pincay, Qin Wu, and Carlos Simmerling. ff19SB: Amino-Acid-Specific Protein Backbone Parameters Trained against Quantum Mechanics Energy Surfaces in Solution. *Journal of Chemical Theory and Computation*, 16(1):528–552, January 2020.

[278] Ilian T. Todorov, William Smith, Kostya Trachenko, and Martin T. Dove. Dlpoly3: new dimensions in molecular dynamics simulations via massive parallelism. *Journal of Materials Chemistry*, 16(20):1911, 2006.

[279] Oliver Treutler and Reinhart Ahlrichs. Efficient molecular numerical integration schemes. *The Journal of Chemical Physics*, 102(1):346–354, 01 1995.

[280] Oleg Trott and Arthur J. Olson. Autodock vina: Improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *Journal of Computational Chemistry*, 31(2):455–461, June 2009.

[281] Nicholas J. Turner. Directed evolution drives the next generation of biocatalysts. *Nature Chemical Biology*, 5(8):567–573, August 2009. Number: 8 Publisher: Nature Publishing Group.

[282] Sai Mahit Vadaddi, Qiyuan Zhao, and Brett M Savoie. Graph to activation energy models easily reach irreducible errors but show limited transferability. November 2023.

[283] Valerie Vaissier, Sudhir C. Sharma, Karl Schaettle, Taoran Zhang, and Teresa Head-Gordon. Computational Optimization of Electric Fields for Improving Catalysis of a Designed Kemp Eliminase. *ACS Catalysis*, 8(1):219–227, January 2018. Publisher: American Chemical Society.

[284] Crystal E. Valdez, Amanda Morgenstern, Mark E. Eberhart, and Anastassia N. Alexandrova. Predictive methods for computational metalloenzyme redesign – a test case with carboxypeptidase a. *Physical Chemistry Chemical Physics*, 18(46):31744–31756, 2016.

[285] Santiago Vargas, Shobhit Chaturvedi, and Anastassia Alexandrova. Machine-learning prediction of protein function from the portrait of its intramolecular electric field. June 2024.

[286] Santiago Vargas, Winston Gee, and Anastassia Alexandrova. High-throughput quantum theory of atoms in molecules (qtaim) for geometric deep learning of molecular and reaction properties. *Digital Discovery*, 3:987–998, 2024.

[287] Santiago Vargas, Matthew R. Hennefarth, Zhihao Liu, and Anastassia N. Alexandrova. Machine Learning to Predict Diels–Alder Reaction Barriers from the Reactant State Electron Density. *Journal of Chemical Theory and Computation*, 17(10):6203–6213, October 2021.

[288] Santiago Vargas, Matthew R. Hennefarth, Zhihao Liu, and Anastassia N. Alexandrova. Machine learning to predict diels–alder reaction barriers from the reactant state electron density. *Journal of Chemical Theory and Computation*, 17(10):6203–6213, September 2021.

[289] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph attention networks. In *International Conference on Learning Representations*, 2018.

[290] Oriol Vinyals, Samy Bengio, and Manjunath Kudlur. Order matters: Sequence to sequence for sets, 2016.

[291] Malte Von Arnim and Reinhart Ahlrichs. Performance of parallel turbomole for density functional calculations. *Journal of Computational Chemistry*, 19(15):1746–1757, 1998.

[292] Beatriz von der Esch, Johannes C. B. Dietschreit, Laurens D. M. Peters, and Christian Ochsenfeld. Finding reactive configurations: A machine learning approach for estimating energy barriers applied to sirtuin 5. *Journal of Chemical Theory and Computation*, 15(12):6660–6667, 2019. PMID: 31765138.

[293] S. H. Vosko, L. Wilk, and M. Nusair. Accurate spin-dependent electron liquid correlation energies for local spin density calculations: a critical analysis. *Canadian Journal of Physics*, 58(8):1200–1211, 1980.

[294] Steffen Wagner, Angelika Hofmann, Bettina Siedle, Lothar Terfloth, Irmgard Merfort, and Johann Gasteiger. Development of a structural model for nf-b inhibition of

sesquiterpene lactones using self-organizing neural networks. *Journal of Medicinal Chemistry*, 49(7):2241–2252, February 2006.

[295] Sodiq O. Waheed, Shobhit S. Chaturvedi, Tatyana G. Karabencheva-Christova, and Christo Z. Christov. Catalytic Mechanism of Human Ten-Eleven Translocation-2 (TET2) Enzyme: Effects of Conformational Changes, Electric Field, and Mutations. *ACS Catalysis*, 11(7):3877–3890, April 2021. Publisher: American Chemical Society.

[296] Xianwei Wang and Xiao He. An Ab Initio QM/MM Study of the Electrostatic Contribution to Catalysis in the Active Site of Ketosteroid Isomerase. *Molecules*, 23(10):2410, September 2018.

[297] Zhanfeng Wang, David Danovich, Rajeev Ramanan, and Sason Shaik. Oriented-external electric fields create absolute enantioselectivity in diels–alder reactions: Importance of the molecular dipole moment. *Journal of the American Chemical Society*, 140(41):13350–13359, September 2018.

[298] ThomasR. Ward. Artificial Enzymes Made to Order: Combination of Computational Design and Directed Evolution. *Angewandte Chemie International Edition*, 47(41):7802–7803, September 2008.

[299] A Warshel. Energetics of enzyme catalysis. *Proceedings of the National Academy of Sciences*, 75(11):5250–5254, November 1978.

[300] Arieh Warshel. Electrostatic basis of structure-function correlation in proteins. *Accounts of Chemical Research*, 14(9):284–290, September 1981.

[301] Arieh Warshel. Electrostatic Origin of the Catalytic Power of Enzymes and the Role of Preorganized Active Sites. *Journal of Biological Chemistry*, 273(42):27035–27038, October 1998. Publisher: Elsevier.

[302] Arieh Warshel. Electrostatic origin of the catalytic power of enzymes and the role of preorganized active sites. *Journal of Biological Chemistry*, 273(42):27035–27038, October 1998.

[303] Arieh Warshel, Pankaz K. Sharma, Zhen T. Chu, and Johan Åqvist. Electrostatic Contributions to Binding of Transition State Analogues Can Be Very Different from the Corresponding Contributions to Catalysis: Phenolates Binding to the Oxyanion Hole of Ketosteroid Isomerase. *Biochemistry*, 46(6):1466–1476, February 2007.

[304] Arieh Warshel, Pankaz K. Sharma, Mitsunori Kato, Yun Xiang, Hanbin Liu, and Mats H. M. Olsson. Electrostatic Basis for Enzyme Catalysis. *Chemical Reviews*, 106(8):3210–3235, August 2006.

[305] Arieh Warshel, Pankaz K. Sharma, Mitsunori Kato, Yun Xiang, Hanbin Liu, and Mats H. M. Olsson. Electrostatic Basis for Enzyme Catalysis. *Chemical Reviews*, 106(8):3210–3235, August 2006.

[306] Arieh Warshel and Robert M. Weiss. An empirical valence bond approach for comparing reactions in solutions and in enzymes. *Journal of the American Chemical Society*, 102(20):6218–6226, September 1980.

[307] Michael L. Waskom. seaborn: statistical data visualization. *Journal of Open Source Software*, 6(60):3021, 2021.

[308] Yang Wei, Antonio Tinoco, Viktoria Steck, Rudi Fasan, and Yong Zhang. Cyclopropanations via Heme Carbenes: Basic Mechanism and Effects of Carbene Substituent, Protein Axial Ligand, and Porphyrin Substitution. *Journal of the American Chemical Society*, 140(5):1649–1662, February 2018. Publisher: American Chemical Society.

[309] Florian Weigend and Reinhart Ahlrichs. Balanced basis sets of split valence, triple zeta valence and quadruple zeta valence quality for h to rn: Design and assessment of accuracy. *Phys. Chem. Chem. Phys.*, 7:3297–3305, 2005.

[310] Florian Weigend and Reinhart Ahlrichs. Balanced basis sets of split valence, triple zeta valence and quadruple zeta valence quality for h to rn: Design and assessment of accuracy. *Physical Chemistry Chemical Physics*, 7(18):3297, 2005.

[311] Valerie Vaissier Welborn and Teresa Head-Gordon. Fluctuations of Electric Fields in the Active Site of the Enzyme Ketosteroid Isomerase. *Journal of the American Chemical Society*, 141(32):12487–12492, August 2019.

[312] Valerie Vaissier Welborn, Luis Ruiz Pestana, and Teresa Head-Gordon. Computational optimization of electric fields for better catalysis design. *Nature Catalysis*, 1(9):649–655, September 2018. Number: 9 Publisher: Nature Publishing Group.

[313] Mingjian Wen, Samuel M. Blau, Evan Walter Clark Spotte-Smith, Shyam Dwaraknath, and Kristin A. Persson. Bondnet: a graph neural network for the prediction of bond dissociation energies for charged molecules. *Chemical Science*, 12(5):1858–1868, 2021.

[314] Mingjian Wen, Samuel M. Blau, Xiaowei Xie, Shyam Dwaraknath, and Kristin A. Persson. Improving machine learning performance on small chemical reaction data with unsupervised contrastive pretraining. *Chemical Science*, 13(5):1446–1458, 2022.

[315] Harry Wiener. Structural determination of paraffin boiling points. *Journal of the American Chemical Society*, 69(1):17–20, January 1947.

[316] Timothy R. Wilson, Amanda Morgenstern, Anastassia N. Alexandrova, and M.E. Eberhart. Bond Bundle Analysis of Ketosteroid Isomerase. *The Journal of Physical Chemistry B*, 126(46):9443–9456, November 2022.

[317] W. Todd Wipke and Thomas M. Dyott. Stereochemically unique naming algorithm. *Journal of the American Chemical Society*, 96(15):4834–4842, July 1974.

[318] Richard Wolfenden and Mark J. Snider. The Depth of Chemical Time and the Power of Enzymes as Catalysts. *Accounts of Chemical Research*, 34(12):938–945, December 2001. Publisher: American Chemical Society.

[319] Yufan Wu and Steven G. Boxer. A Critical Test of the Electrostatic Contribution to Catalysis with Noncanonical Amino Acids in Ketosteroid Isomerase. *Journal of the American Chemical Society*, 138(36):11890–11895, September 2016.

[320] Zhenqin Wu, Bharath Ramsundar, Evan N. Feinberg, Joseph Gomes, Caleb Geniesse, Aneesh S. Pappu, Karl Leswing, and Vijay Pande. Moleculenet: a benchmark for molecular machine learning. *Chemical Science*, 9(2):513–530, 2018.

[321] Wen Jun Xie, Mojgan Asadi, and Arieh Warshel. Enhancing computational enzyme design by a maximum entropy strategy. *Proceedings of the National Academy of Sciences*, 119(7):e2122355119, February 2022. Publisher: Proceedings of the National Academy of Sciences.

[322] Zhongyue Yang, Natalia Hajlasz, Adam H. Steeves, and Heather J. Kulik. Quantifying the long-range coupling of electronic properties in proteins with ab initio molecular dynamics**. *Chemistry–Methods*, 1(8):362–373, July 2021.

[323] Andy Hsien-Wei Yeh, Christoffer Norn, Yakov Kipnis, Doug Tischer, Samuel J. Pellock, Declan Evans, Pengchen Ma, Gyu Rie Lee, Jason Z. Zhang, Ivan Anishchenko, Brian Coventry, Longxing Cao, Justas Dauparas, Samer Halabiya, Michelle DeWitt, Lauren Carter, K. N. Houk, and David Baker. De novo design of luciferases using deep learning. *Nature*, 614(7949):774–780, February 2023.

[324] J.M. Yon, D. Perahia, and C. Ghélis. Conformational dynamics and enzyme activity. *Biochimie*, 80(1):33–42, January 1998.

[325] Wonsuk Yoo, Robert Mayberry, Sejong Bae, Karan Singh, Q. He, and James Lillard. A study of effects of multicollinearity in the multivariable analysis. *International journal of applied science and technology*, 4:9–19, 10 2014.

[326] Zhiling Zheng, Oufan Zhang, Christian Borgs, Jennifer T. Chayes, and Omar M. Yaghi. Chatgpt chemistry assistant for text mining and the prediction of mof synthesis. *Journal of the American Chemical Society*, 145(32):18048–18062, August 2023.

[327] Tai-Ping Zhou, Jianqiang Feng, Yongchao Wang, Shengying Li, and Binju Wang. Substrate Conformational Switch Enables the Stereoselective Dimerization in P450 NascB: Insights from Molecular Dynamics Simulations and Quantum Mechanical/Molecular Mechanical Calculations. *JACS Au*, 4(4):1591–1604, April 2024.

[328] Ioanna Zoi, Dimitri Antoniou, and Steven D. Schwartz. Electric Fields and Fast Protein Dynamics in Enzymes. *The Journal of Physical Chemistry Letters*, 8(24):6165–6170, December 2017.

[329] Nihan Çelebi Ölçüm, Daniel H. Ess, Viktorya Aviyente, and K. N. Houk. Lewis acid catalysis alters the shapes and products of bis-pericyclic dielsalder transition states. *Journal of the American Chemical Society*, 129(15):4528–4529, 2007. PMID: 17385868.

[330] Katarzyna Świderek, Sergio Marti, Iñaki Tuñón, Vicent Moliner, and Juan Bertran. Peptide bond formation mechanism catalyzed by ribosome. *Journal of the American Chemical Society*, 137(37):12024–12034, September 2015.

[331] Katarzyna Świderek, Iñaki Tuñón, and Vicent Moliner. Predicting enzymatic reactivity: from theory to design. *WIREs Computational Molecular Science*, 4(5):407–421, October 2013.