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

Santiago Vargas[†], Matthew R. Hennefarth[†], Zhihao Liu[†], Anastassia N. Alexandrova[*,†,‡]

[†]*Department of Chemistry and Biochemistry, University of California, Los Angeles, 607 Charles E. Young Drive East, Los Angeles, CA 90095-1569, USA*

[‡]*California NanoSystems Institute, University of California, Los Angeles, 570 Westwood Plaza, Los Angeles, California 90095-1569, USA*

E-mail: *ana@chem.ucla.edu

Phone: +1 310 825-3769

Abstract

Reaction barriers are key to our understanding of chemical reactivity and catalysis. Certain reactions are so seminal in chemistry, that countless variants, with or without catalysts, have been studied and their barriers have been computed or measured experimentally. This wealth of data represents a perfect opportunity to leverage machine learning models, which could quickly predict barriers without explicit calculations or measurement. Here, we show that the topological descriptors of the quantum mechanical charge density in the reactant state constitute a set that is both rigorous and continuous, and can be used effectively for prediction of reaction barrier energies to a high degree of accuracy. We demonstrate this on the Diels-Alder reaction, highly important in biology and medicinal chemistry, and as such, studied extensively. This reaction exhibits a range of barriers as large as 270 kJ/mol. While we trained our single-objective supervised (labeled) regression algorithms on simpler Diels-Alder reactions in solution, they predict reaction barriers also in significantly more complicated contexts, such a Diels-Alder reaction catalyzed by an artificial enzyme and its evolved variants, in agreement with experimental changes in k_{cat}. We expect this tool to apply broadly to a variety of reactions in solution or in the presence of a catalyst, for screening and circumventing heavily involved computations or experiments.

Introduction

For any reaction, we are typically interested in the transition state (TS), activation energy, and potential energy surface.[1] 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. In catalysis, for example, k_{cat} is usually determined through transition state theory, which relates it to the height of the reaction barrier, $\Delta G^{\ddagger}$, via an Arrhenius relationship.[2–5] Another example is electron transfer reactions, where Marcus Theory can be utilized to calculate the rate.[6–8] To determine $\Delta G^{\ddagger}$ computationally, both the reactant state and the TS structure need to be known. While the reactant state is generally easier to compute via optimization of nuclear coordinates to a local minimum, finding a TS, i.e. a first-order saddle point on the potential energy hypersurface, is exponentially more difficult, particularly for complicated systems, such as enzymes or heterogeneous interfaces. For the various TS search algorithms present[9,10,19–28,11,29–38,12,39–48,13,49–52,14–18] the quality of the output is largely determined by the initial guess at the TS,[53] which can be semi-automated. Unfortunately, automation does not always guarantee a success, in which case the process of TS search turns into a tedious trial-and-error procedure. Regardless of the approach, the scaling of this process with the system size is poor. At the same time, it is often of interest to quickly predict many barriers for many variations of the same reaction, for example in catalyst design. To summarize, 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 lands itself well into 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.[54–58] Here, we propose a direct prediction of the reaction barriers through quantum electronic descriptors of the reactant state: the electron density, $\rho(\mathbf{r})$, and its derived mathematical properties. We are building on the following previous findings: our previous work on the Ketosteroid Isomerase enzyme and its mutants,[59] and the Diels-Alder reaction,[60] 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 methods that utilize $\rho(\mathbf{r})$ to predict changes in chemical parameters such as pKa,[61] as well as reactivity.[62–64] Finally, and centrally, according to the Hohenberg-Kohn theorem,[65] 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.

Results and Discussion

A vast array of scientific literature details 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,[66,67,76–79,68–75] as our case study. 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)[80–82] to generate topological parameters of $\rho(\mathbf{r})$ from our computed reactant state structures. 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 8 mutations (introduced through laboratory directed evolution).[83]

$\rho(\mathbf{r})$ in the reactant state was investigated using QTAIM, 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 density (Eq. 1).

$$\nabla\rho(\mathbf{r}) \cdot \mathbf{n}(\mathbf{r}) = 0 \text{ for all } \mathbf{r} \in S(\Omega) \qquad \text{(Eq. 1)}$$

There are 4 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 CCP is a minimum in all three spatial directions.

The input space for the feature selection/regression algorithms consists of mathematical features at a fixed set of ABs and CPs. This initial pick of features requires some knowledge of the reaction mechanism. For the Diels-Alder reaction, we included 10 features shown in Figure 1: 6 ABs corresponding to atoms participating in bond breaking and forming, and 4 BCPs from the dienophile and diene. From each of the 6 ABs, 9 descriptors were selected including localization/delocalization indices, electrostatic potentials, charge, and electronic energy contributions. We also include the electrostatic potential (Φ) and the electronic (Φ^{e}) and nuclear contributions (Φ^{nuc}) which is evaluated only at the nuclei of the AB. Note that these values are well-defined as they exclude the contribution of the nuclei that at which we are evaluating. For each of the 4 BCPs, 19 descriptors were extracted including values such as ellipticity, density, stress tensor eigenvalues, density hessian eigenvalues, divergence of density, potential energy, delocalization index, and kinetic energy. Since our calculations are performed within DFT, the correlation energy component is missing in our algorithms. In addition to the space of QTAIM features, there is a vector of 73 variables for each input system consisting of other important system statistics, such as spin, and charge. Overall, this amounted to an input vector of 203 variables for feature selection (see supporting information for a full list of variables), and it is independent of the total number of ABs and CPs in the system. This fixed length input is a necessity for most machine learning algorithms.

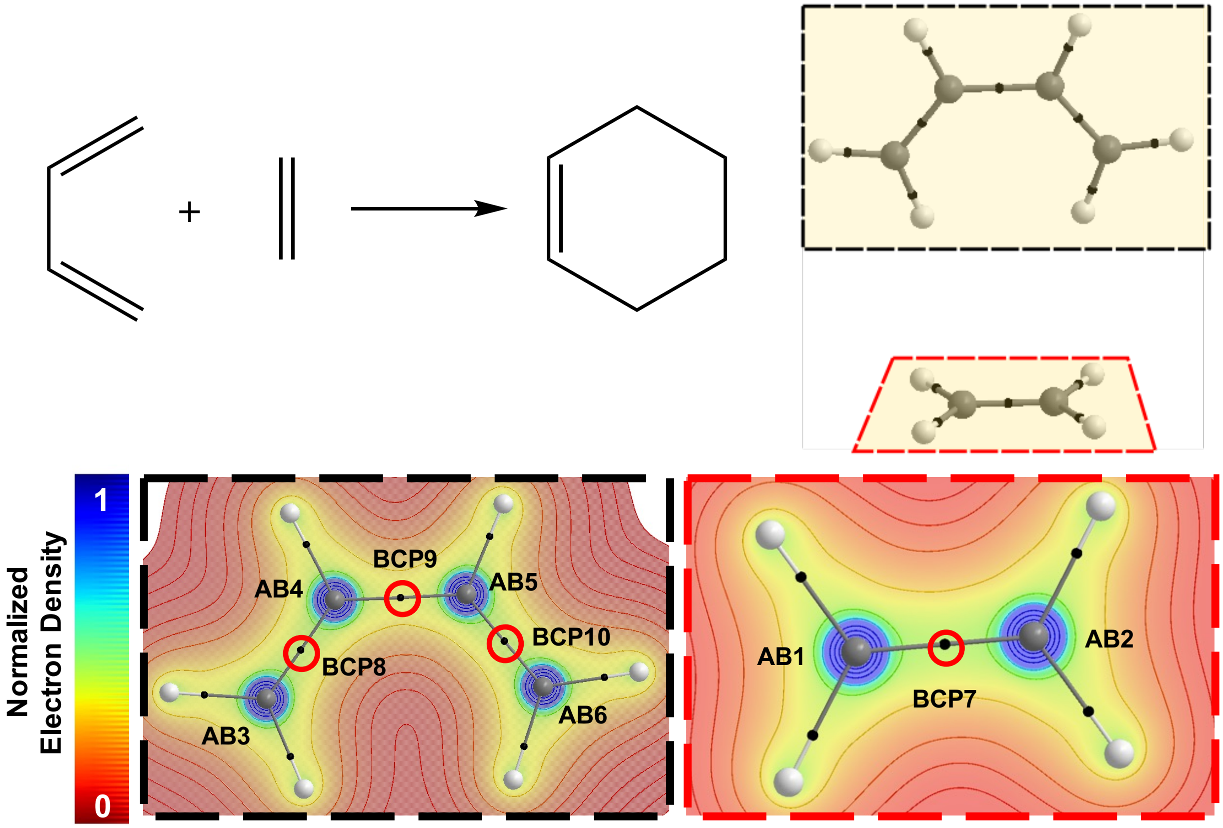
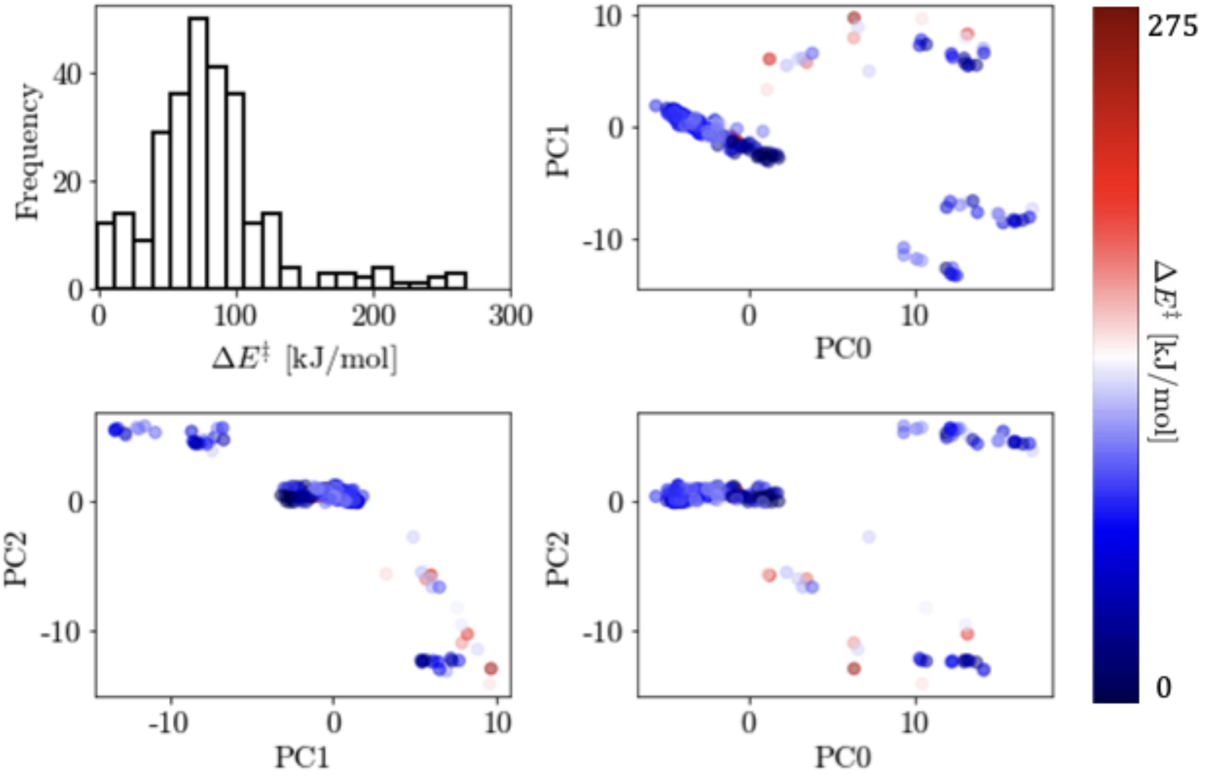


Figure 1. Simplest Diels-Alder reaction between 1,3-butadiene and ethylene to form cyclohexene. Location of ABs and CPs used as the input vector to the ML algorithms. AB 1-6 are the atomic basins defined for those atoms and BCPs are circled in red.

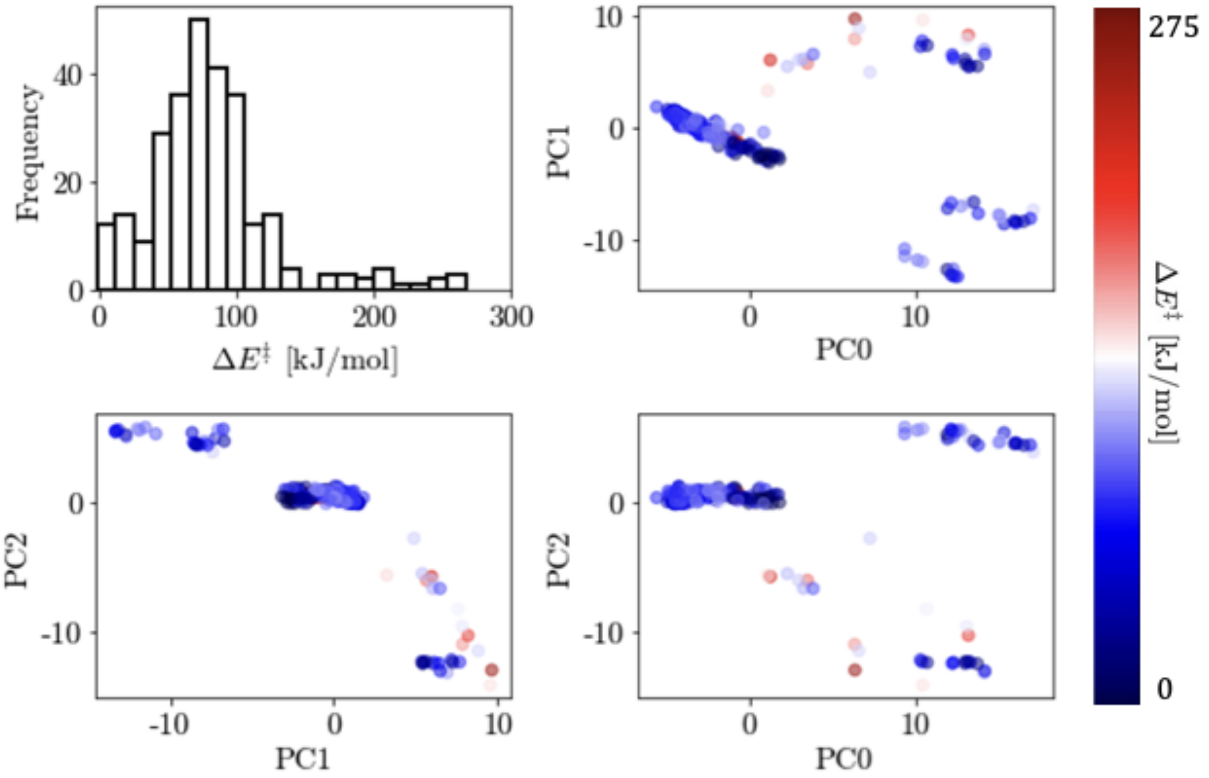
The compiled dataset consisted of 296 Diels-Alder reactions from over a dozen different sources, including reactions with a diverse set of functional groups, sizes, and geometries (Table S1). While the canonical Diels-Alder reaction features the formation of two new C-C bonds with four new stereocenters, 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) (



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

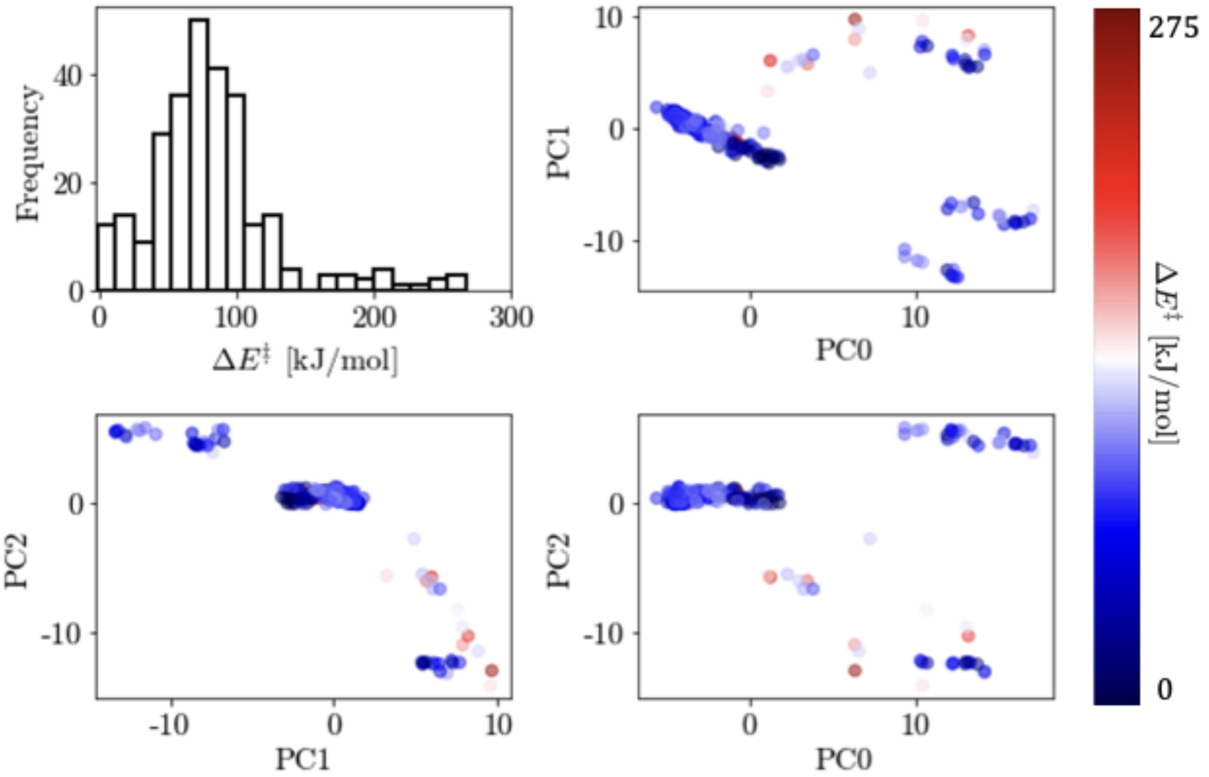
Figure 2). The majority of the reactions have a barrier within the range of 50 to 150 kJ/mol (12 to 35.9 kcal/mol), while higher/lower reaction barriers are underrepresented within the data set. Our dataset only includes Diels-Alder reactions that proceed via a concerted mechanism, and do not include reactions that proceed stepwise.



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

Figure 2. The distribution of computed barrier energies from the dataset. Here we can determine that there is a great degree of variability in the distribution with sparser values for low (0-50 kJ/mol) and high (>150 kJ/mol) barriers. Projection of the first three principal components in the PCA space of the input data illustrate the dataset is not easily linearly separable. The first and second principal components are decomposed into their constituent variables as well.

First, to visualize the input space of this model and understand how variables correlate within the dataset, principal component analysis (PCA) analysis was performed. Along the first three principal component axis, we see that there are no apparent gradients for



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

increasing/decreasing barrier energies (Figure 2). Both high and low $\Delta E^‡$ appear to be spread out throughout the component space implying this data is non-linear 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, 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 (see the Supporting Information for detailed description of each of these methods). Datapoints were divided in an 80-20 split with 5-fold cross validation used to select specific model parameter. 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 (Table *1*) to the ranked list of features from the permutation ranking (Figure 3), Φ (including both Φ^{nuc} and Φ^{e}) and Bader charge (q) appear to be the most physically important set of descriptors from a statistical standpoint. The permutation ranking of features from the physical dataset is shown in Figure 3, and the permutation ranking for features in the full pooled dataset are in Figure S8.

Table 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'.

Feature	Type	Raw Pooled Features	Pooled, Uncorrelated Features	Physical Feature Set
1	AB	$q, E^e, \Phi, \lambda, T, \delta, \delta^{bond}$	q, Φ, E^e, δ	q, Φ, E^e, δ
2	AB	q, Φ, δ^{bond}	q, Φ	q, Φ, E^e, δ
3	AB	$E^e, \Phi, \lambda, T, \delta$	E^e, Φ, δ	q, Φ, E^e, δ
4	AB	$q, E^e, \Phi, T, \delta^{bond}$	q, E^e, Φ	q, Φ, E^e, δ
5	AB	$q, \Phi, \Phi^{nuc}, \lambda, \delta$	$q, \Phi, \Phi^{nuc}, \delta$	$q, \Phi, \Phi^{nuc}, \delta$
6	AB	$\Phi, \lambda, \delta, \delta^{bond}$	Φ, δ	q, Φ, δ
7	BCP	$\epsilon, d', \mathrm{d}, \mathrm{d^{sum}}$	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**

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 (Eq. 2).

$$v[\rho] = v_{ex}(\mathbf{r}) + \int d\tau' \frac{\rho(\mathbf{r}')}{|\mathbf{r} - \mathbf{r}'|} + \frac{\delta E_{ex}[\rho]}{\delta\rho} \qquad \text{(Eq. 2)}$$

$v_{ex}(\mathbf{r})$ 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 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 BCP's was also selected as an important feature (Eq. 3).

$$\epsilon = \frac{\lambda_{\mathbf{H}\rho}^{(1)}}{\lambda_{\mathbf{H}\rho}^{(2)}} - 1 \qquad \text{(Eq. 3)}$$

ϵ is a measure of the elliptical nature of the density within the plane orthogonal to the bond direction. 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.

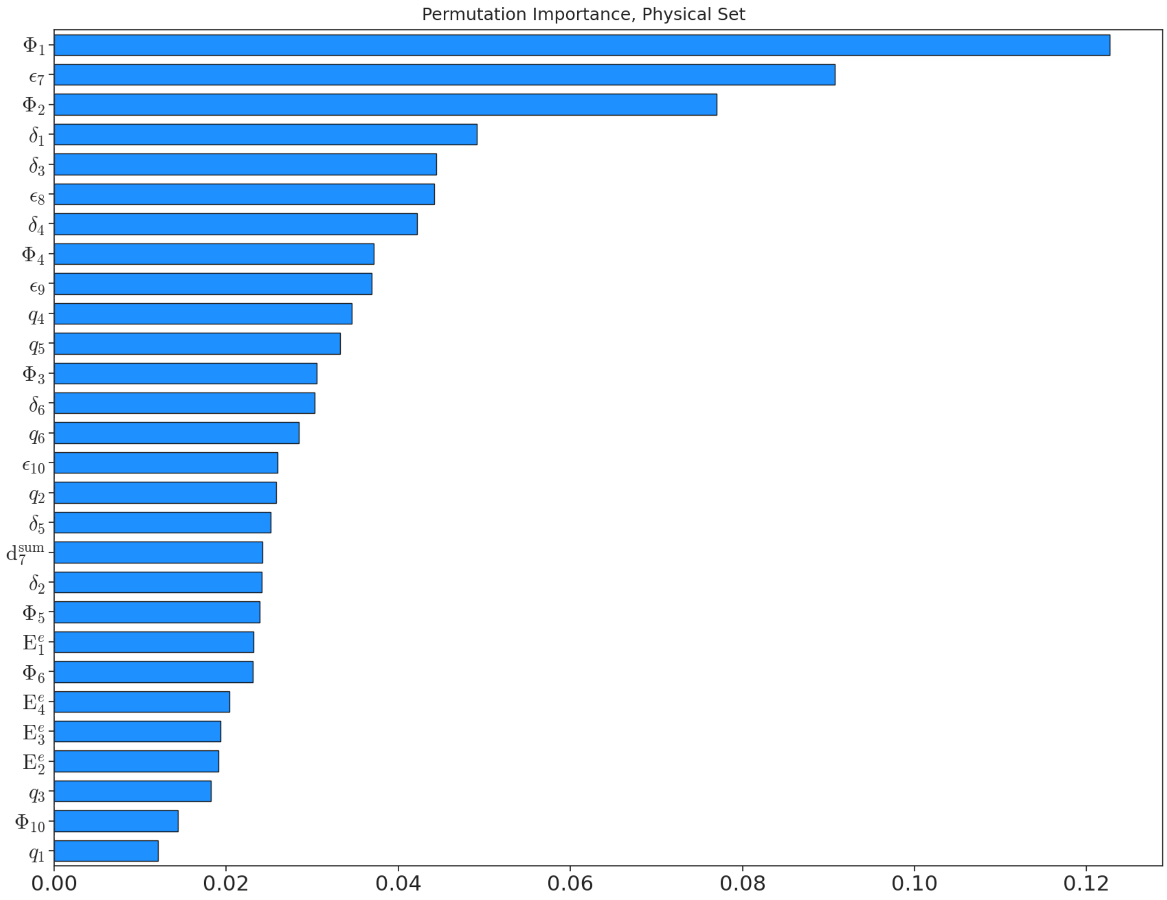


Figure 3. Permutation importance quantifies the importance of each feature relative to each other. Models trained on the compiled dataset show that electrostatic potentials are an important descriptor at almost every CP.

Models were trained using the features selected from the selection algorithms, with an 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 dataset of 38 variables still presented a large input space relative to the size of the dataset; 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 (Figure 4). This reduced dataset (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.[84] 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).

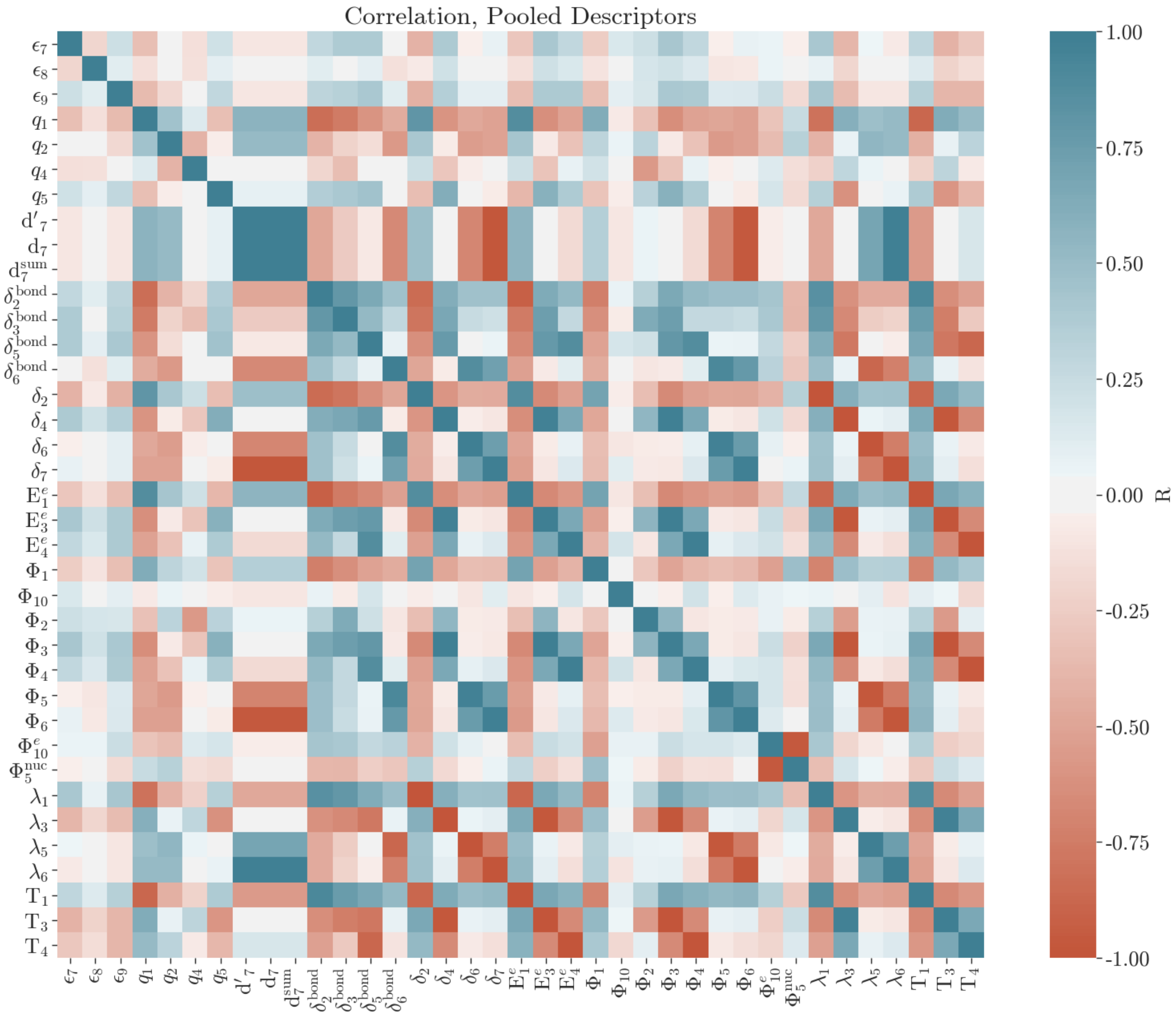


Figure 4. The correlation matrix of the compiled set of descriptors. Features that were heavily correlated with other features were removed.

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 is reported in Table S3. 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 Tree) 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 a baseline metric of guessing the mean barrier energy of the dataset for every instance (

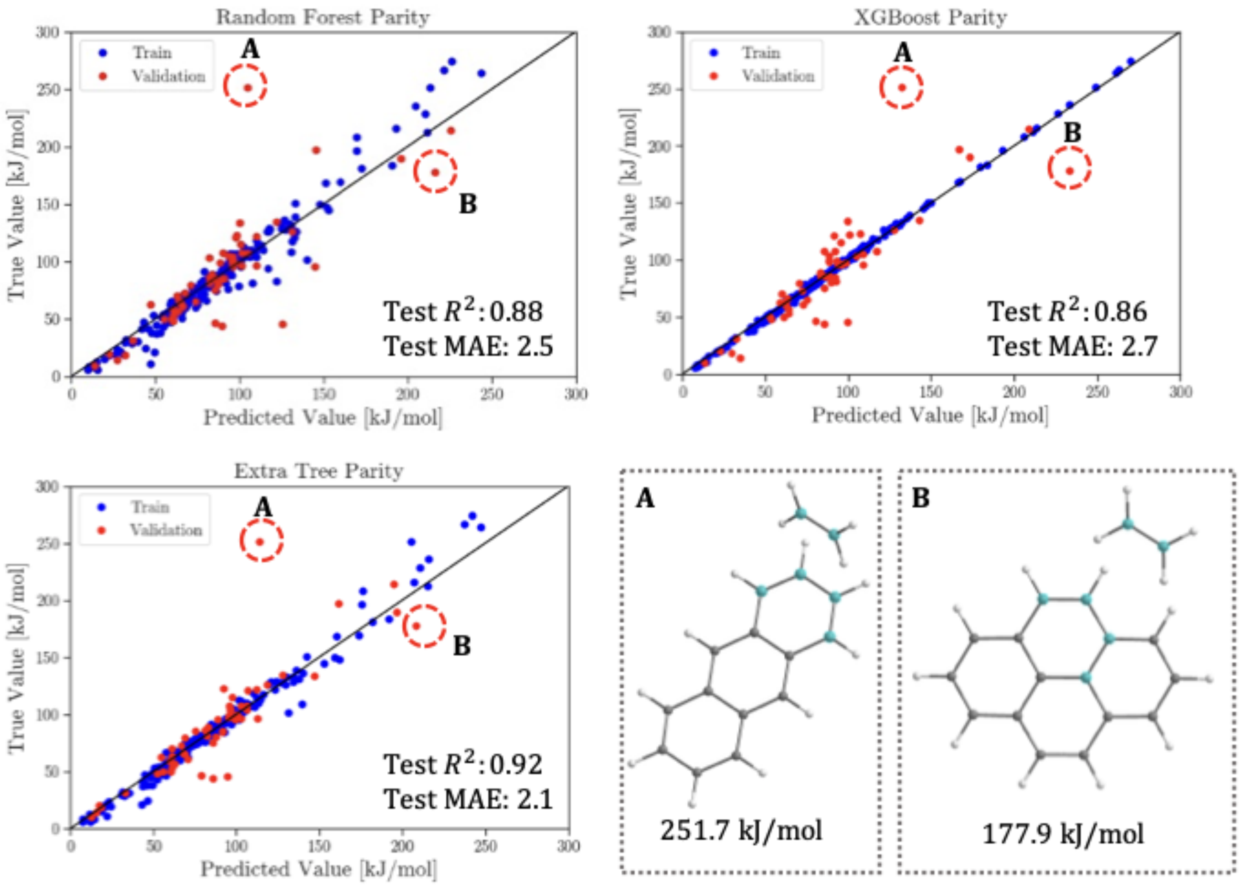


Figure 5).

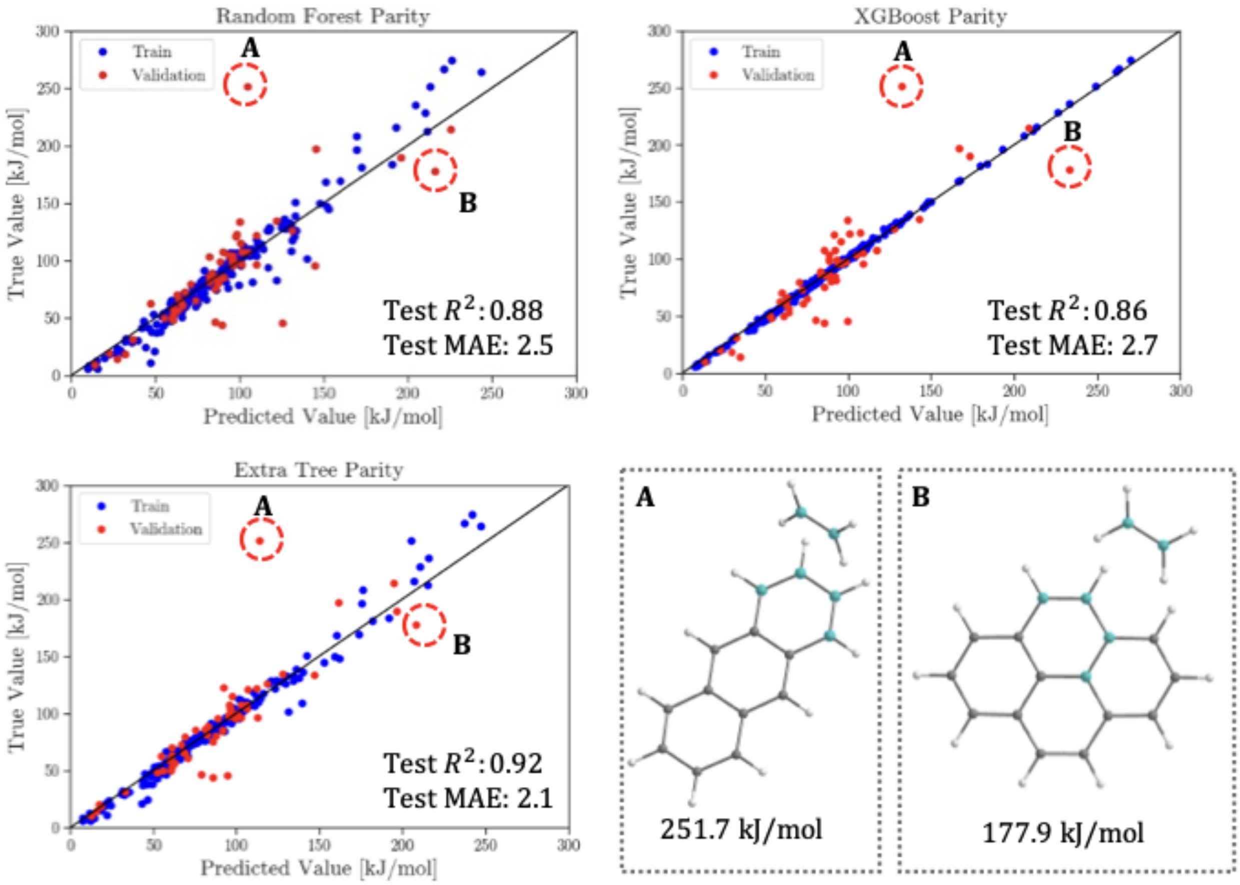


Figure 5. Parity plots of predicted and true dataset values of the reaction barrier for the top performing physical models. Circled in red are the two highest residuals for our testing dataset for the extra tree regressor (structures are shown in the bottom right) The descriptor set utilized here was the physical descriptor set.

Beyond training the best performing model, we wished to create a more general and physically intuitive regression algorithm, for predicting instances outside of our dataset. To do this, we completed sets of physical features labeled as 'Physical Feature Set' in Table 1, by adding back some of the physically meaningful though possibly correlated variables. For example, bond ellipticity, ϵ, was originally selected in 3 of the 4 BCPs as an important feature; in the completed/physical set of variables we included ϵ of all 4 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 dataset and determined that there was almost no loss in performance (Table S3). In general, these best performing algorithms were quick, 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 pairs and thereby predict the preferred reaction product of a Diels-Alder reaction. Our dataset contained a mixture of such reaction pairs but about half of the reactions available did not have the corresponding alternative reaction. In total, our dataset contained 61 endo/exo pairs or 122 compounds. This represents less than half of the total available dataset and therefore the process of training is more difficult. To fully extend this aim, likely more data would be required, but we nonetheless retrained the best model above, Extra Trees, with physical feature set, and an 80-20 train-validation split. Our splitting scheme kept endo/exo reaction pairs in the same dataset 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 Tree regression algorithm was able to correctly predict endo/exo ordering 70% of the time though this figure could likely be improved with more data.

Next, we wish to understand the limitations of our regression models, including regimes where its predictive ability falls short. From the top four regression algorithms, we noted two datapoints with barrier energies of 251.8 kJ/mol (60.23 kcal/mol) and 177.9 kJ/mol (42.56 kcal/mol) (circled outliers in

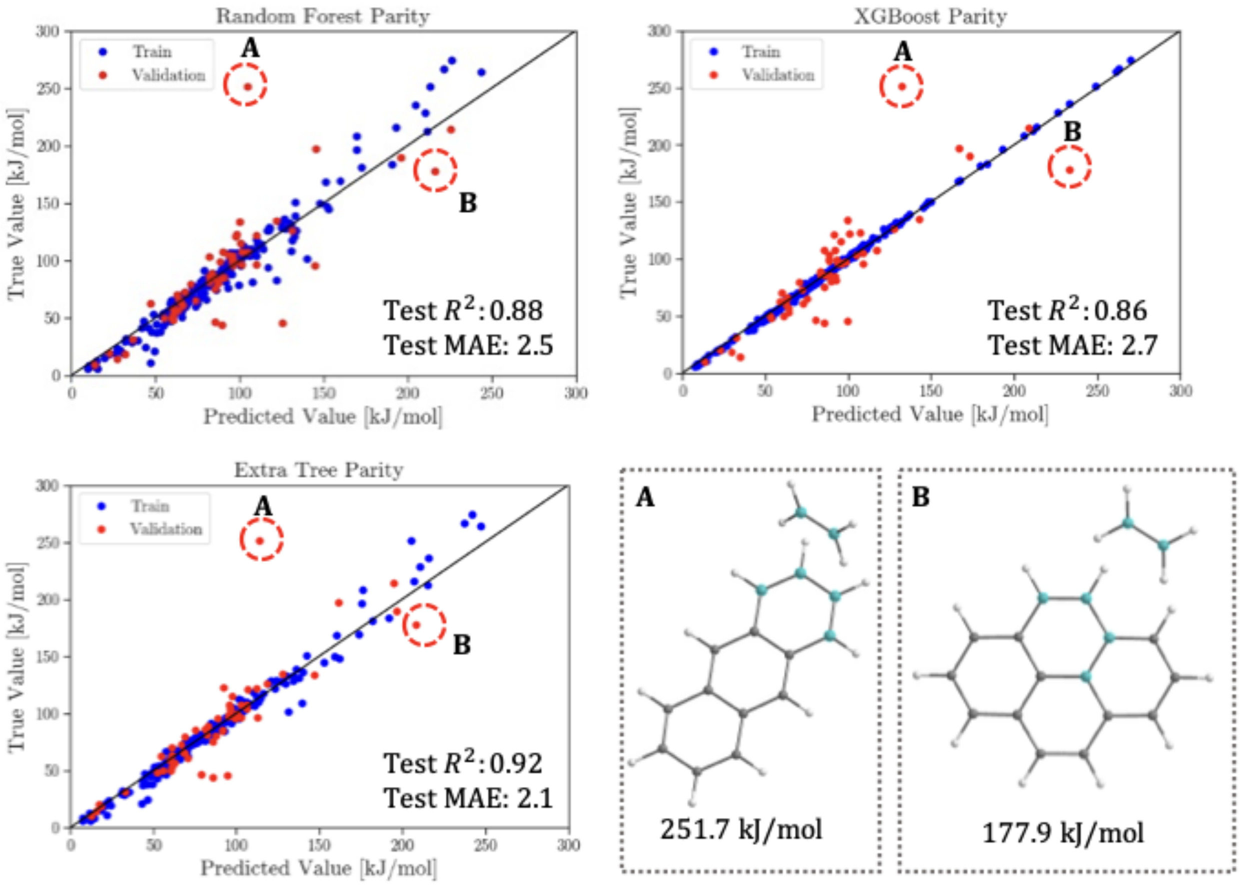


Figure 5) that contributed heavily to training loss in all instances. The consistently large error for predicting these values across different families of algorithms required further probing into the physical reasons yielding such poor performance. Firstly, these data points fall in the underrepresented high-barrier region, where the model might have had insufficient training instances.

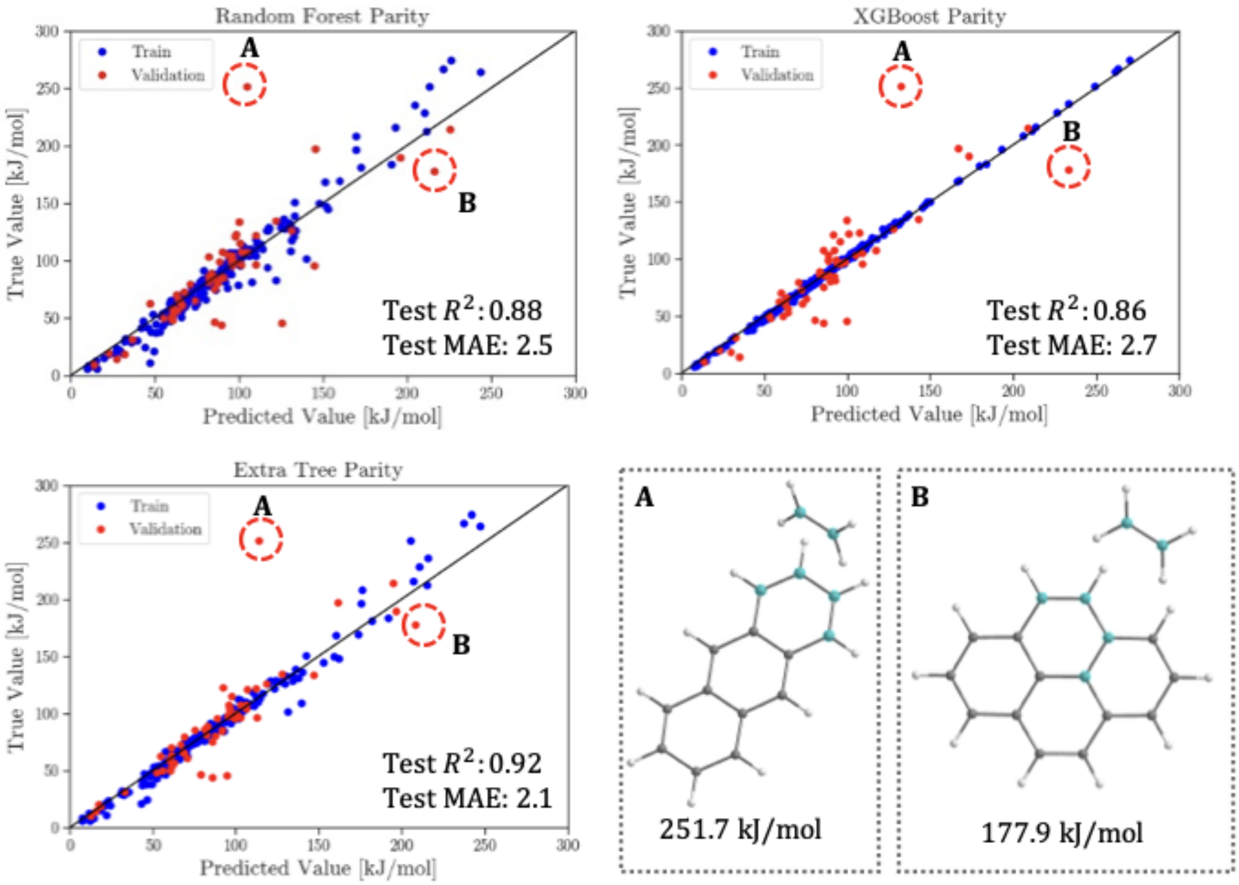


Figure 5 shows the two systems responsible for these two largest testing residuals. Notably, these systems involve dienes with more delocalized π-systems, and thus, the electronic density shifts during the Diels-Alder reaction within these systems extends over the entire conjugated π-system of the diene. 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, both in the training and test set data, but the two outliers feature the greatest extents of π-delocalization. It must be noted that QTAIM properties are computed on optimized reactant geometries, and 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 dataset also includes other regioisomers for the reaction occurring in

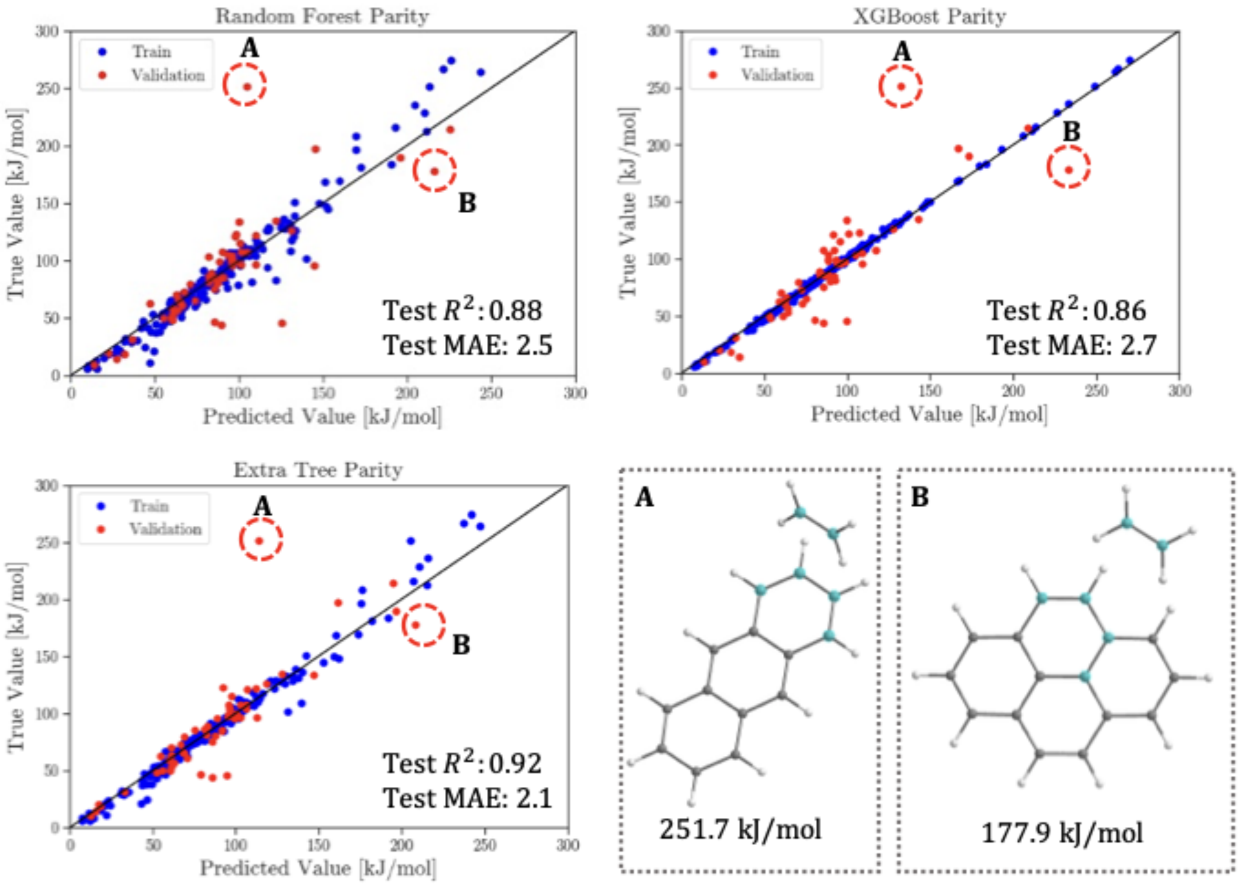


Figure 5A, with the preferred regioisomer being the [5,10] addition and the least preferred being the [12,14] addition (which is the reaction shown in

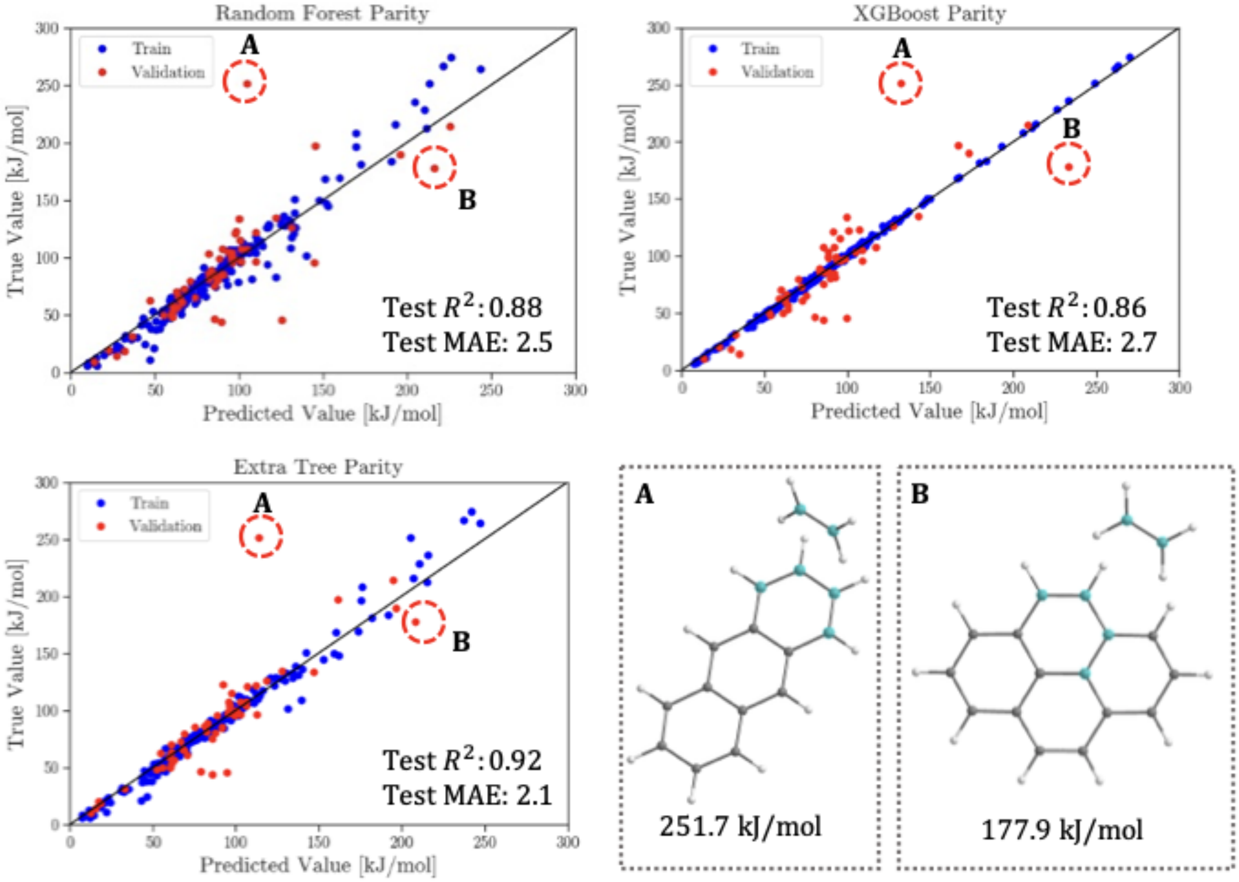
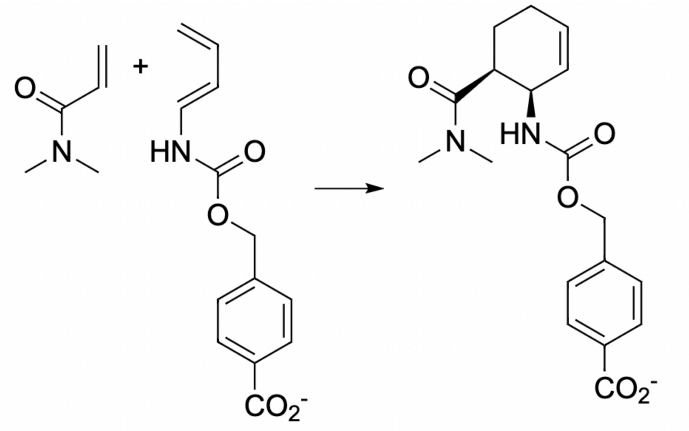


Figure 5A).[66] Upon testing with our best performing algorithm, Extra Trees, we can correctly predict that the [12,14] addition is still least preferred, and the [5,10] addition is most preferred. Hence, our algorithm, while it may not accurately predict the barrier for the [12,14] addition, still predicts the correct regioisomer.

Finally, we put the model to a stringent test, and probe its expandability to 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.[83] These enzymes catalyze the reaction between 4-carboxylbenzyl-*trans*-1,3-butadiene-1-carbamate and *N*,*N*-dimethylacrylamide (Scheme 1). 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 8 mutations between CE11 and CE20 with the majority being within the appended lid-element, and none within the active site (Figure 6). Therefore, these mutations represent realistic, subtle changes to the active site electron density topology brought about by distant point mutations through long-range interactions.



Scheme 1. 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.

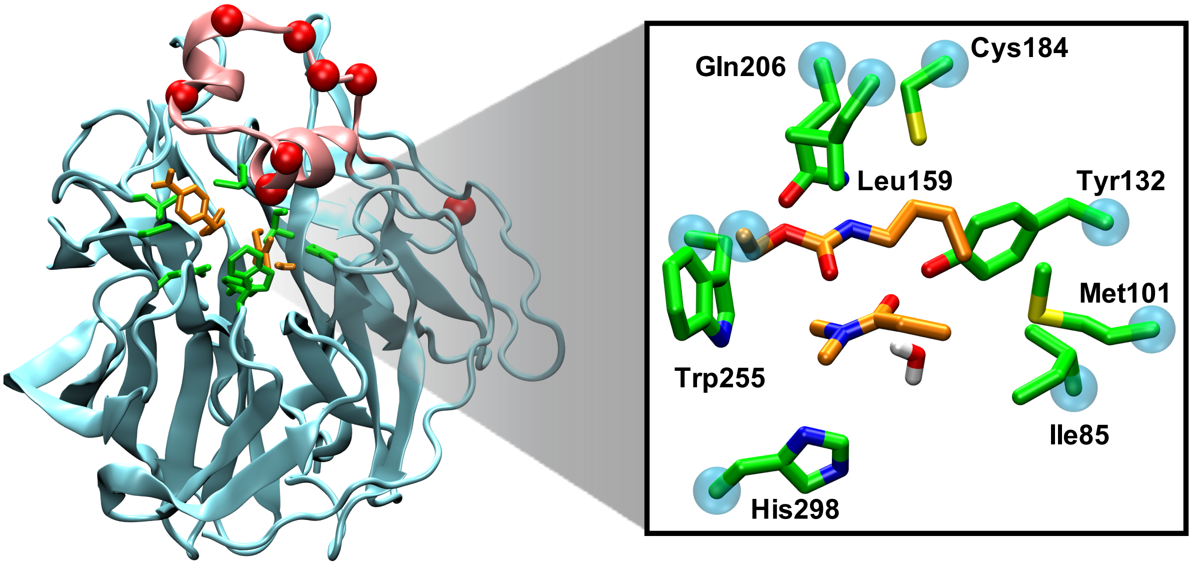


Figure 6. Left: CE20 crystal structure (PDB Code: 4O5T[83]) with the residues included in the QM active site colored green and the substrates in orange. The appended lid element is colored pink with the location of the mutations separating CE11 and CE20 shown as red spheres. The mutations are T43I, K44N, P48L, K53E, S55R, R56S, G57D, E113D. Right: the QM active site; the blue spheres represent C atoms replaced by capping hydrogens and frozen; note that only a part of the diene substrate was treated at the QM level.

We utilized our in-house quantum mechanical/discrete molecular dynamics[85] (QM/DMD) engine to perform sampling of the two protein variants with the bound substrates. The QM active site shown in Figure 6 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 tree 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's 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 was 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.[83] 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.

Conclusions

Here we showed that QTAIM descriptors based on the ground state electron density can be coupled to a supervised machine learning algorithm to effectively regress on predicting reaction barrier energies. Fundamentally, QTAIM appears to be an ideal tool for feature generation for machine learning, because it produces sets of physical and continuous descriptors. As a proof-of-concept, we showed it on the Diels-Alder reactions. We computed reaction barriers of a diverse array of Diels-Alder reactions from 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. Thus, this study alone could serve as a screening filter for experimental and computational studies on the Diels-Alder reaction. We believe that this concept could be extended to other important reactions in chemistry. 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 cheminformatics toolkit.

Future studies may include building classifier algorithms to bin reactions into high/low/middle barrier energies (or any arbitrary number of bins) 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 (Receiver Operating Characteristic) 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 computable 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[86]) 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, whether a CP is present or not provides a chemically significant, bonding information). 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 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 descriptor, 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.

Computational Methods

All QM calculations for the machine learning algorithm were performed in Gaussian 09.[87] Geometries were optimized with the B3LYP functional[88–91] and 6-31G* basis set.[92–94] 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.[95] 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 dataset from these values. QTAIM analysis of the electron density generated from Gaussian was performed using the AIMALL software.[96] Machine Learning analysis was performed using Sci-Learn and parameters were tuned using Skopt Bayesian Parameter optimization.[97,98] Here we opted for 20 different algorithms for each tunable algorithm parameter as well as a 5-fold cross validation for model selection.

A total of 5 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 following reference.[85] CE20 QM/DMD trajectories started from the 4O5T crystal structure.[83] 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)[99,100,109,101–108] with the pure meta-GGA TPSS functional[110] with D3 dispersion correction.[111] All atoms were treated with the double-zeta def2-SVP basis set.[112] The Conductor-like Screening Model (COSMO)[113] with a constant dielectric of 4 was used to approximate the screening and solvation effects from the protein scaffold in this buried active site.[114] πDMD[115,116] was used for DMD within QM/DMD. πDMD uses an implicit solvent along with discretized potentials.

Supporting Information Available

Full code, dataset, feature sets and definitions, detailed computational methods for machine learning algorithms implemented, parity plots, correlation matrices, permutation importance, Diels-Alderase QM/DMD RMSD and energy plots, Diels-Alderase mutations, and structures.

Acknowledgement

This work was supported by the National Science Foundation grant CHE-1903808 and used the Extreme Science and Engineering Discovery Environment (XSEDE) Bridges system at the Pittsburgh Supercomputing Center (PSC). We also acknowledge the UCLA-IDRE cluster Hoffman2 computational resources. SV would also like to thank U.S. Department of Energy, Office of Science, Office of Advanced Scientific Computing Research, Department of Energy Computational Science Graduate Fellowship who supported this work under Award Number DE-SC0021110.

References

(1) Hratchian, H. P.; Schlegel, H. B. Finding Minima, Transition States, and Following Reaction Pathways on Ab Initio Potential Energy Surfaces. In *Theory and Applications of Computational Chemistry*; Elsevier, 2005; pp 195–249. https://doi.org/10.1016/B978-044451719-7/50053-6.

(2) Laidler, K. J.; King, M. C. Development of Transition-State Theory. *J. Phys. Chem.* **1983**, *87* (15), 2657–2664.

(3) Pechukas, P. Transition State Theory. *Annu. Rev. Phys. Chem.* **1981**, *32* (1), 159–177.

(4) Lienhard, G. E. Enzymatic Catalysis and Transition-State Theory. *Science (80-.).* **1973**, *180* (4082), 149–154. https://doi.org/10.1126/science.180.4082.149.

(5) Truhlar, D. G.; Hase, W. L.; Hynes, J. T. Current Status of Transition-State Theory. *J. Phys. Chem.* **1983**, *87* (15), 2664–2682.

(6) Marcus, R. A. On the Theory of Oxidation-Reduction Reactions Involving Electron Transfer. I. *J. Chem. Phys.* **1956**, *24* (5), 966–978. https://doi.org/10.1063/1.1742723.

(7) Marcus, R. A. On the Theory of Oxidation-Reduction Reactions Involving Electron Transfer. II. Applications to Data on the Rates of Isotopic Exchange Reactions. *J. Chem. Phys.* **1957**, *26* (4), 867–871.

(8) Marcus, R. A. On the Theory of Oxidation-Reduction Reactions Involving Electron Transfer. III. Applications to Data on the Rates of Organic Redox Reactions. *J. Chem. Phys.* **1957**, *26* (4), 872–877.

(9) Kästner, J.; Sherwood, P. Superlinearly Converging Dimer Method for Transition State Search. *J. Chem. Phys.* **2008**, *128* (1), 014106. https://doi.org/10.1063/1.2815812.

(10) Galván, I. F.; Field, M. J. Improving the Efficiency of the NEB Reaction Path Finding Algorithm. *J. Comput. Chem.* **2008**, *29* (1), 139–143. https://doi.org/10.1002/jcc.20780.

(11) Quapp, W. Finding the Transition State without Initial Guess: The Growing String Method for Newton Trajectory to Isomerization and Enantiomerization Reaction of Alanine Dipeptide and Poly(15)Alanine. *J. Comput. Chem.* **2007**, *28* (11), 1834–1847. https://doi.org/10.1002/jcc.20688.

(12) E, W.; Ren, W.; Vanden-Eijnden, E. Simplified and Improved String Method for Computing the Minimum Energy Paths in Barrier-Crossing Events. *J. Chem. Phys.* **2007**, *126* (16), 164103. https://doi.org/10.1063/1.2720838.

(13) Reddy, C. K.; Chiang, H.-D. A Stability Boundary Based Method for Finding Saddle Points on Potential Energy Surfaces. *J. Comput. Biol.* **2006**, *13* (3), 745–766. https://doi.org/10.1089/cmb.2006.13.745.

(14) González-García, N.; Pu, J.; González-Lafont, À.; Lluch, J. M.; Truhlar, D. G. Searching for Saddle Points by Using the Nudged Elastic Band Method: An Implementation for Gas-Phase Systems. *J. Chem. Theory Comput.* **2006**, *2* (4), 895–904. https://doi.org/10.1021/ct060032y.

(15) Maeda, S.; Ohno, K. A New Approach for Finding a Transition State Connecting a Reactant and a Product without Initial Guess: Applications of the Scaled Hypersphere Search Method to Isomerization Reactions of HCN, (H2O)2, and Alanine Dipeptide. *Chem. Phys. Lett.* **2005**, *404* (1–3), 95–99. https://doi.org/10.1016/j.cplett.2005.01.068.

(16) Heyden, A.; Bell, A. T.; Keil, F. J. Efficient Methods for Finding Transition States in Chemical Reactions: Comparison of Improved Dimer Method and Partitioned Rational Function Optimization Method. *J. Chem. Phys.* **2005**, *123* (22), 224101. https://doi.org/10.1063/1.2104507.

(17) Carr, J. M.; Trygubenko, S. A.; Wales, D. J. Finding Pathways between Distant Local Minima. *J. Chem. Phys.* **2005**, *122* (23), 234903. https://doi.org/10.1063/1.1931587.

(18) Olsen, R. A.; Kroes, G. J.; Henkelman, G.; Arnaldsson, A.; Jónsson, H. Comparison of Methods for Finding Saddle Points without Knowledge of the Final States. *J. Chem. Phys.* **2004**, *121* (20), 9776–9792. https://doi.org/10.1063/1.1809574.

(19) Ohno, K.; Maeda, S. A Scaled Hypersphere Search Method for the Topography of Reaction Pathways on the Potential Energy Surface. *Chem. Phys. Lett.* **2004**, *384* (4–6), 277–282. https://doi.org/10.1016/j.cplett.2003.12.030.

(20) Peters, B.; Heyden, A.; Bell, A. T.; Chakraborty, A. A Growing String Method for Determining Transition States: Comparison to the Nudged Elastic Band and String Methods. *J. Chem. Phys.* **2004**, *120* (17), 7877–7886. https://doi.org/10.1063/1.1691018.

(21) Trygubenko, S. A.; Wales, D. J. A Doubly Nudged Elastic Band Method for Finding Transition States. *J. Chem. Phys.* **2004**, *120* (5), 2082–2094. https://doi.org/10.1063/1.1636455.

(22) Xie, L.; Liu, H.; Yang, W. Adapting the Nudged Elastic Band Method for Determining Minimum-Energy Paths of Chemical Reactions in Enzymes. *J. Chem. Phys.* **2004**, *120*

(17), 8039–8052. https://doi.org/10.1063/1.1691404.

(23) Govind, N.; Petersen, M.; Fitzgerald, G.; King-Smith, D.; Andzelm, J. A Generalized Synchronous Transit Method for Transition State Location. *Comput. Mater. Sci.* **2003**, *28* (2), 250–258. https://doi.org/10.1016/S0927-0256(03)00111-3.

(24) Alfonso, D. R.; Jordan, K. D. A Flexible Nudged Elastic Band Program for Optimization of Minimum Energy Pathways Usingab Initio Electronic Structure Methods. *J. Comput. Chem.* **2003**, *24* (8), 990–996. https://doi.org/10.1002/jcc.10233.

(25) Crehuet, R.; Field, M. J. A Temperature-Dependent Nudged-Elastic-Band Algorithm. *J. Chem. Phys.* **2003**, *118* (21), 9563–9571. https://doi.org/10.1063/1.1571817.

(26) Chu, J.-W.; Trout, B. L.; Brooks, B. R. A Super-Linear Minimization Scheme for the Nudged Elastic Band Method. *J. Chem. Phys.* **2003**, *119* (24), 12708–12717. https://doi.org/10.1063/1.1627754.

(27) Maragakis, P.; Andreev, S. A.; Brumer, Y.; Reichman, D. R.; Kaxiras, E. Adaptive Nudged Elastic Band Approach for Transition State Calculation. *J. Chem. Phys.* **2002**, *117* (10), 4651–4658. https://doi.org/10.1063/1.1495401.

(28) Bofill, J. M.; Anglada, J. M. Finding Transition States Using Reduced Potential-Energy Surfaces. *Theor. Chem. Acc.* **2001**, *105* (6), 463–472. https://doi.org/10.1007/s002140000252.

(29) Chaudhury, P.; Bhattacharyya, S. P.; Quapp, W. A Genetic Algorithm Based Technique for Locating First-Order Saddle Point Using a Gradient Dominated Recipe. *Chem. Phys.* **2000**, *253* (2–3), 295–303. https://doi.org/10.1016/S0301-0104(00)00010-0.

(30) Chaudhury, P.; Bhattacharyya, S. . Locating Critical Points on Multi-Dimensional Surfaces by Genetic Algorithm: Test Cases Including Normal and Perturbed Argon Clusters. *Chem. Phys.* **1999**, *241* (3), 313–325. https://doi.org/10.1016/S0301-0104(98)00414-5.

(31) Dellago, C.; Bolhuis, P. G.; Csajka, F. S.; Chandler, D. Transition Path Sampling and the Calculation of Rate Constants. *J. Chem. Phys.* **1998**, *108* (5), 1964–1977. https://doi.org/10.1063/1.475562.

(32) Ayala, P. Y.; Schlegel, H. B. A Combined Method for Determining Reaction Paths, Minima, and Transition State Geometries. *J. Chem. Phys.* **1997**, *107* (2), 375–384. https://doi.org/10.1063/1.474398.

(33) Ulitsky, A.; Shalloway, D. Finding Transition States Using Contangency Curves. *J. Chem. Phys.* **1997**, *106* (24), 10099–10104. https://doi.org/10.1063/1.474043.

(34) Quapp, W. A Gradient-Only Algorithm for Tracing a Reaction Path Uphill to the Saddle of a Potential Energy Surface. *Chem. Phys. Lett.* **1996**, *253* (3–4), 286–292. https://doi.org/10.1016/0009-2614(96)00255-2.

(35) Cárdenas-Lailhacar, C.; Zerner, M. C. Searching for Transition States: The Line - Then - Plane (LTP) Approach. *Int. J. Quantum Chem.* **1995**, *55* (6), 429–439. https://doi.org/10.1002/qua.560550602.

(36) Ionova, I. V.; Carter, E. A. Direct Inversion in the Iterative Subspace-induced Acceleration of the Ridge Method for Finding Transition States. *J. Chem. Phys.* **1995**, *103* (13), 5437–5441. https://doi.org/10.1063/1.470579.

(37) Smart, O. S. A New Method to Calculate Reaction Paths for Conformation Transitions of Large Molecules. *Chem. Phys. Lett.* **1994**, *222* (5), 503–512. https://doi.org/10.1016/0009-2614(94)00374-2.

(38) Sun, J.; Ruedenberg, K. Locating Transition States by Quadratic Image Gradient Descent on Potential Energy Surfaces. *J. Chem. Phys.* **1994**, *101* (3), 2157–2167. https://doi.org/10.1063/1.467721.

(39) Ionova, I. V.; Carter, E. A. Ridge Method for Finding Saddle Points on Potential Energy Surfaces. *J. Chem. Phys.* **1993**, *98* (8), 6377–6386. https://doi.org/10.1063/1.465100.

(40) Helgaker, T. Transition-State Optimizations by Trust-Region Image Minimization. *Chem. Phys. Lett.* **1991**, *182* (5), 503–510. https://doi.org/10.1016/0009-2614(91)90115-P.

(41) Nichols, J.; Taylor, H.; Schmidt, P.; Simons, J. Walking on Potential Energy Surfaces. *J. Chem. Phys.* **1990**, *92* (1), 340–346. https://doi.org/10.1063/1.458435.

(42) Ulitsky, A.; Elber, R. A New Technique to Calculate Steepest Descent Paths in Flexible Polyatomic Systems. *J. Chem. Phys.* **1990**, *92* (2), 1510–1511. https://doi.org/10.1063/1.458112.

(43) Berry, R. S.; Davis, H. L.; Beck, T. L. Finding Saddles on Multidimensional Potential Surfaces. *Chem. Phys. Lett.* **1988**, *147* (1), 13–17. https://doi.org/10.1016/0009-2614(88)80215-X.

(44) McDouall, J. J. W.; Robb, M. A.; Bernardi, F. An Efficient Algorithm for the Approximate Location of Transition Structures in a Diabatic Surface Formalism. *Chem. Phys. Lett.* **1986**, *129* (6), 595–602. https://doi.org/10.1016/0009-2614(86)80407-9.

(45) Bell, S.; Crighton, J. S. Locating Transition States. *J. Chem. Phys.* **1984**, *80* (6), 2464–2475. https://doi.org/10.1063/1.446996.

(46) Schlegel, H. B. Optimization of Equilibrium Geometries and Transition Structures. *J. Comput. Chem.* **1982**, *3* (2), 214–218. https://doi.org/10.1002/jcc.540030212.

(47) Cerjan, C. J.; Miller, W. H. On Finding Transition States. *J. Chem. Phys.* **1981**, *75* (6), 2800–2806. https://doi.org/10.1063/1.442352.

(48) Rothman, M. J.; Lohr, L. L. Analysis of an Energy Minimization Method for Locating Transition States on Potential Energy Hypersurfaces. *Chem. Phys. Lett.* **1980**, *70* (2), 405–409. https://doi.org/10.1016/0009-2614(80)85361-9.

(49) Müller, K. Reaction Paths on Multidimensional Energy Hypersurfaces. *Angew. Chemie Int. Ed. English* **1980**, *19* (1), 1–13. https://doi.org/10.1002/anie.198000013.

(50) Müller, K.; Brown, L. D. Location of Saddle Points and Minimum Energy Paths by a Constrained Simplex Optimization Procedure. *Theor. Chim. Acta* **1979**, *53* (1), 75–93. https://doi.org/10.1007/BF00547608.

(51) Halgren, T. A.; Lipscomb, W. N. The Synchronous-Transit Method for Determining Reaction Pathways and Locating Molecular Transition States. *Chem. Phys. Lett.* **1977**, *49* (2), 225–232. https://doi.org/10.1016/0009-2614(77)80574-5.

(52) Poppinger, D. On the Calculation of Transition States. *Chem. Phys. Lett.* **1975**, *35* (4), 550–554. https://doi.org/10.1016/0009-2614(75)85665-X.

(53) Jacobson, L. D.; Bochevarov, A. D.; Watson, M. A.; Hughes, T. F.; Rinaldo, D.; Ehrlich, S.; Steinbrecher, T. B.; Vaitheeswaran, S.; Philipp, D. M.; Halls, M. D.; Friesner, R. A. Automated Transition State Search and Its Application to Diverse Types of Organic Reactions. *J. Chem. Theory Comput.* **2017**, *13* (11), 5780–5797. https://doi.org/10.1021/acs.jctc.7b00764.

(54) Friederich, P.; dos Passos Gomes, G.; De Bin, R.; Aspuru-Guzik, A.; Balcells, D. Machine Learning Dihydrogen Activation in the Chemical Space Surrounding Vaska's Complex. *Chemical Science*. 2020, pp 4584–4601. https://doi.org/10.1039/d0sc00445f.

(55) Jinnouchi, R.; Asahi, R. Predicting Catalytic Activity of Nanoparticles by a DFT-Aided Machine-Learning Algorithm. *J. Phys. Chem. Lett.* **2017**, *8* (17), 4279–4283. https://doi.org/10.1021/acs.jpclett.7b02010.

(56) Meyer, B.; Sawatlon, B.; Heinen, S.; von Lilienfeld, O. A.; Corminboeuf, C. Machine Learning Meets Volcano Plots: Computational Discovery of Cross-Coupling Catalysts. *Chem. Sci.* **2018**, *9* (35), 7069–7077. https://doi.org/10.1039/C8SC01949E.

(57) Singh, A. R.; Rohr, B. A.; Gauthier, J. A.; Nørskov, J. K. Predicting Chemical Reaction Barriers with a Machine Learning Model. *Catal. Letters* **2019**, *149* (9), 2347–2354. https://doi.org/10.1007/s10562-019-02705-x.

(58) von der Esch, B.; Dietschreit, J. C. B.; Peters, L. D. M.; Ochsenfeld, C. Finding Reactive Configurations: A Machine Learning Approach for Estimating Energy Barriers Applied to Sirtuin 5. *J. Chem. Theory Comput.* **2019**, *15* (12), 6660–6667. https://doi.org/10.1021/acs.jctc.9b00876.

(59) Fuller III, J.; Wilson, T. R.; Eberhart, M. E.; Alexandrova, A. N.; Fuller, J.; Wilson, T. R.; Eberhart, M. E.; Alexandrova, A. N. Charge Density in Enzyme Active Site as a Descriptor of Electrostatic Preorganization. *J. Chem. Inf. Model.* **2019**, *59* (5), 2367–2373. https://doi.org/10.1021/acs.jcim.8b00958.

(60) Hennefarth, M. R.; Alexandrova, A. N. Heterogeneous Intramolecular Electric Field as a Descriptor of Diels–Alder Reactivity. *J. Phys. Chem. A* **2021**, acs.jpca.1c00181. https://doi.org/10.1021/acs.jpca.1c00181.

(61) Matta, C. F.; Arabi, A. A. Electron-Density Descriptors as Predictors in Quantitative Structure-Activity/Property Relationships and Drug Design. *Future Medicinal Chemistry*. Future Med Chem June 2011, pp 969–994. https://doi.org/10.4155/fmc.11.65.

(62) Jones, T. E.; Eberhart, M. E.; Imlay, S.; Mackey, C. Bond Bundles and the Origins of Functionality. *J. Phys. Chem. A* **2011**, *115* (45), 12582–12585. https://doi.org/10.1021/jp203013r.

(63) Knoerr, E. H.; Eberhart, M. E. Toward a Density-Based Representation of Reactivity: S N

2 Reaction. *J. Phys. Chem. A* **2001**, *105* (5), 880–884. https://doi.org/10.1021/jp0028711.

(64) Morgenstern, A.; Morgenstern, C.; Miorelli, J.; Wilson, T.; Eberhart, M. E. The Influence of Zero-Flux Surface Motion on Chemical Reactivity. *Phys. Chem. Chem. Phys.* **2016**, *18* (7), 5638–5646. https://doi.org/10.1039/C5CP07852K.

(65) Hohenberg, P.; Kohn, W. Inhomogeneous Electron Gas. *Phys. Rev.* **1964**, *136* (3B), B864–B871. https://doi.org/10.1103/PhysRev.136.B864.

(66) Hayden, A. E.; Houk, K. N. Transition State Distortion Energies Correlate with Activation Energies of 1,4-Dihydrogenations and Diels-Alder Cycloadditions of Aromatic Molecules. *J. Am. Chem. Soc.* **2009**, *131* (11), 4084–4089. https://doi.org/10.1021/ja809142x.

(67) James, N. C.; Um, J. M.; Padias, A. B.; Hall, H. K.; Houk, K. N. 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*. 2013, pp 6582–6592. https://doi.org/10.1021/jo400900x.

(68) Liu, F.; Paton, R. S.; Kim, S.; Liang, Y.; Houk, K. N. 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*. 2013, pp 15642–15649. https://doi.org/10.1021/ja408437u.

(69) Liu, F.; Liang, Y.; Houk, K. N. Theoretical Elucidation of the Origins of Substituent and Strain Effects on the Rates of Diels-Alder Reactions of 1,2,4,5-Tetrazines. *J. Am. Chem. Soc.* **2014**, *136* (32), 11483–11493. https://doi.org/10.1021/ja505569a.

(70) Gordillo, R.; Houk, K. N. Origins of Stereoselectivity in Diels−Alder Cycloadditions Catalyzed by Chiral Imidazolidinones. *Journal of the American Chemical Society*. 2006, pp 3543–3553. https://doi.org/10.1021/ja0525859.

(71) Paton, R. S.; Kim, S.; Ross, A. G.; Danishefsky, S. J.; Houk, K. N. Experimental Diels-Alder Reactivities of Cycloalkenones and Cyclic Dienes Explained through Transition-State Distortion Energies. *Angewandte Chemie*. 2011, pp 10550–10552. https://doi.org/10.1002/ange.201103998.

(72) Gordillo, R.; Dudding, T.; Anderson, C. D.; Houk, K. N. Hydrogen Bonding Catalysis Operates by Charge Stabilization in Highly Polar Diels - Alder Reactions. *Org. Lett.* **2007**, *9* (3), 501–503. https://doi.org/10.1021/ol0629925.

(73) Jones, G. O.; Guner, V. A.; Houk, K. N. 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. *J. Phys. Chem. A* **2006**, *110* (4), 1216–1224. https://doi.org/10.1021/jp052055z.

(74) Paton, R. S.; Mackey, J. L.; Kim, W. H.; Lee, J. H.; Danishefsky, S. J.; Houk, K. N. Origins of Stereoselectivity in the Trans Diels-Alder Paradigm. *J. Am. Chem. Soc.* **2010**, *132* (27), 9335–9340. https://doi.org/10.1021/ja1009162.

(75) Lam, Y.-H.; Cheong, P. H.-Y.; Blasco Mata, J. M.; Stanway, S. J.; Gouverneur, V.; Houk, K. N. Diels-Alder Exo Selectivity in Terminal-Substituted Dienes and Dienophiles: Experimental Discoveries and Computational Explanations. *J. Am. Chem. Soc.* **2009**, *131*

(5), 1947–1957. https://doi.org/10.1021/ja8079548.

(76) Levandowski, B. J.; Houk, K. N. 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*. 2016, pp 16731–16736. https://doi.org/10.1021/jacs.6b10463.

(77) Levandowski, B. J.; Zou, L.; Houk, K. N. Hyperconjugative Aromaticity and Antiaromaticity Control the Reactivities and π-Facial Stereoselectivities of 5-Substituted Cyclopentadiene Diels–Alder Cycloadditions. *The Journal of Organic Chemistry*. 2018, pp 14658–14666. https://doi.org/10.1021/acs.joc.8b02537.

(78) Çelebi-Ölçüm, N.; Ess, D. H.; Aviyente, V.; Houk, K. N. Lewis Acid Catalysis Alters the Shapes and Products of Bis-Pericyclic Diels - Alder Transition States. *J. Am. Chem. Soc.* **2007**, *129* (15), 4528–4529. https://doi.org/10.1021/ja070686w.

(79) Pieniazek, S. N.; Houk, K. N. The Origin of the Halogen Effect on Reactivity and Reversibility of Diels–Alder Cycloadditions Involving Furan. *Angewandte Chemie International Edition*. 2006, pp 1442–1445. https://doi.org/10.1002/anie.200502677.

(80) Bader, R. F. W. *The Quantum Theory of Atoms in Molecules*; Matta, C. F., Boyd, R. J., Eds.; Wiley, 2007. https://doi.org/10.1002/9783527610709.

(81) Bader, R. F. W. Quantum Topology of Molecular Charge Distributions. III. The Mechanics of an Atom in a Molecule. *J. Chem. Phys.* **1980**, *73* (6), 2871–2883. https://doi.org/10.1063/1.440457.

(82) Bader, R. F. W. Atoms in Molecules. *Acc. Chem. Res.* **1985**, *18* (1), 9–15. https://doi.org/10.1021/ar00109a003.

(83) Preiswerk, N.; Beck, T.; Schulz, J. D.; Milovnik, P.; Mayer, C.; Siegel, J. B.; Baker, D.; Hilvert, D. Impact of Scaffold Rigidity on the Design and Evolution of an Artificial Diels-Alderase. *Proc. Natl. Acad. Sci.* **2014**, *111* (22), 8013–8018. https://doi.org/10.1073/pnas.1401073111.

(84) Yoo, W.; Mayberry, R.; Bae, S.; Singh, K.; Peter He, Q.; Lillard, J. W. A Study of Effects of MultiCollinearity in the Multivariable Analysis. *Int. J. Appl. Sci. Technol.* **2014**, *4* (5), 9–19.

(85) Sparta, M.; Shirvanyants, D.; Ding, F.; Dokholyan, N. V.; Alexandrova, A. N. Hybrid Dynamics Simulation Engine for Metalloproteins. *Biophys. J.* **2012**, *103* (4), 767–776. https://doi.org/10.1016/j.bpj.2012.06.024.

(86) Thom, R. Stabilité Structurelle et Morphogénèse–Essai d'une Théorie Générale Des Modèles. *Read. Mass* **1972**.

(87) Frisch, M.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Mennucci, B.; Petersson, Ga. Gaussian 09, Revision D. 01. Gaussian, Inc., Wallingford CT 2009.

(88) Becke, A. D. Density-functional Thermochemistry. III. The Role of Exact Exchange. *J. Chem. Phys.* **1993**, *98* (7), 5648–5652. https://doi.org/10.1063/1.464913.

(89) Lee, C.; Yang, W.; Parr, R. G. Development of the Colle-Salvetti Correlation-Energy Formula into a Functional of the Electron Density. *Phys. Rev. B* **1988**, *37* (2), 785–789. https://doi.org/10.1103/PhysRevB.37.785.

(90) Vosko, S. H.; Wilk, L.; Nusair, M. Accurate Spin-Dependent Electron Liquid Correlation Energies for Local Spin Density Calculations: A Critical Analysis. *Can. J. Phys.* **1980**, *58* (8), 1200–1211. https://doi.org/10.1139/p80-159.

(91) Stephens, P. J.; Devlin, F. J.; Chabalowski, C. F.; Frisch, M. J. Ab Initio Calculation of Vibrational Absorption and Circular Dichroism Spectra Using Density Functional Force Fields. *J. Phys. Chem.* **1994**, *98* (45), 11623–11627. https://doi.org/10.1021/j100096a001.

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

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

(94) Francl, M. M.; Pietro, W. J.; Hehre, W. J.; Binkley, J. S.; Gordon, M. S.; DeFrees, D. J.; Pople, J. A. Self-consistent Molecular Orbital Methods. XXIII. A Polarization-type Basis Set for Second-row Elements. *J. Chem. Phys.* **1982**, *77* (7), 3654–3665. https://doi.org/10.1063/1.444267.

(95) Ess, D. H.; Jones, G. O.; Houk, K. N. Conceptual, Qualitative, and Quantitative Theories of 1,3-Dipolar and Diels–Alder Cycloadditions Used in Synthesis. *Adv. Synth. Catal.* **2006**, *348* (16–17), 2337–2361. https://doi.org/10.1002/adsc.200600431.

(96) Keith, T. A. AIMALL, TK Gristmill Software. *Overl. Park KS, USA,(aim. tkgristmill. com)* **2016**.

(97) Head, T.; Kumar, M.; Nahrstaedt, H.; Louppe, G.; Shcherbatyi, I. Scikit-Optimize/Scikit-Optimize. **2020**. https://doi.org/10.5281/zenodo.4014775.

(98) Garreta, R.; Moncecchi, G. *Learning Scikit-Learn: Machine Learning in Python*; Packt Publishing Ltd, 2013.

(99) Ahlrichs, R.; Bär, M.; Häser, M.; Horn, H.; Kölmel, C. Electronic Structure Calculations on Workstation Computers: The Program System Turbomole. *Chem. Phys. Lett.* **1989**, *162* (3), 165–169. https://doi.org/10.1016/0009-2614(89)85118-8.

(100) Häser, M.; Ahlrichs, R. Improvements on the Direct SCF Method. *J. Comput. Chem.* **1989**, *10* (1), 104–111. https://doi.org/10.1002/jcc.540100111.

(101) Treutler, O.; Ahlrichs, R. Efficient Molecular Numerical Integration Schemes. *J. Chem. Phys.* **1995**, *102* (1), 346–354. https://doi.org/10.1063/1.469408.

(102) Eichkorn, K.; Weigend, F.; Treutler, O.; Ahlrichs, R. Auxiliary Basis Sets for Main Row Atoms and Transition Metals and Their Use to Approximate Coulomb Potentials. *Theor. Chem. Accounts Theory, Comput. Model. (Theoretica Chim. Acta)* **1997**, *97* (1–4), 119–

124. https://doi.org/10.1007/s002140050244.

(103) Eichkorn, K.; Treutler, O.; Öhm, H.; Häser, M.; Ahlrichs, R. Auxiliary Basis Sets to Approximate Coulomb Potentials. *Chem. Phys. Lett.* **1995**, *240* (4), 283–290. https://doi.org/10.1016/0009-2614(95)00621-A.

(104) Weigend, F. Accurate Coulomb-Fitting Basis Sets for H to Rn. *Phys. Chem. Chem. Phys.* **2006**, *8* (9), 1057. https://doi.org/10.1039/b515623h.

(105) Sierka, M.; Hogekamp, A.; Ahlrichs, R. Fast Evaluation of the Coulomb Potential for Electron Densities Using Multipole Accelerated Resolution of Identity Approximation. *J. Chem. Phys.* **2003**, *118* (20), 9136–9148. https://doi.org/10.1063/1.1567253.

(106) Deglmann, P.; May, K.; Furche, F.; Ahlrichs, R. Nuclear Second Analytical Derivative Calculations Using Auxiliary Basis Set Expansions. *Chem. Phys. Lett.* **2004**, *384* (1–3), 103–107. https://doi.org/10.1016/j.cplett.2003.11.080.

(107) Von Arnim, M.; Ahlrichs, R. Performance of Parallel TURBOMOLE for Density Functional Calculations. *J. Comput. Chem.* **1998**, *19* (15), 1746–1757. https://doi.org/10.1002/(SICI)1096-987X(19981130)19:15<1746::AID-JCC7>3.0.CO;2-N.

(108) von Arnim, M.; Ahlrichs, R. Geometry Optimization in Generalized Natural Internal Coordinates. *J. Chem. Phys.* **1999**, *111* (20), 9183–9190. https://doi.org/10.1063/1.479510.

(109) Ahlrichs, R. Efficient Evaluation of Three-Center Two-Electron Integrals over Gaussian Functions. *Phys. Chem. Chem. Phys.* **2004**, *6* (22), 5119. https://doi.org/10.1039/b413539c.

(110) Hellweg, A.; Grün, S. A.; Hättig, C. Benchmarking the Performance of Spin-Component Scaled CC2 in Ground and Electronically Excited States. *Phys. Chem. Chem. Phys.* **2008**, *10* (28), 4119. https://doi.org/10.1039/b803727b.

(111) Grimme, S.; Antony, J.; Ehrlich, S.; Krieg, H. A Consistent and Accurate Ab Initio Parametrization of Density Functional Dispersion Correction (DFT-D) for the 94 Elements H-Pu. *J. Chem. Phys.* **2010**, *132* (15), 154104. https://doi.org/10.1063/1.3382344.

(112) Weigend, F.; Ahlrichs, R. Balanced Basis Sets of Split Valence, Triple Zeta Valence and Quadruple Zeta Valence Quality for H to Rn: Design and Assessment of Accuracy. *Phys. Chem. Chem. Phys.* **2005**, *7* (18), 3297. https://doi.org/10.1039/b508541a.

(113) Klamt, A.; Schüürmann, G. 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* **1993**, *0* (5), 799–805. https://doi.org/10.1039/P29930000799.

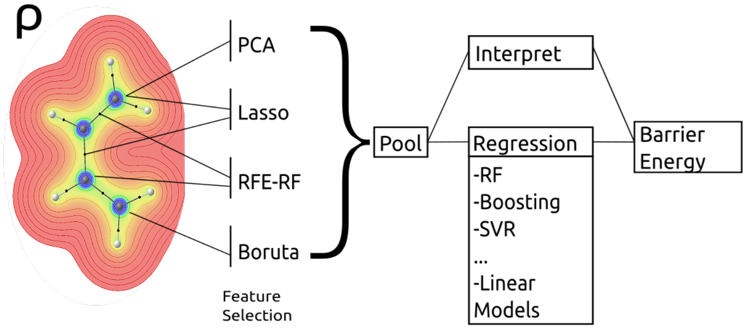
(114) Sahakyan, A. B. Computational Studies of Dielectric Permittivity Effects on Chemical Shifts of Alanine Dipeptide. *Chem. Phys. Lett.* **2012**, *547*, 66–72. https://doi.org/10.1016/j.cplett.2012.07.069.

(115) Shirvanyants, D.; Ding, F.; Tsao, D.; Ramachandran, S.; Dokholyan, N. V. Discrete

Molecular Dynamics: An Efficient And Versatile Simulation Method For Fine Protein Characterization. *J. Phys. Chem. B* **2012**, *116* (29), 8375–8382. https://doi.org/10.1021/jp2114576.

(116) Proctor, E. A.; Ding, F.; Dokholyan, N. V. Discrete Molecular Dynamics. *WIREs Comput. Mol. Sci.* **2011**, *1* (1), 80–92. https://doi.org/10.1002/wcms.4.

Graphical TOC


ρ
PCA
Lasso
RFE-RF
Boruta
Feature Selection
Pool
Interpret
Regression
-RF
-Boosting
-SVR
...
-Linear Models
Barrier Energy