Machine Learning Analysis of Dynamic-Dependent Bond Formation in Trajectories with Consecutive Transition States

Jesse Melville, Cal Hargis, Michael T. Davenport, R. Spencer Hamilton, and Daniel H. Ess*

Department of Chemistry and Biochemistry, Brigham Young University, Provo, Utah, 84602, USA

Email for corresponding author: *dhe@chem.byu.edu

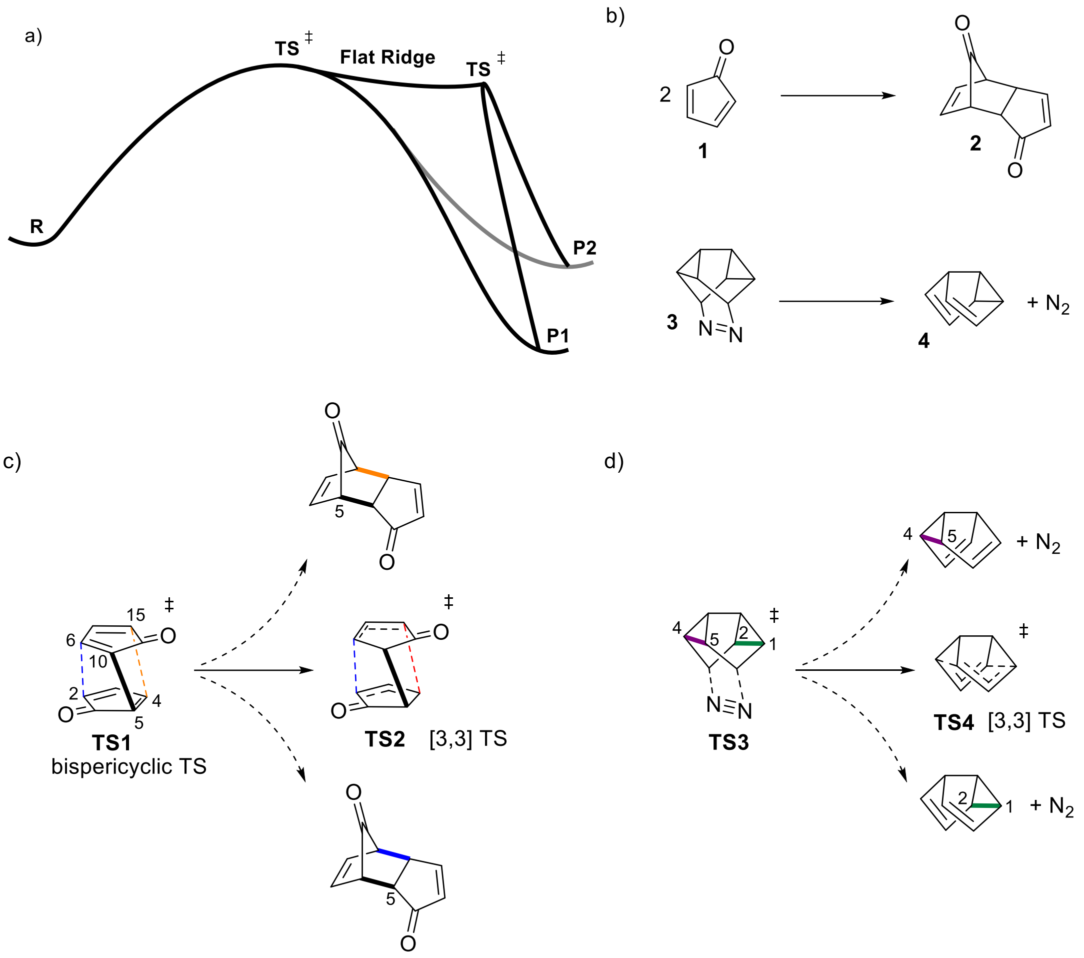
Special Issue in Honor of Prof. Barry Carpenter

Abstract

Dynamic motion often controls selectivity in reactions featuring two consecutive potential-energy transition states. Here we report DFT-based direct dynamics trajectories and machine learning classification analysis for cyclopentadienone dimerization and a N_2 extrusion reaction leading to semibullvalene. These reactions have consecutive transition states and there is dynamic selectivity that determines which of two possible C-C bonds is formed after the first transition state. For cyclopentadienone dimerization with a bispericyclic first transition state, machine learning analysis using transition-state based features provided >90% trajectory classification accuracy, but only using AdaBoost and random forest algorithms. Many other relatively sophisticated machine learning algorithms showed poor accuracy despite the obvious motion responsible for selectivity. Feature importance analysis confirmed that the sigmatropic rearrangement vibrational motion in the bispericyclic transition state provides prediction of which of the second C-C bonds is dynamically formed. For the reaction leading to semibullvalene, machine learning analysis provides solid accuracy for classifying trajectories and predicting which C-C bond is formed and which C-C bond is broken immediately after N_2 ejection. Like the cyclopentadienone dimerization reaction, machine learning feature importance analysis showed that the sigmatropic rearrangement vibrational motion in the N_2 extrusion transition state determines which C-C bond is formed and which is broken. Surprisingly, machine learning struggles to predict which trajectories undergo a subsequent [3,3] sigmatropic rearrangement process, which isomerizes equivalent forms of semibullvalene.

Introduction

In reactions with two consecutive transition states the potential energy surface has a relatively flat ridge region that divides two products, which is illustrated in Scheme 1a.[1,2] This type of energy landscape became prominent after Caramella described the energy surfaces of cyclopentadiene[3] and cyclopentadienone[4] dimerization reactions, which revealed consecutive bispericyclic cycloaddition and sigmatropic rearrangement transition states. Subsequently, this general type of consecutive transition-state bifurcated energy landscape was proposed for a variety of addition, [5,6,7,8] substitution,[9,10,11,12] pericyclic,[13,14] rearrangement,[15,16,17,18,19,20,21] and radical[22,23,24,25,26,27] reactions. Importantly, the selectivity between products (or intermediates) in these types of reactions is not easily determined with typical statistical theories,[28,29] especially selectivity models that utilize independent transition states leading to each product. Instead, dynamic atomic motion is responsible for selectivity and direct/ab initio dynamics trajectories have emerged as the dominant tool to provide qualitative or even sometimes quantitative treatment for selectivity in these reactions.[30,31,32]



Scheme 1. a) Qualitative energy landscape with consecutive transition states (**TS**). Reactants (**R**) transform into either product 1 (**P1**) or product 2 (**P2**) and selectivity between these products is determined by dynamic motion through and past **TS1**. b) Cyclopentadienone dimerization and N_2 extrusion reactions with consecutive transition states examined in this work using quasiclassical direct dynamics trajectories and machine learning classification. c) Outline of Caramella's consecutive bispericyclic cycloaddition (**TS1**) and [3,3] sigmatropic rearrangement (**TS2**) transition states for cyclopentadienone dimerization. The dotted arrows represent dynamic pathways after **TS1** leading to one of the cycloadducts. d) Outline of N_2 extrusion (**TS3**) and [3,3] sigmatropic rearrangement (**TS4**) consecutive transition states. The dotted arrows represent dynamic pathways after **TS3** leading to semibullvalene.

Despite the growing number of organic reactions that display consecutive transition states and dynamic selectivity, the origin of selectivity is often not directly analyzed or only analyzed in a qualitative manner. There has been a recent emphasis on predicting dynamic reaction selectivity using only a few key points on an energy landscape or transition-state partial bond lengths.[33,34] This general type of approach was outlined by Carpenter in 1992.[35] More recently, Truhlar proposed a method that does not require propagation of trajectories.[36] The general emphasis on qualitative analysis arises because quantitative analysis is complicated by the large amount of information contained in a large ensemble of trajectories propagated along a complex multi-dimensional energy landscape.

Machine learning provides one type of approach to quantitively analyze dynamics trajectories and the opportunity to reveal the origin of dynamic selectivity in reactions with consecutive transition states by identifying patterns of motion related to selectivity. Complete sampling often requires hundreds or thousands of trajectories to be propagated and therefore machine learning is a powerful general technique because it can handle large quantities of chemical reaction information in a straightforward manner, and many different algorithms are readily available.[37,38] Throughout chemistry, machine learning has become prominently used as a sophisticated regression tool to predict properties,[39] but its use for classification has been significantly less explored. Previously, we used machine learning to analyze quasiclassical direct dynamics trajectories for the thermal deazetization of 2,3-diazabicyclo[2.2.1]hept-2-ene,[40] which was inspired by the previous work of Carpenter who examined this reaction using experiments, ab initio calculations, and dynamics trajectories based on a semiempirical potential-energy surface method.[41,42] Our work showed that at the transition state machine learning could qualitatively, but not quantitatively, identify Carpenter's proposal of methylene bridge out-of-plane bending as the origin of nonstatistical endo-exo product selectivity. More recently, we used machine learning to determine the origin of IRC versus non-IRC motion in cyclopropyl radical ring opening, which revealed that there are two key vibrational modes, and that their directional combination provides correlation and prediction of the trajectory motion.[43] However, at the transition state only two machine learning models (random forest and logistic regression classifier) provided quantitative accuracy above 80% and several machine learning models were close to the baseline accuracy of 50%, which corresponds to random assignment of one of two outcomes.

In this work we wanted to examine reactions where selectivity is likely only controlled by the motion of two competing bonds, which could potentially be identified through vibrational mode information. Specifically, we wanted to examine reactions with two consecutive transition states where the second transition state is a [3,3] sigmatropic rearrangement. Therefore, we decided to run quasiclassical density functional theory (DFT) direct (Born-Oppenheimer) dynamics trajectories and perform machine learning classification analysis on cyclopentadienone (**1**) dimerization and N_2 extrusion from **3** that results in semibullvalene (Scheme 1b). These reactions were selected for analysis because previous computational studies revealed consecutive transition states with the second transition state corresponding to a [3,3] sigmatropic rearrangement process.[4,44] In cyclopentadienone dimerization the dynamic selectivity determines which of two possible second C-C bonds is formed (Scheme 1c).[45,46] In the N_2 extrusion reaction dynamic selectivity controls which bicyclo C-C bond is fully formed and which C-C bond is fully broken (Scheme 1d).[47] Another reason why we examined cyclopentadienone dimerization is that Singleton previously reported cyclopentadiene dimerization trajectories and found a straightforward correlation (about 85%) between the second vibrational mode of the transition-state structure and the trajectory outcome.[45]

As will be presented, for cyclopentadienone dimerization, machine learning analysis using transition-state based features provided >90% trajectory classification accuracy using AdaBoost[48] and random forest algorithms.[49] However, and perhaps surprisingly, several other machine learning algorithms that are generally considered to be robust significantly struggled to provide trajectory classification accuracy higher than 60%, which is only slightly higher than random assignment. As expected, for cyclopentadienone trajectories, feature importance analysis revealed the expected transition-state [3,3] sigmatropic rearrangement directional vibrational motion as controlling selectivity. For the N_2 extrusion reaction leading to semibullvalene, machine learning at the first transition state (**TS3**) provides high accuracy (>90%) for classifying trajectories and predicting which C-C bond is immediately formed. However, and surprisingly, machine learning struggles to predict the outcome of the [3,3] sigmatropic rearrangement process that takes place after the first C-C bond is formed, which isomerizes the equivalent forms of semibullvalene. Zero-point energy flow into the reaction coordinate was examined as a possible origin of this rearrangement and poor machine learning performance. Trajectories with 75% of the zero-point energy showed nearly identical dynamic motion but trajectories with 50% of the zero-point energy showed less rearrangement.

Results and Discussion

Cycloadditions have become the most prominent example of reactions with two consecutive transition states. This was principally the result of Caramella's foundational reports that described the consecutive relationship of the endo bispericyclic transition states and a [3,3] sigmatropic rearrangement transition states in cyclopentadiene and cyclopentadienone dimerization.[3,4] Outlined in Scheme 1c, Caramella showed that the first transition state, **TS1** (called bispericyclic), for dimerization of cyclopentadienone involves a symmetrical process with one advanced partial C-C bond and two lagging C-C bonds that are portrayed with orange and blue dashed lines. The [3,3] sigmatropic rearrangement transition state, **TS2**, has a related symmetrical structure with one formed C-C bond and similar blue and orange partial C-C bonds. In Caramella's original report, for **TS1**, B3LYP/6-31G* gave partial bond lengths of 2.17 Å and 2.80 Å. Our M06-2X/6-31G**[50] first-order saddle structure **TS1** ($<S^2> = 0$) has very similar distances of 2.11 Å and 2.72 Å (Figure 1). Caramella previously demonstrated that this endo bispericyclic transition state is lower in energy than exo and diradical transition states.[4] Caramella also outlined that the consecutive transition states **TS1** and **TS2** result in a symmetrical potential energy surface with a ridge region as depicted in Scheme 1a.[4] This has the result of atomic momenta likely inducing the reaction pathway to divert and fall off the ridge and form either C2-C6 (blue) or C4-C15 (orange) prior to progressing to the [3,3] sigmatropic rearrangement transition state, and this is depicted by dotted reaction arrows in Scheme 1c.

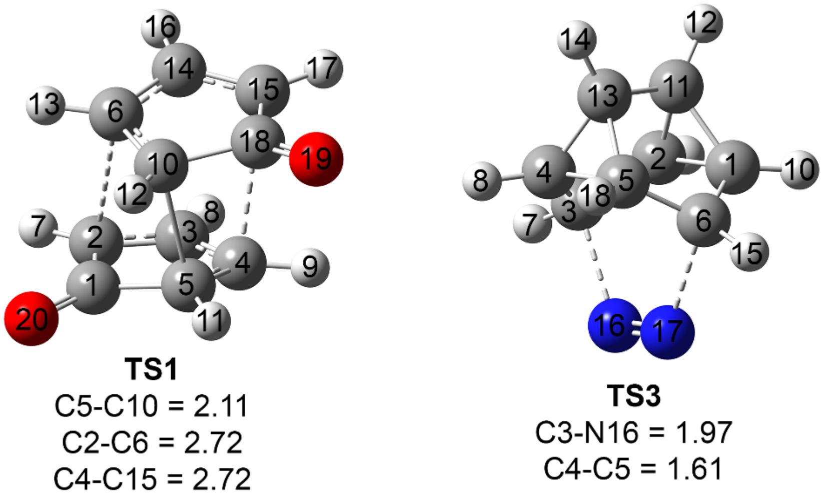


Figure 1. M06-2X/6-31G** transition-state structures and atomic labels. Distances reported in Å.

Based on Caramella's energy surface description that the endo bispericyclic transition state should dynamically lead to the cycloadduct product by going down the ridge before reaching the [3,3] sigmatropic rearrangement transition state, we calculated quasiclassical direct dynamics trajectories starting at **TS1** and tracked C2-C6 versus C4-C15 bond formation. M06-2X/6-31G** trajectories were generated with local mode and thermal sampling at 298 K in Gaussian 16 using the BOMD method,[51] which creates a kinetic energy and potential energy ensemble (distorted geometry) of **TS1**. Trajectories were propagated using the gradient and updated, but not fully calculated, Hessian at each step. Trajectories were propagated with an approximate timestep of 0.5 femtoseconds (fs) using the default predictor-corrector type integration algorithm. The chemically forward direction was followed for approximately 400 fs. Reverse trajectories were created by inversion of all mass-weighted atomic velocities. The reverse direction was followed until cyclopentadienones were separated by several Ångstroms. Approximately 20% of forward trajectories showed recrossing and return to separated cyclopentadienones. These trajectories were removed from our data set and not analyzed with machine learning methods.

As expected, due to the symmetrical potential-energy surface we found the nearly 1:1 ratio of C2-C6/blue versus C4-C15/orange bond formation in trajectories. Figure 2 shows snapshots of a representative trajectory from **TS1** that rapidly leads to forming the cycloadduct in about 50 timesteps (about 25 fs). In this example trajectory, the C5-C10 has a distance of 2.09 Å and decreases to 1.90 Å and 1.43 Å at 25 and 50 timesteps, respectively. At timestep 1, the motion favoring the C2-C6 bond (2.42 Å) and not forming the C4-C15 (2.92 Å) bond is already evident, which is a function of the atomic motion passing through the transition-state region. Progression to timestep 25 shows very little change in the C2-C6 and C4-C15 indicating most of the initial motion is formation of the leading C-C bond and then after timestep 25 to timestep 50 there is decrease in the C2-C6 distance while the C4-C15 distance is relatively unaffected.

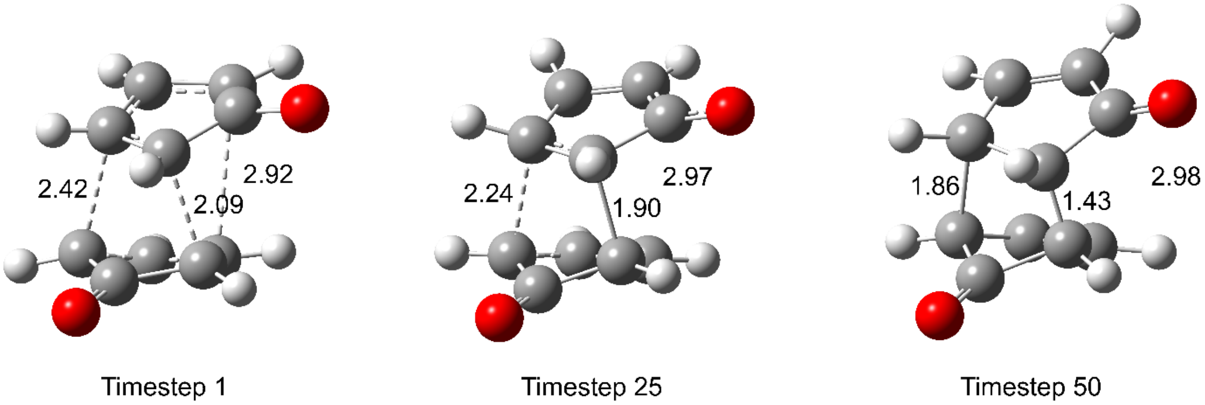
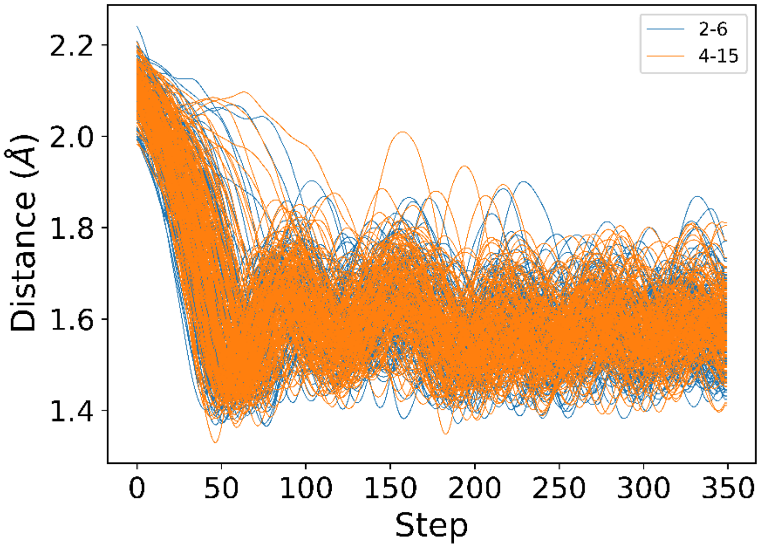
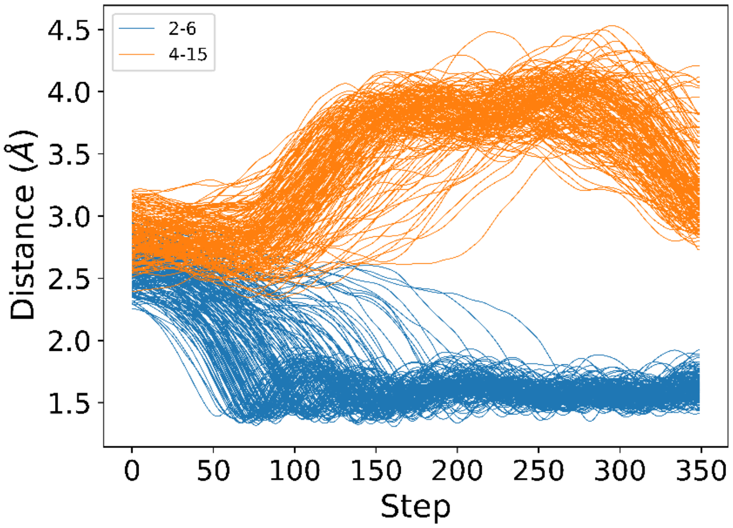


Figure 2. Example trajectory beginning at **TS1** and leading to cycloadduct **2** by formation of the C2-C6 bond.

Figure 3 plots trajectory steps versus C-C bond distance for 336 trajectories propagated from **TS1**. Blue color coding corresponds to trajectories that fully form the C2-C6 bond (170 total trajectories) and orange color coding refers to trajectories that fully form the C4-C15 (166 total trajectories). The top plot in Figure 2 shows that in all trajectories the C5-C10 bond is fully formed by about 25 fs (50 steps) after **TS1**. The middle and bottom plots show that the lagging C2-C6 or C4-C15 bond is formed as early as 25 fs after the transition state or as late as about 100 fs after the transition state with average being about 60 fs. On average, at 298 K, there is about 35 fs between formation of the first and second C-C bonds. This time gap between leading and lagging bonds is in the vicinity to the time gap for cyclopentadiene dimerization.[46] While this time gap is larger than in highly synchronous cycloaddition reactions, it is likely an insufficient amount of time to provide significant intramolecular vibrational energy redistribution[52] to generate a sustained diradical or zwitterionic intermediate.





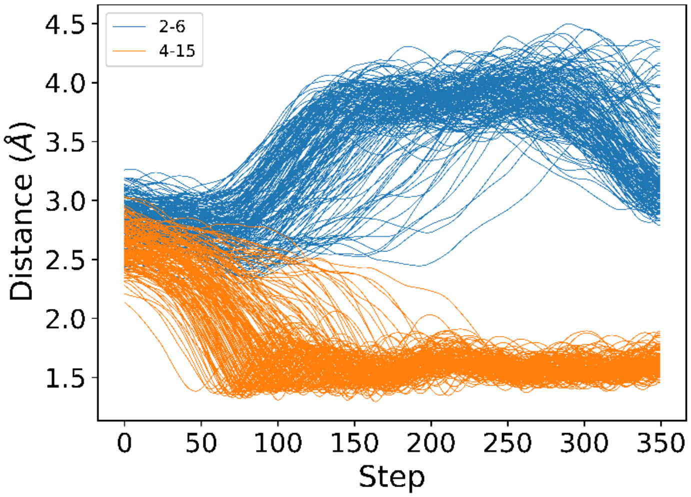


Figure 3. Top: Plot of trajectory steps versus the C5-C10 distance. Middle: Plot of trajectory steps versus the C2-C6 distance. Bottom: Plot of trajectory steps versus the C4-C15 distance. Blue trajectories are classified as forming the C2-C6 bond and orange trajectories are classified as forming the C4-C15 bond. Trajectory steps are approximately 0.5 fs. Distances plotted in Å.

With the generation and classification of trajectories as either forming the C2-C6/blue or C4-C15/orange bond we then extracted transition-state features for machine learning analysis. The features were chosen based on fundamental physical differences between individual trajectories. We extracted about 100 transition-state features that include vibrational mode quanta, mass-weighted vibrational mode atomic displacements, mass-weighted atomic velocities, geometry distances, angles, and dihedral angles. For each machine learning model created, we used an equal number of C2-C6 versus C4-C15 bond forming trajectories (i.e., we randomly left out four C2-C6 bond forming trajectories) and similar to our previous approach,[40,43] we used the Scikit-Learn[53,54] Python library to set up and train classifiers with a 10-fold cross validation to determine the classification accuracy of each machine learning model. This was done by dividing the sampled data set into 10-equally sized subsets, training the model on 9 subsets, and then evaluating the predictive accuracy using the left-out subset. This analysis was performed 10 times with a different subset withheld in each instance. For training, trajectory classification labels were based on geometry analysis described above. Representative source code illustrating the creation, training, and use of classifiers can be found in the Supporting Information (SI). The reported accuracy of each machine learning model is the mean value of all iterations. Because there is binary classification, random assignment of trajectories would give 50% accuracy.

There are several types of supervised machine learning algorithms implemented in the scikit-learn Python library. Broadly, a supervised machine learning algorithm uses labeled data during the training process which then allows a model-based prediction for unlabeled data. Within the tent of supervised classification methods there are linear models, kernel ridge regression, support vector machines, stochastic gradient descent, nearest neighbors, Gaussian processes, and neutral network models. Because it is difficult to know which algorithms will perform best for a specific chemical system, we examined several of these algorithms. Details of the algorithms can be found in References 53 and 54. For the best performing adaptive boosting and random forest models, which will be discussed later, we used GridSearchCV to optimize the features. GridSearch is an exhaustive search of parameter combinations, selecting the optimal parameters using cross-validation scores. For random forest we optimized the number of estimators, criterion, max_depth, and max features. For adaptive boosting we examined different base estimators, including a random forest model as the base estimator, number of estimators, and learning rate. Typically GridSearchCV was used several times, each time shrinking the boundaries of the parameter grid to narrow the parameters to specific values and this was done until no more model improvement was found.

Because of Caramella's potential-energy surface description,[4] Singleton's previous report on the dynamics of cyclopentadiene dimerization,[45] and the visualization of a few representative trajectories (see Figure 2), it seemed likely that for cyclopentadienone dimerization C2-C6 versus C4-C15 bond formation should be heavily influenced by the [3,3] rearrangement process and machine learning models should be able to provide high accuracy classification. We examined the accuracy for predicting whether C2-C6 or C4-C15 is formed in each trajectory based on transition-state features using eight supervised classification algorithms (Figure 4). To our surprise, despite the relatively straightforward motion of these trajectories, several machine learning methods that are typically robust and effective struggled to forecast classification based on transition-state features. For example, logistic regression that predicts the target variable probability using binary classification only showed 66% accuracy with a baseline random assignment accuracy being 50%. Similarly, a support-vector machine method, which is typically useful for high dimensional problems,[53] showed only 60% accuracy. While logistic regression and support-vector machine methods showed significant inaccuracy for classification, an unoptimized random forest model gave 84% accuracy and an optimized random forest model gave 88% accuracy. Even better, a hyperparameter optimized adaptive boosting method resulted in a model with 91% accuracy. For this adaptive boosting algorithm, using only 200 rather than 332 trajectories in the dataset only resulted in accuracy decreasing by only a few percent, which suggests that this number of trajectories was sufficient to give a saturated result. Overall, ensemble learning methods outperform other types of machine learning methods.

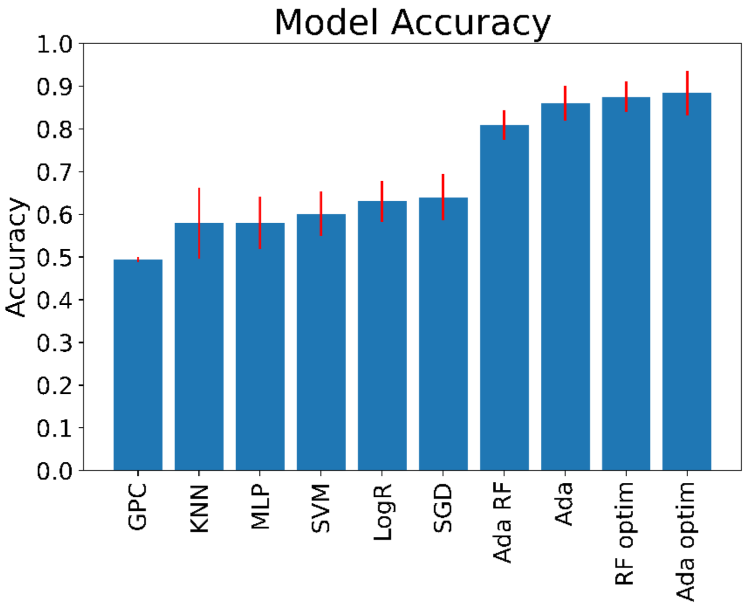


Figure 4. Plot of cyclopentadienone dimerization trajectory classification accuracy with several popular machine learning algorithms. The accuracy for each machine learning model is the mean accuracy of 10 iterations where the accuracy is defined as the number of correct predictions divided by the total number of predictions. Red bars represent a 10-fold cross validation 95% confidence interval. GPC = gaussian process classification, KNN = K-nearest neighbor, MLP = Multilayer perceptron, SVM = support vector machine, SGD = stochastic gradient descent, LogR = logistic regression, Ada RF = adaptive boosting combined with random forest, Ada = adaptive boosting, RF Opt = hyperparameter optimized random forest, Ada optim = hyperparameter optimized adaptive boosting.

This relatively high accuracy for adaptive boosting classification is important because it provides the ability to analyze the importance of physical features in the model and provide the origin of dynamic selectivity. The top of Figure 5 plots the 13 most important features found in the adaptive boosting model. As expected based on the example trajectory displayed in Figure 2 and Singleton's previous analysis of cyclopentadiene dimerization,[45] this analysis revealed that the two most important features correspond to the mass-weighted displacement and the mass-weighted velocities for the transition-state structure normal vibration mode 2. Vibrational mode zero-point and thermal energies are separated into potential and kinetic energies during the sampling process. Normal mode 1 is the negative vibrational frequency corresponding to the reaction coordinate. Mode 2 is the 74 cm^{-1} vibrational frequency that corresponds to the asymmetric rocking motion that develops into the subsequent negative vibrational mode of the [3,3]-sigmatropic transition state **TS2**. Figure 6 illustrates **TS1** vibrational mode 2 motion and negative and positive coordinate displacement.

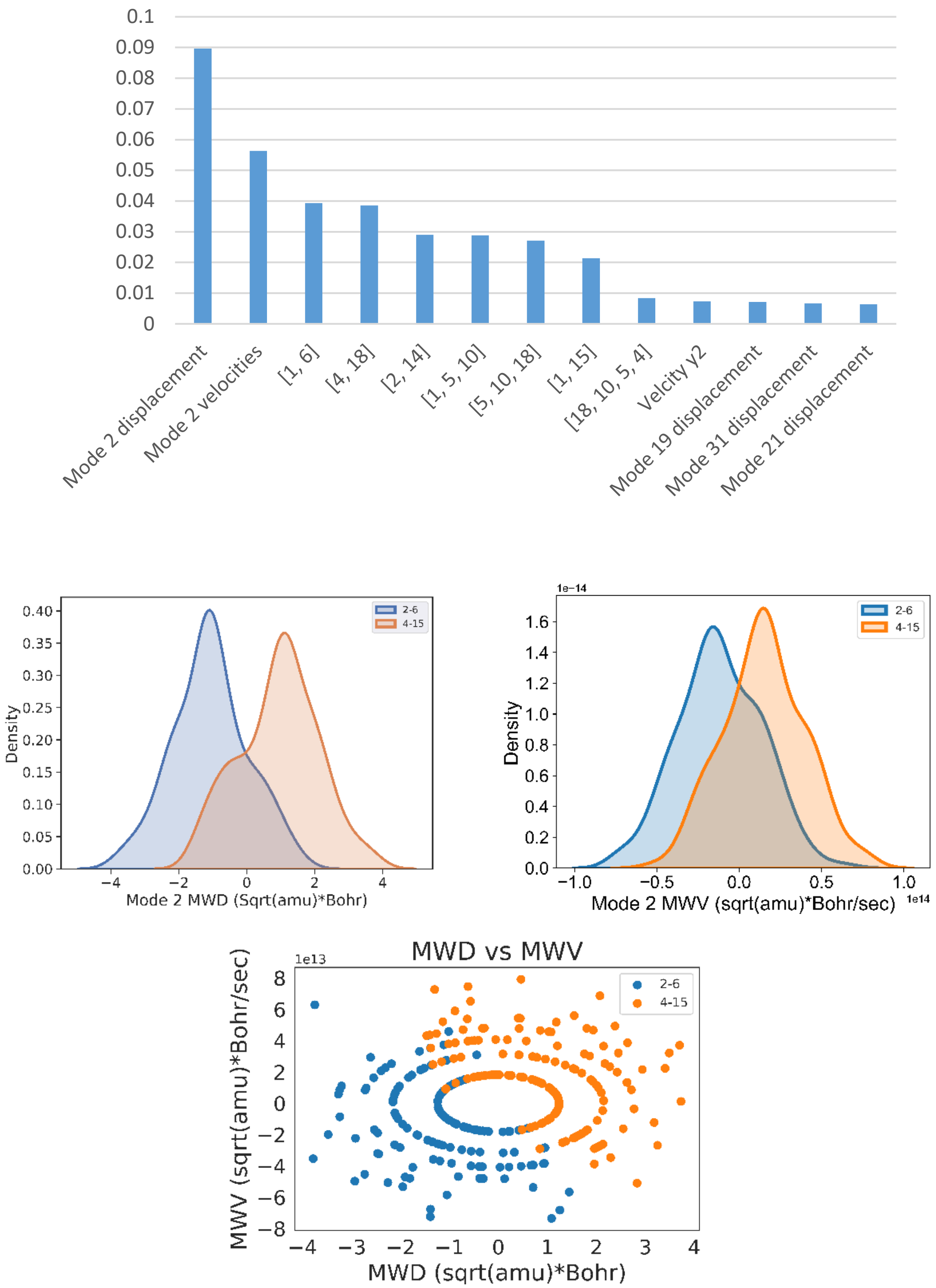


Figure 5. Cyclopentadienone dimerization trajectory classification analysis. Top: Weighted relative feature importance. The mode 2 mass-weighted displacement feature corresponds to the first positive transition-state normal mode vibration. [1, 6], [4, 18], [2, 14], and [1, 15] correspond to distance features. [1, 5, 10] and [5, 10, 18] are angle features. [18, 10, 5, 4] is a dihedral angle feature. Velocity y2 is the y component of the velocity for atom 2. Atom numbers are given in Figure 1. Middle: Plot of trajectory density versus the trajectory value for mode 2 mass-weighted displacement (MWD) and plot of trajectory density versus the trajectory value for mode 2 mass-weighted velocities (MWV). Bottom: Plot of mode 2 mass-weighted displacement values versus mode 2 mass-weighted velocity values. Blue corresponds to trajectories that form the C2-C6 bond and orange corresponds to trajectories that form the C4-C15 bond.

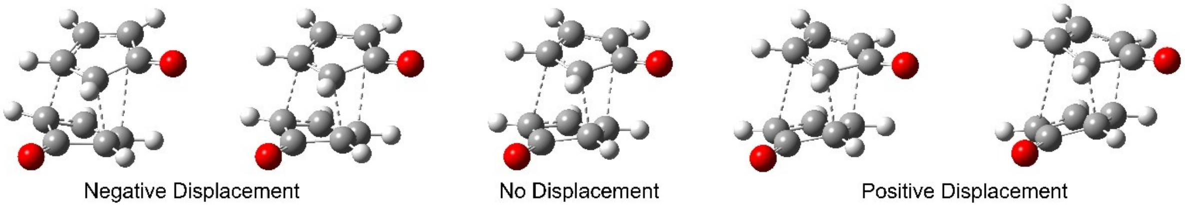


Figure 6. Graphical depiction of motion for vibrational mode 2.

The next seven most important features correspond to geometric distances, angles, and dihedral angles. However, their relative importance to the model is much smaller than mode 2 features. Features 10-13 while non-zero are dramatically smaller than the first seven features. While the importance of mode 2 may be considered obvious because this reaction potential-energy surface has been previously well described by Caramella,[4] this analysis demonstrates that machine learning can identify physical features in the transition state that control dynamic bond forming selectivity. It is interesting to highlight that in cyclopentadienone trajectories there are four possible combinations of vibrational mode 2 displacement and vibrational mode 2 velocities creating four different vibrational value combinations and machine learning handles these dependent features. The middle of Figure 5 plots the density of trajectories with corresponding mode 2 displacement and velocity values. The bottom of Figure 5 plots mode 2 mass-weighted displacement values versus mode 2 mass-weighted velocity values. These plots qualitatively show that when both the displacement and velocities are negative the C2-C6 bond is formed and when the displacement and velocities are both positive the C4-C15 bond is formed. In contrast, when the displacements and velocities have opposite signs there is overlap of the features, and it is likely that this is where the adaptive boosting machine learning model provides some inaccurate predictions.

Because several trajectories showed overlap when the mass-weighted displacement values and mass-weighted velocities were plotted we analyzed the trajectories that are most difficult to classify. Therefore, we ran the adaptive boosting model 240 times and identified all trajectories that were correctly classified less than 48 times out of 240 times. Each model has different subgroups of training and testing during the cross-validation process. Figure 7 plots the mode 2 mass-weighted displacement values and density for the mode 2 mass-weighted velocities for these incorrectly forecasted trajectories. In accord with our speculation, the inaccurately predicted trajectories show values that heavily overlap. For example, comparison of the right-hand plot of Figure 7 with mass-weighted velocities shows orange and blue colored values that are heavily overlaid, and this contrasts with the plot in the middle of Figure 5 where there is clear separation between blue and orange values.

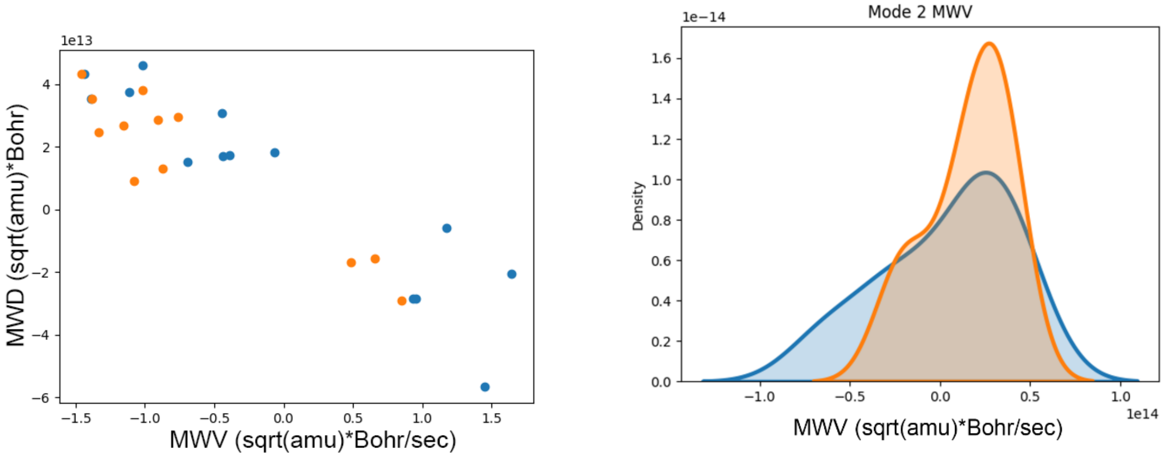


Figure 7. Left: Plot of mode 2 mass-weighted displacement values for or trajectories that were correctly forecasted less than 20% of the time. Right: Plot of mass-weighted velocity values for trajectories that were correctly forecasted less than 20% of the time. Blue corresponds to trajectories that form the C2-C6 bond and orange corresponds to trajectories that form the C4-C15 bond. MWD = Mass-weighted displacement. MWV = Mass-weighted velocities.

With the success of classifying and understanding the cyclopentadienone dimerization reaction, we wanted to examine another reaction with consecutive transition states where the second transition state also invovles a [3,3] sigmatropic rearrangement. Therefore, we examined N_2 extrusion from **3** that results in semibullvalene. Based on Sauer’s synthesis of semibullvalenes,[55] Birney examined the energy landscape for N_2 extrusion converting **3** to **4** (Scheme 1b).[44] In this work, Birney demonstrated with B3LYP DFT calculations that **TS3** (Scheme 1d) leads to a ridge region and the [3,3] sigmatropic rearrangement transition state **TS4**.[44] Inspection of **TS3** indicates that both C1-C2 and C4-C5 bonds, green and purples bonds shown in Scheme 1d, are significantly stretched and depending on which direction a trajectory will follow will lead to either formation of C1-C2 and complete cleavage of C4-C5 or formation of C4-C5 and complete cleavage of C1-C2. Datta later used trajectories to confirm this dynamic bond control and to analyze the lifetime of the transition-state zone and possible heavy atom ^{13}C kinetic isotope effects.[47] It was found that the C-N bonds were broken within about 15 fs and product **4** could be identified on average shortly after 100 fs.

For this N_2 extrusion reaction, starting with a **TS3** ensemble created using the same methodology as the cyclopentadienone dimerization trajectories, we calculated more than 800 reactive and non-recrossing trajectories, but only 643 were used for machine learning analysis (see below). For all trajectories, Figure 8 plots the C1-C2 distance versus trajectory steps starting from the transition state. This plot shows the expected separation of the two possible semibullvalenes formed at 125 steps after the transition state. We classified trajectories as either C1-C2 forming/green or C4-C5 forming/purple at 125 steps (about 63 fs) This classification was done by determining if the C1-C2 distance is less than 1.7 Å and the C4-C5 greater than 2.5 Å. Figure 8 also shows the [3,3] rearrangement process that occurs between 150-300 steps (crossover of purple and green lines) after **TS3** and after the first semibullvalene structure is formed, which is likely due to the low barrier for rearrangement combined with a lack of significant internal vibrational energy redistribution.

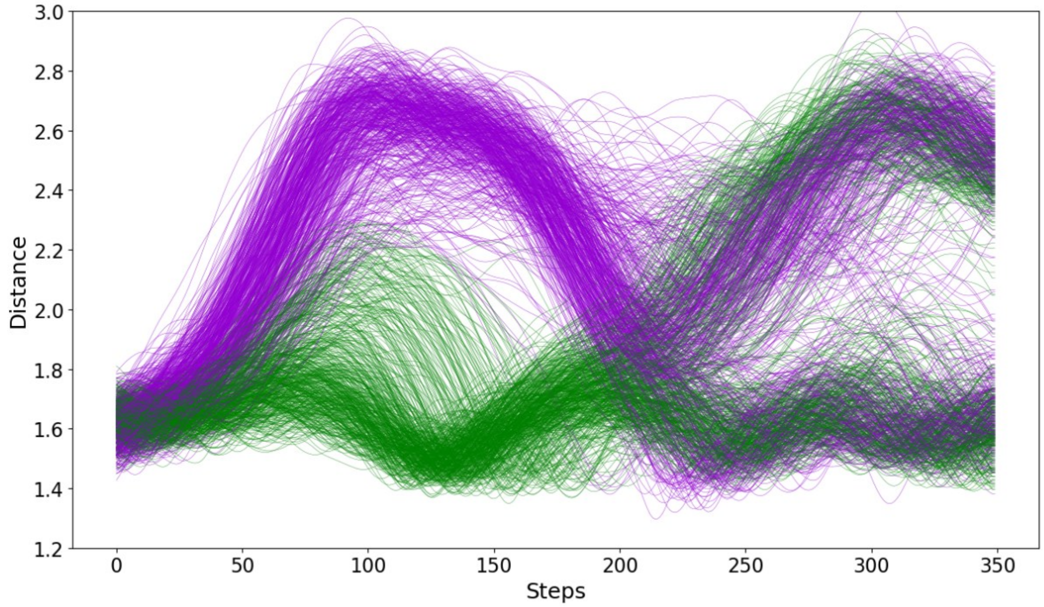
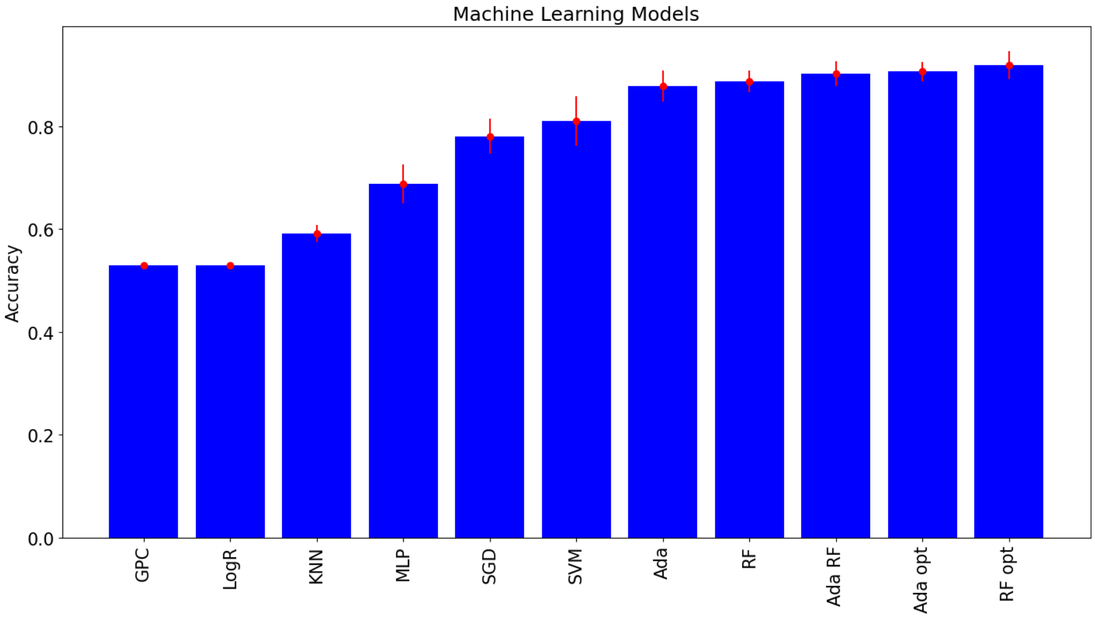
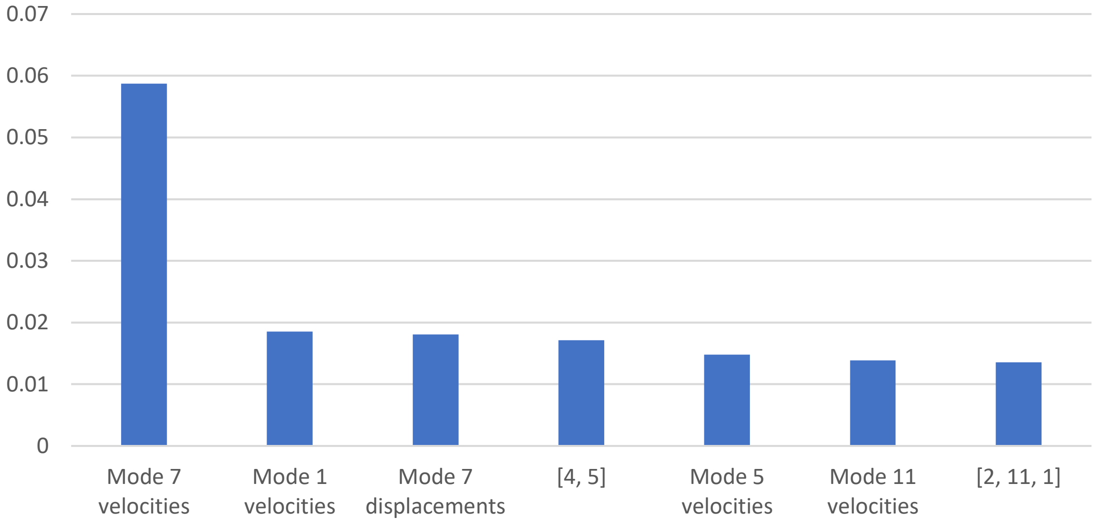


Figure 8. Plot of trajectory steps versus the C1-C2 distance (in Å) starting at **TS3** and leading to semibullvalene. Trajectories are colored green if the C1-C2 distance is less than 2.2 Å at step 125. Trajectories are colored purple if the C1-C2 distance is greater than 2.2 Å at step 125.

In the machine learning analysis, we used 303 C1-C2 bond-forming and 340 C4-C5 bond-forming trajectories, which provides a total of 643 analyzed trajectories. We did not include trajectories where at step 125 the structure showed the C1-C2 distance to be greater than 1.7 Å and the C4-C5 distance to be less than 2.5 Å, which structures do not provide definitive classification at this trajectory step and still in the process of forming a definitive semibullvalene structure or could be considered a short-lived diradical/zwitterionic intermediate. Similar to the cyclopentadienone dimerization analysis we used a 10-fold cross validation strategy to determine the classification accuracy. The top of Figure 9 plots the classification accuracy for these semibullvalene forming trajectories. Several models gave between 88%-95% accuracy and like the cyclopentadienone dimerization reaction (see Figure 4), adaptive boosting and random forest type models performed best. As expected, using velocity and geometry features at 50 steps after the transition state (no vibrational mode features) improves accuracy well above 90%.


Machine Learning Models
Accuracy
0.0
0.2
0.4
0.6
0.8
GPC
LogR
KNN
MLP
SGD
SVM
Ada
RF
Ada RF
Ada opt
RF opt


0.07
0.06
0.05
0.04
0.03
0.02
0.01
0
Mode 7 velocities
Mode 1 velocities
Mode 7 displacements
[4, 5]
Mode 5 velocities
Mode 11 velocities
[2, 11, 1]

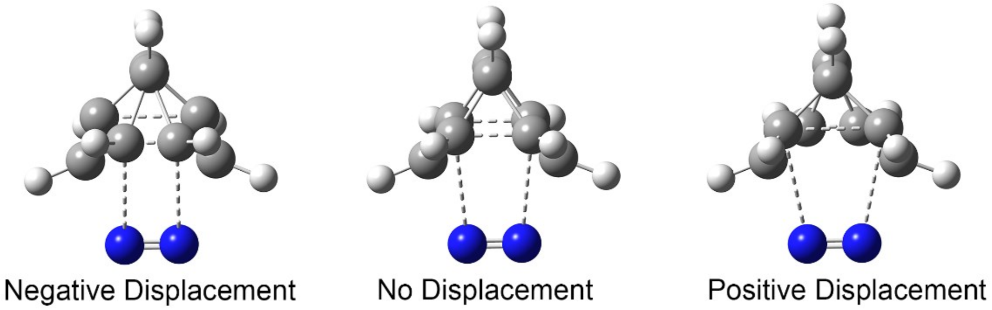
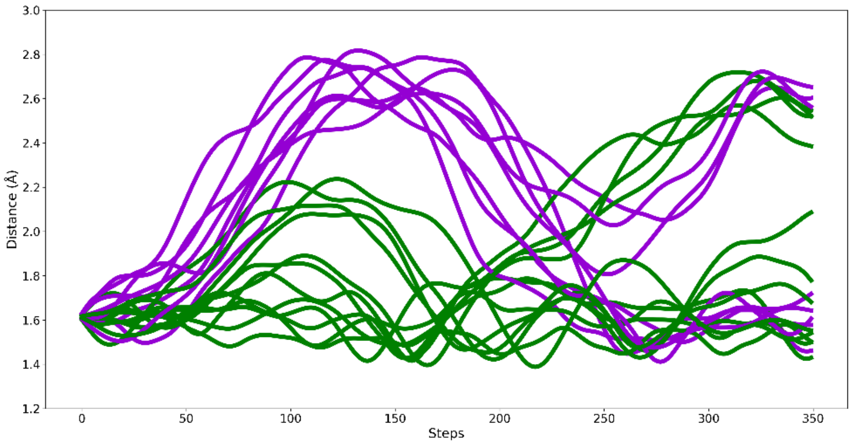

Negative Displacement
No Displacement
Positive Displacement

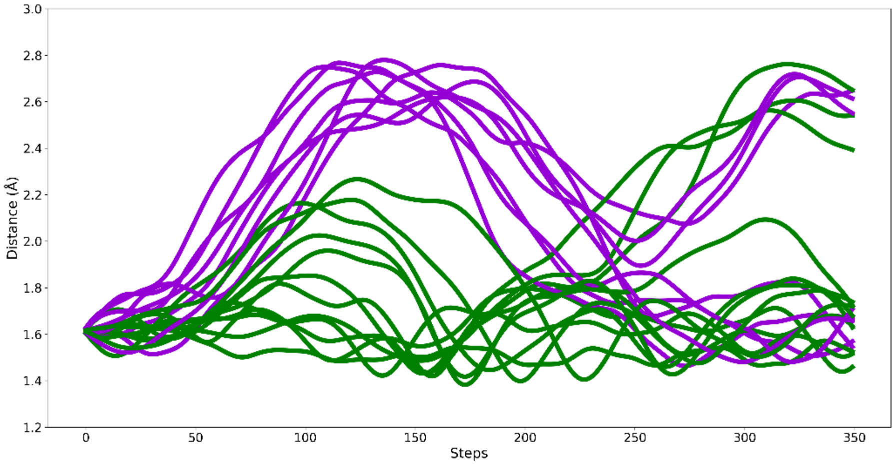
Figure 9. Top: Plot of N_2 extrusion reaction trajectory classification accuracy with several popular machine learning algorithms. The accuracy for each machine learning model is the mean accuracy of 10 iterations where the accuracy is defined as the number of correct predictions divided by the total number of predictions. Red bars represent a 10-fold cross validation 95% confidence interval. GPC = gaussian process classification, LogR = logistic regression, KNN = K-nearest neighbor, MLP = Multilayer perceptron, SGD = stochastic gradient descent, SVM = support vector machine, Ada = adaptive boosting, RF = random forest, Ada RF = adaptive boosting combined with random forest, Ada optim = hyperparameter optimized adaptive boosting, RF Opt = hyperparameter optimized random forest. Middle: Weighted relative feature importance for the N_2 extrusion reaction. The mode 7 mass-weighted displacement feature corresponds to the sixth positive transition-state normal mode vibration. [4, 5] corresponds to a distance feature between carbons 4 and 5. [2, 11, 1] is an angle feature. Bottom: Graphical depiction of motion for vibrational mode 7.

Using the optimized adaptive boosting model, we analyzed the feature importance of the trajectories at the transition state. The middle of Figure 9 displays the top seven features. Mode 7 velocities is significantly more important than the other features. The bottom Figure 7 shows the normal mode 7 displacement. This normal mode displacement involves [3,3] sigmatropic rearrangement type motion that is very similar to mode 2 in the cyclopentadienone dimerization transition state. Again, this demonstrates that machine learning can identify chemically meaningful motion that determines and predicts the outcome of trajectories.

This semibullvalene reaction presents a challenge for machine learning classification that the cyclopentadienone dimerization reaction did not. As noted above, after semibullvalene is formed from **TS3** many trajectories undergo a [3,3] sigmatropric rearrangement, which isomerizes the semibullvalene and is shown by the crossover of green and lines in Figure 8. We wanted to determine if machine learning could accurately classify the outcome of both semibullvalene formation and subsequent [3,3] sigmatropic rearrangement using **TS3** features. Therefore, using classification at step 300 step rather than step 125 we examined the performance of machine learning models. Surprisingly, all machine learning models resulted in severely poor performance with no models giving classification prediction better than 60%. Because of this very poor prediction using transition-state features for classification at step 300 we extracted features at step 200, which is just prior to the [3,3] sigmatropic rearrangement process. In this analysis the machine learning models all showed revival of performance. For example, optimized random forest gave 85% prediction and the optimized adaptive boosting model gave 88% accuracy. Overall, this indicates that machine learning is effective for classification, but only when the features used are close enough in time to the desired structure prediction.

It is possible that machine learning struggles to predict the outcome of the [3,3] sigmatropic rearrangement process because during the trajectory zero-point energy flows into the reaction coordinate. Doubleday demonstrated this type of zero-point energy leakage in the Bergman cyclization reaction where an endothermic diradical intermediate undergoes recrossing of a relatively large energy barrier.[56] To test if zero-point energy leakage into the semibullvalene intermediate is the origin of the isomerization process between 150-300 steps we carried out trajectories compare a variable amount of total zero-point energy but retaining the same reaction coordinate velocity, which is similar to how Doubleday examined this issue in the Bergman cyclization.[56] Figure 10 plots the trajectories starting at **TS3** with 100%, 75% and 50% of each vibrational mode zero-point energy. The trajectories with 75% of zero-point energy have motion and rearrangement very similar to the trajectories with 100% zero-point energy. Trajectories with only 50% of the zero-point energy still show rearrangement, but the number of trajectories that undergoes this process is dampened and therefore this could be in part the reason for poor machine learning performance but unlikely the only cause.


3.0
2.8
2.6
2.4
2.2
2.0
1.8
1.6
1.4
1.2
Distance (Å)
0
50
100
150
200
250
300
350
Steps


3.0
2.8
2.6
2.4
2.2
2.0
1.8
1.6
1.4
1.2
Distance (Å)
0
50
100
150
200
250
300
350
Steps

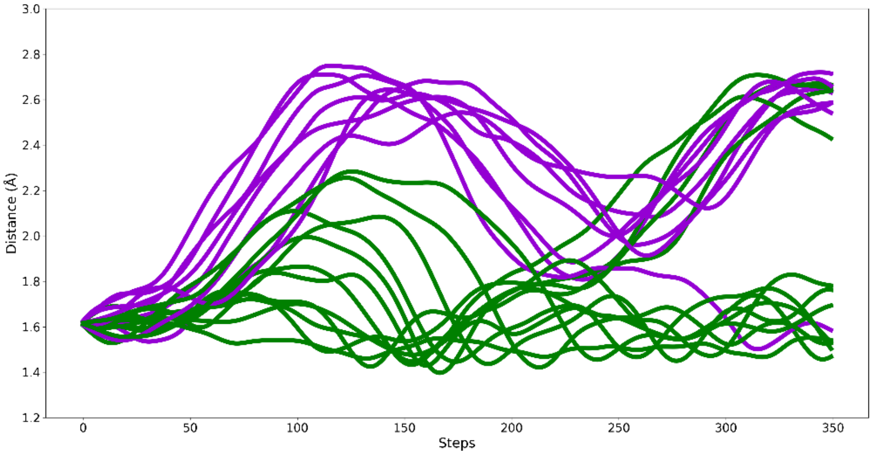

3.0
2.8
2.6
2.4
2.2
2.0
1.8
1.6
1.4
1.2
Distance (Å)
0
50
100
150
200
250
300
350
Steps

Figure 10. Plot of same 20 trajectories starting at **TS3** and leading to semibullvalene. Plots show trajectory steps versus the C1-C2 distance (in Å). Top: Trajectories with 100% zero-point energy. Middle: Trajectories with 75% of zero-point energy. Bottom: Trajectories with 50% of zero-point energy. Trajectories are colored green if the C1-C2 distance is less than 2.2 Å at step 125. Trajectories are colored purple if the C1-C2 distance is greater than 2.2 Å at step 125.

Conclusions

This work demonstrated that machine learning analysis of transition-state features provides a platform to predict the outcome of quasiclassical trajectories for reactions featuring two sequential transition states. This analysis was performed for cyclopentadienone dimerization and a N_2 extrusion reaction forming semibullvalene where dynamic selectivity determines which of two possible C-C bonds is formed. For cyclopentadienone dimerization, only a few specific machine learning algorithms provided >90% trajectory classification accuracy (AdaBoost and random forest type algorithms). In our opinion it is surprising that several other generally reliable and robust machine learning algorithms showed less than 60% classification accuracy (50% is the floor). This is especially surprising since AdaBoost transition-state feature analysis revealed the expected [3,3] sigmatropic rearrangement vibrational mode velocity and displacement features as providing straightforward correlation for classification. For the N_2 extrusion reaction leading to semibullvalene, again, only a few machine learning algorithms provided reasonable classification accuracy using **TS3** transition state features. Like the cyclopentadienone dimerization reaction, the [3,3] sigmatropic rearrangement vibrational motion correlates/determines trajectory outcomes. Different than the cyclopentadienone dimerization reaction, machine learning cannot easily predict the outcome of the subsequent [3,3] sigmatropic rearrangement process that occurs after initial semibullvalene formation.

Supplementary Information

Initial mass-weighted velocities for classified trajectories.

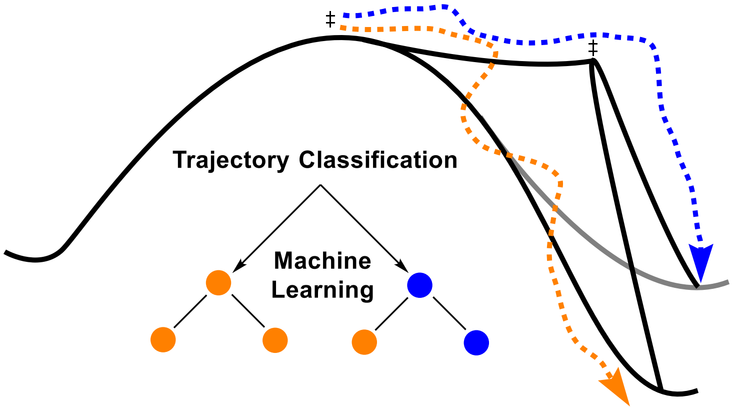
Data Availability

Data (Jupyter notebooks and data sets) available on request from the authors.

Acknowledgements

We thank the BYU Office of Scientific Computing and the Fulton Supercomputer Lab for computational resources. D.H.E acknowledges the United States National Science Foundation Chemical Structure, Dynamics, and Mechanisms B (CSDM-B) Program for support under award CHE 1952420. J.M., C.H., and R.S.H. thank the BYU Department of Chemistry and Biochemistry for undergraduate research awards. We thank Spencer Yu for helpful conversations.

Table of Contents Graphic



References

1. S. R. Hare, D. J. Tantillo, Post-transition state bifurcations gain momentum – current state of the field. *Pure Appl. Chem.* 2017, **89**, 679–698.

2. Ess, D. H.; Wheeler, S. E.; Iafe, R. G.; Xu, L.; Celebi-Olcum, N.; Houk, K. N. Bifurcations on Potential Energy Surfaces of Organic Reactions. *Angew. Chem. Int. Ed.* **2008**, *47*, 7592–7601.

3. Caramella, P.; Quadrelli, P.; Toma, L. An Unexpected Bispericyclic Transition Structure Leading to 4+2 and 2+4 Cycloadducts in the *Endo* Dimerization of Cyclopentadiene. *J. Am. Chem. Soc.* **2002**, *124*, 1130–1131.

4. Quadrelli, P.; Romanao, S.; Toma, L. Caramella, P. A Bispericyclic Transition Structure Allows for Efficient Relief of Antiaromaticity Enhancing Reactivity and Endo Stereoselectivity in the Dimerization of the Fleeting Cyclopentadienone. *J. Org. Chem.* **2002**, *68*, 6035–6038.

5. Y. Oyola, D. A. Singleton, Dynamics and the Failure of Transition State Theory in Alkene Hydroboration. *J. Am. Chem. Soc.* 2009, **131**, 3130–3131.

6. J. O. Bailey, D. A. Singleton, Failure and Redemption of Statistical and Nonstatistical Rate Theories in the Hydroboration of Alkenes. *J. Am. Chem. Soc.* 2017, **139**, 15710–15723.

7. Z. Chen, Y. Nieves-Quinones, J. R. Waas, D. A. Singleton, Isotope Effects, Dynamic Matching, and Solvent Dynamics in a Wittig Reaction. Betaines as Bypassed Intermediates. *J. Am. Chem. Soc.* 2014, **136**, 13122–13125.

8. H. R. Aziz, D. A. Singleton, Concert along the Edge: Dynamics and the Nature of the Border between General and Specific Acid-Base Catalysis. *J. Am. Chem. Soc.* 2017, **139**, 5965–5972.

9. X. S. Bogle, D. A. Singleton, Dynamic Origin of the Stereoselectivity of a Nucleophilic Substitution Reaction. *Org. Lett.* 2012, **14**, 2528–2531.

10. J. G. Lopez, G. Vayner, U. Lourderaj, S. V. Addepalli, S. Kato, W. A. de Jong, T. L. Windus, W. A. Hase, A Direct Dynamics Trajectory Study of F^- + CH_3OOH Reactive Collisions Reveals a Major Non-IRC Reaction Path. *J. Am. Chem. Soc.* 2007, **129**, 9976–9985.

11. J. Xie, R. Otto, J. Mikosch, J. Zhang, R. Wester, W. L. Hase, Identification of Atomic-Level Mechanisms for Gas-Phase X^- + CH_3Y S_N2 Reactions by Combined Experiments and Simulations. *Acc. Chem. Res.* 2014, **47**, 2960–2969.

12. P. Manikandan, J. Zhang, W. L. Hase, Chemical Dynamics Simulations of $X^- + CH_3Y \rightarrow XCH_3 + Y^-$ Gas-Phase S_N2 Nucleophilic Substitution Reactions. Nonstatistical Dynamics and Nontraditional Reaction Mechanisms. *J. Phys. Chem. A* 2012, **116**, 3061–3080.

13. Z. Wang, J. S. Hirschi, D. A. Singleton, Recrossing and Dynamic Matching Effects on Selectivity in a Diels-Alder Reaction. *Angew. Chem. Int. Ed. Engl.* 2009, **48**, 9156–9159.

14. P. Yu, T. Q. Chen, Z. Yang, C. Q. He, A. Patel, Y.-h. Lam, C.-Y. Liu, K. N. Houk, Mechanisms and Origins of Periselectivity of the Ambimodal [6 + 4] Cycloadditions of Tropone to Dimethylfulvene. *J. Am. Chem. Soc.* 2017, **139**, 8251–8258.

15. B. Biswas, D. A. Singleton, Controlling Selectivity by Controlling the Path of Trajectories. *J. Am. Chem. Soc.* 2015, **137**, 14244–14247.

16. B. Biswas, S. C. Collins, D. A. Singleton, Dynamics and a Unified Understanding of Competitive [2,3]- and [1,2]-Sigmatropic Rearrangements Based on a Study of Ammonium Ylides. *J. Am. Chem. Soc.* 2014, **136**, 3740–3743.

17. S. R. Hare, A. Li, D. J. Tantillo, Post-transition state bifurcations induce dynamical detours in Pummerer-like reactions. *Chem. Sci.* 2018, **9**, 8937–8945.

18. R. P. Pemberton, D. J. Tantillo, Lifetimes of carbocations encountered along reaction coordinates for terpene formation. *Chem. Sci.* 2014, **5**, 3301–3308.

19. Y. J. Hong, D. J. Tantillo, Biosynthetic consequences of multiple sequential post-transition-state bifurcations. *Nat. Chem.* 2014, **6**, 104–111.

20. M. R. Siebert, P. Manikandan, R. Sun, D. J. Tantillo, W. L. Hase, Gas-Phase Chemical Dynamics Simulations on the Bifurcating Pathway of the Pimaradienyl Cation Rearrangement: Role of Enzymatic Steering in Abietic Acid Biosynthesis. *J. Chem. Theory Compu.* 2012, **8**, 1212–1222.

21. M. R. Siebert, J. Zhang, S. V. Addepalli, D. J. Tantillo, W. L. Hase, The need for enzymatic steering in abietic acid biosynthesis: Gas-phase chemical dynamics simulations of carbocation rearrangements on a bifurcating potential energy surface. *J. Am. Chem. Soc.* 2011, **133**, 8335–8343.

22. D. R. Glowacki, S. Marsden, M. J. Pilling, Significance of Nonstatistical Dynamics in Organic Reaction Mechanisms: Time-Dependent Stereoselectivity in Cyclopentyne−Alkene Cycloaddition. *J. Am. Chem. Soc.* 2009, **131**, 13896–13897.

23. C. Doubleday, C. P. Suhrada, K. N. Houk, Dynamics of the Degenerate Rearrangement of Bicyclo[3.1.0]hex-2-ene. *J. Am. Chem. Soc.* 2006, **128**, 90–94.

24. T. Bekele, C. F. Christian, M. A. Lipton, D. A. Singleton, “Concerted” Transition State, Stepwise Mechanism. Dynamics Effects in C2-C6 Enyne Allene Cyclizations. *J. Am. Chem. Soc.* 2005, **127**, 9216–9223.

25. C. Doubleday, G. Li, W. L. Hase, Dynamics of the biradical mediating vinylcyclopropane-cyclopentene rearrangement. *Phys. Chem. Chem. Phys.* 2002, **4**, 304–312.

26. C. Doubleday, C. P. Suhrada, K. N. Houk, Dynamics of the degenerate rearrangement of bicyclo[3.1.0] hex-2-ene. *J. Am. Chem. Soc.* 2006, **128**, 90–94.

27. C. Doubleday, M. Nendel, K. N. Houk, D. Thweatt, M. Page, Direct Dynamics Quasiclassical Trajectory Study of the Stereochemistry of the Vinylcyclopropane-Cyclopentene Rearrangement. *J. Am. Chem. Soc.* 1999, **121**, 4720–4721.

28. J. Rehbein, B. K. Carpenter, Do we fully understand what controls chemical selectivity?. *Phys. Chem. Chem. Phys.* 2011, **13**, 20906–20922.

29. B. K. Carpenter, Nonstatistical Dynamics in Thermal Reactions of Polyatomic Molecules. *Annu. Rev. Phys. Chem.* 2005, **56**, 57–89.

30. H. Yamataka, Molecular dynamics simulations and mechanism of organic reactions: non-TST behaviors. *Adv. Phys. Org. Chem.* 2010, **44**, 173–222.

31. X.-S. Xue, C. S. Jamieson, M. Garcia-Borras, X. Dong, Z. Yang, K. N. Houk, Ambimodal Trispericyclic Transition State and Dynamic Control of Periselectivity. *J. Am. Chem. Soc.* 2019, **141**, 1217–1221.

32. U. Lourderaj, K. Park, W. L. Hase, Classical trajectory simulations of post-transition state dynamics. *Int. Rev. Phys. Chem.* 2008, **27**, 361–403.

33. S. Lee, J. M. Goodman, Rapid Route-Finding for Bifurcating Organic Reactions. *J. Am. Chem. Soc.* 2020, **142**, 9210–9219.

34. Z. Yang, X. Dong, Y. Yu, P. Yu, Y. Li, C. Jamieson, K. N. Houk, Relationships between Product Ratios in Ambimodal Pericyclic Reactions and Bond Lengths in Transition Structures. *J. Am. Chem. Soc.* 2018, **140**, 3061–3067.

35. T. H. Peterson, B. K. Carpenter, Estimation of dynamic effects on product ratios by vectorial decomposition of a reaction coordinate. Application to thermal nitrogen loss from bicyclic azo compounds. *J. Am. Chem. Soc.* 1992, **114**, 766–767

36. J. Zheng, E. Papajak, D. G. Truhlar, Phase Space Prediction of Product Branching Ratios: Canonical Competitive Nonstatistical Model. *J. Am. Chem. Soc.* 2009, **131**, 15754–15760.

37. Meuwly, M. Machine Learning for Chemical Reactions. *Chem. Rev.* **2021**, *121*, 10218–10239.

38. Keith, J. A.; Vassilev-Galindo, V.; Cheng, B.; Chmiela, S.; Gastegger, M.; Müller, K-R.; Tkatchenko, A. Combining Machine Learning and Computational Chemistry for Predictive Insights Into Chemical Systems. *Chem. Rev.* **2021**, *121*, 9816–9872.

39. T. F. G. G. Cova, A. A. C. C. Pais, Deep Learning for Deep Chemistry: Optimizing the Prediction of Chemical Patterns. *Front. Chem.* 2019, **7**, 11–22.

40. N. Rollins, S. L. Pugh, S. M. Maley, B. O. Grant. R. S. Hamilton, M. S. Teynor, R. Carlsen, J. R. Jenkins, D. H. Ess, Machine Learning Analysis of Direct Dynamics Trajectory Outcomes for Thermal Deazetization of 2,3-Diazabicyclo[2.2.1]hept-2-ene. *J. Phys. Chem. A* 2020, **124**, 4813–4826.

41. Lyons, B. A.; Pfeifer, J. Peterson, T. H.; Carpenter, B. K. Dynamic models for the thermal deazetization of 2,3-diazabicyclo[2.2.1]hept-2-ene. *J. Am. Chem. Soc.* **1993**, *115*, 2427–2437.

42. Sorescu, D. C.; Thompson, D. L.; Raff, L. M. Molecular dynamics studies of the thermal decomposition of 2,3-diazabicyclo(2.2.1)hept-2-ene. *J. Chem. Phys.* **1995**, *20*, 7910–7924.

43. Maley, S. M.; Melville, J.; Yu, S.; Teynor, M. S.; Carlsen, R.; Hargis, C.; Hamilton, R. S.; Grant, B. O.; Ess, D. H. Machine Learning Classification of Disrotatory IRC and Conrotatory Non-IRC Trajectory Motion for Cyclopropyl Radical Ring Opening. *Phys. Chem. Chem. Phys.* **2021**, 23, 12309–12320.

44. Zhou, C.; Birney, D. M. Sequential Transition States and the Valley−Ridge Inflection Point in the Formation of a Semibullvalene. *Org. Lett.* **2002**, *4*, 3279–3282.

45. Kelly, K. K.; Hirschi, J. S.; Singleton, D. A. Newtonian Kinetic Isotope Effects. Observation, Prediction, and Origin of Heavy-Atom Dynamic Isotope Effects. *J. Am. Chem. Soc.* **2009**, *131*, 8382–8383.

46. Yang, Z.; Zou, L.; Yu, Y.; Liu, F.; Dong, X.; Houk, K. N. *Chem. Phys.* **2018**, *514*, 120–125.

47. Mandal, N.; Datta, A. Dynamical Effects along the Bifurcation Pathway Control Semibullvalene Formation in Deazetization Reactions. *J. Phys. Chem. B* **2018**, *122*, 1239–1244.

48. https://scikit-learn.org/stable/modules/ensemble.html#adaboost (accessed 1/1/2021).

49. https://scikit-learn.org/stable/modules/ensemble.html#forests-of-randomized-trees (accessed 1/1/2021).

50. Y. Zhao, D. G. Truhlar, *J. Chem. Phys.* 2006, **125**, 194101.

51. Gaussian 16, Revision B.01, 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, Inc., Wallingford CT, 2016.

52. Verhoeven, J. W. Glossary of terms used in photochemistry (IUPAC Recommendations 1996). *Pure and Applied Chemistry*, vol. 68, no. 12, 1996, pp. 2223-2286.

53. See https://scikit-learn.org/stable/ (accessed 1/1/2021).

54. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, b.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D. Scikit-learn: Machine Learning in Python, *J. Machine Learning Res.* **2011**, *12*, 2825-2830.

55. Sauer, J.; Bäuerlein, P.; Ebenbeck, W.; Schuster, J.; Sellner, I.; Sichert, H.; Stimmelmayr, H. An One-Pot Synthesis of Semibullvalenes and Its Mechanism *Eur. J. Org. Chem.* **2002**, *2002*, 791– 801.

56. Doubleday, C.; Boguslav, M.; Howell, C.; Korotkin, S. D.; Shaked, D. Trajectory Calculations for Bergman Cyclization Predict H/D Kinetic Isotope Effects Due to Nonstatistical Dynamics in the Product. *J. Am. Chem. Soc.* **2016**, *138*, 7476–7479.