Machine Learning Classification of Disrotatory IRC and Conrotatory Non-IRC Trajectory Motion for Cyclopropyl Radical Ring Opening

Steven M. Maley, Jesse Melville, Spencer Yu, Matthew S. Teynor, Ryan Carlsen, Cal Hargis, R. Spencer Hamilton, Benjamin O. Grant, and Daniel H. Ess *

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

*dhe@chem.byu.edu

Abstract

Quasiclassical trajectory analysis is now a standard tool to analyze non-minimum energy pathway motion of organic reactions. However, due to the large amount of information associated with trajectories, quantitative analysis of the dynamic origin of selectivity is complex. For the electrocyclic ring opening of the cyclopropyl radical, more than 4000 trajectories were run showing that the allyl radicals are formed through a mixture of disrotatory intrinsic reaction coordinate (IRC) motion as well as conrotatory non-IRC motion. Geometric, vibrational mode, and atomic velocity transition-state features from these trajectories were used for supervised machine learning analysis with classification algorithms. Accuracy >80% with a random forest model enabled quantitative and qualitative assessment of transition-state trajectory features controlling disrotatory IRC versus conrotatory non-IRC motion. This analysis revealed that there are two key vibrational modes where their directional combination provides prediction of IRC versus non-IRC motion.

Introduction

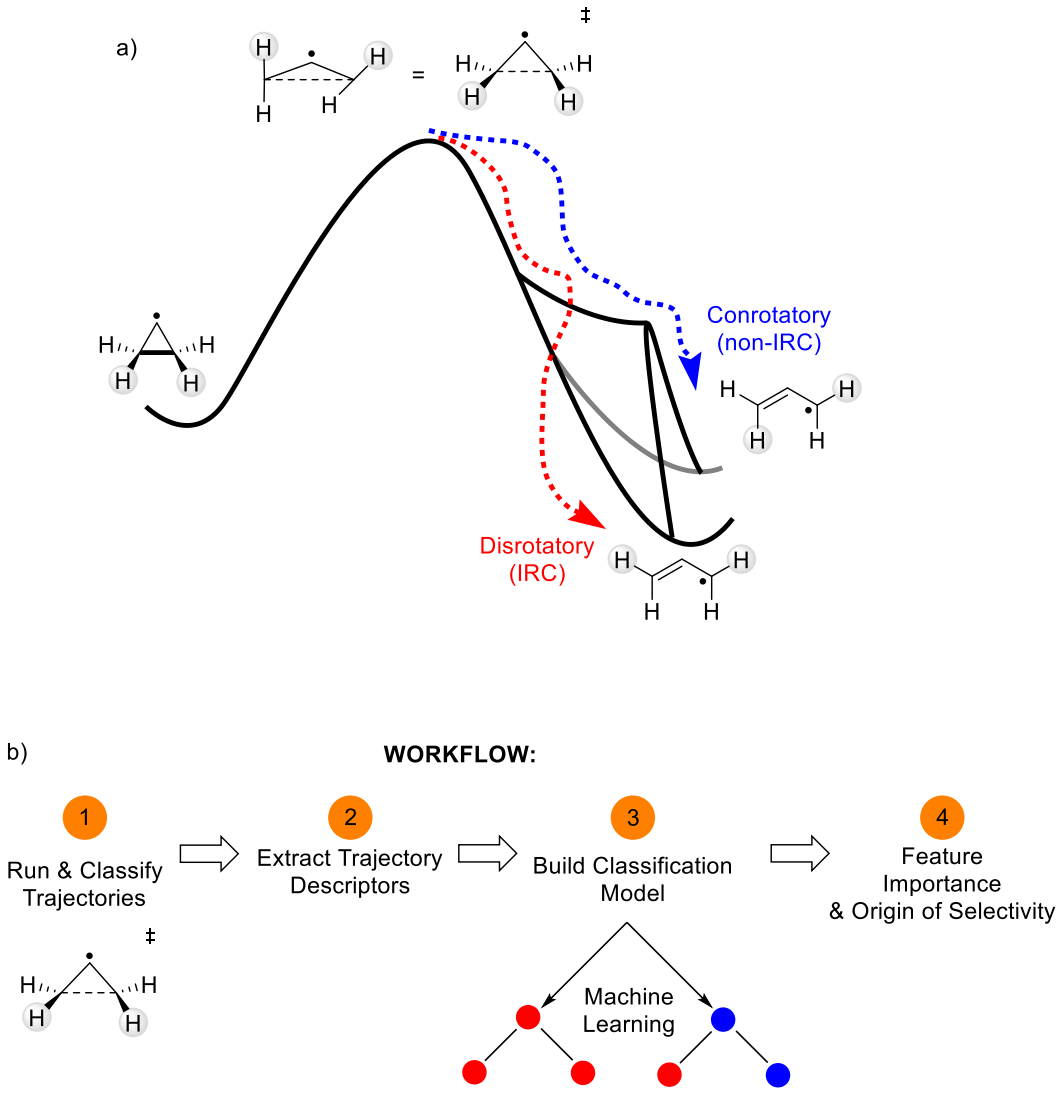
For many addition,[1,2] substitution,[3,4,5,6] pericyclic,[7,8] rearrangement,[9,10,11,12,13,14,15] and radical[16,17,18,19,20,21] organic reactions it has been established that transition state and other statistical theories sometimes do not provide adequate quantitative or qualitative treatment of reaction selectivity,[22,23] especially when there is post-transition state valley-ride inflection point.[24,25,26] In many of these cases, quasiclassical direct dynamics trajectories descending from the transition state provide quantitative treatment of the nonstatistical selectivity,[27,28,29] capturing either non-minimum energy pathway/non-intrinsic reaction coordinate (non-IRC) motion or incomplete intramolecular vibrational redistribution (IVR).[30]

While these organic reactions display dynamic selectivity, the origin of selectivity is generally not quantitatively analyzed due to the necessity of comparing hundreds or thousands of trajectories that each have a different starting vibrational-sampled atomic velocity configuration and structure as well as being propagated along a complex multi-dimensional energy landscape. Recently, there has been interest in developing qualitative protocols to predict dynamical reaction selectivity using only a few key points on an energy landscape or transition-state partial bond lengths,[31,32] and this type of approach was broadly introduced by Carpenter several years ago.[33] Truhlar has also proposed a quantitative method for assessment of nonstatistical effects without requiring trajectories.[34] While these approaches may speed up analysis, they neither provide direct analysis of trajectory selectivity nor quantitative assessment of the dynamic origin of selectivity.

Because of the large amount of information associated with quasiclassical direct dynamics simulations, machine learning is potentially well suited for quantitative and qualitative analysis. While machine learning has emerged as a popular tool in chemistry, most uses have focused on regression analysis to predict properties, such as reactivity.[35] Machine learning can also be used for classification, but this has been significantly less explored. Previously, we used classification-based machine learning to analyze and predict the outcome of quasiclassical trajectories for the thermal deazetization of 2,3-diazabicyclo[2.2.1]hept-2-ene,[36] which results in either the exo bicyclo product or a diradical intermediate.

Due to the relative complexity of this reaction, and chaotic trajectory behavior, supervised classification algorithms only provided poor (~60%) classification accuracy based on transition-state vibrational quanta and energy features, and only ~70% classification accuracy based on transition-state atomic velocities and atomic positions. Significantly better accuracy was only achieved (85-95%) using features from later trajectory time steps, and feature importance analysis showed the key predictive feature is the methylene bridge out-of-plane bending.

Based on what we learned from our classification analysis of deazetization trajectories, we wanted to identify and analyze a reaction where machine learning can provide quantitative classification accuracy at the transition state to enable analysis of dynamic selectivity based on transition-state vibrational mode features that provide a clear physical understanding. We chose to analyze the ring opening of cyclopropyl radical to allyl radical (Scheme 1) because: i) This ring opening is directly related to experimental cyclopropyl radical ring openings[37,38,39,40] and is an example within the general pericyclic electrocyclization reaction class. ii) This reaction has only three heavy atoms and we could calculate thousands of DFT direct dynamics trajectories. This was important because we worried that the relatively low number of trajectories was a possible reason for the poor machine-learning classification of the deazetization reaction. iii) Most important, despite a disrotatory ring opening IRC route, Hase previously reported 120 reactive direct dynamics trajectories that showed only a relatively small preference (57%) for disrotatory ring opening compared to conrotatory non-IRC motion (43%, Scheme 1a, compare red and blue dotted arrows).[41] Therefore, this system provides an opportunity to analyze IRC versus non-IRC motion due to dynamic reaction pathway branching that can only generally be understood to result from the post-transition state valley-ridge inflection point,[42,43,44,45] which creates in a dividing ridge separating disrotatory and conrotatory formed allyl radicals.[46,47]



Scheme 1. a) Energy landscape for cyclopropyl radical ring opening to allyl radical. The IRC pathway ring-opening transition state leads to the disrotatory formed allyl radical. The red dotted arrow represents trajectory motion along the IRC pathway. The blue dotted arrow represents trajectory motion that leads from the ring opening transition state to the allyl radical by conrotatory non-IRC motion. b) Outline of workflow that involves running and classifying quasiclassical direct dynamics trajectories, extracting trajectory descriptors at the transition state (e.g. vibrational mode quanta and atomic velocities), quantitively using supervised machine learning to classify trajectories based on transition-state features, and performing a feature importance analysis.

Here we report unrestricted M06-2X DFT quasiclassical direct dynamics trajectories of the cyclopropyl radical ring opening. More than 4000 trajectories were run showing that the allyl radicals are formed through a mixture of disrotatory IRC motion as well as conrotatory non-IRC motion, and the DFT ratio is very similar to the previous CASSCF ratio reported.[41] Geometric, vibrational mode, and atomic

velocity transition-state features from these trajectories were used for supervised machine learning analysis with classification algorithms (see workflow in Scheme 2b). Classification accuracy well above 80% was achieved with popular algorithms, which enabled quantitative and qualitative assessment of transition-state trajectory features controlling disrotatory IRC versus conrotatory non-IRC motion. This analysis revealed that there are two key vibrational modes where their directional combination provides prediction and possible physical control of IRC versus non-IRC motion. This analysis also showed that a subset of trajectories is nearly impossible for machine learning to classify, which are disproportionately conrotatory trajectories.

Results and Discussion

Structure and trajectory details

All energies, optimized structures, and trajectories were calculated with unrestricted M06-2X/6-31G**.[48] All structures have unrestricted SCF solutions with $<S^2>$ values close to 0.75 indicating very little spin contamination. Vibrational frequencies were computed to ensure the allyl radical ring opening transition-state, **TS1**, had only one imaginary frequency corresponding to the reaction coordinate and that the reactant and product had no imaginary frequencies.

Trajectories[49] were initialized and propagated from **TS1** in Gaussian 16.[50] Initialization of the trajectories was done using local mode and thermal sampling at 447 K, which includes zero-point energy (ZPE). Figure 1 shows an overlay and geometrical histograms of the starting positions for the 4149 trajectories. Trajectories were propagated in both forward and reverse directions for about 1200 fs in mass-weighted Cartesian velocities with an approximate step of 0.25 fs. In this procedure, the trajectory ensemble is initiated as a combination of kinetic and potential energy, which provides the ability to analyze both atomic velocities/momenta and geometry features using machine learning.

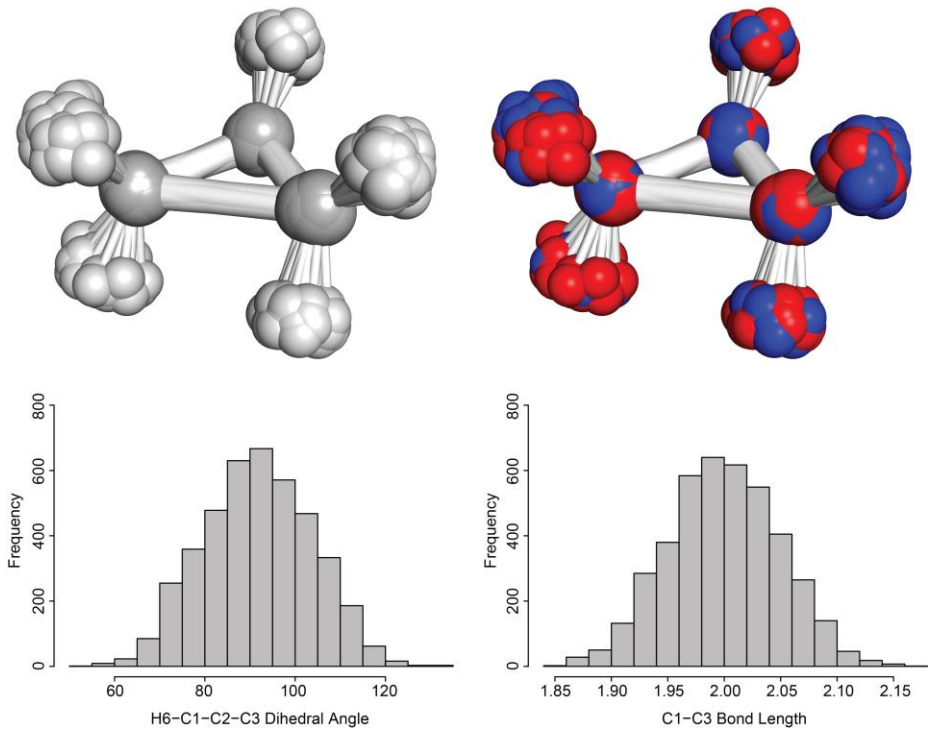


Figure 1. Top left: Overlay of trajectory starting geometries. Top right: Overlay of disrotatory (red) and conrotatory (blue) trajectory starting geometries. Bottom: For all trajectories, histograms of starting H6-C1-C2-C3 dihedral angles (in degrees) and C1-C3 bond lengths (Å).

Transition state, IRC pathway, and quasiclassical trajectories

Our M06-2X transition state for cyclopropyl radical ring opening, **TS1** (qualitatively shown in Scheme 1a), is geometrically extremely similar to previously reported transition states based on CASSCF and DFT methods. In this unsymmetrical transition state, the breaking C-C bond is severely stretched and diradical like (1.99 Å) and there is significant twisting of one methylene group, but very little twisting of the second methylene group. Importantly, we have previously shown that UM06-2X provides an accurate treatment of diradicals.[51]

As expected from inspection of the **TS1** geometry and previous qualitative and semiempirical analyses,[52,53,54,55,56,57,58,59,60] Olivella's early calculations showed the IRC route, which has infinitesimal velocity, from **TS1** descending to the allyl radical occurs through to very asynchronous rotation of the two methylene groups.[61] Along the IRC pathway, one methylene group rotates first followed by lagged rotation of the second methylene group to complete the planar allylic system. This highly asynchronous rotation of

methylene groups was interpreted by Olivella as meaning that **TS1** is a common transition state for disrotatory and conrotatory formed allyl radicals. However, because of the unsymmetrical transition state, Carpenter's B3LYP calculations confirmed that the IRC connection occurs between **TS1** and the disrotatory formed allyl radical.[62] Similarly, using ab initio methods, Liu also showed that the IRC connects **TS1** to the allyl radical through disrotatory ring motion.[63] We also confirmed that the UM06-2X IRC connects **TS1** with disrotatory ring opening motion. However, consistent with Olivella's initial suggestion, there is a valley-ridge inflection very close to, but not along the IRC pathway.

The valley-ridge inflection point very close to the IRC pathway, combined with the steep energy drop from the transition state to the allyl radical (>50 kcal/mol), suggests the possibility of significant non-IRC motion. Indeed, Hase reported 120 reactive CASSCF quasiclassical trajectories where 68/120 (57%) followed disrotatory IRC motion and 52/120 (43%) followed non-IRC conrotatory motion to the allyl radical. In these trajectories there was about 30-50 fs delay between rotation of each methylene group. Somewhat germane to this work, Hase suggested that a larger reaction coordinate translational energy can favor disrotatory motion. Mann showed through DFT/MM quasiclassical trajectories that only at very high condensed phase density does the solvent environment significantly impact the disrotatory versus conrotatory motion,[64,65] but it does inhibit rotations of the methylene groups after allyl radical formation. Kramer, Carpenter, and Wiggins also examined the dynamics of cyclopropyl radical ring opening using a reduced dimensional potential-energy surface containing the valley-ridge inflection point.[66] They found that the "decision" between disrotatory or conrotatory mechanisms occurs upon passage over the ridge structure on the potential surface. Additionally, they found that large amplitude motions of the allyl structure, such as the wag of the perpendicular methylene group and local mode bending of the perpendicular methylene hydrogens are important to pathway control.

Using UM06-2X, we completed 4149 fully connected reactive trajectories starting from **TS1**. Figure 2 provides snapshots for representative disrotatory and conrotatory trajectories. The example disrotatory trajectory is consistent with IRC motion. The methylene group labeled with H7 has significant motion prior to motion of the methylene group labeled with H6. This latter methylene group does not begin

twisting motion until about 20 fs beyond the transition state. In the example conrotatory trajectory, similar to the disrotatory trajectory, the H7 methylene twists clockwise into the plane of the carbon atoms well before twisting of the H6 methylene group. From the transition state, the allyl radical was generally formed between 50-100 fs.

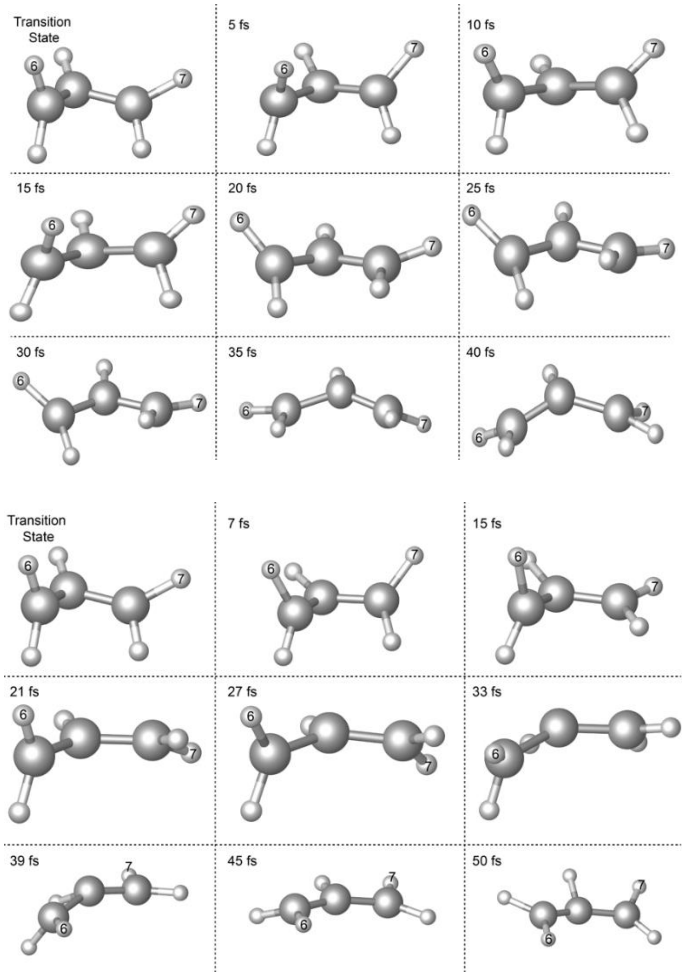


Figure 2. Top: Snapshots of an example disrotatory trajectory beginning from **TS1**. Bottom: Snapshots of an example conrotatory trajectory beginning from **TS1**.

2323/4149 (56%) trajectories led to the allyl radial through disrotatory motion and 1826/4149 (44%) trajectories led to the allyl radical through non-IRC conrotatory motion. The disrotatory:conrotatory ratio of 1.3:1 is nearly identical to Hase's ratio obtained with CASSCF(3,3)/6-31G* trajectories, which indicates that UM06-2X method is capable of capturing the key determining forces on the energy landscape. To confirm that this ratio is not initiation or propagation algorithm dependent, we also initiated and

propagated 979 trajectories with our DynSuite program,[36] which provides quasiclassical initiation with only atomic velocities (i.e. no geometric displacement) and Verlet integration. For these trajectories, 560 (57%) resulted in disrotatory motion and 419 (43%) resulted in conrotatory motion.

To visualize the disrotatory and conrotatory classes of trajectories, Figure 3 plots trajectory steps versus the H6-C1-C2-C3 dihedral angle. The red disrotatory trajectories show a decrease in the H6-C1-C2-C3 dihedral angle while the blue conrotatory trajectories shows an increase in this dihedral angle. This plot shows that by about 100 fs most trajectories have clearly separated into disrotatory and conrotatory motion. While this plot shows a simple classification method at and beyond 100 fs from the transition state until about 50 fs this single geometric value cannot easily distinguish between these two classes of trajectories.

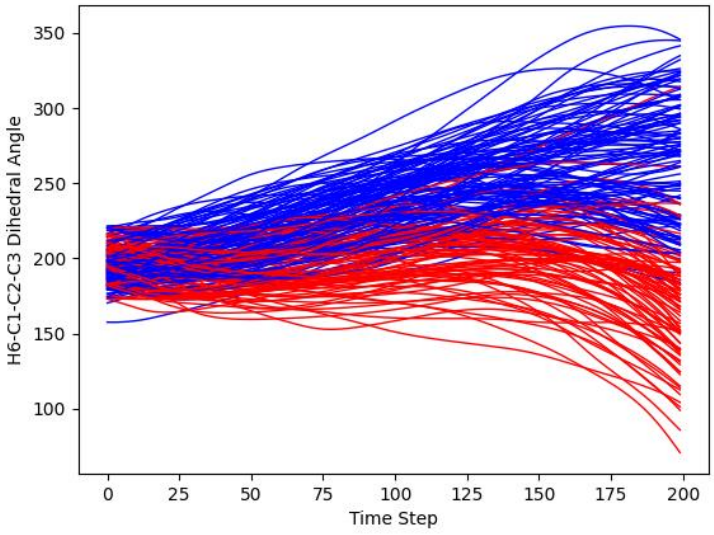


Figure 3. Plot of trajectory steps (fs) versus H6-C1-C2-C3 dihedral angle (in degrees). Red trajectories are classified as disrotatory motion and blue trajectories are classified as conrotatory motion.

Machine learning analysis

While dynamic selectivity of organic reactions has been recognized for several decades, and trajectory studies have been carried out ranging from parameterized semi-empirical methods to more modern DFT-based quasiclassical trajectories, the interpretation and analysis of trajectory results has typically been highly qualitative and not based on fundamental quantities, such as vibrational mode quanta

and directionality. This is because in a typical reaction the combination of vibrational modes, excitation quanta, and directionality results in an extraordinarily large number of starting trajectory configurations.

Machine learning provides a possible approach to analyze the importance of the very large number of features impacting the outcomes of trajectories.[67,68] Specifically, we wanted to perform classification prediction based on transition-state features. After running 4149 trajectories, we extracted 85 transition-state features, based on vibrational mode quanta, atomic velocities, and geometries. More specifically, the features extracted were bond lengths, bond angles, dihedral angles, directional atomic velocity components, vibrational mode mass-weighted velocities, and mass-weighted atomic displacements. These features were chosen as machine learning inputs because vibrational and atomic velocity sampling differentiate individual trajectories of the ensemble. X, y, and z component atomic velocities rather than component momenta were used because we earlier showed that momenta are a relative mass scalar and give identical machine learning results to velocities.[36]

During each machine learning analysis an equal number of disrotatory and conrotatory trajectories was maintained in the data set by random sampling and iteration, which results in a baseline accuracy of 50%. A 20-fold cross validation was used at each iteration to determine the classification accuracy of each model. This is done by dividing the sampled data set into 20-equally sized subsets, training the model on 19 of these, and then evaluating the predictive accuracy using the withheld subset. This was performed 20 times with a different subset withheld at each iteration. The reported accuracy of each model is the mean accuracy of all iterations where the accuracy is defined as the number of correct predictions divided by the total predictions.

The Scikit-Learn Python Library was used to set up and train classifiers. Source code detailing our machine learning workflow can be found in the Supplementary Information (SI). Seven supervised machine learning classification algorithms were selected: random forest, multilayer perceptron, gaussian process classifier, stochastic gradient descent, support vector machine, and logistic regression classifier (Figure 4). As mentioned earlier, the accuracy for each model was evaluated using cross validation averaged across sampling iterations. Random forest (accuracy: 81.9%) and logistic regression (accuracy: 80.5%) provided

the highest accuracy. K-nearest neighbor, multiLayer perceptron and stochastic gradient descent had mediocre performance ranging from 69.2-71.6% and gaussian process classifier and support vector machine both had very poor classification accuracy of 56.0%. The GridSearchCV method from the Scikit-Learn library was used to perform hyperparameter optimization for the random forest model. This method tested permutations of different parameters and used five-fold cross validation to determine the set of hyperparameters that maximized classification accuracy. This optimized random forest model was then applied to the data set for validation and showed statistically real, but only slight accuracy improvement to 82.9%. The more modern XGBoost algorithm showed the highest accuracy at 84.5%. The >80% accuracy prediction using only transition-state features is a major improvement over our previous results for the deazetization of 2,3-diazabicyclo[2.2.1]hept-2-ene that generally showed between 65-69% accuracy with the best performing machine learning methods. More importantly, this greater than 80% accuracy provides the possibility to quantitatively analyze features that contribute to the machine learning model and control disrotatory versus conrotatory motion.

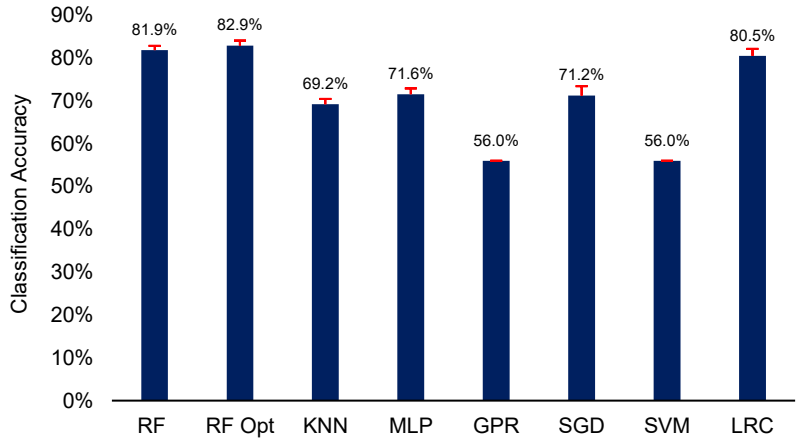


Figure 4. Plot of classification accuracy of random forest (RF), optimized random forest (RF Opt), K-nearest neighbour (KNN), multiLayer perceptron (MLP), gaussian process regression (GPR), stochastic gradient descent (SGD), support vector machine (SVR) and logistic regression classifier (LRC). The accuracy values from each machine learning model is the mean accuracy of all 20 data set iterations where the accuracy is defined as the number of correct predictions divided by the total number of predictions.

Before analyzing the importance of features contributing to the random forest machine learning model, it is useful to make a few comments on why near 100% accuracy was not obtained. First, with >4000

trajectories it is unlikely that addition of more trajectories would significantly increase accuracy. Stated another way, the 4000 trajectories calculated in this work likely represents a reasonable statistically sampled set for the complete set of possible trajectories, especially the most probable trajectories. Consistent with this thinking, we also determined the accuracy for smaller subsets of trajectories. The optimized random forest model was applied to datasets with 200, 400, 600, 800, 1000, and 1200 total trajectories. The accuracy of the model was only ~70% with 200 trajectories and increased to ~84% accuracy for analysis of 800 trajectories. The precision of the random forest model increases as the number of trajectories analyzed was increased. Second, inherent in classical trajectory propagation, although sampled using quasiclassical method, is the possibility of chaos, which is the sensitivity of the outcome to initial conditions. In this reaction, **TS1** has a very elongated C-C bond and there is a ridge region that divides the allyl radical structures, which potentially induces chaos. It is generally viewed that popular machine learning methods struggle to accurately make predictions of chaotic systems. Last, there is also the possibility that some trajectories are inherently difficult to classify due to indistinguishable feature values.

Machine learning feature importance

A significant advantage of decision forest ensemble learning for classification is that once a relatively high accuracy model is obtained, features can be analyzed to determine which features are most significant in classification. Random forest gave an accuracy prediction of ~83% which is sufficient for feature importance analysis. The relative importance of features can be determined by random forest models by replacing their values with random values and observing the change in root mean square error (RMSE). If replacing feature values with a random value has little or no impact on the RMSE then the feature has low importance in the model. Conversely, if there is a large change in the RMSE then the feature has a large importance for prediction. Figure 5 plots the weights of the importance versus geometric, atomic velocity, and vibrational mode features used in the random forest model. Strikingly, this plot of feature importance clearly shows that vibrational mode 2 and vibrational mode 4 velocities are the critical features in determining disrotatory and conrotatory trajectory outcomes. The weights of vibrational modes 2 and 4 to

the random forest model are many times greater than all other features. Mode 1 is the transition-state reaction coordinate negative vibrational mode, and the kinetic energy from this motion is not highly important. Importantly, the relative feature importance does not change if a smaller trajectory set is used for analysis (see SI). As another confirmation that modes 2 and 4 are highly important features, removal of these two features showed a decrease in model prediction accuracy, and as expected, velocities for atoms 6 and 8 become the most important feature. Removal of all velocity features results in a model with accuracy of less than 60%.

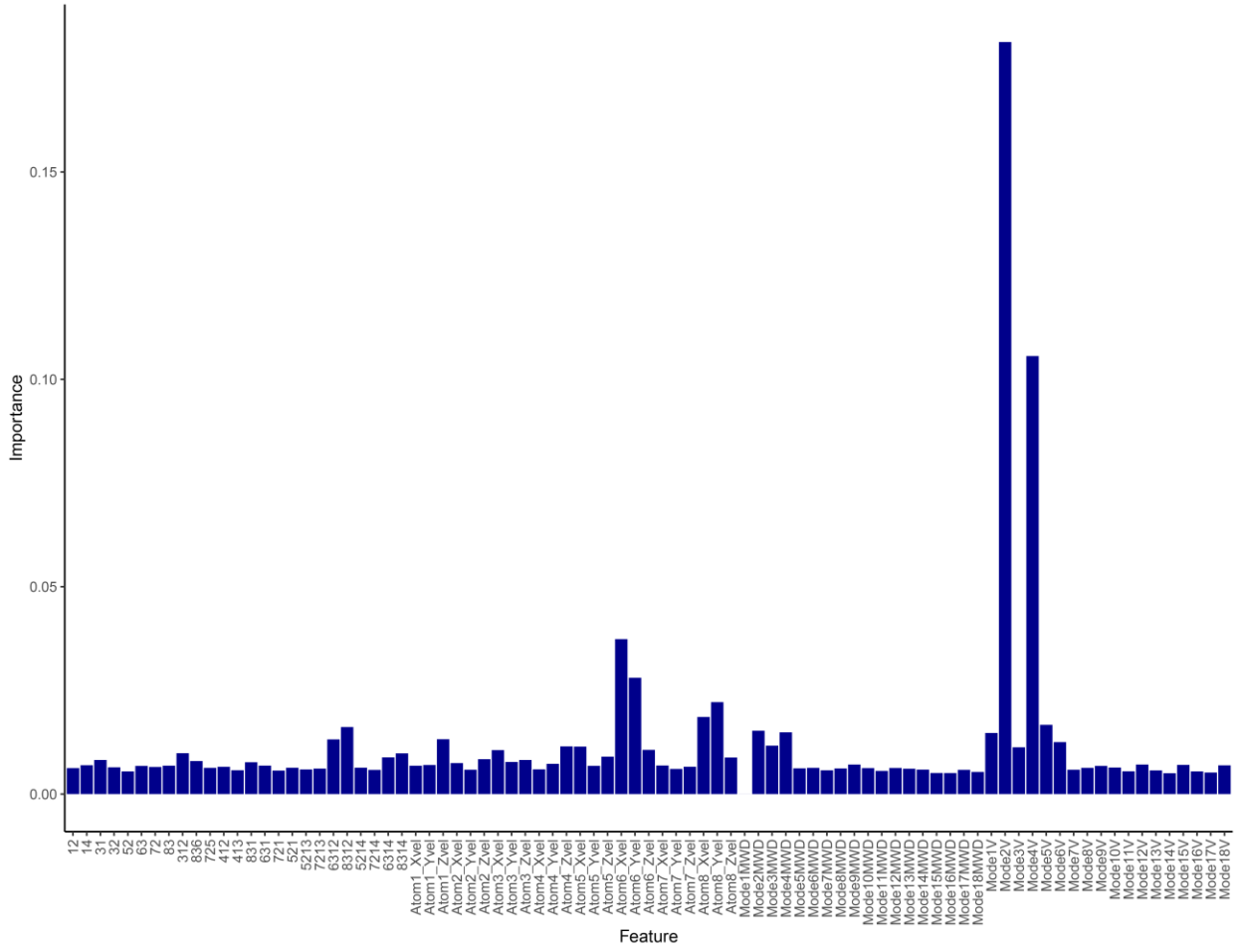


Figure 5. Classification model contributions/feature importance (weighted from 0.0 to 1.0) for the optimized random forest model. Two-digit numbers refer to atom bond lengths, three-digit numbers refer to angles, and four-digit numbers refer to dihedral angles. Atomic velocities are coded as: Atom#_Xvel, Atom#_Yvel, and Atom#_Zvel. Vibrational mode mass-weighted atomic displacements are coded as: Mode#MWVD. Vibrational mode mass-weighted velocities are coded as: Mode#V. Vibrational mode 1 is the transition-state structure negative vibrational mode.

Figure 6 displays vector representations of the motion associated with **TS1** vibrational modes 2 and 4. Importantly, the velocities associated with modes 2 and 4 correspond to methylene pyramidalization and twisting, which from a physical/chemical perspective are reasonably expected to impact disrotatory versus conrotatory motion. Consistent with the importance of the velocities from these vibrational modes, the x and y velocity components of hydrogens 6 and 8, are the most important velocity features in the data set, although their importance is significantly less than the vibrational mode velocities. Vibrational mode velocities are useful features because they encode both vibrational quanta/energy as well as direction. A random forest model using only vibrational quanta provides very poor accuracy prediction, which highlights the need to encode mode directionality.

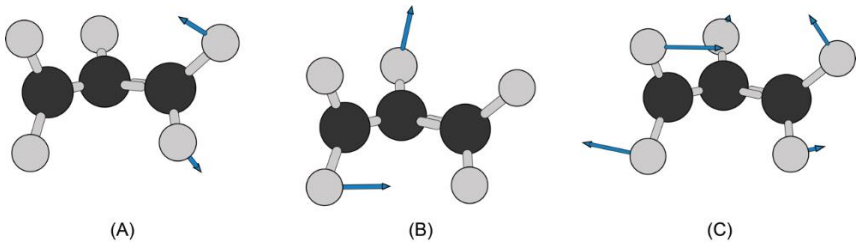


Figure 6. Qualitative vector representation of **TS1** (A) reaction coordinate motion, (B) vibrational mode 2, and (C) vibrational mode 4.

To further analyze the features identified as important to the random forest model, Figure 7 displays density plots showing the distribution of feature values and color coding for disrotatory and conrotatory trajectories. As expected, due to the smaller model contribution by the x and y components of the atom 6 and 8 velocities, there is significant overlap, but distinguishable peaks, of the red and blue distributions. In contrast, for the **TS1** velocities of vibrational modes 2 and 4 there is very clear separation of density distributions. The red and blue peaks and significant sections of the shoulders are clearly separated. However, there is some minor overlap of red and blue, which is consistent with less than perfect accuracy attained by the random forest model.

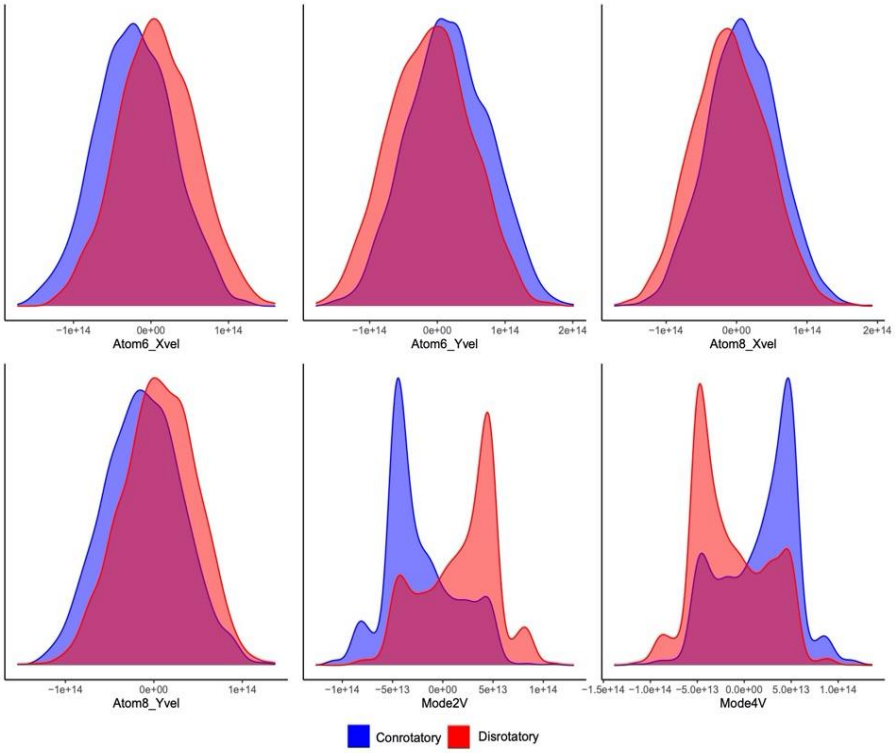


Figure 7. Density plots for the six most important transition-state features identified by the optimized Random Forest model. The velocity features have significant overlap whereas the mode 2 and mode 4 velocities show clear peak separation.

The mode velocity distribution plots in Figure 7 also reveal the importance of directionality associated with disrotatory versus conrotatory ring opening motion. For example, for mode 2 (labeled as Mode2V), the highest density with positive direction values predicts disrotatory motion while the highest density with negative values predicts conrotatory motion. Opposite to mode 2, mode 4 (labeled as Mode4V) predicts conrotatory motion for the highest density with positive values and disrotatory motion with the highest density of negative values. In a similar analysis, Figure 8 plots the relative percentage of disrotatory and conrotatory trajectories versus the directionality and combination of vibrational mode 2 and 4 velocities. For example, when mode 2 and mode 4 both have positive directional values, there is about 60%:40% disrotatory trajectories to conrotatory non-IRC trajectories, which is close to the overall disrotatory to conrotatory ratio. When mode 2 is negative and mode 4 is negative there is a minor shift from disrotatory motion to conrotatory motion. Figure 8 also shows when mode 2 is positive and mode 4 is

negative there is a large preference for disrotatory motion and when mode 2 is negative and mode 4 is positive there is a large preference for conrotatory motion. This suggests there is a matching and antimatching impact from the directionality of modes 2 and 4. This is consistent with the vibrational vector modes displayed in Figure 6 where mode 2 and 4 have opposite directionality for the hydrogen on the left-hand methylene group.

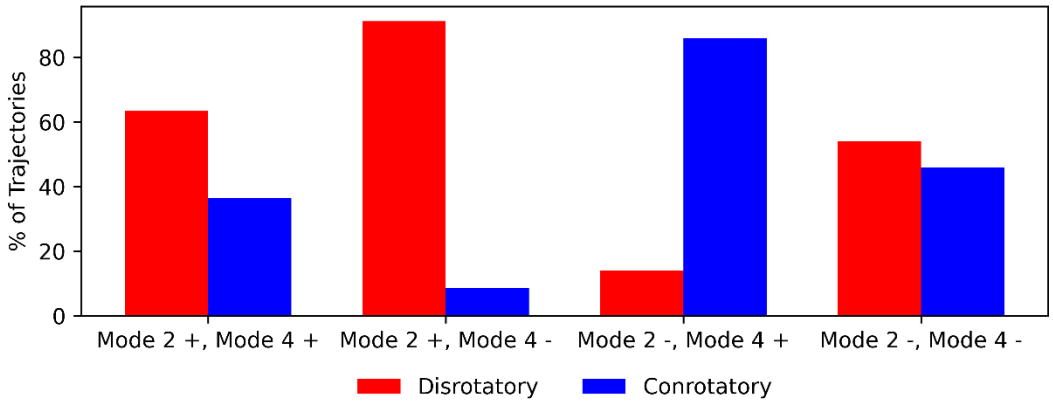


Figure 8. Relative percentages of disrotatory and conrotatory trajectories with directional combination of vibrational mode 2 and 4 velocities. The total percentage for each comparison equals 100. Mode2 +, Mode 4 + refers to vibrational modes 2 and 4 both having positive directional velocity. Mode2 +, Mode 4 - refers to vibrational mode 2 with positive directional velocity and mode 4 with negative vibrational mode velocity. Mode2 -, Mode 4 + refers to vibrational mode 2 with positive directional velocity and mode 4 with negative vibrational mode velocity. Mode2 -, Mode 4 - refers to vibrational modes 2 and 4 both having negative directional velocity. Importantly, the definition of positive and negative velocity is consistent for all trajectories.

To confirm the relationship between mode 2 and mode 4 velocity directionality and trajectory outcome, a small set of trajectories were run where the mode quanta and directionality were manually controlled. For a specific trajectory starting configuration with identical reaction coordinate energy, the velocity magnitude of modes 2 and 4 were increased from zero-point energy (ZPE) to two times and four times ZPE in both the positive and negative. This resulted in 25 different trajectories for each starting configuration. This was done for a conrotatory and disrotatory trajectory from each combination shown in Figure 8.

The IRC disrotatory motion corresponds to a clockwise twisting of the H7 methylene group followed by counterclockwise twisting of the H6 methylene group. Motion of mode 2 in the positive direction and mode 4 in the negative direction both contribute to counterclockwise motion of the H6 methylene group, which both match the IRC motion. In contrast, mode 2 motion in the negative direction and mode 4 motion in the positive direction should result in clockwise twisting of the H6 methylene group, which should generally dimmish IRC and enhance conrotatory motion. Consistent with this analysis, for the trajectories where we manually controlled mode quanta and directionality, when mode 4 is positive and mode 2 is negative there was only conrotatory trajectory outcomes. When mode 2 is positive all trajectories, except those where mode 4 is positive and two times or four times ZPE, resulted in disrotatory motion. This could indicate that selectivity for conrotatory in Figure 8 with positive directions for modes 2 and 4 arises because mode 4 is dominated by trajectories with only ZPE.

Trajectories that are difficult to classify

With the less than 100% accuracy by the random forest model, we wondered if there were specific trajectories that are inherently difficult to classify. To examine this issue, using an adaboost-type random forest model, the entire data set was iterated over 100 times and at each iteration the data set was split in 10 equally sized subsets. The random forest model was trained on 8 of the subsets and predictions were made for the individual trajectories of the remaining two data sets. The subsets used in the training and testing sets were permuted and the process was repeated until each subset appeared in the testing set twice. Therefore, after 100 iterations, each trajectory was classified 200 times. Table 1 reports the number of disrotatory and conrotatory trajectories binned into accuracy categories of 0%, 1-20% 20-40% 40-70%, 70-90%, 90-99%, and 100%. Stated another way, these bins display the percentage of accuracy a specific trajectory was accurately predicted out of the 200 times it was evaluated.

Table 1. Number and accuracy of conrotatory and disrotatory trajectories correctly classified after 200 machine learning predictions.

Accuracy (%)	Disrotatory Trajectories	Conrotatory Trajectories	Total Trajectories (% of the total trajectories)
0	58	189	247 (5.9%)
1-20	79	170	249 (6.0%)
20-40	70	86	156 (3.7%)
40-70	92	141	233 (5.6%)
70-90	121	146	267 (6.4%)
90-99	292	211	503 (12.1%)
100	1610	883	2493 (60.1%)

Expected from the overall >80% accuracy of the random forest model, the majority of trajectories (60.1%) were classified correctly 100% of the time. Interestingly, 5.9% of trajectories were incorrectly classified 200 times. The relative number of conrotatory trajectories in the 0% accuracy group is approximately twice that of the 100% accuracy group. The higher accuracy groups generally have fewer relative conrotatory trajectories. Importantly, while the high accuracy (90-99% and 100%) groups generally have more disrotatory than conrotatory trajectories, they also contain a significantly larger proportion of the total number of conrotatory trajectories in the entire trajectory data set. For example, in the 100% group, there are 883 of the total conrotatory trajectories.

To demonstrate why the random forest model struggles to predict the trajectories in the 0% group, Figure 9 shows mode 2 and mode 4 velocity density plots. Comparison of the top plots for 0% accuracy versus the bottom plots for 100% accuracy shows that despite feature separation the values are completely inverted in their directionality. For example, in the 100% accuracy predictions mode 2 velocities are positive for disrotatory trajectories are negative for conrotatory trajectories. In the 0% accuracy group, the disrotatory trajectories have negative velocities while the conrotatory trajectories have positive velocities. For accuracies between 20-70% the majority of the red and blue densities are heavily overlapped.

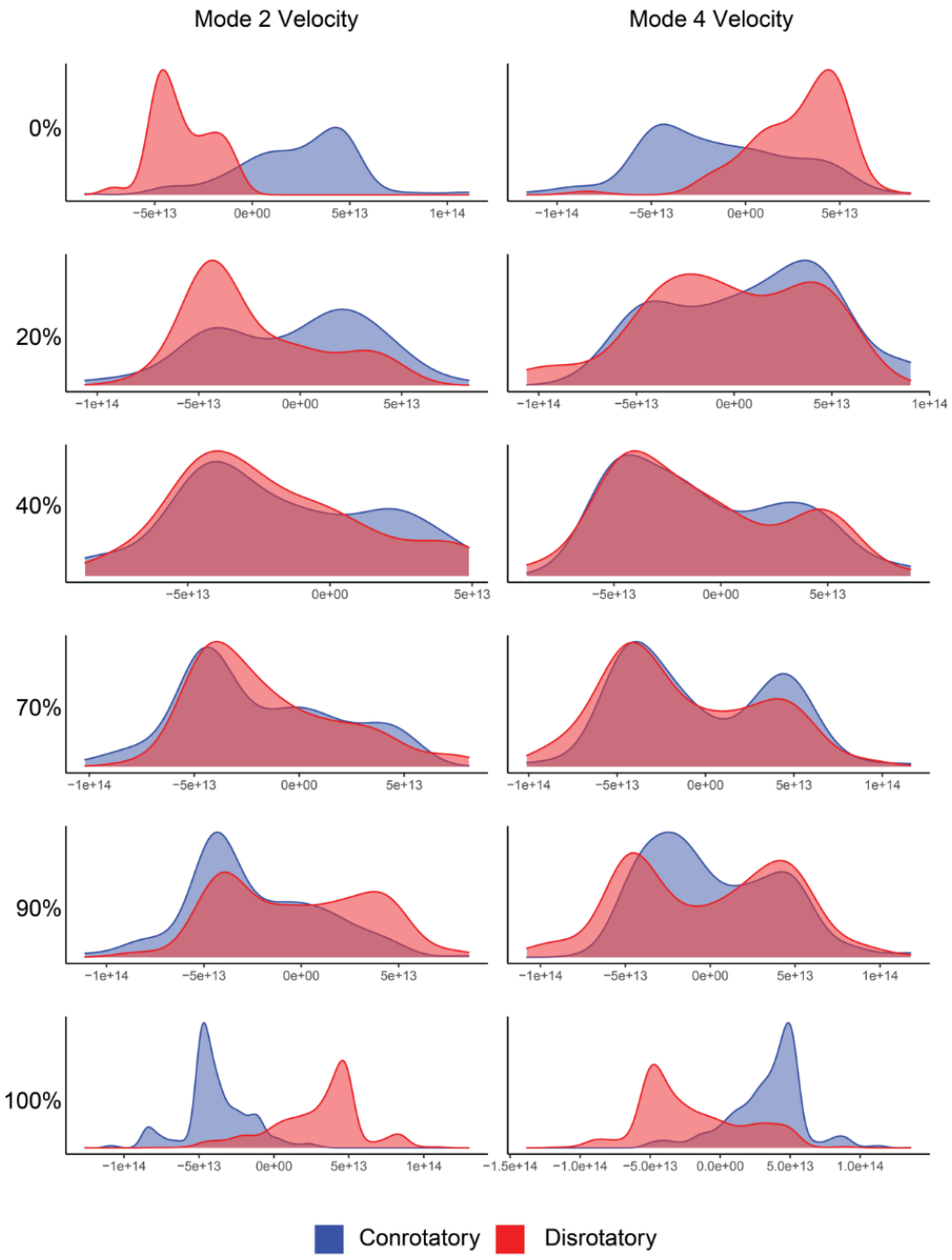


Figure 9. Density plots for mode 2 and mode 4 velocities for 0%, 20%, 40%, 70%, 90% and 100% accuracy groupings.

Conclusions

Greater than 4000 M06-2X quasiclassical trajectories for the electrocyclic ring-opening of the cyclopropyl radical provided a data set for machine learning classification of disrotatory versus conrotatory

ring opening motion. Using geometric, velocity, and vibrational features of the transition state, supervised decision forest classification algorithms provided 83% prediction accuracy. This relatively high accuracy enabled the quantitative assessment of feature importance to reveal that the velocities and direction of transition-state vibrational mode 2 and mode 4 have significant influence on disrotatory versus conrotatory motion. This is consistent with the vibrational vectors of modes 2 and 4 that have opposite directionality for motion of one of the methylene groups. The ability of these two vibrational modes to be used as a predictor of trajectory outcome, and physical origin of relative ring opening motion, was demonstrated by manually controlling mode quanta and directionality for a set of trajectories. Despite the random forest machine learning model providing relatively high accuracy for prediction of disrotatory versus conrotatory motion, we found that nearly 250 trajectories (out of 4000) could never be classified correctly, and these are disproportionately conrotatory trajectories. This is potentially the result of the random forest model too strongly correlating prediction based on vibrational modes 2 and 4.

Supplementary Information

Example data sets and Jupyter notebooks with code for machine learning model training and testing.

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., S. Y., M. S. T., C. H., R. S. H., and B. O. G. thank the BYU Department of Chemistry and Biochemistry for undergraduate research awards.

References

1. 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.

2. 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.

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

4. 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.

5. 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.

6. 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.

7. 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.

8. 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.

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

10. 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.

11. 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.

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

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

14. 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.

15. 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.

16. 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.

17. 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.

18. 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.

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

20. 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.

21. 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.

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

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

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

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

26. 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.

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

28. 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.

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

30. S. Karmakar, S. Keshavamurthy, Intramoleuclar vibrational energy redistribution and the quantum ergodicity transition: a phase space perspective. *Phys. Chem. Chem. Phys.* 2020, **22**, 11139-11173

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

32. 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.

33. 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

34. 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.

35. 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.

36. 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.

37. J. A. Kerr, A. Smith, A. F. Trotman-Dickenson, Reactions of Cyclopropyl Radicals in the Methyl-Initiated Decomposition of Cyclopropanecarbaldehyde. *J. Chem. Soc. A* 1969, 1400–1403.

38. R. P. Corbally, M. J. Perkins, A. Elnitski, *J. Chem. Soc. Perkin I* 1979, 793–798.

39. G. Greig, J. C. J. Thynne, Reactions of Cyclic Alkyl Radicals. Part 2. Photolysis of Cyclopropane Carboxaldehyde. *Trans. Faraday Soc.* 1967**, 63**, 1369–1374.

40. S. Sustmann, C. Ruechardt, A. Bieberbach, G. Boche, Stereochemistry and rate of electrocyclic ring opening reactions of cyclopropyl radicals. I. *Tet. Lett.* 1972, **47**, 4759–4764.

41. D. J. Mann, W. L. Hase, Ab Initio Direct Dynamics Study of Cyclopropyl Radical Ring-Opening. *J. Am. Chem. Soc.* 2002, **124**, 3208–3209.

42. W. Quapp, J. M. Bofill, A. Aguilar-Mogas, Exploration of cyclopropyl radical ring opening to allyl radical by Newton trajectories: importance of valley-ridge inflection points to understand the topography. *Theor. Chem. Acc.* 2011, **129**, 803–821.

43. W. Quapp, J. M. Bofill, Topography of cyclopropyl radical ring opening to allyl radical on the CASSCF(3,3) surface: valley-ridge inflection points by Newton trajectories. *J. Math. Chem.* 2012, **50**, 2061–2085.

44. W. Quapp, How does a path branching take place? A classification of bifurcation events. *J. Mol. Struct.* 2004, **695-696**, 95–101.

45. W. Quapp, M. Hirsch, D. Heidrich, Bifurcation of reaction pathways: the set of valley ridge inflection points of a simple three-dimensional potential energy surface. *Theor. Chem. Acc.* 1998, **100**, 285–299.

46. P. Collins, B. K. Carpenter, G. S. Ezra, S. Wiggins, Nonstatistical dynamics on potentials exhibiting reaction path bifurcations and valley-ridge inflection points. *J. Chem. Phys.* 2013, **139**, 154108.

47. P. Collins, Z. C. Kramer, B. K. Carpenter, G. S. Ezra, S. Wiggins, Nonstatistical dynamics on the caldera. *J. Chem. Phys.* 2014, **141**, 034111.

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

49. All 4149 trajectories are fully connected from cyclopropyl radical to allyl radical. All recrossing trajectories were removed from this set.

50. 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.

51. D. H. Ess, T. C. Cook, Unrestricted Prescriptions for Open-Shell Singlet Diradicals: Using Economical Ab Initio and Density Functional Theory to Calculate Singlet-Triplet Gaps and Bond Dissociation Curves. *J. Phys. Chem. A* 2012, **116**, 4922-4929.

52. R. B. Woodward, R. Hoffmann, Stereochemistry of Electrocyclic Reactions. *J. Am. Chem. Soc.* 1965, **87**, 395–397.

53. G. Szeimies, G. Boche, Electrocyclic Reactions of Radicals. *Angew. Chem., Int. Ed. Engl.* 1971, **10**, 911–912.

54. M. J. S. Dewar, S. Kirschner, MINDO/2 Study of Aromatic ("Allowed") Electrocyclic Reactions of Cyclopropyl and Cyclobutene. *J. Am. Chem. Soc.* 1971, **93**, 4290-4291.

55. J. R. Bews, C. Glidwell, J. C. Walton, Homolytic ring fission reactions of cycloalkylmethyl and bicycloalykl radicals. *J. Chem. Soc., Perkin Trans. 2* 1982, 1447-1453.

56. S. Beran, R. Zahradnik, MO study of the reactivity of cyclopropane, its radical and radical ions, and models of its transition metal complexes. *Collect. Czech. Chem. Commun.* 1976, **41**, 2303-2319.

57. M. J. S. Dewar, S. Kirschner, Nature of the Transition States in "Forbidden" Electrocyclic Reactions. *J. Am. Chem. Soc.* 1974, **96**, 5244–5246.

58. L. Farnell, W. G. Richards, *Ab initio* calculations on the electrocyclic transformation of the cyclopropyl radical to the allyl radical. *J. Chem. Soc., Chem. Commun.* 1973, 334–335.

59. P. Merlet, S. D. Peyerimhoff, R. J. Buenker, S. Shih, Ab initio SCF and CI [configuration interaction] study of the electrocyclic transformations of cyclopropyl and allyl systems. *J. Am. Chem. Soc.* 1974, **96**, 959–969.

60. L. Farnell, W. G. Richards, Ab initio Calculations on the Electrocyclic Transformation of the Cyclopropyl Radical to the Allyl Radical. *J. Chem. Soc. Chem. Comm.* 1973, 334–335.

61. S. Olivella, A. Solé, J. M. A. Bofill, Theoretical Investigation of the Thermal Ring Opening of Cyclopropyl Radical into Allyl Radical. Evidence for a Highly Nonsymmetric Transition State. *J. Am. Chem. Soc.* 1990, **112**, 2160–2167.

62. P. A. Arnold, B. K. Carpenter, Computational studies on the ring openings of cyclopropyl radical and cyclopropyl cation. *Chem. Phys. Lett.* 2000, 328, 90–96.

63. K. Liu, H-M. Zhao, S-Y. Ma, Z-H. Li, Computational study of the thermal ring opening of cyclopropyl radical, cation and anion. *J. Mol. Struct. (Theochem)* 2004, **672**, 209–213.

64. D. J. Mann, M. D. Halls, Ring-opening of the cyclopropyl radical in the condensed phase: A combined density functional theory/molecular mechanics quasiclassical trajectory study. *Phys. Chem. Chem. Phys.* 2002, **4**, 5066–5071.

65. B. K. Carpenter, J. N. Harvey, A. J. Orr-Ewing, The Study of Reactive Intermediates in Condensed Phases. *J. Am. Chem. Soc.* 2016, **138**, 4695–4705.

66. Z. C. Kramer, B. K. Carpenter, G. S. Ezra, S. Wiggins, Reaction Path Bifurcation in an Electrocyclic Reaction: Ring-Opening of the Cyclopropyl Radical. *J. Phys. Chem. A* 2015, **119**, 6611–6630.

67. B. K. Carpenter, G. S. Ezra, S. C. Farantos, Z. C. Kramer, S. Wiggins, Empirical Classification of Trajectory Data: An Opportunity for the Use of Machine Learning in Molecular Dynamics. *J. Phys. Chem. B* 2018, **122**, 3230–3241.

68. F. Häse, I. F. Galván, A. Aspuru-Guzik, R. Lindh, M. Vacher, M. How machine learning can assist the interpretation of ab initio molecular dynamics simulations and conceptual understanding of chemistry. *Chem. Sci.* 2019, **10**, 2298–2307.