Machine learning for analysing ab initio molecular dynamics simulations

Florian Häse,[1,2,3,4] Ignacio Fdez. Galván,[5] Alán Aspuru-Guzik,[2,3,4,6] Roland Lindh[5] and Morgane Vacher[7]

[1] Department of Chemistry and Chemical Biology, Harvard University, Cambridge, MA 02138, USA
[2] Department of Chemistry, University of Toronto, Toronto, ON M5S 3H6, Canada
[3] Department of Computer Science, University of Toronto, Toronto, ON M5S 3H6, Canada
[4] Vector Institute for Artificial Intelligence, Toronto, ON M5S 1M1, Canada
[5] Department of Chemistry – BMC, Uppsala University, Box 576, 75123 Uppsala, Sweden
[6] Canadian Institute for Advanced Research (CIFAR) Senior Fellow, Toronto, ON M5S 1M1, Canada
[7] Department of Chemistry – Ångström, The Theoretical Chemistry Programme, Uppsala University, Box 538, 751 21 Uppsala, Sweden

E-mail: `morgane.vacher@kemi.uu.se`

Abstract. Post-calculation analyses are often required to extract physical insights from ab initio molecular dynamics simulations. In the present work, we use different machine learning classifiers to take a new perspective on the decomposition reaction of dioxetane. Upon thermally activated decomposition, dioxetane can form products in an electronically excited state and can thus chemiluminesce. Simulated dynamics trajectories exhibit both successful and frustrated dissociations. As an exhaustive and systematic study of the decomposition mechanism "by hand" is beyond feasibility, machine learning models have been employed to study the relevant nuclear distortions governing molecular dissociation. According to all classifiers used in the study, the two sets of geometries differ by the in-phase planarisation of the two formaldehyde moieties. New insights are obtained from this analysis: if both moieties are not planar enough when the dissociation is attempted, it is frustrated and the molecule remains trapped. The postponing of the decomposition reaction by the so-called entropic trap enhances the chemiexcitation efficiency.

1. Introduction

Ab initio molecular dynamics simulations provide a complete picture of the evolution of the electrons and the nuclei along a chemical reaction. However, post-calculation analyses are often required to understand *why* a molecule reacts the way it does [1, 2, 3]. Recently, we have studied the thermally-activated decomposition of dioxetane [4, 5, 1]. The reaction occurs in two steps leading to two formaldehyde fragments: first the O–O bond breaks and then the C–C bond breaks (Figure 1). This reaction is responsible for the non-adiabatic population of electronic excited states, a process called chemiexcitation [6, 7]. The subsequent radiative relaxation is called chemiluminescence, or bioluminescence when occurring in living organisms [8].

It was suggested that an "entropic trap" regulates the outcome of the dissociation (instead of a transition state) [9]. By delaying the exothermic ground-state dissociation, the entropic

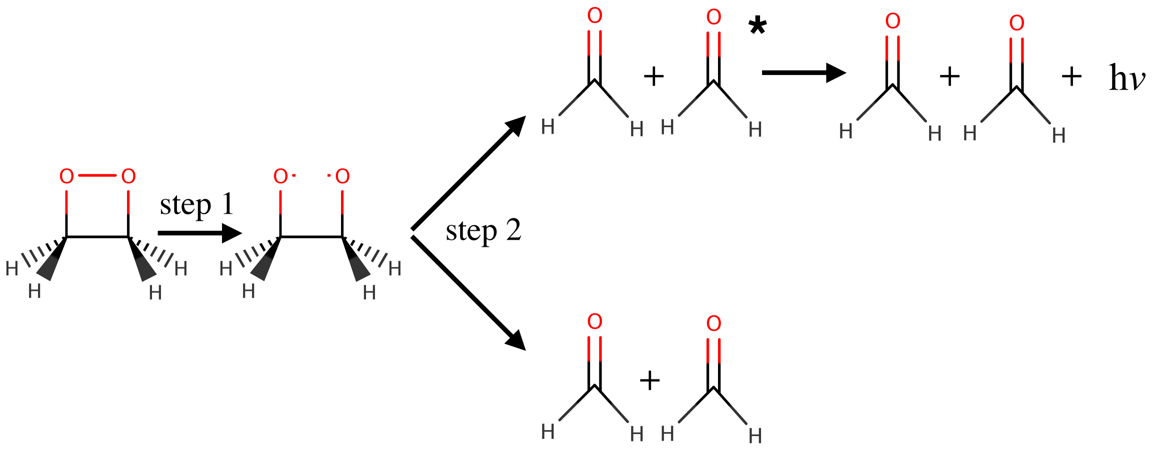

Figure 1. Chemiluminescent (top) and dark (bottom) decomposition reactions of 1,2-dioxetane into two formaldehydes.

trap would give the molecule time to explore other routes and access electronic excited states. More precisely, it was recently demonstrated that the yield of chemiexcitation is determined by the time scale of the decomposition reaction [5]: the slower the reaction, the higher the chemiexcitation yield. Ab initio molecular dynamics trajectories that dissociate late in time exhibit several attempts to dissociate before the final successful dissociation. We call these attempts *frustrated dissociations* [4]. Understanding the yield of chemiexcitation thus relates to understanding why some dissociations are frustrated and others are successful.

In the present work, we use statistical models frequently employed in machine learning to analyse the data produced from ab initio molecular dynamics simulations. Based on our analyses, physical insights into the dissociation mechanism, the entropic trap and thus the chemiexcitation yield are extracted.

2. Theoretical methods

We will below in some details describe the specifics of the ab initio molecular dynamics that produced the data, and the subsequent machine learning algorithms for the post-simulation analysis.

2.1. Data

The analysis is performed on an ensemble of 250 Born-Oppenheimer trajectories, initiated around the first transition state (corresponding to the O–O bond breaking) and heading towards the products [10, 4]. The vibrational ground state is reproduced by sampling initial nuclear positions and velocities (along all nuclear coordinates except the reaction coordinate) from a Wigner distribution [11, 12]. The electronic structure is described using the complete active space self-consistent field (CASSCF) method [13] with an active space of 12 electrons in 10 orbitals, and the ANO-RCC-VTZP basis set [14]. The dynamics simulations are performed using the OpenMolcas package [15].

Geometries along the trajectories which correspond to local maxima in the time evolution of the C–C bond length are assigned the label *frustrated*; the largest local maximum reaches 1.76 Å. *Successful* geometries are geometries at which the C–C bond length reaches 1.76 Å and keeps on increasing monotonically until successful dissociation occurs. Among the ensemble of trajectories, 420 frustrated and 250 successful geometries are extracted, respectively.

2.2. Machine learning algorithms

We aim to identify the nuclear distortions that distinguish best the frustrated and the successful geometries. To do that, we suggest the use of standard machine learning classifiers to elucidate the relevant nuclear distortions that enable the maximal distinction of frustrated and successful geometries. In this work, we employ three different statistical models: linear discriminant analysis (LDA), linear support vector classification (SVC) and support vector machines (SVM) with a linear kernel.

LDA assumes that the probabilities $\mathrm{p}(x|y=0)$ and $\mathrm{p}(x|y=1)$ follow normal distributions with identical covariances, and aims to identify a predictor for y by maximizing the likelihood ratios between the two distributions [16]. Under these assumptions, the decision criterion for x belonging to either $y=0$ or $y=1$ is based on a linear combination of the input features.

SVC and SVM [17] with linear kernels provide a more flexible approach to separating the space. These models create classifiers by constructing linear combinations of the input features and minimizing the quadratically smoothed hinge loss [18]. The two approaches used in this study differ in their regularization and penalty terms. SVC were implemented with an L2 regularization on the learned coefficients [19] while no further penalty terms were added to SVM classification.

The linearity of all three models simplifies the interpretation of the trained classifiers as learned feature transformations are based on linear combinations of the provided input features.

3. Results

In this section, we will present the results on the two sets of successful and frustrated geometries obtained from the ab initio molecular dynamics simulations. First, we will present how the two sets compare in normal mode coordinates. Then, we will show how the machine learning algorithms perform in separating the two sets of geometries, before finally interpreting the new suggested nuclear coordinate.

3.1. Normal modes as starting points for nuclear coordinates

The geometries are expressed in terms of normal modes (calculated at the initial transition state structure), which form a complete set of linearly independent nuclear coordinates. The normal modes are numbered from 0 to 17, where 0 refers to the reaction coordinate. The relevant normal modes are represented in a simplified manner in Figure 2.

Figure 3 presents the two sets of successful and frustrated geometries in normal mode coordinates. The upper subplot shows the average position with the standard deviation represented by the black line. First, we see that both sets of geometries do not show significant average distortions along modes 2, 3, 5, 9, 10, 12, 14 and 17. Then, along most other modes, the successful geometries show larger average distortions than the frustrated ones. The difference between both sets of geometries is quantified by the Kolmogorov–Smirnov (KS) statistics [20] plotted in the lower part of the figure. The modes that exhibit the largest differences between the two sets of geometries are modes 0, 8 and 13, with a KS statistic larger than 0.5, 0.6 and 0.5, respectively. The distribution of successful and frustrated geometries along normal mode 8 (which exhibits the highest KS statistics) is presented in Figure 4a. Normal mode 8 corresponds mainly to the planarisation of the formaldehyde moieties, with only the “inner” hydrogen atoms moving (and not the “outer” ones). Normal mode 0 corresponds mainly to the O-C-C-O dihedral angle while normal mode 13 corresponds mainly to the in-phase bending of the two H–C–H angles (Figure 2).

3.2. Performance of the machine learning analysis

The three different classifiers used suggest new combined coordinates to maximally separate the two sets of geometries. The composition of these new coordinates in terms of normal modes is

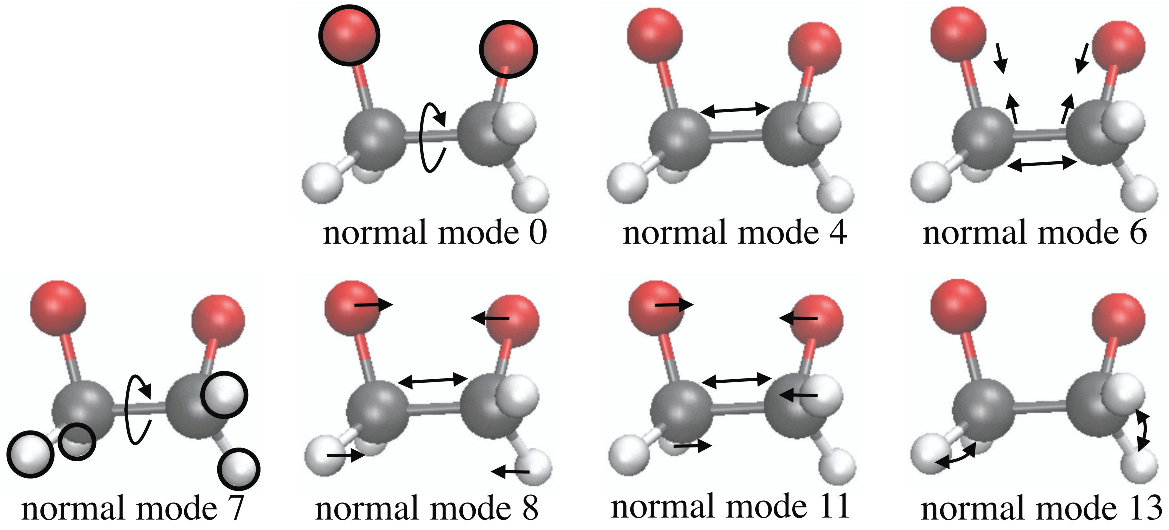

Figure 2. Simplified representation of the relevant normal mode coordinates (calculated at the O–O bond breaking transition state structure). Normal mode 0 corresponds mainly to the OCCO dihedral angle while normal mode 7 corresponds mainly to the HCCH dihedral angles.

represented in Figure 5. The distributions of successful and frustrated geometries along these combined coordinates are presented in Figures 4b-d, where the KS statistics is also indicated for each case. Regarding the ability to separate the two sets of geometries, all three algorithms perform very well since they reach a KS statistics higher than 0.99, indicating that the two sets of geometries barely overlap along these new nuclear coordinates.

3.3. Interpreting the machine learning analysis

What new nuclear coordinate has been identified for the separation? Looking at the coefficients in Figure 5, the new coordinates suggested by the three classifiers are quite similar. The angle of the LDA coordinate vector is only 11 and 10 degrees with the linear SVC and SVM (with a linear kernel) coordinate vectors, respectively; the angle between the two SVC and SVM coordinate vectors is only 4 degrees. Based on a visual inspection of the new coordinate, it mainly corresponds to the in-phase planarisation of the two formaldehyde moieties.

In all classifier cases, the normal mode with the largest amplitude is normal mode 11. Normal mode 11 corresponds mainly to the planarisation of the formaldehyde moieties, with only the “outer” hydrogen atoms moving and not the “inner” ones as in normal mode 8 (Figure 2). It is noted that along normal mode 11 alone, the two sets of geometries do not differ much (KS $<$ 0.2, see Figure 3). The second most important normal mode is mode 8, already identified as a good separative coordinate (Figure 3).

Taking the linear SVC classifier for instance, the five most important modes (11, 8, 6, 4 and 7) out of the 18 normal modes account for more than 90% of the suggested new nuclear coordinate. Using only their coefficients in creating the new coordinate (and setting the others to zero) gives a KS of 0.896. When re-optimising these five coefficients, a KS of 0.989 is obtained.

To test the robustness of the analysis procedure, we repeated the training of the linear SVC classifier with 1.8 Å and 1.9 Å for the threshold value used to detect the successful geometries. The angle between the two new coordinate vectors is only 7 degrees. It is on the same order of magnitude as the angle obtained for different classifiers and it is satisfactorily small. It is also noted that, although it might be interesting to study in detail, we do not expect the results to depend much on the choice of initial basis of nuclear coordinates, for example when using

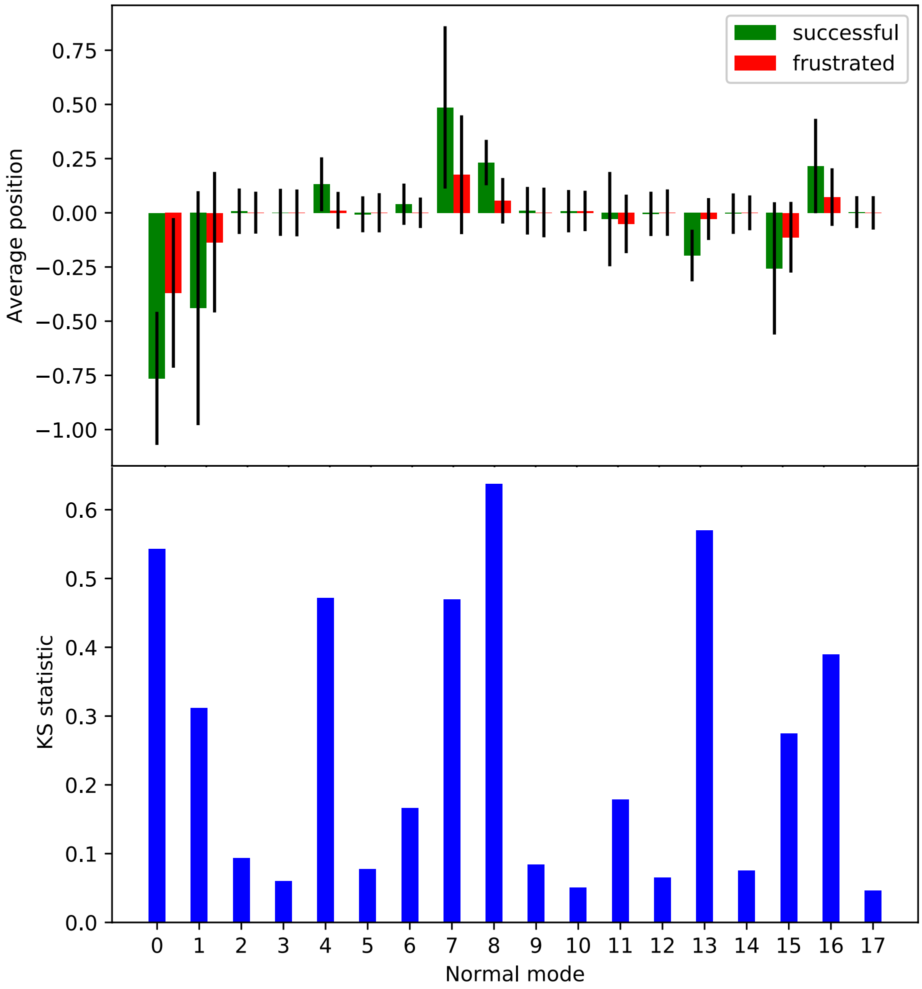

Figure 3. Frustrated (red) and successful (green) geometries in normal mode coordinates. The upper subfigure shows the average positions and standard deviations. The lower one shows the KS statistics which quantifies the difference between both sets of geometries.

localised modes [21] instead of normal modes.

4. Conclusions and discussions

In this work, we have used three different machine learning classifiers to suggest a nuclear coordinate that distinguishes between frustrated and successful dissociations (identified along ab initio molecular dynamics simulations). The three algorithms perform very well in separating the two sets of geometries and all three suggest a similar new nuclear coordinate. The latter mainly corresponds to the in-phase planarisation of the two formaldehyde moieties.

We note that it was previously suggested that the nuclear distortions responsible for the frustration of dissociations were a small O-C-C-O dihedral angle and large O-C-C angles [4]. The quantification of how much the two sets of geometries differ along those previously suggested coordinates gives a KS of slightly larger than 0.2 and slightly lower than 0.5 along the O-C-C angle and O-C-C-O dihedral angle, respectively. The machine learning classifiers are thus much better at finding separative nuclear coordinates than humans. Also, it is interesting to observe that none of these coordinates seem to be directly identified as important by the machine learning classifiers. (Normal mode 7 corresponds mainly to the H-C-C-H dihedral angles, and

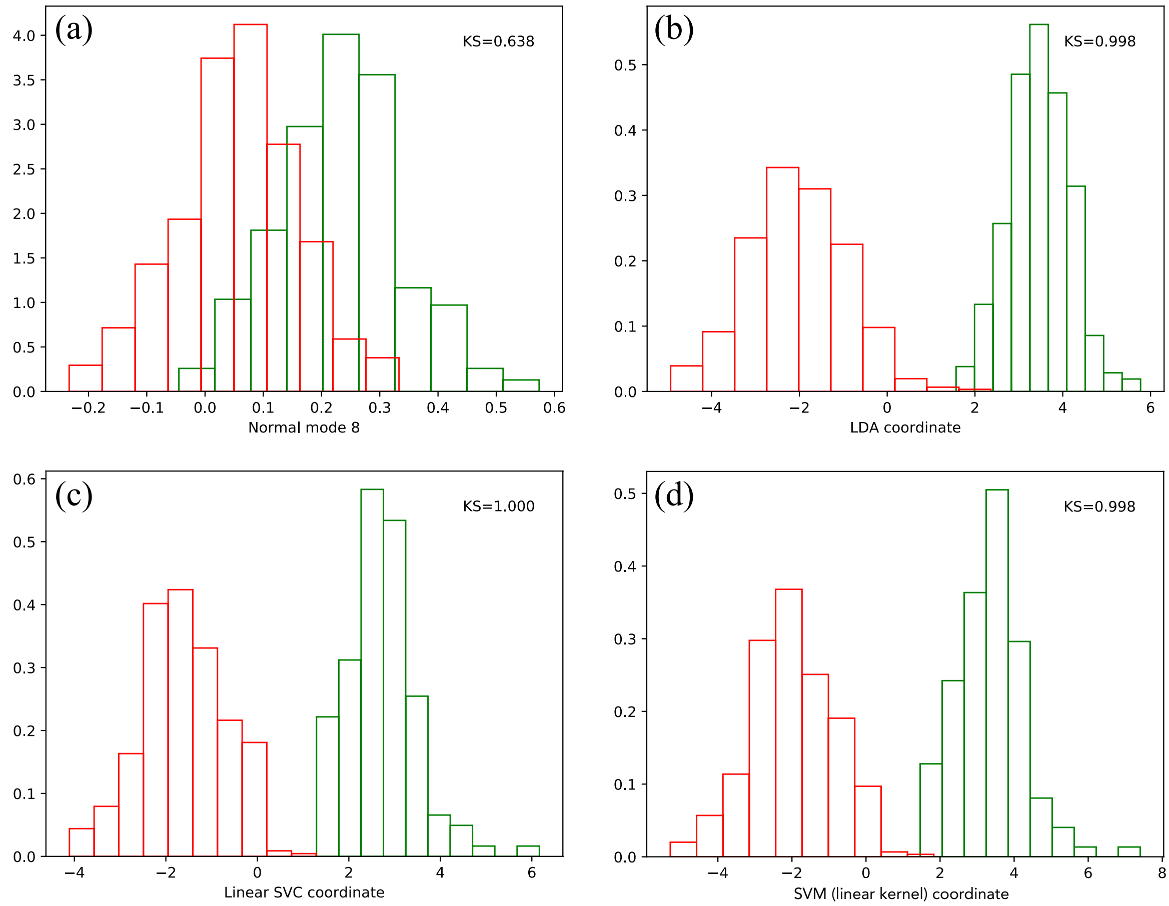

Figure 4. Frustrated (red) and successful (green) geometries along (a) normal mode 8, and the coordinates optimised with (b) LDA, (c) linear SVC and (d) SVM with a linear kernel. The KS statistics is indicated to quantify the difference between both sets of geometries.

not the O-C-C-O angle.) The planarisation distortions must indirectly imply the reduction of the O-C-C angles though.

The present analysis of the ab initio molecular dynamics simulations brings a fresh and more accurate look onto the dissociation mechanism and the entropic trap. Contrary to what was previously suggested [4], the geometrical condition that is necessary for a trajectory to escape the entropic trap and for dissociation to be successful is the in-phase planarisation of the two formaldehyde moities. If both moities are not planar enough when the dissociation is attempted, the molecule remains trapped. This way, the so-called entropic trap leads to frustrated dissociations, postponing the decomposition reaction. A chemical distortion that would hamper the in-phase planarisation of the two formaldehyde moities may thus enhance the efficiency of chemiexcitation. Performing the same type of analysis on excited state dynamics might reveal complementary relevant nuclear coordinates.

References

[1] Häse F, Fdez Galván I, Aspuru-Guzik A, Lindh R and Vacher M 2019 *Chem. Sci.* **10**(8) 2298–2307 URL `http://dx.doi.org/10.1039/C8SC04516J`

[2] Grazioli G, Roy S and Butts C T 2019 *Journal of Chemical Information and Modeling* **59**

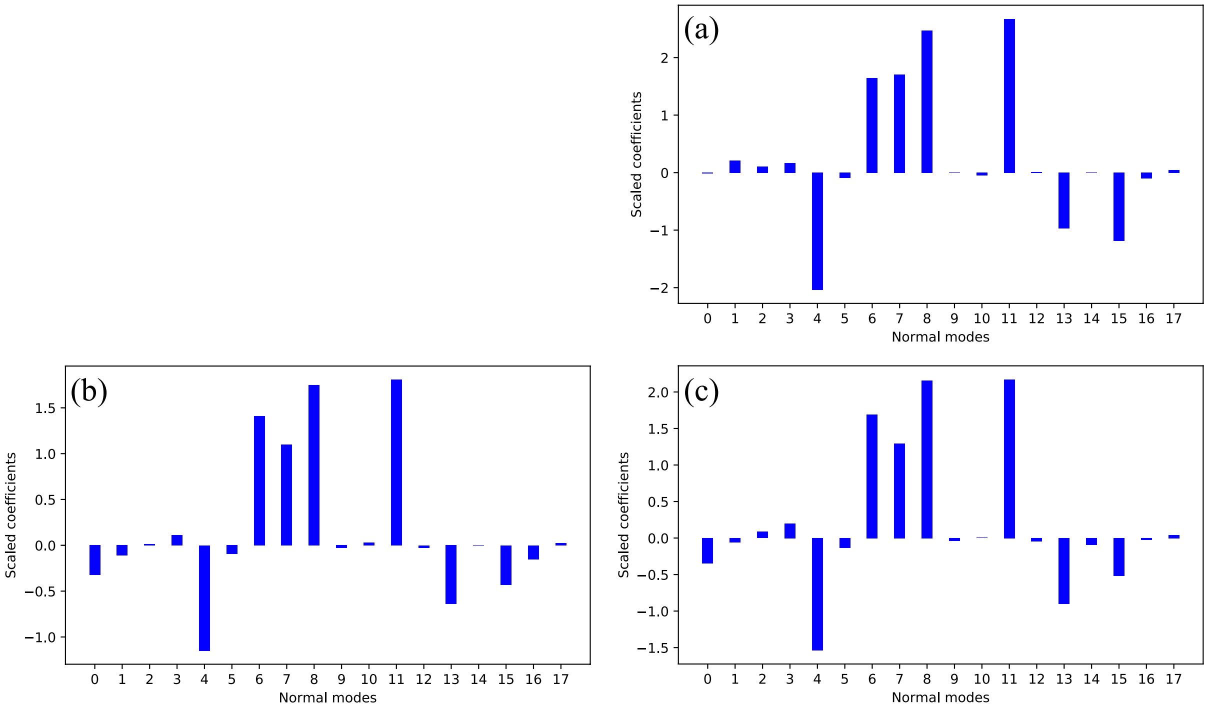

Figure 5. Combined coordinates (expressed in terms of normal modes) suggested by (a) LDA, (b) linear SVC and (c) SVM with a linear kernel.

2753–2764 pMID: 31063694 (*Preprint* `https://doi.org/10.1021/acs.jcim.9b00134`) URL `https://doi.org/10.1021/acs.jcim.9b00134`

[3] Berishvili V P, Perkin V O, Voronkov A E, Radchenko E V, Syed R, Venkata Ramana Reddy C, Pillay V, Kumar P, Choonara Y E, Kamal A and Palyulin V A 0 *Journal of Chemical Information and Modeling* **0** null pMID: 31276400 (*Preprint* `https://doi.org/10.1021/acs.jcim.9b00135`) URL `https://doi.org/10.1021/acs.jcim.9b00135`

[4] Vacher M, Brakestad A, Karlsson H O, Fdez Galván I and Lindh R 2017 *J. Chem. Theory Comput.* **13** 2448–2457 pMID: 28437611 (*Preprint* `https://doi.org/10.1021/acs.jctc.7b00198`) URL `https://doi.org/10.1021/acs.jctc.7b00198`

[5] Vacher M, Farahani P, Valentini A, Frutos L M, Karlsson H O, Fdez Galván I and Lindh R 2017 *J. Phys. Chem. Lett.* **8** 3790–3794 pMID: 28749694 (*Preprint* `https://doi.org/10.1021/acs.jpclett.7b01668`) URL `https://doi.org/10.1021/acs.jpclett.7b01668`

[6] FdezGalván I, Gustafsson H and Vacher M 2019 *ChemPhotoChem* **3** 957–967 (*Preprint* `https://onlinelibrary.wiley.com/doi/pdf/10.1002/cptc.201800232`) URL `https://onlinelibrary.wiley.com/doi/abs/10.1002/cptc.201800232`

[7] Vacher M, Fdez Galván I, Ding B W, Schramm S, Berraud-Pache R, Naumov P, Ferré N, Liu Y J, Navizet I, Roca-Sanjuán D, Baader W J and Lindh R 2018 *Chem. Rev.* **118** 6927–6974 pMID: 29493234 (*Preprint* `https://doi.org/10.1021/acs.chemrev.7b00649`) URL `https://doi.org/10.1021/acs.chemrev.7b00649`

[8] Navizet I, Liu Y J, Ferré N, Roca-Sanjuán D and Lindh R 2011 *ChemPhysChem* **12** 3064–3076 ISSN 1439-7641 URL `http://dx.doi.org/10.1002/cphc.201100504`

[9] Feyter S D, Diau E W G, Scala A A and Zewail A H 1999 *Chemical Physics Letters* **303** 249 – 260 ISSN 0009-2614 URL `http://www.sciencedirect.com/science/article/pii/S0009261499002638`

[10] Lourderaj U, Park K and Hase W L 2008 *International Reviews in Physical Chemistry* **27** 361–403 (*Preprint* `http://dx.doi.org/10.1080/01442350802045446`) URL `http://dx.doi.org/10.1080/01442350802045446`

[11] Wigner E 1932 *Phys. Rev.* **40**(5) 749–759 URL `http://link.aps.org/doi/10.1103/PhysRev.40.749`

[12] Barbatti M, Granucci G, Lischka H, Persico M and Ruckenbauer M 2006 Newton-x: a package for newtonian dynamics close to the crossing seam, version 0.11b www.univie.ac.at/newtonx

[13] Roos B O, Taylor P R and Siegbahn P E 1980 *Chemical Physics* **48** 157 – 173 ISSN 0301-0104 URL `http://www.sciencedirect.com/science/article/pii/0301010480800450`

[14] Roos B O, Lindh R, Malmqvist P Å, Veryazov V and Widmark P O 2004 *The Journal of Physical Chemistry A* **108** 2851–2858 (*Preprint* `http://dx.doi.org/10.1021/jp031064+`) URL `http://dx.doi.org/10.1021/jp031064+`

[15] Fdez Galván I, Vacher M, Alavi A, Angeli C, Aquilante F, Autschbach J, Bao J J, Bokarev S I, Bogdanov N A, Carlson R K, Chibotaru L F, Creutzberg J, Dattani N, Delcey M G, Dong S S, Dreuw A, Freitag L, Frutos L M, Gagliardi L, Gendron F, Giussani A, González L, Grell G, Guo M, Hoyer C E, Johansson M, Keller S, Knecht S, Kovačević G, Källman E, Li Manni G, Lundberg M, Ma Y, Mai S, Malhado J P, Malmqvist P Å, Marquetand P, Mewes S A, Norell J, Olivucci M, Oppel M, Phung Q M, Pierloot K, Plasser F, Reiher M, Sand A M, Schapiro I, Sharma P, Stein C J, Sørensen L K, Truhlar D G, Ugandi M, Ungur L, Valentini A, Vancoillie S, Veryazov V, Weser O, Wesołowski T A, Widmark P O, Wouters S, Zech A, Zobel J P and Lindh R *Journal of Chemical Theory and Computation* PMID: 31509407 (*Preprint* `https://doi.org/10.1021/acs.jctc.9b00532`) URL `https://doi.org/10.1021/acs.jctc.9b00532`

[16] Franklin J 2005 *The Mathematical Intelligencer* **27** 83–85 ISSN 0343-6993 URL `https://doi.org/10.1007/BF02985802`

[17] Aizerman M A 1964 *Automation and remote control* **25** 821–837

[18] Rosasco L, Vito E D, Caponnetto A, Piana M and Verri A 2004 *Neural Computation* **16** 1063–1076 (*Preprint* `https://doi.org/10.1162/089976604773135104`) URL `https://doi.org/10.1162/089976604773135104`

[19] Ng A Y 2004 *Proceedings of the Twenty-first International Conference on Machine Learning* ICML '04 (New York, NY, USA: ACM) pp 78– ISBN 1-58113-838-5 URL `http://doi.acm.org/10.1145/1015330.1015435`

[20] 2008 *Kolmogorov–Smirnov Test* (New York, NY: Springer New York) pp 283–287 ISBN 978-0-387-32833-1 URL `https://doi.org/10.1007/978-0-387-32833-1_214`

[21] Jacob C R and Reiher M 2009 *The Journal of Chemical Physics* **130** 084106 (*Preprint* `https://doi.org/10.1063/1.3077690`) URL `https://doi.org/10.1063/1.3077690`