Thousands of reactants and transition states for competing E2 and SN2 reactions

IOP Publishing Mach. Learn.: Sci. Technol. 1 (2020) 045026 https://doi.org/10.1088/2632-2153/aba822

MACHINE LEARNING
Science and Technology

OPEN ACCESS

RECEIVED
31 May 2020

REVISED
7 July 2020

ACCEPTED FOR PUBLICATION
21 July 2020

PUBLISHED
28 October 2020

PAPER

Thousands of reactants and transition states for competing E2 and S_N2 reactions

Guido Falk von Rudorff, Stefan N Heinen, Marco Bragato and O Anatole von Lilienfeld

Institute of Physical Chemistry and National Center for Computational Design and Discovery of Novel Materials (MARVEL), Department of Chemistry, University of Basel, Klingelbergstrasse 80, CH-4056 Basel, Switzerland

E-mail: anatole.vonlilienfeld@unibas.ch

Keywords: reactions, chemical space, competing reaction channels

Abstract

Reaction barriers are a crucial ingredient for first principles based computational retro-synthesis efforts as well as for comprehensive reactivity assessments throughout chemical compound space. While extensive databases of experimental results exist, modern quantum machine learning applications require atomistic details which can only be obtained from quantum chemistry protocols. For competing E2 and S_N2 reaction channels we report 4,466 transition state and 143,200 reactant complex geometries and energies at MP2/6-311G(d) and single point DF-LCCSD/cc-pVTZ level of theory, respectively, covering the chemical compound space spanned by the substituents NO_2, CN, CH_3, and NH_2 and early halogens (F, Cl, Br) and hydrogen as nucleophiles and early halogens as leaving groups. Reactants are chosen such that the activation energy of the competing E2 and S_N2 reactions are of comparable magnitude. The correct concerted motion for each of the one-step reactions has been validated for all transition states. We demonstrate how quantum machine learning models can support data set extension, and discuss the distribution of key internal coordinates of the transition states.

1. Introduction

Reactions are the very core of chemistry and their understanding is crucial for molecular design problems: Even if a compound has been identified to be interesting for a certain application, a reaction pathway has to be found to connect abundant compounds to the desired target molecule. Large experimental databases of reaction paths with associated barriers and yields have been compiled to that end [1] and have been proven to be useful in the design of reaction steps [2, 3] or for the optimization of reaction environments [4].

These databases however, rely on careful experimental work and would benefit from a computational perspective, since their extension relies on manual work. As a consequence, they are of limited detail and size when compared to chemical space. High-throughput calculations are one way of obtaining reaction paths, but pose another complex problem: Finding the relevant transition state geometries is technically difficult, in particular if the reaction pathway is not known beforehand, since it requires finding the saddle points on the potential energy surface [5–7]. As a consequence, previous computational work reporting on transition state configurations covered only a modest number of cases, and employed a wide range of levels of theory [8–17]. Additionally, an accurate representation of the Minimum Energy Path requires knowledge of the conformational space spanned by the reactant and products, a challenging task by itself [18, 19]. Furthermore, not all established quantum chemistry methods are suitable for yielding accurate potential energies of reactive processes [8, 18]. Direct comparison of calculated energy barriers to experiment in itself is often impracticable since the relevant barriers require the calculation of ensemble-averaged free energies in explicit solvent. This task on its own is already challenging just for a single molecule [20] and might be computationally prohibitive for large numbers of reactions. In the reverse picture, gas-phase reaction experiments are particularly challenging but possible in some cases [21, 22].

With recent successes of machine learning models in the context of exploration of chemical space [23] e.g. non-covalent interactions [24], response properties [25], and molecular forces [26], it would be desirable

to also explore reaction space. Some initial work in this direction has been done already [27–35]. For any machine learning approach, consistent data sets are of high value for training and validation. Typically, a single study in literature gives about five (experimental) to fifty (computational) transition state geometries or energies. This is insufficient for the training of converged and meaningful quantum machine learning models. Furthermore, atomistic details (geometries) are often lacking in the case of experimental data, while level of theory used in the case of theoretical studies can often no longer be considered to be state of the art. While it is possible to merge reaction data from different sources or to learn their respective differences in the potential energy surface by means of Delta machine learning (Δ-ML) [36], multi-fidelity machine learning models [37], multi-level combination grid technique [38] or transfer learning [39], the resulting multilevel approaches require at least part of the data to be evaluated in many different sources. Thus there is considerable need for one large consistent data set which subsequently could be used as a basis for multilevel machine learning models and their application in reaction design. When assessing possible reactions from a given reactant, it is not always sufficient to be able to quantify just one particular pathway. Rather, several competing reaction channels need to be estimated at the same time to decide which reactions will occur with which weight. To enable such modeling, a homogeneous data set for competing reactions is desirable.

Starting from the lowest lying conformers of the organic molecules listed in the GDB-7 [40] data set, Grambow *et al* [41] have just recently generated 12k transition state geometries using the double-hybrid ωB97X-D3 density functional approximation, allowing for any feasible reaction mechanisms. In contrast, we here focus on the narrow reaction space obtained for typical substitutions and attacking and leaving groups of the competing textbook reactions E2 and S_N2 with the specific intent to enable more thorough, systematic and comprehensive explorations of the nature of the corresponding chemical compound space. Often, S_N2 was used as a benchmark reaction due to its iconic, well established mechanism [42–46], and having the advantage of a less complex transition state over its competing reaction E2 [47]. Even though the overall reaction mechanisms are well understood, their competition in terms of exploring the chemical compound space defined by specific combinations of substituents, leaving groups, and nucleophiles has not yet been studied in a systematic manner—to the best of our knowledge.

We include geometries of reactant and product conformers, reactant complexes, and transition state geometries. For our calculations, we chose the MP2/6-311G(d) [48–52] level of theory since benchmark studies have found this level to be a good compromise between accuracy and computational effort for the reactions under investigation in particular with regards to geometries [8, 53, 54]. DFT methods have been found to exhibit significant deviations for both energies and geometries [55]. Even for hybrid functionals, it is known for a long time that their share of exact exchange should be different for reactants and for saddle points in order to yield best accuracy [56] which renders them inapplicable for activation energies. MP2 has been shown to be more accurate for saddle point geometries, all else being equal [56, 57]. For e.g. nucleophilic substitution, the MP2 error in energies is nearly half the error of typical DFT methods [43]. In order to further improve on the accuracy of the MP2 energies, we also performed single-point DF-LCCSD calculations for every transition state geometry, as well as for their reactants.

We see the main use case of this data set in the context of assessing competing reactions with machine learning methods. This is key to chemical synthesis design where competing reactions could have a strong impact on the yield. With most existing data sets focused on (near-)equilibrium geometries and associated properties, the current work offers access to a larger part of potential energy surfaces. This is particularly challenging as the ideal machine learning model would only require the reaction type and reactant information to estimate the transition state geometry or its energy, since an explicit search for each transition state geometry is expensive (as shown below). This requires strategies to estimate a property at a different point of the potential energy surface than the explicit query configuration. To develop such strategies, this data set might prove particularly useful. Moreover, the reaction data set is directly applicable to cases where the low ambient temperature renders the potential energy dominant for reaction barriers, e.g. interstellar environments. For these cases, a list of potential reactions taking place can be derived directly from the activation energy data in this work.

2. Methods and computational details

In our database, we have considered all 7,500 reactant molecules that can be built from ethane with the substituents listed in table 1 using the positions shown in figure 1.

These substituents were selected for their following properties: i) electronic effects should be maximized and ii) steric hindrance minimized. More precisely, while being as small as possible in order to make the reaction center sufficiently accessible to the nucleophile, electron donating groups and withdrawing groups should cover weak as well as strong inductive effects.

Table 1. Chemical space for our reaction database: substituents R, leaving groups X and the nucleophiles Y^-. Molecular skeleton is ethane, see also figure 1. The letters refer to the labels in our data set files.

	A	B	C	D	E
R*k*	H	NO_2	CN	CH_3	NH_2
X	F	Cl	Br		
Y	H	F	Cl	Br	

2.1. Machine learning

In this study we used delta machine learning (Δ-ML) in kernel ridge regression (KRR) implemented in the QMLcode [58]. Kernel based methods were introduced in the 1950 s by Kriging *et al* [59]. KRR uses as input a kernel function with the feature vector $\mathbf{x}$ to learn a mapping function to a property $y_q^{\text{est}}(\mathbf{x}_j)$ given a training set of N reference pairs $\{\mathbf{x}_i, y_i\}^N$:

$$y_q^{\text{est}}(\mathbf{x}_j) = \sum_i^N \alpha_i k(\mathbf{x}_i, \mathbf{x}_j) \tag{1}$$

where α is the regularization coefficient and $k(\mathbf{x}_i, \mathbf{x}_j)$ a Gaussian kernel element:

$$k(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(-\frac{||\mathbf{x}_i - \mathbf{x}_j||_2^2}{2\sigma^2}\right) \tag{2}$$

A more detailed discussion of the KRR method employed in this work and pertinent references can be found in Heinen *et al* [60]. In the context of Δ-ML, the procedure stays the same and only the property (y) changes from a molecular property to a difference in properties, e.g. from $y^{\text{est}} \mathrel{\widehat{=}} E_a$ to $y^{\text{est}} \mathrel{\widehat{=}} \Delta E_a$.

The feature vector or representation $\mathbf{x}$ we used is one-hot encoding [61], which is a bit vector. For every substitution site R*k*, nucleophile Y and leaving group X, we denote presence of a given combination with ones. In our case, this means that for any transition state, six out of the 27 entries of the representation vector are ones, the rest zeros.

2.2. Reactants and products

We started from the unsubstituted case fluoroethane optimized with openbabel [62] using the universal force field (UFF) [63] and functionalized the substituent sites R*k* in figure 1 using the C++ interface of openbabel. Again, each resulting structure was optimized with UFF to remove potential bad contacts. Using the Experimental-Torsion Knowledge Distance Geometry (ETKDG) method as implemented in RDKit [64], we searched for 1,000 conformer geometries. They subsequently were ordered by UFF energy. Starting from the most stable conformer, all those configurations were included in the followings steps if and only if their root mean squared difference (RMSD) to the previously accepted configuration was at least 0.01 Å or the energy difference between the two was at least 0.1 kcal mol^{-1}.

The resulting conformer candidate configurations were relaxed at MP2/6-311G(d) level with ORCA 4.0.1 [48–50, 65–67] to be compatible with the level of theory to be employed for the transition state search. For each of these minimized configurations, all possible nucleophiles given in table 1 were placed along the expected axis of the CH bond in figure 3. With the nucleophile being constrained to that axis, the geometries were optimized to obtain an estimate of the reactant complex geometry.

For each of these reactant complexes, we subsequently lifted the constraint and relaxed further. This was helpful as the potential energy landscape around the reactant complex is comparably shallow and therefore direct optimization to the free reactant complex was often ineffective.

Each unconstrained reactant complex was validated using a variety of geometrical criteria to ensure that the more than 100,000 minimum energy geometries represented meaningful configurations. The overall procedure is shown in figure 2. First, we required the reactant complex to constitute two fragments based on the topology obtained from MDAnalysis [68] where one fragment needed to be of exactly one atom, i.e. the nucleophile. This is to avoid erroneous fragmentation where e.g. a proton is abstracted from the reactant. In the case of E2 reactions, we required that the angle C–H$\cdots$Y must not be smaller than 178 degrees since configurations with larger angles indicate trapping of the nucleophile by other hydrogen atoms of the reactant not involved in this particular reaction channel.

For S_N2, more validations are required.

- The C$\cdots$Y distance had to be at least 1.14 Å, 1.41 Å, 1.86 Å, and 2.04 Å for hydrogen, fluorine, chlorine, and bromine, respectively. This avoids configurations that are actually product complexes. Due to the low

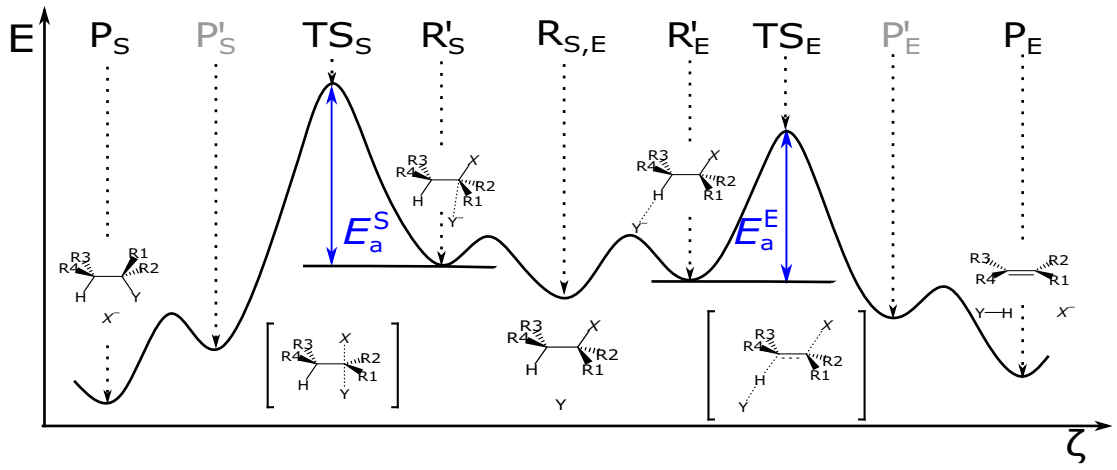

Figure 1. Energy diagram of the competing E2 and S_N2 reactions, exemplifying kinetic vs. thermodynamic control, respectively. Reactant conformers ($R_{S,E}$) are shared between the reactions, while transition states ($TS_{S/E}$), product conformers ($P_{S/E}$), reactant complexes ($R'_{S/E}$) ,and product complexes ($P'_{S/E}$) are specific to each reaction. For each reaction, the energy difference between transition state and reactant complex is the activation energy $E_a^{S/E}$.

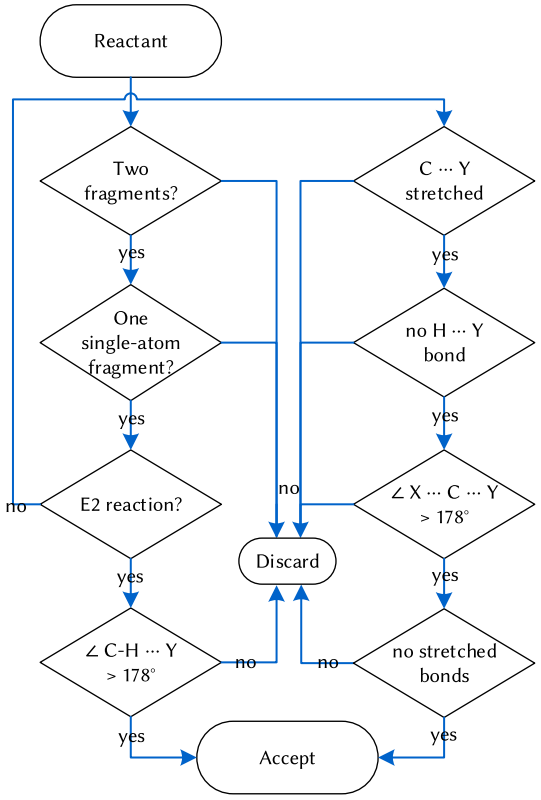

Figure 2. Validation procedure for reactant geometries starting from a candidate reactant structure to the decisions whether to accept or discard that candidate geometry.

activation energy for many such cases, a geometry optimization can end up in a product complex minimum from a reactant complex initial guess.

- To avoid trapping of the nucleophile by reactant hydrogen atoms, the distance between the nucleophile and the closest hydrogen of the reactant is required to be at least 0.78 Å, 0.96 Å, 1.33 Å, and 2.48 Å for hydrogen, fluorine, chlorine, and bromine, respectively.
- Since the S_N2 reaction requires nearly planar bonds for the reaction center, we require that the angle X···C···Y must be at least 178 degrees.
- We avoid artificially stretched geometries by requiring no carbon-carbon distance to be within 1.65–2 Å and no nitrogen-oxygen distance to be within 1.5–2.5 Å.

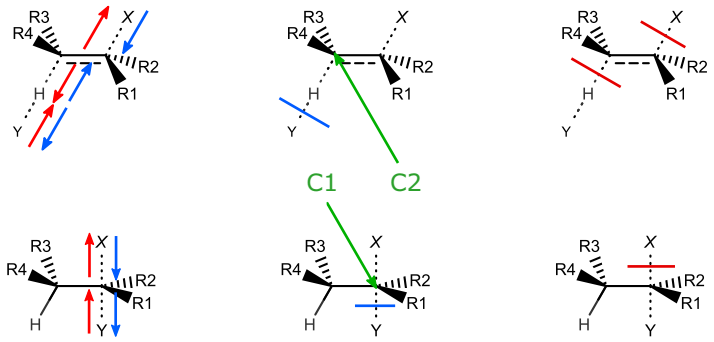

Figure 3. Illustration of validation procedures for generating E2 (top) and S_N2 (bottom) geometries. Normal mode requirements for transition states (left column) show concerted motions which are characteristic for the reaction in question (red arrows point towards product, blue arrows towards reactant). Bond cleavages tested for reactant complexes and product complexes are shown in the mid and right column, respectively. Blue perpendicular lines correspond to removal of Y^-, YH and X^- for E2 and X^- or Y^- for S_N2 leading to infinite separation as shown in figure 1). Bond cleavage indicated by red perpendicular lines corresponds to product formation.

Whenever these validation steps were successful, the lowest such minimum from all conformers investigated is considered to represent the reactant complex. Otherwise, the lowest energy configuration from the constrained optimization is taken as an approximation of the reactant complex. In the latter case, Δ-ML [36] was employed to estimate the residual relaxation energy between the constrained and unconstrained reactant complex.

Duplicate reactant and product conformer geometries were identified using the FCHL19 [26] representation. By that measure, only unique geometries are retained. This test was not applied to reactant complexes as their local minima energies and geometries can be very similar yet distinct.

2.3. Transition states

Using Gaussian09 [69] for an initial transition state geometry with B3LYP/6-31G* [70–75] and subsequently ORCA 4.0.1 for a final transition state with MP2, we first found the transition state of the unsubstituted case with chloride as nucleophile. Functionalization followed the same procedure as for the reactants. Using these starting geometries, transition states were obtained via eigen mode following as implemented in ORCA. After a transition state was found, the local Hessian matrix was obtained from a numerical frequency calculation by finite displacements as implemented in ORCA.

Once a transition state was found for a combination of the four substituents, this geometry was employed as starting geometry for further transition state searches for missing cases where exactly one out of the four substituents was different from the case where a validated transition state has been found. This scheme was used only for those molecules where the substituent that was to be replaced did not have the same functional group as the neighbouring substituent on the same carbon atom. For some cases, this procedure was employed several times in a row, each time resulting in an additional set of transition states which served as starting guesses. Similarly, the nucleophile of validated transition states was replaced to obtain promising starting geometries for the transition state search.

Once the transition state geometry has been found for any potential reaction target, the Hessian was evaluated to ascertain that the geometry in fact is a transition state with exactly one imaginary frequency. We only included a transition state in our data set if this frequency was at least 400 cm^{-1} and that the resulting motion corresponding with this one normal mode was as shown in figure 3 (left column). The ethane skeleton features two carbon atoms C*k*, where the one with substituents R1 and R2 is numbered C1. For the E2 transition state, X, Y and the hydrogen atom were displaced along the normal mode and checked if the distances C2-H as well as C1-X were larger and the C2-Y distance was smaller compared to the non-displaced geometry. In the S_N2 transition state, the nucleophile and leaving group were displaced along the normal mode and C1-X was compared to C1-Y.

While the investigation of the normal modes alone ensures that the vibrational motion belongs to the main configurational change the molecule undergoes during each reaction, it is not a sufficient criterion that this particular transition state geometry actually connects reactant and product. We use the intrinsic reaction coordinate (IRC) [76] as final criterion to ensure that the transition state indeed connects a valid reactant complex with a valid product for the reaction in question. The IRC is commonly employed to find a reaction pathway starting from a transition state. The Cartesian IRC is given by the steepest descent path in forward

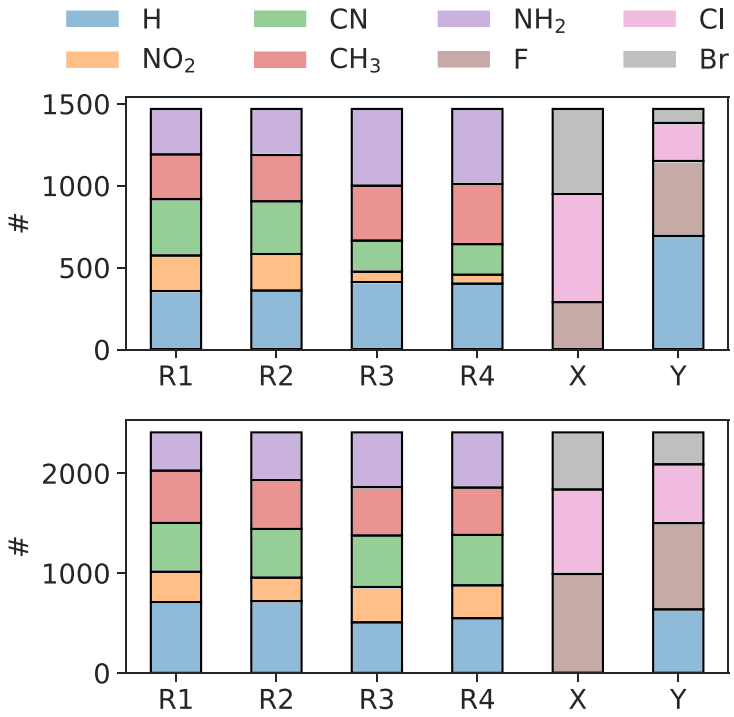

Figure 4. Distribution of substituents R*k*, leaving groups X, and nucleophiles Y for all activation energies in the data set. Top: distribution for E2, bottom: distribution for S_N2.

and backward direction of the reaction. We use steepest descent as implemented in ORCA to trace the Cartesian IRC. If the energy curvature near the transition state and along the reaction coordinate is small, steepest descent paths can become subject to numerical instabilities. To avoid this issue, we approximate the IRC close to the transition state by a line scan in either direction based on the normal mode displacement of imaginary frequency. From the final point of the line scan, a regular steepest descent is followed until a local minimum has been reached.

Since the sign of the normal mode of imaginary frequency is not fixed with respect to the direction of the reaction, we analyze the minimum energy endpoints of the IRC to classify them as either close to reactant or close to product based on the bond length as shown in figure 3. If and only if exactly one of the endpoints is found to be close in geometry to a reactant configuration and the other is found to be close in geometry to the product configuration, the corresponding transition state is included in our data set. To test whether the configurations are close to reactant or product, we measured C2-H distances for the E2 case and C1-X and C1-Y distances for the S_N2 reaction to ensure bonds have been broken as shown in figure 3.

For cases where several validated transition states for the same reaction have been found, we consider the lowest one for the reaction barrier.

Finally, we performed single-point DF-LCCSD/cc-pVTZ calculations, as implemented in Molpro2018 [77–83] using the extremal geometries as obtained with MP2/6-311G(d). All in all, the complete generation of the data set took about 2.8 million core hours.

3. Results

3.1. Data

Our resulting data set contains 4,466 validated transition state geometries, of which 2,785 are for S_N2 (TS_S) and 1,681 for E2 (TS_E). Based on 26,997 reactant conformers ($R_{S,E}$), we identified 81,950 constrained reactant complexes for E2 (R'_E) and 57,642 constrained reactant complexes for S_N2 (R'_S) which in turn have been refined to yield 2,030 unconstrained reactant complexes for E2 (R'_E) and 1,532 unconstrained reactant complexes for S_N2 (R'_S). Finally, we have found 15,706 S_N2 product conformers (P_S) and 9,588 E2 product conformers (P_E). All geometries are calculated at MP2/6-311G(d) level of theory and given as XYZ files in this work. Two additional files specify all individual energies and activation energies, respectively. The labels in the text files relate to the labels in table 1.

All data is available in the materials cloud (https://doi.org/10.24435/materialscloud:sf-tz).

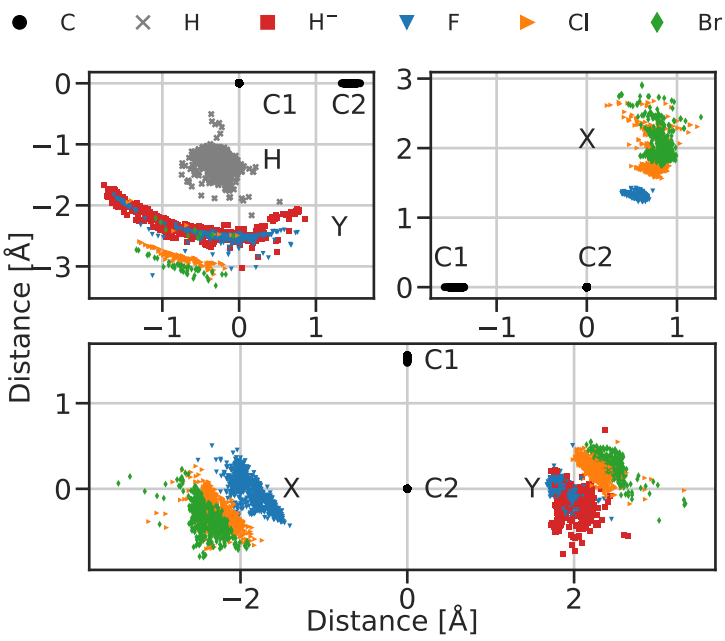

Figure 5. Overview scatter plot of atomic positions of scaffold carbon atoms and nucleophiles and leaving groups for all transition states for E2 (top row) and S_N2 (bottom). Transition state geometries have been translated such that C1 (top left) or C2 (top right and bottom) are in the origin. Additionally, they have been rotated such that all atoms shown with the exception of the hydrogen sites in the top left panel are planar. Coordinates of other atoms have been projected into the figure plane.

3.2. Geometries

As shown in figure 4, we were able to find many transition states for a variety of substituents, nucleophiles and leaving groups. This means that we have reached a substantial coverage of the chemical space in question, which is key for machine learning. The challenge here is the low success rate of the transition state search which might have been the key reason why such data sets have not yet been published earlier. In particular, machine learning models will benefit from the comparably low noise in the data set coming from our validation procedure. Moreover, the data set features many different combinations of substituents such that there is considerable promise that their interplay for the competing reactions can be analysed and understood.

As in any iterative optimization scheme, convergence thresholds influence the final results. This is the case for a transition state search as well and might potentially give rise to some small noise in the transition state geometries. Since we calculated the explicit Hessian matrix, we know that the transition state geometries reported in this data set are indeed saddle points, and that their mass-weighted normal modes represent the concerted rearrangement expected for E2 and S_N2 reactions. Together with the tight convergence criteria required for transition state optimization, this means that our data set contains only highly compatible transition state geometries for all the validated combinations of substituents, nucleophiles and leaving groups.

This is demonstrated in figure 5 which shows a scatter plot of the most important internal coordinates for the transition states. The reduction of dimensionality from the more complex 3D geometry is obtained by placing one of the two central carbon atoms in the origin and then aligning the carbon-carbon bond along one Cartesian axis. The other markers then show the position of one atom for each transition state found. For E2 reactions, the transition state geometry has been rotated such that all three points shown in the corresponding panels are exactly within one plane. For the S_N2 case, this is not possible, as the four atoms in question are not necessarily exactly in a plane even though they are very close to that. For this panel, the projection on the fitted plane through all four points is shown. For the internal carbon-carbon bond, the variance of the bond length is significantly higher for E2 than for S_N2, as shown in figure 5. This can be explained by the nature of the two reactions: While E2 consists of a concerted action on both carbons, S_N2 happens only at one of the two carbons. We also see that each element for the nucleophile and leaving group has its own distribution of positions relative to the two central carbon atoms. This distribution reflects the impact of the different substituents on each transition state geometry. It is interesting that fluorine atoms exhibit much less spatial variation as leaving group than other halogens for E2 while this is not at all the case for the role of fluorine as nucleophile in the very same reaction. This is likely attributed to the comparably short bond distance of fluorine for the leaving group, since in the case of the nucleophile this distance is

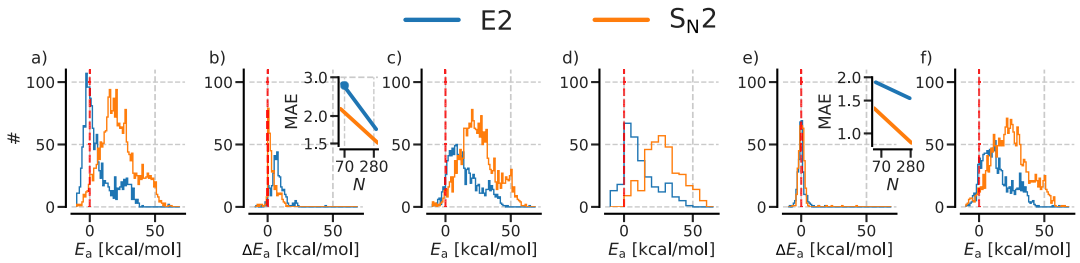

Figure 6. (a) MP2 energies of constrained geometries. (b) ΔE_a of constrained to unconstrained geometries as obtained from quantum chemistry calculations (training data for the ML models). Inset: Learning curves for the Δ-ML models (constrained to unconstrained energies) illustrating test errors (MAE in kcal/mol) vs training set size (N). (c) ML shifted MP2 energies. (d) LCCSD energies on unconstrained MP2 geometries. (e) ΔE_a of MP2 to LCCSD energies as obtained from quantum chemistry calculations (training data for the ML models). Inset: Learning curves for the Δ-ML models (MP2 to LCCSD energies) illustrating test errors (MAE in kcal/mol) vs training set size (N). (f) ML shifted LCCSD energies.

increased due to one intermediate hydrogen atom between the central carbon and the nucleophile. The reduced distance in the former case then would lead to a more pronounced Coulombic interaction with the molecule, effectively restraining the fluorine atom to a smaller volume of configurational space.

The centers of the positional distributions of the three halogens as leaving group increase with the period of the element, which is in line with typical bond radii for these elements. This is more pronounced in the case of the nucleophiles in E2 reactions where the intermediate hydrogen atom reduces the interaction between nucleophile and molecule. The result is that the nucleophile positions are spread out on arcs around the central carbon with most of the positional freedom captured by the intermediate hydrogen atom. Again, the radii of the halogen arcs follow the period of the elements, while a hydrogen as nucleophile is most flexible in regards to its distance from the central carbon.

For the distribution of internal coordinates for the S_N2 reaction in figure 5, two features are most striking: the triangular domain of the positions of halogenic nucleophiles and the bimodal distribution of hydrogens in the same case which in turn is mirrored in a bimodal distribution for the leaving group positions for all elements.

The triangular domain for halogenic nucleophiles in figure 5 can be explained by their electrostatic interaction with the reactant molecule in the gas phase. For the transition state to be a saddle point, all but one degrees of freedom must yield an increase of energy. At the tip of the triangular domain, there are three bounds to observe. First, if the distance to the carbon forming the reaction center would decrease, then the binding energy gain would become dominant, so this distance needs to be slightly above the equilibrium bond length. Secondly, the direction towards the planar substituents R1 and R2 would reduce the distance between the partially negatively charged nucleophiles and the partially positively charged hydrogen atoms of the substituents. This Coulombic interaction is more pronounced in gas-phase and restricts the possible geometries for transition states in this direction. Finally, pushing more towards the other carbon atom of the reactant skeleton (upwards in figure 5), would be unfavourable in the sp^2 hybridisation of the reaction center. Only for larger distances of the nucleophiles to the reaction center, deviations from the last two constraining factors become possible, hence the triangular shape of the domain for each element.

The bimodal distribution of the hydrogen nucleophiles for S_N2 as shown in figure 5 correlates with the leaving group in the corresponding reaction. Only leaving groups of chlorine and bromine allow a distance C2-H larger than 2 Å. This could be linked to the substantially higher electronegativity of fluorine, pulling more of the C2 electron cloud towards the leaving group, allowing for a shorter distance to the H^-.

Results such as the triangular domains and the bimodal distribution can be easily identified in large homogeneous data sets such as this one and can be interesting test cases for machine learning models for phenomena resulting from the complex interplay of competing physical interactions.

3.3. Energies

Based on the conformational search for the reactant geometries and the validated transition states, we could calculate activation energies for both reactions. Figure 6 shows the broad distribution of said activation energies which span about 50 kcal/mol. In general, E2 activation energies are lower than S_N2 activation energies. Since the activation energies are defined as the difference in energy between the transition state and the reactant complex, the nature of the reactant complex is highly relevant. This is exemplified by the significant portion of negative activation energies if we consider the constrained approximation of the reactant complex alone (panel (a) in figure 6).

These spurious negative activation energies result from two aspects: the finite number of conformers tested as potential reactant complex geometry and the constraint enforcing the characteristic alignment of the nucleophiles with the molecule when forming a reactant complex. To alleviate the impact of the former effect, we searched for more conformers until the number of negative activation energies could not be reduced any further despite testing of additional conformers. Here, the small size of the molecular skeleton was helpful, as only a few conformers can be realized for each molecule in our chemical space. We dealt with the second reason for negative activation energies by removing the constraint for the characteristic alignment of the nucleophiles with the molecule. This constraint was needed initially to ensure that the relaxation (described in the Methods section above) did not converge to an irrelevant reaction complex where the nucleophiles would be trapped by the partially positively charged hydrogen atoms of the substituents. Since the minimum of the reaction complex is only shallow, this initial constraint drastically improved the success rate of finding reactant complexes matching the reaction mechanism.

Relaxing the reactant complexes further without the constraint again bears the risk of the substituents trapping the nucleophiles. Consequently, many but not all reactant complexes could be refined this way: 301 and 348 targets for E2 and S_N2, respectively. We expect that turning the constraint into a restraint that subsequently is reduced during the minimization until the unconstrained minimum is found could be one route to identify the correct relaxation energy for all reactant complexes in our data set. However, this would be extremely costly and is subject to many degrees of freedom, like the speed at which the restraint is removed such that this route is not feasible for the thousands of reactant complexes we have in our data set. Therefore, we trained a one-hot-encoding KRR machine learning model to take the explicit relaxation energies we have found and to predict the relaxation energies for the remaining compounds. These relaxation energies span about 15 kcal/mol. We could machine learn the relaxation energy down to prediction errors of 1.5 and 1.8 kcal/mol (for 280 randomly chosen training instances) for two separate models for S_N2 and E2 reactions, respectively (see inset panel (b) of figure 6). This is much less than the expected error of the quantum chemistry method that we use, MP2. We do expect that more sophisticated machine learning methods could possibly improve upon this accuracy.

Panel (b) in figure 6 shows the activation energies for those barriers where we were able to find the explicit minimum geometry for the unconstrained reactant complex. The fact that this exhibits nearly no negative activation energy is in line with our observation that searching for additional conformers as basis for the reactant complex did not yield any further change to the activation energies. Using the explicitly calculated relaxed reactant complexes where available and including a machine learned relaxation energy in the activation energy for all other reactions, we obtain our final MP2/6-311G(d) and ML corrected MP2/6-311G(d) numbers for the activation energy, shown in panel (c) of figure 6 which now span 60 kcal/mol for S_N2 reactions and 50 kcal/mol for E2 reactions.

Comparing panels a) and c) in figure 6 shows how the number of negative activation energies has been greatly reduced by removing the constraint on the reactant complex, confirming that this was the main reason for negative activation energies in the initial case of panel (a). Calculating activation energies directly from the reactants at infinite separation is no substitute for this complicated procedure of correcting for the constraint impact. This is due to the significant interaction energy gained in forming the reactant complex in the E2 case where the negatively charged nucleophile approaches a hydrogen. For S_N2, in few cases the reactant complex might be higher in energy than the reactants at infinite separation.

Given the documented quality of MP2 geometries for substitution reactions [8], the main difference to higher level of theory than MP2 is expected to come from higher-quality energies for MP2 geometries. Since higher level of theory calculations are not affordable in the context of the geometry optimizations for this many configurations, additional single points on top of MP2 geometries recover at least a substantial part of the difference in the potential energy landscape. For those cases where we have both the transition state and the unconstrained reactant complex, we performed DF-LCCSD/cc-pVTZ calculations. The explicit data is shown in panel (d) of figure 6. The difference to the MP2 data however is more interesting and shown in panel (e) of the same figure. While the distribution of the corrections is centered around zero, the typical correction is on the order of a few kcal/mol with only few substantially larger values.

Explicit calculations of the LCCSD energy are only accessible for cases where we have an explicit molecular geometry. If the unconstrained geometry optimization did not successfully find the shallow minimum of the reactant complex, then this explicit molecular geometry is not available. To extend the coverage of the LCCSD correction which improves the accuracy of the activation energy data, we built a one-hot encoding machine learning model that predicts the LCCSD energy for the missing geometries. This Δ-ML approach exhibits learning with an error of less than 1 kcal/mol (S_N2) and 1.5 kcal/mol (E2) after training on 280 instances. After this second step, we obtain our final activation energies which have a slightly broader distribution than before, shown in panel (f) of figure 6. It is interesting to note that this final activation energy distribution of the E2 is dramatically more skewed towards very small values than the S_N2

which appears to be more normally distributed. This could be due to the symmetry in the case of the S_N2 (as also shown in figure 5) where one covalent bond is broken as the other is formed. The E2 reaction is less symmetric, effectively breaking one single bond while forming a double bond. The structural lack of symmetry is also on display in figure 5.

We also note that the learning curves for the activation energy of E2 display a higher off-set than for S_N2 even though, the E2 data has a smaller magnitude and variance. This latter aspect could be due to some extreme outliers in the E2 data set for which values larger than 50 kcal/mol have been observed, introducing severe bias in the mean absolute error. A median error measure might be better tempered for such a data set.

Panel (f) of figure 6 shows some remaining negative activation energies. For S_N2, there are 43 such negative energies, all but one of which are from machine learning predictions only. For E2, there are 120 such negative energies in total, 79 of which come from machine learning predictions. Therefore, for the majority of cases the machine learning model needs improvement, possibly by adding more explicit unconstrained reactant complexes. The cases where the explicitly calculated activation energies are still negative likely come from a finite search of conformer geometries, meaning that some unconstrained reactant complex minima have not been found. In our data set, we include these negative activation energies such that future machine learning models correcting e.g. the constrained to unconstrained relaxation can test whether they improve upon our approach.

4. Conclusion

We present a large comprehensive data set of key geometries for the two competing E2 and S_N2 reactions. We report energies and geometries obtained in a consistent and systematic manner such that this data set can serve as a playing ground for machine learning models dealing with competing reaction channels for a broad range of substituent combinations. The substituents have been chosen to reflect a substantial chemical diversity over a wide range of electron donating and electron withdrawing effect strengths. We have used the internal consistency of the data set to discuss the distribution of structural effects in transition state geometries. This was only made possible due to the large chemical space covered by our calculations. We have shown how simple machine-learning models can be used to reduce the computational cost and to curate and extend (imputation) the data set in such high-throughput efforts. The entire data set including geometries and energies at DF-LCCSD/cc-pVTZ//MP2/6-311G(d) and MP2/6-311G(d) level of theory is available as part of this publication.

Acknowledgments

We acknowledge support by the Swiss National Science foundation (No. PP00P2_138932, 407540_167186 NFP 75 Big Data, 200021_175747, NCCR MARVEL) and from the European Research Council (ERC-CoG grant QML). This work was supported by a grant from the Swiss National Supercomputing Centre (CSCS) under project ID s848. Some calculations were performed at sciCORE (http://scicore.unibas.ch/) scientific computing core facility at University of Basel.

ORCID iDs

Guido Falk von Rudorff https://orcid.org/0000-0001-7987-4330
Stefan N Heinen https://orcid.org/0000-0001-9382-2342
Marco Bragato https://orcid.org/0000-0001-5633-3101
O Anatole von Lilienfeld https://orcid.org/0000-0001-7419-0466

References

[1] Warr W A 2014 *Mol. Inform.* **33** 469
[2] Schneider N, Lowe D M, Sayle R A and Landrum G A 2015 *J. Chem. Inf. Model.* **55** 39
[3] Baylon J L, Cilfone N A, Gulcher J R and Chittenden T W 2019 *J. Chem. Inf. Model.* **59** 673
[4] Gao H, Struble T J, Coley C W, Wang Y, Green W H and Jensen K F 2018 *ACS Cent. Sci.* **4** 1465
[5] Henkelman G, Jóhannesson G and Jónsson H 2002 *Theoretical Methods in Condensed Phase Chemistry* (Berlin: Springer) pp 269–302
[6] Henkelman G, Uberuaga B P and Jónsson H 2000 *J. Chem. Phys.* **113** 9901
[7] Henkelman G and Jónsson H 2000 *J. Chem. Phys.* **113** 9978
[8] Zheng J, Zhao Y and Truhlar D G 2009 *J. Chem. Theory Comput.* **5** 808
[9] Bento A P and Bickelhaupt F M 2008 *J. Org. Chem.* **73** 7290

[10] Yi R, Basch H and Hoz S 2002 *J. Org. Chem.* **67** 5891
[11] Liu S, Hu H and Pedersen L G 2010 *J. Phys. Chem.* A **114** 5913
[12] Bickelhaupt F M 1999 *J. Comput. Chem.* **20** 114
[13] Wu X-P, Sun X-M, Wei X-G, Ren Y, Wong N-B and Li W-K 2009 *J. Chem. Theory Comput.* **5** 1597
[14] Zhao Y and Truhlar D G 2010 *J. Chem. Theory Comput.* **6** 1104
[15] Villano S M, Eyet N, Lineberger W C and Bierbaum V M 2009 *J. Am. Chem. Soc.* **131** 8227
[16] Safi B, Choho K and Geerlings P 2001 *J. Phys. Chem.* A **105** 591
[17] Gronert S, Fagin A E, Okamoto K, Mogali S and Pratt L M 2004 *J. Am. Chem. Soc.* **126** 12977
[18] Grimme S 2011 *Wiley Interdiscip. Rev.: Comput. Mol. Sci.* **1** 211
[19] Schwabe T and Grimme S 2008 *Acc. Chem. Res.* **41** 569
[20] Laio A and Parrinello M 2002 *Proc. Natl. Acad. Sci.* **99** 12562
[21] Gronert S, Pratt L M and Mogali S 2001 *J. Am. Chem. Soc.* **123** 3081
[22] Villano S M, Kato S and Bierbaum V M 2006 *J. Am. Chem. Soc.* **128** 736
[23] von Lilienfeld O A 2018 *Angew. Chem. Int. Ed.* **57** 4164
[24] Mezei P D and von Lilienfeld O A 2020 *J. Chem. Theory Comput.* **16** 2647
[25] Christensen A S and von Lilienfeld O A 2019 *Chimia* **73** 1028
[26] Christensen A S, Bratholm L A, Faber F A and von Lilienfeld O A 2020 *J. Chem. Phys.* **152** 044107
[27] Coley C W, Barzilay R, Jaakkola T S, Green W H and Jensen K F 2017 *ACS Cent. Sci.* **3** 434
[28] Sanchez-Lengeling B and Aspuru-Guzik A 2018 *Science* **361** 360
[29] Segler M H S and Waller M P 2017 *Chem. Eur.* J. **23** 5966
[30] Fabrizio A, Meyer B, Fabregat R and Corminboeuf C 2019 *Chimia* **73** 983
[31] Coley C W, Jin W, Rogers L, Jamison T F, Jaakkola T S, Green W H, Barzilay R and Jensen K F 2019 *Chem. Sci.* **10** 370
[32] Kammeraad J A, Goetz J, Walker E A, Tewari A and Zimmerman P M 2020 *J. Chem. Inf. Model.* **60** 1290
[33] Sadowski P, Fooshee D, Subrahmanya N and Baldi P 2016 *J. Chem. Inf. Model.* **56** 2125
[34] Brandt S, Sittel F, Ernst M and Stock G 2018 *J. Phys. Chem. Lett.* **9** 2144
[35] Singh A R, Rohr B A, Gauthier J A and Nørskov J K 2019 *Catal. Lett.* **149** 2347
[36] Ramakrishnan R, Dral P O, Rupp M and von Lilienfeld O A 2015 *J. Chem. Theory Comput.* **11** 2087
[37] Pilania G, Gubernatis J E and Lookman T 2017 *Comput. Mater. Sci.* **129** 156
[38] Zaspel P, Huang B, Harbrecht H and von Lilienfeld O A 2018 *J. Chem. Theory Comput.* **15** 1546
[39] Smith J S, Nebgen B T, Zubatyuk R, Lubbers N, Devereux C, Barros K, Tretiak S, Isayev O and Roitberg A E 2019 *Nat. Commun.* **10** 2903
[40] Ruddigkeit L, van Deursen R, Blum L C and Reymond J-L 2012 *J. Chem. Inf. Model.* **52** 2864
[41] Grambow C A, Pattanaik L and Green W H 2020 *Sci. Data* **7** 1
[42] Gonzales J M, Cox R S, Brown S T, Allen W D and Schaefer H F 2001 *J. Phys. Chem.* A **105** 11327
[43] Zhao Y, González-García N and Truhlar D G 2005 *J. Phys. Chem.* A **109** 2012
[44] Stei M, Carrascosa E, Kainz M A, Kelkar A H, Meyer J, Szabó I, Czakó G and Wester R 2015 *Nat. Chem.* **8** 151
[45] Hamlin T A, Swart M and Bickelhaupt F M 2018 *ChemPhysChem* **19** 1315
[46] Unke O T and Meuwly M 2019 *J. Chem. Theory Comput.* **15** 3678
[47] Gronert S 1991 *J. Am. Chem. Soc.* **113** 6041
[48] Krishnan R, Binkley J S, Seeger R and Pople J A 1980 *J. Chem. Phys.* **72** 650
[49] Curtiss L A, McGrath M P, Blaudeau J-P, Davis N E, Binning R C and Radom L 1995 *J. Chem. Phys.* **103** 6104
[50] McLean A D and Chandler G S 1980 *J. Chem. Phys.* **72** 5639
[51] Frisch M J, Pople J A and Binkley J S 1984 *J. Chem. Phys.* **80** 3265
[52] Clark T, Chandrasekhar J, Spitznagel G W and Schleyer P V R 1983 *J. Comput. Chem.* **4** 294
[53] Fast P L and Truhlar D G 2000 *J. Phys. Chem.* A **104** 6111
[54] Baker J and Pulay P 2002 *J. Chem. Phys.* **117** 1441
[55] Schenker S, Schneider C, Tsogoeva S B and Clark T 2011 *J. Chem. Theory Comput.* **7** 3586
[56] Lynch B J and Truhlar D G 2001 *J. Phys. Chem.* A **105** 2936
[57] Xu X, Alecu I M and Truhlar D G 2011 *J. Chem. Theory Comput.* **7** 1667
[58] Christensen A S, Faber F A, Huang B, Bratholm L A, Tkatchenko A, Müller K-R and von Lilienfeld O A 2017
[59] Krige D G 1951 *J. Chem. Metall. Min. Soc. South Afr.* **52** 119
[60] Heinen S, Schwilk M, von Rudorff G F and von Lilienfeld O A 2020 *Mach. Learn.: Sci. Technol.* **1** 025002
[61] Murphy K P 2012 *Machine Learning: A Probabilistic Perspective* Adaptive Computation and Machine Learning (Cambridge, MA: The MIT Press)
[62] OBoyle N M, Banck M, James C A, Morley C, Vandermeersch T and Hutchison G R 2011 *J. Cheminformatics* **3** 33
[63] Rappe A K, Casewit C J, Colwell K S, Goddard W A and Skiff W M 1992 *J. Am. Chem. Soc.* **114** 10024
[64] Riniker S and Landrum G A 2015 *J. Chem. Inf. Model.* **55** 2562
[65] Neese F 2011 *Wiley Interdiscip. Rev.: Comput. Mol. Sci.* **2** 73
[66] Francl M M, Pietro W J, Hehre W J, Binkley J S, Gordon M S, DeFrees D J and Pople J A 1982 *J. Chem. Phys.* **77** 3654
[67] Valeev E F 2020 Libint: A library for the evaluation of molecular integrals of many-body operators over Gaussian functions http://libint.valeyev.net/ version 2.7.0-beta.5
[68] Michaud-Agrawal N, Denning E J, Woolf T B and Beckstein O 2011 *J. Comput. Chem.* **32** 2319
[69] Frisch M J *et al* 2009 Gaussian 09 RevisionD.01 Gaussian Inc. Wallingford CT
[70] Becke A D 1993 *J. Chem. Phys.* **98** 5648
[71] Lee C, Yang W and Parr R G 1988 *Phys. Rev.* B **37** 785
[72] Stephens P J, Devlin F J, Chabalowski C F and Frisch M J 1994 *J. Phys. Chem.* **98** 11623
[73] Ditchfield R, Hehre W J and Pople J A 1971 *J. Chem. Phys.* **54** 724
[74] Hehre W J, Ditchfield R and Pople J A 1972 *J. Chem. Phys.* **56** 2257
[75] Hariharan P C and Pople J A 1973 *Theor. Chim. Acta* **28** 213
[76] Fukui K 1981 *Acc. Chem. Res.* **14** 363
[77] Werner H-J and Schütz M 2011 *J. Chem. Phys.* **135** 144116
[78] Hampel C, Peterson K A and Werner H-J 1992 *Chem. Phys. Lett.* **190** 1
[79] Schütz M and Manby F R 2003 *Phys. Chem. Chem. Phys.* **5** 3349

[80] Dunning T H 1989 *J. Chem. Phys.* **90** 1007
[81] Kendall R A, Dunning T H and Harrison R J 1992 *J. Chem. Phys.* **96** 6796
[82] Wilson A K, Woon D E, Peterson K A and Dunning T H 1999 *J. Chem. Phys.* **110** 7667
[83] Woon D E and Dunning T H 1993 *J. Chem. Phys.* **98** 1358