The Journal of Chemical Physics

Toward the design of chemical reactions: Machine learning barriers of competing mechanisms in reactant space

Cite as: J. Chem. Phys. **155**, 064105 (2021); https://doi.org/10.1063/5.0059742
Submitted: 11 June 2021 • Accepted: 22 July 2021 • Published Online: 10 August 2021

Stefan Heinen, Guido Falk von Rudorff and O. Anatole von Lilienfeld

The Journal of Chemical Physics ARTICLE scitation.org/journal/jcp

Toward the design of chemical reactions: Machine learning barriers of competing mechanisms in reactant space

Cite as: J. Chem. Phys. 155, 064105 (2021); doi: 10.1063/5.0059742
Submitted: 11 June 2021 • Accepted: 22 July 2021 •
Published Online: 10 August 2021

Stefan Heinen,[1,2] Guido Falk von Rudorff,[1,2] and O. Anatole von Lilienfeld[1,2,a)]

AFFILIATIONS

[1] Faculty of Physics, University of Vienna, Kolingasse 14-16, AT-1090 Wien, Austria
[2] 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

Note: This paper is part of the JCP Special Topic on Chemical Design by Artificial Intelligence.
[a)] **Author to whom correspondence should be addressed:** anatole.vonlilienfeld@univie.ac.at

ABSTRACT

The interplay of kinetics and thermodynamics governs reactive processes, and their control is key in synthesis efforts. While sophisticated numerical methods for studying equilibrium states have well advanced, quantitative predictions of kinetic behavior remain challenging. We introduce a reactant-to-barrier (R2B) machine learning model that rapidly and accurately infers activation energies and transition state geometries throughout the chemical compound space. R2B exhibits improving accuracy as training set sizes grow and requires as input solely the molecular graph of the reactant and the information of the reaction type. We provide numerical evidence for the applicability of R2B for two competing text-book reactions relevant to organic synthesis, E2 and S_N2, trained and tested on chemically diverse quantum data from the literature. After training on 1–1.8k examples, R2B predicts activation energies on average within less than 2.5 kcal/mol with respect to the coupled-cluster singles doubles reference within milliseconds. Principal component analysis of kernel matrices reveals the hierarchy of the multiple scales underpinning reactivity in chemical space: Nucleophiles and leaving groups, substituents, and pairwise substituent combinations correspond to systematic lowering of eigenvalues. Analysis of R2B based predictions of ~11.5k E2 and S_N2 barriers in the gas-phase for previously undocumented reactants indicates that on average, E2 is favored in 75% of all cases and that S_N2 becomes likely for chlorine as nucleophile/leaving group and for substituents consisting of hydrogen or electron-withdrawing groups. Experimental reaction design from first principles is enabled due to R2B, which is demonstrated by the construction of decision trees. Numerical R2B based results for interatomic distances and angles of reactant and transition state geometries suggest that Hammond's postulate is applicable to S_N2, but not to E2.

I. INTRODUCTION

To accelerate robotic experimental materials synthesis, design, and discovery,[2,3] a reliable operating system that can deploy robust virtual models of alternative chemical reaction channels is necessary. Rapid yet accurate predictions of the kinetic control of reaction outcomes for given reactants and competing reaction channels, however, are still an unsolved problem. Considerable efforts in quantum chemistry were already directed at the development of automated transition state (TS) searches and chemical reaction paths. However, calculation of the relevant parts of potential energy surfaces remains a difficult challenge under active research.[4] To this end, many TS search algorithms have been introduced, which can be grouped into single or double ended methods.[5,6] An example of the former is the single-ended growing string method,[7] which uses only the reactant as the starting point and then searches minimum energy paths and transition states. Double-ended methods, such as nudged elastic band[8,9] or the two-sided growing string method,[10] employ both reactant and product geometries to obtain a TS geometry. While successful, both approaches are computationally demanding and, in practice, often limited to small systems with mostly single step reactions.[11] Recent advances in synthesis planning and modern

machine learning techniques hold the promise for dramatic acceleration of such numerical challenges.[12,13] Already, several artificial neural networks to predict reaction outcomes were introduced (see Ref. 14 for a recent review), including work based on molecular orbital interactions of reactive sites,[15] molecular fingerprints (template based),[16] reaction site identifiers (template free),[17,18] scoring functions in search trees,[19] sequence to sequence maps,[20] and multiple fingerprint features.[21] However, all these machine learning models rely on experimental records, meaning that they are agnostic of the underlying kinetics, which are known to be crucial for reliably predicting reaction outcomes. Neglecting the energetics of chemical reactivity can be problematic, however, due to the reaction rate's exponential dependency on the activation energy (cf. the Arrhenius equation).

To use machine learning to go beyond experimental data records and toward more reliable virtual predictions of reaction outcomes for new chemistries, reaction conditions, catalysts, or solvents, access to substantial and systematic relevant training data of fundamental energetics, e.g., encoding kinetic or thermodynamic effects, is required.[22] Very recent first steps in the direction of quantum machine learning applied to reactivity included the prediction of H_2 activation barriers of Vaska's complexes,[23] the effect of nucleophilic aromatic substitution to reaction barriers,[24] the temperature dependency of coupled reaction rates,[25] or the prediction of enantioselectivity in organocatalysts.[26]

In this work, we demonstrate how the reactant-to-barrier (R2B) model effectively unifies the two directions (yield vs energy) in order to deliver robust predictions of reaction outcomes of competing mechanisms. We show how R2B can be used to predict and discriminate competing reaction channels among two of the most famous text-book reactions in chemistry, S_N2 vs E2[27] (see Fig. 1), using a quantum dataset from the literature encoding thousands of transition states obtained from high-level quantum chemistry.[28] Using our R2B model, we complete the dataset for undocumented combinations for which transition state optimizers did not converge. We also demonstrate how decision trees based on R2B give actionable suggestions for experiments on how to control which reaction channel dominates and thus the reaction outcome. On the synthetic chemistry side, an analysis of the predicted activation energies as well as transition state and reactant complex geometries based on our models suggests that Hammond's postulate is not applicable to E2.

II. METHODS

A. Kernel ridge regression

Ridge regression belongs to the family of supervised learning methods where the input space is mapped to a feature space within which fitting is performed. The transformation to the feature space is unknown *a priori* and computationally expensive. To circumvent this problem, the "kernel trick"[29] is applied where the inner product $\langle \mathbf{x}_i, \mathbf{x}_j \rangle$ of the representations of the two compounds i and j is replaced by the so-called kernel function $k(\mathbf{x}_i, \mathbf{x}_j)$. This results in kernel ridge regression (KRR). A kernel is a measurement of

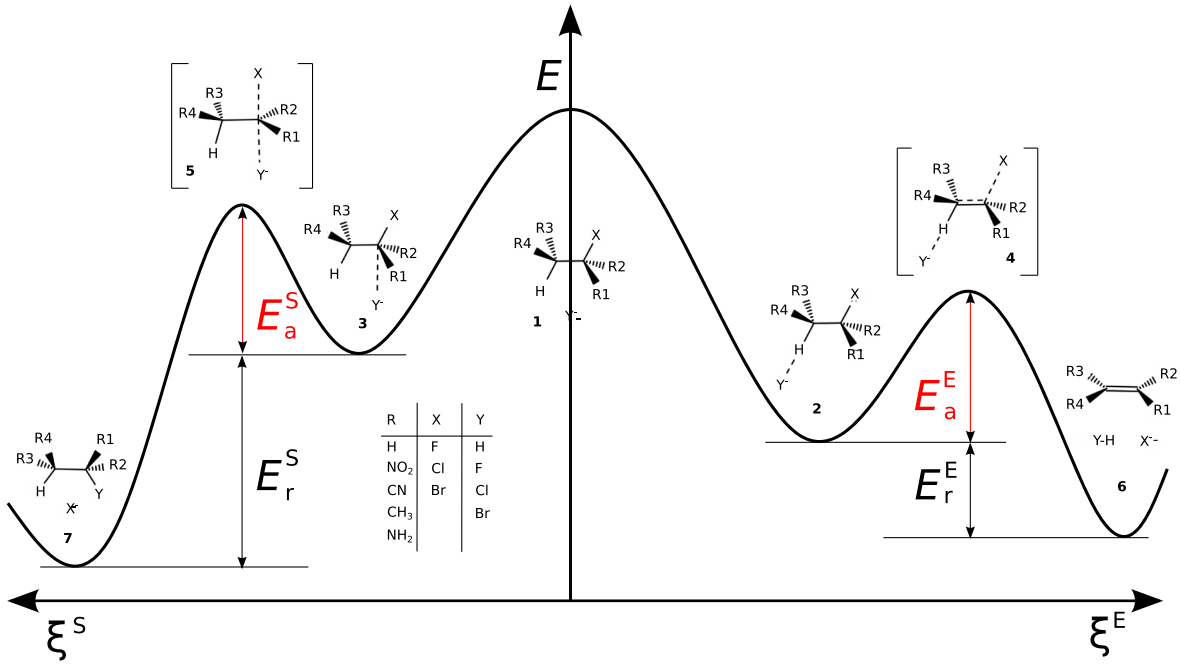

FIG. 1. Scheme for competing reactions E2 vs S_N2 with transition states E2 (**4**) and S_N2 (**5**). Reactant and nucleophile at infinite separation (**1**): in the gas phase, the energy of the transition state often lies lower than the energy of the reactants at infinite separation.[1] Product geometries at infinite separation (**6** and **7**) and reactant complexes (**2** and **3**). Properties of interest for this work are activation energies E_a^E and E_a^S, reactants, reactant complexes, and transition states. The table shows substituents R, leaving groups X, and nucleophiles Y.

similarity between two input vectors $\mathbf{x}_i$ and $\mathbf{x}_j$. In this work, we used the Gaussian kernel

$$k(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|_2^2}{2\sigma^2}\right), \quad (1)$$

with the length scale hyperparameter σ and representation $\mathbf{x}$. Using the representation of a molecule as input space, KRR learns a mapping function to a property $y_q^{\text{est}}(\mathbf{x}_q)$, given a training set of N reference pairs $(\mathbf{x}_i, y_i)$. The representation FCHL19 was optimized for the Gaussian kernel and currently represents state of the art for energy predictions within KRR based ML models. The property $y_q^{\text{est}}(\mathbf{x}_q)$ can be expanded in a kernel-basis set series centered on all the N training instances i,

$$y_q^{\text{est}}(\mathbf{x}_q) = \sum_i^N \alpha_i k(\mathbf{x}_i, \mathbf{x}_q), \quad (2)$$

where α_i is the ith component of the regression coefficient vector $\boldsymbol{\alpha}$ which is obtained as:

$$\boldsymbol{\alpha} = (\mathbf{K} + \lambda \mathbf{I})^{-1}\mathbf{y}, \quad (3)$$

with the regularization strength λ, the identity matrix $\mathbf{I}$, and the kernel matrix $\mathbf{K}$ with kernel elements $k(\mathbf{x}_i, \mathbf{x}_j)$ for all training compounds. The kernel ($\mathbf{K}$) within a representation stays the same for both reactions and the difference in the R2B models ($\boldsymbol{\alpha}$) enters in the change of the label ($\mathbf{y}$).[30]

B. Representations

Here, we have selected four representations of varying complexity: the Bag of Bonds (BoB),[31] spectrum of London[32] and Axilrod–Teller–Muto[33,34] (SLATM) potentials, FCHL19,[35] and one-hot encoding.[29]

BoB uses the nuclear Coulomb repulsion terms from the Coulomb matrix representation (CM[36]) and groups them into different bins (so-called bags) for the different elemental atom pair combinations. SLATM[37] uses London dispersion contributions as the two body term (rather than coulomb repulsion) and the Axilrod–Teller–Muto potential as the three body term. While the FCHL18 parameterization accounts for one-body effects in terms of the position of the element in the Periodic Table (group and period),[38] FCHL19 limits itself to two- and three-body terms for the sake of computational efficiency.[35] Its two-body terms contain interatomic distances R scaled by R^{-4}, and the three-body terms account for the angular information among all atom triples scaled by R^{-2}.

All three geometry-based representations have been tested extensively on close-to-equilibrium structures. Since reactive processes, by definition, deal with out of equilibrium structures, we have also included a simple geometry free representation, namely, one-hot encoding. This representation has also been used to encode amino acids in peptides for artificial neural networks.[39,40] In one-hot encoding, the representation is a vector of zeros and ones (i.e., a bit vector) where only one entry is non-zero per feature. To describe the molecules, we used a bit vector for every substitution site R$i \in \{1, 2, 3, 4\}$ and one for the nucleophiles (Y) and the leaving group (X), respectively. This results in a combined vector containing 6 bit vectors of total length of 27 bits.

C. Training and testing: Learning curves

To train our R2B models, the dataset was split into a training set and a test set to optimize the hyperparameters and evaluate the model, respectively. To get the optimal hyperparameters, we used k-fold cross validation.[29] We divide the training data into k folds and, for each fold, we trained on all but one fold, which was used for evaluating the model. This procedure was done in an iterative fashion over all the folds. We then calculated the averaged error over these folds. This was done for different combinations of hyperparameters σ and λ.

The input for all the geometry based R2B models was the reactants at infinite separation (Fig. 1, compound **1**). For each reaction, different reactant conformers (yielding different reactant complexes; Fig. 1, compounds **2** and **3**) have been reported in the dataset.[28] To obtain a uniquely defined problem for the ML models, we canonicalized the reactant complexes by always choosing the lowest-lying one from the source database. Using compound **1**, the kernel for both reaction channels is the same (K^{tot}), which contains two kernels: one for the molecule (M and M') and one for the attacking group (Y and Y'), as shown in Eq. (4). Therefore, for both reactions, the same kernel can be used, and the difference in the training enters by the activation energy ($\mathbf{y}$) in Eq. (3),

$$K^{\text{tot}} = K(Y, Y') \circ K(M, M'). \quad (4)$$

Since one-hot encoding does not depend on the geometry, the kernel can be calculated directly for the entire system.

In order to measure the accuracy of our R2B models, we picked the best set of hyperparameters and trained the model using different training set sizes N and plotted the mean absolute errors (MAEs) vs N (in a log–log plot), resulting in learning curves. Using learning curves allowed us to see the learning behavior of our R2B models and compare different representations. The error ϵ of a consistently improving ML model should decrease following a power law for increasing training set sizes N[41] in a logarithmic scale,

$$\log(\epsilon) = \log(a) - b \cdot \log(N) + \cdots, \quad (5)$$

where a is the offset (an indicator of how well the selected basis functions fit reality) and b is the slope of the learning curve that describes the speed at which the accuracy increases using larger training set sizes. Higher order terms ($\cdots$) were neglected in this work, as commonly done.

D. Data and scripts

The data extracted from QMrxn20[28] are available on GitHub.[42] The scripts used to optimize the hyperparameters and to generate the learning curves are also available in the same Git repository.

The dataset QMrxn20[28] contains 1286 E2 and 2361 S_N2 machine learned LCCSD activation barriers (ΔE_a). From these reactions, 529 are overlapping reactions, meaning that they start from the same reactant (**1**) and go over different reactant complexes (E2: **2** and S_N2: **3**) toward the corresponding transition states (E2: **4** and S_N2: **5**). All geometries in the dataset had been optimized with MP2/6-311G(d),[43–47] and subsequently, DF-LCCSD/cc-TZVP single point calculations (as implemented in Molpro2018) were performed.[48–54] The backbone scaffold of all reactants is an ethane molecule, which is substituted by functional groups and a leaving

group. The system also contains the nucleophile (attacking group). The chemical composition of the reactant complexes is shown in the table in Fig. 1 and contains the functional groups –H, $–NO_2$, –CN, $–CH_3$, and $–NH_2$; the leaving groups –F, –Cl, and –Br; and the nucleophiles H^-, F^-, Cl^-, and Br^-. The molecular system (e.g., the reactant complex) is negatively charged and contains at most 21 atoms (including hydrogen atoms) or 16 heavy atoms (non-hydrogen atoms). To ensure the data source[28] did not contain duplicated reactions, we calculated the L^2 norm of all pairwise differences between training and test compounds of the corresponding FCHL19 representations and identified only 3 out of the 3647 cases where that norm is very close to zero. We have inspected these three cases, and they correspond to systems that only differ in the location of the same set of substituents. As such, they are distinct but are, due to their similarity, mapped to very similar regions in feature space. In any case, since they amount to only less than one per mille of the data points, we chose to work on the original dataset for better comparison to the literature.

III. RESULTS AND DISCUSSION

A. Learning barriers

Conventionally, the first principles based prediction of activation energies requires the use of sophisticated search-algorithms that iteratively converge toward relevant transition state geometries that satisfy the potential energy saddle-point criterion.[8,10,55] The activation energy is then obtained as the energy difference between reactant and transition state geometry. By contrast, our R2B models solely rely on reactant information as input. We trained them using aforementioned geometry based representations BoB,[31] SLATM,[37] FCHL19,[35] and one-hot-encoding to predict activation energies solely based on reactants at infinite separation as input geometries (compound **1** in Fig. 1). The resulting learning curves in Fig. 2 indicate systematically improving activation energy predictions with increasing training set size N for E2 and S_N2. For both mechanisms, the most data-efficient R2B models (one-hot-encoding) reach prediction errors of 3 kcal/mol with respect to Coupled-Cluster Singles Doubles (CCSD) reference, i.e., on par with the deviation of MP2 from CCSD, already for less than 300 training instances. For 2000 training instances, the prediction error approaches 2 kcal/mol. Moreover, the lack of convergence suggests that chemical accuracy (1 kcal/mol) could be reached if several thousand training data points had been available. The insets of Fig. 2 show true (E_a^{ref}) vs predicted (E_a^{est}) activation barriers for both reactions. Barriers in the range of 0–50 kcal/mol are predicted with decent correlation coefficients (0.89 and 0.94 for E2 and S_N2, respectively). In short, after training on reference activation energies obtained for explicit transition state geometries (taken from the QMrxn20 dataset[28]), the learning curves in Fig. 2 amount to overwhelming evidence that it is possible to circumvent the necessity for explicit transition state structural search when predicting activation energies for out-of-sample reactants.

The trends among learning curves in Fig. 2 are consistent with literature results for equilibrium structures: The accuracy improves when going from BoB to SLATM and FCHL19 for a given training set size.[56] Most surprisingly, however, all R2B models based on geometry dependent representations are less accurate than one-hot encoding. While still unique (a necessary requirement for functional

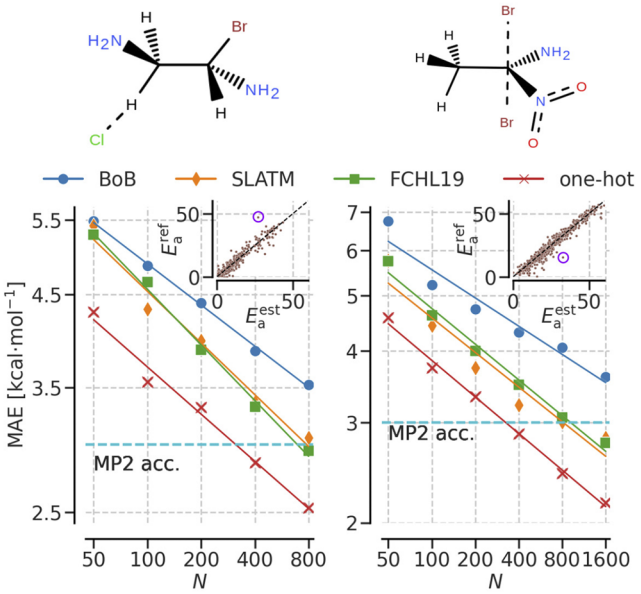

FIG. 2. Learning curves: activation energy prediction errors (out-of-sample) as a function of training set size N for activation barriers (E_a) of E2 (left) and S_N2 (right) using reactant information as inputs only. Results are shown for four representations (BoB, SLATM, FCHL19, and one-hot) used within KRR models. The training data reference level of theory corresponds to DF-LCCSD/cc-pVTZ//MP2/6-311G(d), and the estimated MP2 error is denoted as a blue dashed horizontal line. Insets: reference vs estimated activation barriers using one-hot-based predictions and R^2 values being 0.89 and 0.94 for E2 (left) and S_N2 (right), respectively.

R2B models[57,58]), one-hot encoding is devoid of any structural information, and its outstanding performance is therefore in direct conflict with the commonly made conclusion that a physics inspired functional form of the representation is crucial for the performance of R2B models.[56,59,60] Relying only on the period and group information in the Periodic Table to encode composition, other geometry-free representations have also been applied successfully to the study of elpasolite[61] or perovskite[62] crystal structures. Here, by contrast, one-hot encoding provides the compositional information for a fixed scaffold.

One can speculate about the reasons for the surprising relative performance of one-hot encoding. Due to its inherent lack of resolution, which prohibits the distinction between reactant and transition state geometries, it could be that one-hot encoding represents a more efficient basis, which effectively maps onto a lower dimensionality with superior learning performance. In particular, the inductive effect (practically independent of specific geometric details) is known to dominate barrier heights for the types of reactions under consideration,[63] and it is explicitly accounted for through one-hot encoding without imposing the necessity to differentiate it from the configurational degrees of freedom.

Figure 2 shows one outlier per reaction. For the E2 case, the molecule closest in one-hot encoding to the failed prediction (only differs in X and Y) has a much smaller barrier of 12 kcal/mol. Similarly, for the S_N2 reaction, the closest molecule (only differs in X and Y) has a barrier of 24 kcal/mol. As such, this scarcity of training instances in close vicinity to the outlier might be at the origin

for such relatively large prediction errors. To get an idea of the inner workings of the one-hot encoding model, we performed a principal component analysis (PCA) of the kernel matrix of the predictions, which can go either way, i.e., E2 or S_N2. For this subset, it is the difference in activation energy that will determine the kinetically stabilized product. Color coding the first two components by the difference in reference activation barrier labels for the two reactions results in the graphic featured in Fig. 3. Confidence ellipsoids of the covariance using Pearson correlation coefficients encode intuitive clusters corresponding to leaving-group/nucleophile combinations and suggest that substituents have less significant effect on trends in activation energies. However, the eigenvalue spectrum of the PCA in Fig. 3 decays rapidly only after the 21st eigenvalue, which indicates the number of effective dimensions of the model and implies that the substituents, although smaller, still have an effect on the activation barrier. This is consistent with the dimensionality of the one-hot encoding representation: the vector length is 27 (3 X's, 4 Y's, and 4 · 5 R's), which is overdetermined, meaning that the X part of the representation vector consists of three elements F (1, 0, 0), Cl (0, 1, 0), or Br (0, 0, 1). This could also be uniquely defined with F (0, 0), Cl (1, 0), Br (0, 1), which leads to a dimension of 21 and is in agreement with the dimensionality of the representation. To further investigate the R2B model, we looked at the training set selection. It is known that for clustered data (see Fig. 3), random splits used in this work tend to perform better than splits along a cluster, even though random splits are more congruent with the nature of the reaction space under investigation. For comparison, in a first model, we excluded the functional group NO_2 at position R2 and in a second model at two positions R2 and R3 from the training to see how one-hot encoding and FCHL19 perform for known functional groups but unknown positions in the test set. Figure 4 shows the learning curves for both cases. Although there is still learning, one-hot encoding does not perform as well as a structural representation (FCHL19). For FCHL19 in the E2 case, the learning is not affected at all compared to random training set selection and the model reaches a similar MAE for 800 training instances. FCHL19 is able to infer the missing functional group at position R2 from training compounds where this functional group is present at the neighboring position R1, since the corresponding representation vectors are similar. In addition, one-hot encoding shows learning, but it is not the dominant model anymore. In this case, learning is possible because the functional groups contribute additively to the activation energy as described in Ref. 63. This means that all the other functional groups improve, except NO_2 at positions R2 and R3, since it has no corresponding training data. For S_N2, both models perform worse when excluding a functional group, especially for the position at R2, which is closer to the reaction center and therefore contributes more to the barrier. This also explains why the models perform better if two functional groups are missing in the S_N2 reaction. The second functional group at position R3 adds more barriers to the test set with a smaller impact on the barrier (farther away from the reaction center), which makes the learning problem easier. For larger molecules, not all combinations of functional groups are present in the training data, rendering a cluster split a more realistic scenario. In those cases, one-hot encoding will be less applicable and likely outperformed by scalable approaches, e.g., Amons.

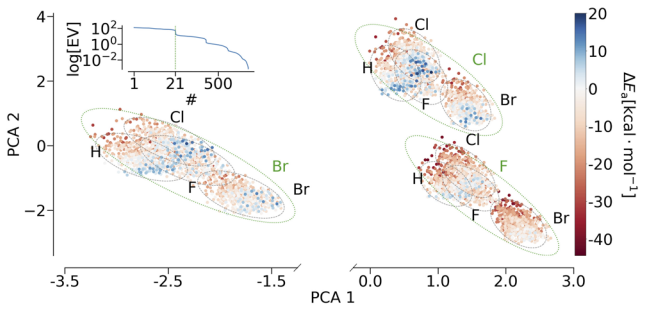

FIG. 3. Kernel PCA of the training set. Kernel PCA of one-hot encoding colored by the energy difference of activation energies of the two reactions $\Delta E_a = E_a^E - E_a^S$. Inset: eigenvalues of the kernel PCA. Clusters represent most frequent combinations of leaving groups X (green) and nucleophiles Y (black).

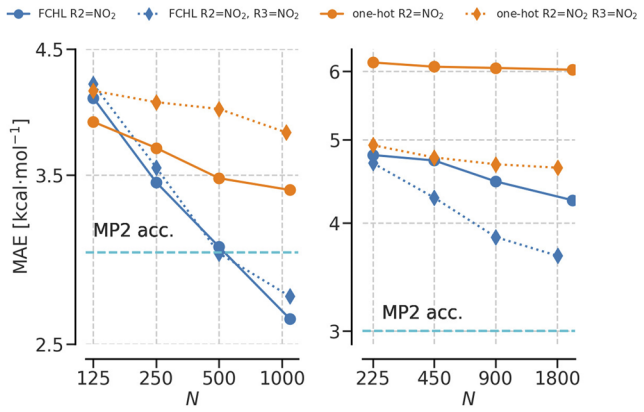

FIG. 4. Learning curves across cluster test error (MAE) vs training set size (*N*) excluding NO_2 from training at position R2 (spheres) and at positions R2 and R3 (diamonds) for both reactions E2 (left) and S_N2 (right). The test set only contains compounds with NO_2 at position R2 (spheres) or at positions R2 and R3 (diamonds).

B. New barrier estimates

Using one-hot encoding (leading to the most performing model), we have trained two models, corresponding to the 1286 and 2361 activation energies of E2 and S_N2 transition state geometries, respectively. Subsequently, these two models were used to predict 11 353 E2 and S_N2 activation barriers for which conventional transition state search methods had failed within the protocol leading up to the training dataset.[28] A comparison of the Rogers–Tanimoto distances (see the supplementary material) between the QMrxn20 dataset and the missing data points showed that the dissimilarity within the QMrxn20 dataset is comparable to the one of QMrxn20 vs the missing data points. Together with the learning curves shown above, this suggests that our model is applicable to the missing data points from QMrxn20. A summary of the difference in these predicted activation barriers is presented in Fig. 5, where the *x*-axis corresponds to the nucleophiles Y and the *y*-axis to the leaving groups X. For every combination of X and Y, there are 5 · 5 squares for the functional groups at position R1 and R2. Within these, there are again 5 · 5 squares belonging to R3 and R4. Each

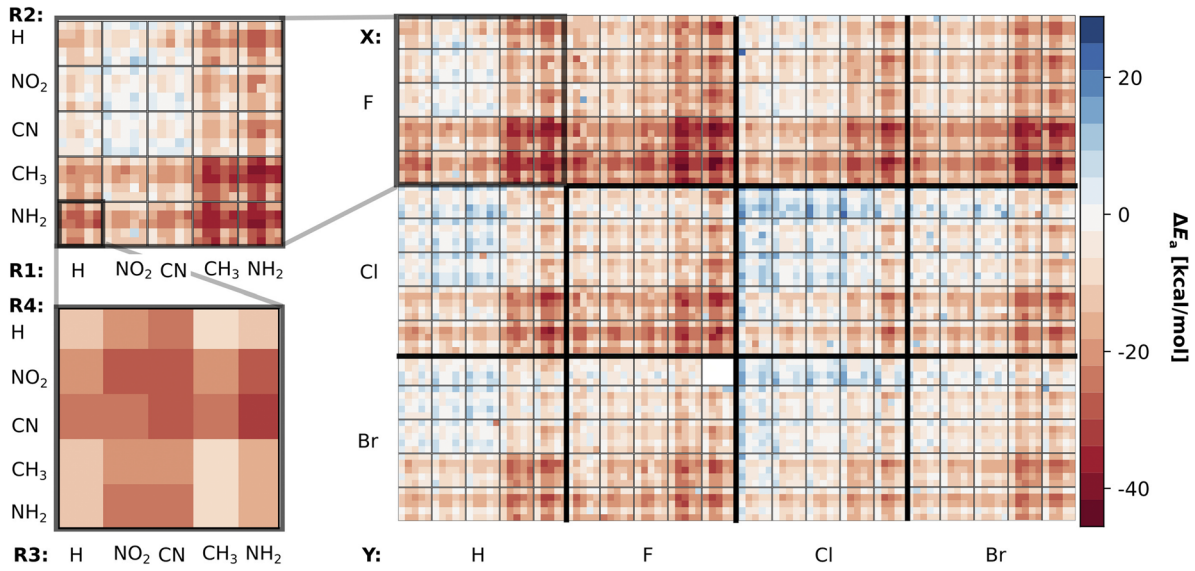

FIG. 5. Completion of the dataset using predictions of R2B models. Differences in activation energies ($\Delta E_a = E_a^E - E_a^S$) for all 7500 reactions (calculated and predicted). Every square stands for a combination of R1-4, X, and Y shown in Fig. 1. Positive values denote compounds that undergo a S_N2 reaction, and negative values lead toward an E2 reaction.

of the squares represents one reaction for a given combination of R1-4, X, and Y. Simple heuristic reactivity rules emerge from inspection of these results: If the nucleophile and the leaving group are Cl, the preferred reaction is S_N2. If the nucleophile and the leaving group are F, the preferred reaction is E2. The functional groups at positions R1 and R2 favor E2 due to their electron donating properties, which disfavor a nucleophilic backside attack in the S_N2 reaction. A comprehensive overview is shown in Fig. 5. The same rules can be observed in Fig. 6, which shows the distribution of the differences in the activation barrier (ΔE^a) of the training, predicted, and total datasets. The molecules of the extreme cases, largest difference in activation energies, are shown for both reactions, E2 (left) and S_N2 (right). Figure 6 shows a favorization of the E2 reaction of a rate of roughly 75%. These results have to be taken with caution since this shift in E2 can also have occurred due to the composition of the molecules in the training set, as well as the choice of small functional groups that minimizes steric effects.

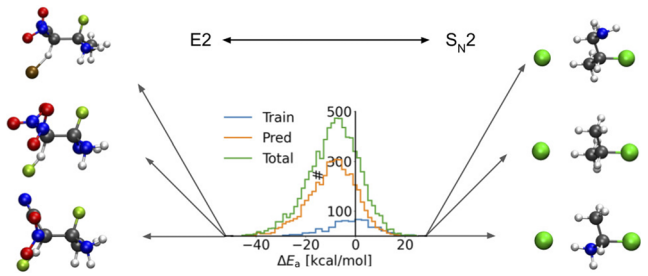

FIG. 6. Histogram of energy distribution of ΔE_a. Differences in activation energies ($\Delta E_a = E_a^E - E_a^S$) of 529 overlapping training instances (blue), 11k predictions (orange), and all 7500 reactions (green). Molecules of the three highest (respectively, lowest) barrier differences are shown as molecules.

A more detailed discussion of the training and the dataset completion with the R2B model can be found in the supplementary material.

C. Design rule extraction

So far, most studies based on artificial neural networks aimed at predicting chemical reactions using experimental data do not account for the kinetics of reactions. It is well known, however, that activation barriers are crucial for chemical synthesis and retrosynthesis planning. This is exemplified by a decision tree for the competing reactions E2 and S_N2 in Fig. 7. The goal of such trees is to improve the search for better reaction pathways (lower activation barriers) by showing the estimated change in energy when changing functional groups, leaving groups, or nucleophiles. To extract such rules for the design problem, a large and consistent reaction dataset is needed. After completing the dataset,[28] we are now able to identify (given a desired product) the estimated changes in the activation barrier when substituting specific functional groups, leaving groups, or nucleophiles. This way, the yield of chemical reactions can be optimized by getting insights into the effects that functional groups have on a certain molecule. Furthermore, this insight could be used to direct reactions toward the desired product. Figure 7 shows such a possible decision tree to determine the change in barriers while exchanging substituents. Starting from the total dataset (left energy level), the first decision considers the functional group NH_2 at position R1. Going down the tree means accepting the suggested change and the respective compounds, while going up means declining and removing these compounds from the data. Depending on which product is sought after, hints to improve the energy path can be found while constantly accepting (going down) or declining (going up) the tree. For example, if the desired reaction is E2, then the best way is to go down the tree (decision accepted),

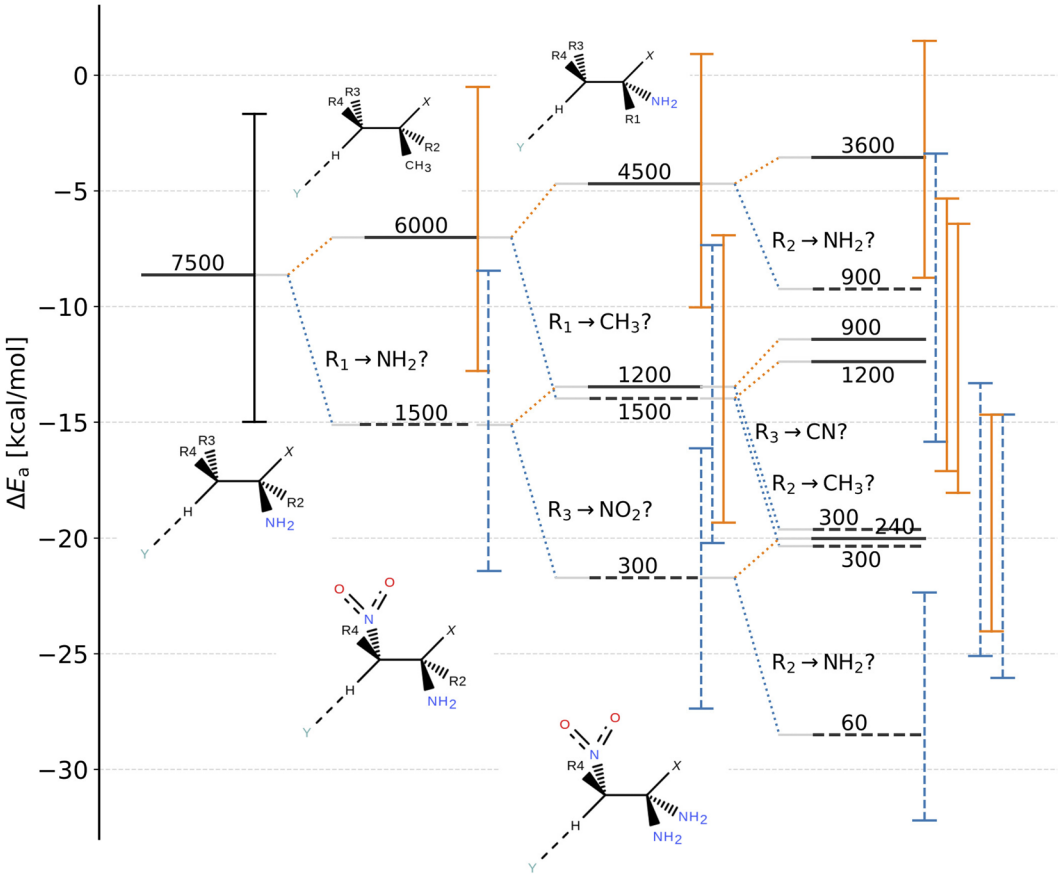

FIG. 7. Decision tree using extracted rules and design guidelines. Decision tree using the R2B estimated activation barriers to predict changes in barrier heights by starting at all reactions (first energy level on the left) and subsequently applying changes by substituting functional groups, leaving groups, and nucleophiles with E2 as an example. Blue dotted lines refer to an accepted change, meaning that only compounds containing these substituents at the position are considered. Orange dotted lines refer to substitution declined, meaning that all compounds except the decision are kept. Vertical lines on the right of energy levels denote the minimum first (lower limit) and the third (upper limit) quartile of a box plot over the energy range. Numbers above energy levels correspond to the number of compounds left after the decision. Lewis structures resemble the decision in question.

which adds electron withdrawing groups to the R3 and R4 positions, as well as electron donating groups to R1 and R2. In Fig. 7, the first decision redirects the barrier toward E2 of about ~8 kcal/mol by adding an electron withdrawing group (NO_2) on the α-carbon. On the other hand, electron donating group at the β-carbon favor the E2 reaction because they facilitate the abstraction of the leaving group, which is shown in the second and the third decision, where NH_2 was added in both positions, R1 and R2. In addition to the R2B predictions, which estimate the outcome of a specific combination of one reaction, a decision tree gives simple rules as a coarsened aggregation that can be used in reaction design to achieve a desired outcome.

D. Estimates of reactant and transition state geometries

Additional to barriers, we analyzed the geometries of the transition states as well as the geometries of the reactant complexes.[28] Choosing key geometrical parameters, such as distances, angles, and dihedrals, we were able to train R2B models to learn these properties using the one-hot encoding as the representation. These parameters were extracted from the ethylene scaffold defining the key positions of the substituents, leaving groups, and nucleophiles shown in Fig. 8, compounds **2** and **3** for the E2 and S_N2 reaction, respectively.

The parameters for the E2 reaction are the C–X distance d_x, the C–Y distance d_y, the X–C–C angle α, the C–C–Y angle β, and the X–C–C–Y dihedral θ. Similarly, for S_N2, we have the C–X distance d_x, the C–Y distance d_y, and the X–C–Y angle α. For every parameter, a separate model was trained using the one-hot representation. Although this representation does not contain any geometrical information, learning was achieved for every parameter. Figure 8 shows the learning curves and as horizontal dashed lines the null model, which uses the mean of the training set for predictions. In the same way as for the transition state geometries, we also trained a model for the reactant complexes. Figure 8 shows the learning

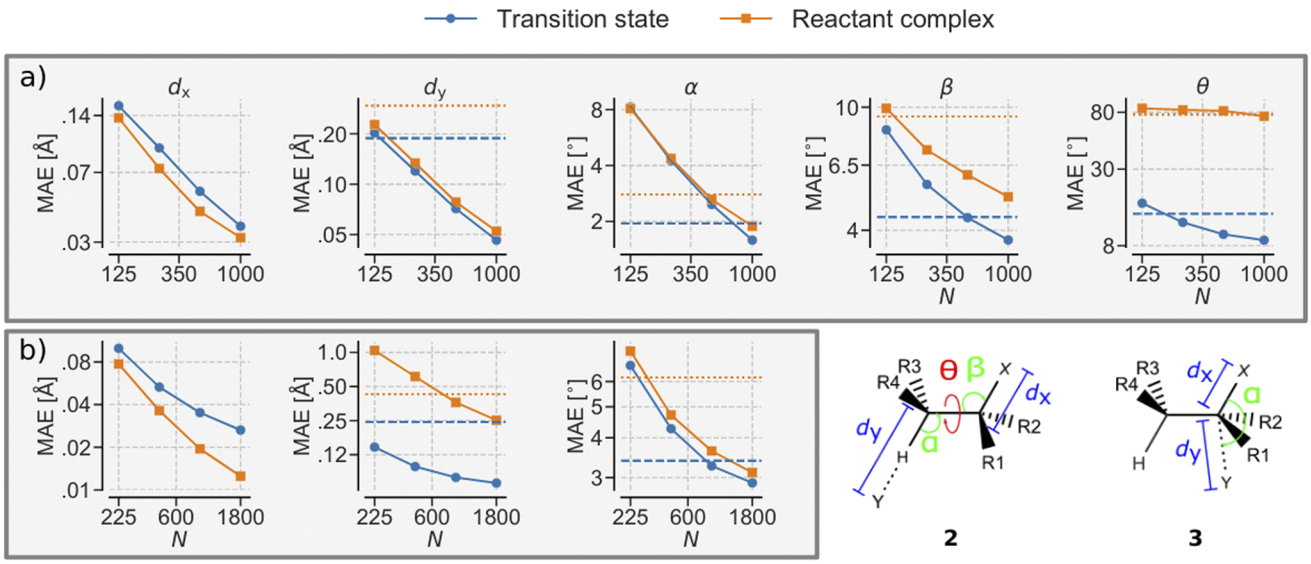

FIG. 8. Model evaluation of geometrical properties using learning curves. Test errors (MAE) of distances $d_{x,y}$, angles α and β, and dihedrals θ for both reactions E2 (a) and S_N2 (b). Horizontal lines correspond to the null model, which uses the mean value of the training set for predictions. Compounds (**2** and **3**) illustrate the learned properties of the E2 reaction (**2**) and the S_N2 reaction (**3**) for reactant complexes and transition states.

curves for both transition states and reactant complexes. The results for both geometries are similar except for the dihedral of the reactant complexes. The poor performance results from the conformer search of the reactants. As opposed to bond distances, dihedrals have multiple local minima, which lead to larger differences between the reactant and transition state structures. The variance of the dihedrals is significantly higher, which makes the learning task much harder. The one-hot representation does not contain any geometrical information and therefore is not able to learn the different geometries only using information about the constitution (R's, X's, and Y's) of the reactant complexes. The poor performance of the model on angles and especially on dihedrals renders the one-hot encoding impractical for 3D geometry predictions. The recently published Graph to Structure (G2S) quantum machine learning model[64] seems to be more suitable for the 3D coordinate prediction problem in QMrxn20.

E. Hammond's postulate

To investigate Hammond's postulate, we took the difference in the predicted geometries (d_x and d_y) for all 7500 reactions for the E2 and the S_N2 reaction, respectively. Then, we plotted these values against the activation energies of both reactions E_a^E and E_a^S (Fig. 9). The distances Δd_x correlate well with the energies with R^2 values of 0.87 and 0.80 for E2 and S_N2, respectively. This is explained by the leaving group that is bonded to the carbon atom in the reactant complex and only small changes in distance happen moving toward the transition state geometry. For the S_N2 reaction, the backside attack

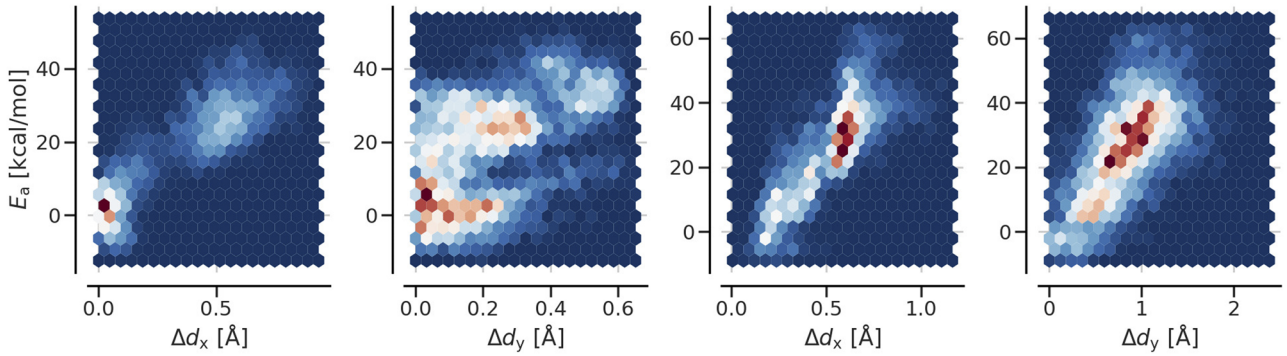

FIG. 9. Applicability of Hammond's postulate. Frequency heat map of activation energies projected onto structural differences in distances (d_x and d_y) between reactant complex conformers and transition states for both reactions E2 (first two plots) and S_N2 (last two plots). For S_N2, a good linear correlation (R^2 are 0.8 and 0.65 for d_x and d_y, respectively) in agreement with Hammond's postulate can be observed, while for E2, only d_x shows good correlation ($R^2 = 0.86$), whereas d_y lacks correlation ($R^2 = 0.5$).

of the nucleophile does not allow a broad distribution of distances and angles in the reactant complex and the transition state. Moreover, the changes in geometry between the reactant complex and the transition state are modest. Therefore, the parameter Δd_y for the S_N2 correlates well with the activation energy E_a^S, which results in an R^2 value of 0.65. The attack of the nucleophile on the hydrogen atom (E2 reaction) allows for a much broader distribution of the position of the nucleophile in the transition state. This makes the learning problem more difficult, especially for a representation not including geometrical information. These higher degrees of freedom result in an R^2 value of 0.50.

Hammond's postulate typically holds for the end points of an intrinsic reaction coordinate (IRC) calculation,[65–67] which leads to a local minimum close to the transition state. Therefore, the reactant only needs a few reorganizations toward the transition state. For geometries that are farther away from the transition state (such as in our E2), Hammond's postulate cannot hold anymore. This means that even though more reorganization steps toward a transition state have to be made, the activation energy is not affected anymore. As a consequence, Hammond's postulate is no longer applicable.

IV. CONCLUSION

We have introduced a new machine learning model dubbed Reactant-To-Barrier (R2B) to predict activation barriers using reactants as input only. This approach renders the model useful in practice, as the dependency on the transition state geometry is only implicitly obtained at the training stage and not explicitly required for querying the model. We find that one-hot-encoding, the trivial geometry free based representation, yields even better results than geometry based representations designed for equilibrium structures. As such, our results indicate that accounting only for the combinations of functional groups, leaving groups, and nucleophiles of the reaction is sufficient for promising data-efficiency of the model. Using R2B predictions, we completed the reaction space of QMrxn20.[28] Future work could include delta ML[68] to improve these results even further, as corroborated by the preliminary results in Ref. 28, further improvements on the representation (as recently found to lead to improved barrier predictions for enantioselectivity in metal-organic catalysts[26]), or catalytic or solvent effects.[69]

Using R2B predicted activation barriers, we have also built a decision tree, enabling the design and discrimination of either reaction channel encoded in the data. Such trees systematically extract the information hidden in the data and the model regarding the combinatorial many-body effects of functional groups, leaving groups, and nucleophiles, which result in one chemical reaction being favored over the other. As such, they enable the control of chemical reactions in the design space spanned by reactants. Finally, we also report on geometries of the reactant complexes consisting of different conformers as well as on R2B based transition state geometry predictions. Using these results, we discuss the limitations of Hammond's postulate, which does not hold for the E2 reactant complexes stored in the QMrxn20 dataset.[28]

SUPPLEMENTARY MATERIAL

The supplementary material contains the results used to generate the learning curves for barrier learning (Tables 1 and 2) and geometry learning (Tables 3 and 4). It also gives a brief explanation on how the models were trained and shows a heat map for a hyperparameter scan of sigmas and lambdas containing the training errors (Fig. 1). Additionally, we added more learning curves (barrier learning) using different geometries as input for the representations. Finally, we added Fig. 3, which compares the Rogers–Tanimoto coefficients between the training and the test set.

ACKNOWLEDGMENT

This project has received funding from the European Union's Horizon 2020 research and innovation program under Grant Agreement Nos. 952165 and 957189. This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (Grant Agreement No. 772834). This work only reflects the author's view and the EU is not responsible for any use that may be made of the information it contains. This work was partly supported by the NCCR MARVEL, funded by the Swiss National Science Foundation.

DATA AVAILABILITY

The data that support the findings of this study are openly available at http://doi.org/10.5281/zenodo.4925938.

REFERENCES

[1] G. Vayner, K. N. Houk, W. L. Jorgensen, and J. I. Brauman, J. Am. Chem. Soc. **126**, 9054 (2004).
[2] J. M. Granda, L. Donina, V. Dragone, D.-L. Long, and L. Cronin, Nature **559**, 377 (2018).
[3] M. M. Flores-Leonar, L. M. Mejía-Mendoza, A. Aguilar-Granda, B. Sanchez-Lengeling, H. Tribukait, C. Amador-Bedolla, and A. Aspuru-Guzik, Curr. Opin. Green Sustainable Chem. **25**, 100370 (2020).
[4] T. A. Young, J. J. Silcock, A. J. Sterling, and F. Duarte, Angew. Chem., Int. Ed. **60**, 4266 (2020).
[5] A. L. Dewyer, A. J. Argüelles, and P. M. Zimmerman, WIREs Comput. Mol. Sci. **8**, e1354 (2017).
[6] C. W. Coley, N. S. Eyke, and K. F. Jensen, Angew. Chem., Int. Ed. **59**, 22858 (2020).
[7] P. M. Zimmerman, J. Comput. Chem. **36**, 601 (2015).
[8] G. Henkelman, B. P. Uberuaga, and H. Jónsson, J. Chem. Phys. **113**, 9901 (2000).
[9] G. Henkelman, G. Jóhannesson, and H. Jónsson, in *Theoretical Methods in Condensed Phase Chemistry* (Springer, 2002), pp. 269–302.
[10] P. M. Zimmerman, J. Chem. Phys. **138**, 184102 (2013).
[11] C. A. Grambow, L. Pattanaik, and W. H. Green, Sci. Data **7**, 137 (2020).
[12] S. Szymkuć, E. P. Gajewska, T. Klucznik, K. Molga, P. Dittwald, M. Startek, M. Bajczyk, and B. A. Grzybowski, Angew. Chem., Int. Ed. **55**, 5904 (2016).
[13] P. M. Pflüger and F. Glorius, Angew. Chem., Int. Ed. **59**, 18860 (2020).
[14] F. Strieth-Kalthoff, F. Sandfort, M. H. S. Segler, and F. Glorius, Chem. Soc. Rev. **49**, 6154 (2020).
[15] M. A. Kayala, C.-A. Azencott, J. H. Chen, and P. Baldi, J. Chem. Inf. Model. **51**, 2209 (2011).
[16] J. N. Wei, D. Duvenaud, and A. Aspuru-Guzik, ACS Cent. Sci. **2**, 725 (2016).
[17] W. Jin, C. Coley, R. Barzilay, and T. Jaakkola, in *Advances in Neural Information Processing Systems 30*, edited by I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Curran Associates, Inc., 2017), pp. 2607–2616.
[18] D. Fooshee, A. Mood, E. Gutman, M. Tavakoli, G. Urban, F. Liu, N. Huynh, D. Van Vranken, and P. Baldi, Mol. Syst. Des. Eng. **3**, 442 (2018).

[19]M. H. S. Segler, M. Preuss, and M. P. Waller, Nature **555**, 604 (2018).
[20]P. Schwaller, T. Gaudin, D. Lányi, C. Bekas, and T. Laino, Chem. Sci. **9**, 6091 (2018).
[21]F. Sandfort, F. Strieth-Kalthoff, M. Kühnemund, C. Beecks, and F. Glorius, Chem **6**, 1379 (2020).
[22]B. Huang and O. A. von Lilienfeld, "Ab initio machine learning in chemical compound space," arXiv:2012.07502 [physics.chem-ph] (2020).
[23]P. Friederich, G. dos Passos Gomes, R. De Bin, A. Aspuru-Guzik, and D. Balcells, Chem. Sci. **11**, 4584 (2020).
[24]K. Jorner, T. Brinck, P.-O. Norrby, and D. Buttar, Chem. Sci. **12**, 1163 (2021).
[25]E. Komp and S. Valleau, J. Phys. Chem. A **124**, 8607 (2020).
[26]S. Gallarati, R. Fabregat, R. Laplaza, S. Bhattacharjee, M. D. Wodrich, and C. Corminboeuf, Chem. Sci. **12**, 6879 (2021).
[27]N. E. Schore and K. P. C. Vollhardt, *Organische Chemie* (John Wiley & Sons, 2011).
[28]G. F. von Rudorff, S. N. Heinen, M. Bragato, and O. A. von Lilienfeld, Mach. Learn.: Sci. Technol. **1**, 045026 (2020).
[29]K. P. Murphy, *Machine Learning: A Probabilistic Perspective (Adaptive Computation and Machine Learning)* (The MIT Press, 2012).
[30]R. Ramakrishnan and O. A. von Lilienfeld, CHIMIA Int. J. Chem. **69**, 182 (2015).
[31]K. Hansen, F. Biegler, R. Ramakrishnan, W. Pronobis, O. A. von Lilienfeld, K.-R. Müller and A. Tkatchenko, J. Phys. Chem. Lett. **6**, 2326 (2015).
[32]J. E. Jones, Proc. R. Soc. Lond. A **106**, 463 (1924).
[33]B. M. Axilrod and E. Teller, J. Chem. Phys. **11**, 299 (1943).
[34]Y. Muto, J. Phys. Soc. Jpn. **17**, 629 (1943).
[35]A. S. Christensen, L. A. Bratholm, F. A. Faber, and O. A. von Lilienfeld, J. Chem. Phys. **152**, 044107 (2020).
[36]M. Rupp, A. Tkatchenko, K.-R. Müller, and O. A. von Lilienfeld, Phys. Rev. Lett. **108**, 058301 (2012).
[37]B. Huang and O. A. von Lilienfeld, Nat. Chem. **12**, 945 (2020).
[38]F. A. Faber, A. S. Christensen, B. Huang, and O. A. von Lilienfeld, J. Chem. Phys. **148**, 241717 (2018).
[39]A. T. Müller, J. A. Hiss, and G. Schneider, J. Chem. Inf. Model. **58**, 472 (2018).
[40]S. Spänig and D. Heider, BioData Min. **12**, 7 (2019).
[41]V. N. Vapnik, *The Nature of Statistical Learning Theory* (Springer, 2000).
[42]S. Heinen, G. F. von Rudorff, and A. von Lilienfeld (2021). "Towards the design of chemical reactions: Machine learning barriers of competing mechanisms in reactant space," Zenodo. https://doi.org/10.5281/zenodo.4925938
[43]R. Krishnan, J. S. Binkley, R. Seeger, and J. A. Pople, J. Chem. Phys. **72**, 650 (1980).
[44]L. A. Curtiss, M. P. McGrath, J. P. Blaudeau, N. E. Davis, R. C. Binning, and L. Radom, J. Chem. Phys. **103**, 6104 (1995).
[45]A. D. McLean and G. S. Chandler, J. Chem. Phys. **72**, 5639 (1980).
[46]M. J. Frisch, J. A. Pople, and J. S. Binkley, J. Chem. Phys. **80**, 3265 (1984).
[47]T. Clark, J. Chandrasekhar, G. W. Spitznagel, and P. V. R. Schleyer, J. Comput. Chem. **4**, 294 (1983).
[48]H.-J. Werner and M. Schütz, J. Chem. Phys. **135**, 144116 (2011).
[49]C. Hampel, K. A. Peterson, and H.-J. Werner, Chem. Phys. Lett. **190**, 1 (1992).
[50]M. Schütz and F. R. Manby, Phys. Chem. Chem. Phys. **5**, 3349 (2003).
[51]T. H. Dunning, J. Chem. Phys. **90**, 1007 (1989).
[52]R. A. Kendall, T. H. Dunning, and R. J. Harrison, J. Chem. Phys. **96**, 6796 (1992).
[53]A. K. Wilson, D. E. Woon, K. A. Peterson, and T. H. Dunning, J. Chem. Phys. **110**, 7667 (1999).
[54]D. E. Woon and T. H. Dunning, J. Chem. Phys. **98**, 1358 (1993).
[55]G. Henkelman and H. Jónsson, J. Chem. Phys. **113**, 9978 (2000).
[56]O. A. von Lilienfeld, K.-R. Müller, and A. Tkatchenko, Nat. Rev. Chem. **4**, 347 (2020).
[57]O. A. von Lilienfeld, R. Ramakrishnan, M. Rupp, and A. Knoll, Int. J. Quantum Chem. **115**, 1084 (2015).
[58]B. Parsaeifard, D. S. De, A. S. Christensen, F. A. Faber, E. Kocer, S. De, J. Behler, A. von Lilienfeld, and S. Goedecker, Mach. Learn.: Sci. Technol. **2**, 015018 (2021).
[59]A. K. Rappe, C. J. Casewit, K. S. Colwell, W. A. Goddard, and W. M. Skiff, J. Am. Chem. Soc. **114**, 10024 (1992).
[60]M. F. Langer, A. Goeßmann, and M. Rupp, arXiv:2003.12081.pdf (2021).
[61]F. A. Faber, A. Lindmaa, O. A. von Lilienfeld, and R. Armiento, Phys. Rev. Lett. **117**, 135502 (2016).
[62]J. Schmidt, J. Shi, P. Borlido, L. Chen, S. Botti, and M. A. L. Marques, Chem. Mater. **29**, 5090 (2017).
[63]M. Bragato, G. F. von Rudorff, and O. A. von Lilienfeld, Chem. Sci. **11**, 11859 (2020).
[64]D. Lemm, G. von Rudorff, and O. von Lilienfeld, Nature Commun. **12**, 4468 (2021).
[65]J. Peiró-García and I. Nebot-Gil, ChemPhysChem **4**, 843 (2003).
[66]R. Q. Zhang, W. C. Lu, Y. L. Zhao, and S. T. Lee, J. Phys. Chem. B **108**, 1967 (2004).
[67]A. Shiroudi, M. S. Deleuze, and S. Canneaux, Phys. Chem. Chem. Phys. **17**, 13719 (2015).
[68]R. Ramakrishnan, P. O. Dral, M. Rupp, and O. A. von Lilienfeld, J. Chem. Theory Comput. **11**, 2087 (2015).
[69]J. Weinreich, N. J. Browning, and O. A. von Lilienfeld, J. Chem. Phys. **154**, 134113 (2021).