IOP Publishing Mach. Learn.: Sci. Technol. 3 (2022) 045005 https://doi.org/10.1088/2632-2153/ac8f1a

MACHINE LEARNING Science and Technology

OPEN ACCESS

RECEIVED
10 June 2022

REVISED
8 August 2022

ACCEPTED FOR PUBLICATION
1 September 2022

PUBLISHED
18 October 2022

PAPER

Physics-based representations for machine learning properties of chemical reactions

Puck van Gerwen[1,2], **Alberto Fabrizio**[1,3], **Matthew D Wodrich**[1,2] **and Clemence Corminboeuf**[1,2,3,*]

[1] Laboratory for Computational Molecular Design, Institute of Chemical Sciences and Engineering, École Polytechnique Fédérale de Lausanne, Lausanne, CH-1015, Switzerland

[2] National Centre for Competence in Research-Catalysis (NCCR-Catalysis), École Polytechnique Fédérale de Lausanne, Lausanne, CH-1015, Switzerland

[3] National Centre for Computational Design and Discovery of Novel Materials (MARVEL), École Polytechnique Fédérale de Lausanne, Lausanne CH-1015, Switzerland

[*] Author to whom any correspondence should be addressed.

E-mail: clemence.corminboeuf@epfl.ch

Keywords: chemical reactions, quantum machine learning, physics-based representation, reaction-based representation

Supplementary material for this article is available online

Abstract

Physics-based representations constructed using only atomic positions and nuclear charges (also known as quantum machine learning, QML) allow for the reliable and efficient inference of molecular properties from training data. Chemistry is a science rooted in chemical reactions, naturally involving multiple molecular species. Here, we extend QML's capabilities to include the prediction of reaction properties by defining *reaction representations*: representations taking as input multiple molecules participating in a reaction, each represented by their corresponding atomic charges and three-dimensional coordinates. Several reaction representations are constructed from established molecular ones and benchmarked on four datasets representative of thermodynamic or kinetic reaction properties. One of these, the Hydroform-22-TS dataset (2350 energy barriers), is introduced as part of this work. The relevant ingredients for a high-performing reaction representation are extracted and used to construct the Bond-Based Reaction Representation (B^2R^2) for the prediction of quantum-chemical properties of chemical reactions. Finally, variations of B^2R^2 with varying representation size *vs.* performance are provided.

1. Introduction

Physics-based or quantum machine learning (QML) [1–4] representations form a comprehensive class of chemical fingerprints that are inspired by fundamental laws of physics and basic laws of symmetry. These representations rely on the fact that all (static) information about a chemical system is uniquely encoded into the system-specific parameters that fix the electronic Schrödinger equation: nuclear charges (Z_I) and positions (R_I). Because QML representations are rooted in foundational laws of nature, they are extremely transferable and do not need to be adapted to each specific learning task. Given their transferability, generality, and deep connection to electronic targets, QML representations have been the forefront of machine learning applied to solve chemical problems [2–7].

Despite their conceptual and mathematical differences, existing QML representations always focus on encoding either an entire molecule ('global') or a molecule as a set of atomic environments ('local'). These representations can then accurately and efficiently predict molecular and atomic properties, such as atomisation energies [1, 8, 9], forces [10–13], potential energy surfaces [14–16], excited state properties [17], polarisabilities [18, 19] and electron densities [20, 21]. Free energies can be predicted with a Boltzmann-weighted ensemble of molecular representations [22]. While atoms and molecules are the instruments of chemistry, chemical reactions are its orchestra. The prediction of reaction properties such as reaction energies, activation barriers, changes in dipole moments, catalytic yields and turnover, are not yet routine within the framework of QML. As opposed to single molecule properties, reaction properties always include a notion of transformation: quantities change from reactants to products. These changes may be subtle, but they dictate chemistry. To be valuable within this context, a representation should capture both transformations from reactants and products, as well as differences between reactions.

Outside the domain of QML, the development of predictive models for reaction properties has been an active field for the last few years. Descriptors derived from 2D molecular graphs of reactants and products [23, 24] are the standard choice for the prediction of reaction properties that are not highly sensitive to subtle changes in three-dimensional molecular structure. An alternative is expert-selected (*ab-initio*) descriptors [25, 26], which often correlate well with reaction properties, but typically rely on a mechanistic understanding of the reaction, and are not transferable across reaction classes. In some cases, simple one-hot encoded descriptors [27, 28] perform equally well to *ab-initio* descriptors for significantly lower computational cost. Finally, representations derived from graphs in deep learning models [29, 30] have shown promising performance on several reaction properties, but are expensive to train, and tend to perform well only for large dataset sizes.

Compared to the above descriptors, QML reaction representations offer a greater degree of generality (*vide supra*). Recently [31], it has been demonstrated that reaction representations derived from molecular ones can be modified to better describe the transformational nature of chemical reactions. Inspired by this initial work, here we analyse the essential characteristics of accurate and efficient QML reaction representations in greater depth. Leveraging three key design principles, we propose the Bond-Based Reaction Representation (B^2R^2)—a specialised fingerprint for reaction property prediction.

2. Methods

2.1. Datasets

To assess the robust performance of the reaction representations, we use four datasets of chemical reactions which are categorised according to their reaction type, property and year of publication. For example, the Proparg-21-TS dataset corresponds to propargylation reactions (Proparg), was published in 2021 (21), and is labelled with barriers (TS). The only set that is not associated with barriers, the SN2-20, correspondingly is missing the -TS. The four sets are SN2-20, GDB7-20-TS, Proparg-21-TS and Hydroform-22-TS. All datasets are available alongside with the source code at https://github.com/lcmd-epfl/b2r2-reaction-rep. The data is also available separately at https://zenodo.org/record/6937747.

2.1.1. SN2-20

This dataset is adapted from the original set of reactants and transition states published as QMrxn by von Rudorff *et al* [32]. The reactants consists of (i) variations of ethane functionalised with four substituents and (ii) a nucleophile. The products consist of (iii) the corresponding substituted product and (iv) the leaving group. The lowest energy conformation of (i) and (iii), along with their corresponding energy were extracted from QMrxn. We then computed the energies of (ii) and (iv) at the MP2 [33]/6-311G(d) [34–36] using ORCA 4.0.1 [37, 38], as per the procedure in QMrxn. The eventual dataset consists of 2670 reactions with corresponding reactant and product structures, and reaction energies. While barriers were published as part of the original work, these correspond to barriers between the reactant complex and transition state, rather than isolated reactants and transition state. Correspondingly, we focus on thermodynamics as a target, and consider kinetic properties in the other three datasets.

2.1.2. GDB7-20-TS

This dataset is taken from [39]. The dataset consists of 11 961 diverse organic reactions constructed from the GDB7 dataset [40–42], with corresponding energy barriers computed at the ωB97X-D3/def2-TZVP level. Unlike in the previous deep learning model [29], we do not mix forward and backward reactions to double the dataset size. As noted by Heid and Green [30], this practice results in misleading low prediction errors if, for example, a forward reaction is in the training set and a backward reaction in the test set.

2.1.3. Proparg-21-TS

This dataset is taken from [43]. It contains 760 structures of intermediates before and after the enantioselective transition state of the propargylation of benzaldehyde, as well as the barriers computed at the B97D/TZV(2p, 2d) level. As in our previous work [31], we target the energy barriers, which could eventually further been used to derive enantiomeric excess (*e.e.*) values.

2.1.4. Hydroform-22-TS

As part of this work, we built the Hydroform-22-TS dataset. It consists of 2350 structures of intermediates before and after the alkene insertion transition state in the catalytic cycle of olefin hydroformylation [44],

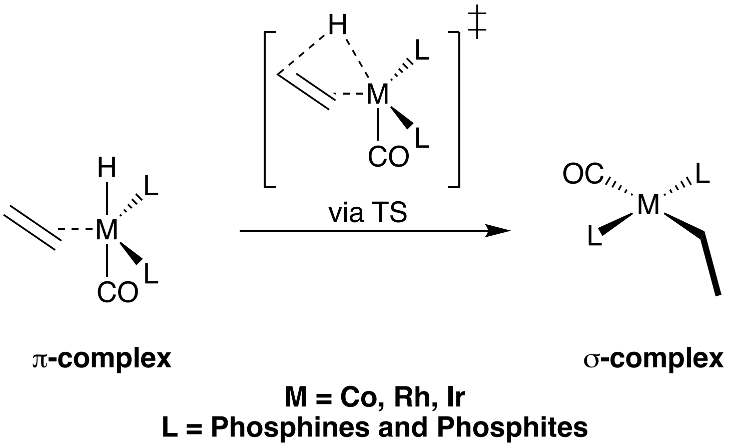

Figure 1. The Hydroform-22-TS dataset contains transition-metal complexes involved in the hydroformylation reaction of olefin. The dataset is built using combinations of three transition metals (Co, Rh, Ir) and two type of ligands (phosphine, phosphite anions).

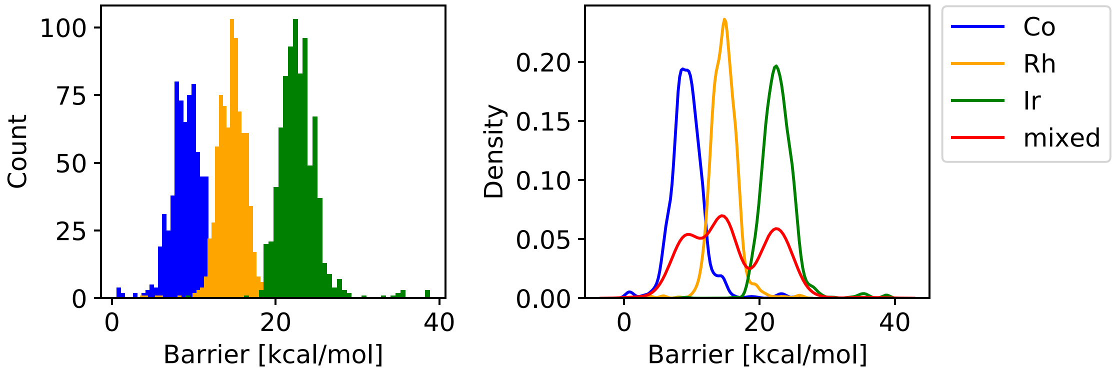

Figure 2. Left: the distribution of the activation barriers of Hydroform-22-TS for each metal. Right: probability density of the separate metal subsets and the combined set, as given by kernel density estimation (KDE).

with associated barriers. The initial set of catalysts were designed from combinations of three group 9 metals (Co, Rh, Ir) with different phosphine/phosphite ligands. The ligands were designed using 29 different substituents appended in an $R_1/R_2/R_2$ fashion to form monodentate phosphine/phosphite ligands as in the original paper [44]. The step of the hydroformylation reaction of interest, shown in figure 1, converts an olefin-bound π-complex to a σ-complex via a 1,2-insertion process. Transition state geometries were generated from a model template through functionalisation with the phosphine/phosphite ligands using the AARON program [45, 46]. The generated structures were then optimised in the gas phase at the PBE0 [47, 48]-D3(BJ) [49, 50]/def2-SVP [51] level followed by single points at the PBE0-D3(BJ)/def2-TZVP level including solvation (in benzene) using the SMD model [52] in Gaussian16 [53]. The π- and σ-intermediate complexes were optimised following the imaginary vibrational mode of the transition state, followed by geometry optimisation at the PBE0-D3(BJ)/def2-SVP level. Free energy corrections (using the def2-SVP basis set) for all species were determined using the rigid-rotor harmonic oscillator model [54] as implemented in the GoodVibes program, version 3.0.1 [55]. Default settings were used.

Finally, additional filters were applied. Some structures had missing hydrogen atoms after optimisation, which were removed. Additionally, many structures relaxed to the *cis* configuration between the CO and hydride group. The *cis* structures are not the relevant ones in the chemical process (thermodynamically higher than *trans*) but are local minima. These were removed such that only the *trans* structures remain. The final dataset is composed of cobalt (726), iridium (809), and rhodium (815) complexes for which the final barrier distribution is illustrated in figure 2.

2.1.5. General remarks

The SN2-20 and GDB7-20-TS datasets focus on small molecules (up to seven heavy atoms per molecule), whereas the Proparg-21-TS and Hydroform-22-TS sets contain larger complexes (up to 52 and 67 heavy atoms per molecule, respectively). The SN2-20 dataset, as many other examples in the literature [29, 30, 56], consists of textbook chemical reactions that are readily interpretable. However, robust reaction representations should also be capable of describing the chemistry of large molecules. While the GDB7-20-TS set is made of small molecules, it spans a broader range of chemical reactions with an associated large range of activation energies (0–200 kcal mol^{-1}), making it the most challenging dataset of the four.

2.2. Representations

2.2.1. Molecular representations

The tested molecular representations are separated into three categories: (i) Coulomb Matrix (CM) [8, 9] and Bag of Bond (BoB) [57]; (ii) SLATM [58] and FCHL19 [59, 60] and (iii) SOAP [18, 61]. The CM relies upon pairwise interactions using Coulomb potential terms [8, 9]. The BoB takes CM terms and organises the elements of the CM into atom-pairwise types 'bags' [57]: e.g. all interactions between pairs of C atoms are organised into a C–C bag. SLATM and FCHL19 append higher-order interaction terms (between triplets of atoms) with different potentials [58–60]. SOAP rather considers atoms in molecules according to their neighbouring atom density. While SOAP was introduced in the context of its kernel, the power spectrum is often treated as a representation and can be fed into any arbitrary kernel functions [3, 62].

2.2.2. Reaction representations

Representations are constructed either using only representations of reactants ($\mathbf{X}_r$), products ($\mathbf{X}_p$) or combinations of reactants and products ($\mathbf{X}_d$ and $\mathbf{X}_{rp}$). CM and BoB use an internal sorting of features by row-norm, which makes these representations non-additive. Representations of individual molecules are therefore concatenated to give reaction representations $\mathbf{X}_r$, $\mathbf{X}_p$ and $\mathbf{X}_{rp}$ where the latter is a concatenation of the first two. For the remaining representations, $\mathbf{X}_r$ is a summation of the N reactant representations:

$$\mathbf{X}_r = \sum_i^N \mathbf{X}^{(i)}. \tag{1}$$

$\mathbf{X}_p$ is a summation of the M product representations:

$$\mathbf{X}_p = \sum_j^M \mathbf{X}^{(j)} \tag{2}$$

and $\mathbf{X}_d$ is a difference in the summed product and reactant representations:

$$\mathbf{X}_d = \mathbf{X}_p - \mathbf{X}_r \tag{3}$$

where $\mathbf{X}$ denotes the molecular representation.

SOAP was generated using the `dscribe` python package [63]. The other molecular representations were generated using the `QML` python package [64]. The default parameters are used for all representations. Our B^2R^2 representation (*vide infra*) is available as part of the code repository. The most important parameter, the cut-off radius R_{cut}, is optimised for each dataset on a grid. Optimal R_{cut} values are provided in the supplementary information.

2.3. Machine learning models

In all cases, Kernel-Ridge Regression (KRR) models were used. While other models, like deep neural networks, might also be suitable, data-efficient KRR has historically dominated the physics-based machine learning landscape [3, 65] with benchmark studies [66] demonstrating its superior performance on prototypical quantum chemistry datasets. For this reason, we choose to initially benchmark our reaction representations using KRR models only.

All KRR models are trained with a Gaussian kernel K. The predicted property p for a reaction $\mathbf{X}$, either the reaction energy or barrier, is then given by:

$$p(\mathbf{X}) = \sum_{i=1}^N \alpha_i K(\mathbf{X}, \mathbf{X}') \tag{4}$$

$$K(\mathbf{X}, \mathbf{X}') = \exp\left(-\frac{||\mathbf{X}-\mathbf{X}'||_2^2}{2\sigma^2}\right) \tag{5}$$

where the coefficients α are learned from the training set:

$$\alpha = (\mathbf{K} + \lambda\mathbf{I})^{-1}\mathbf{p}^{\text{train}}. \tag{6}$$

The kernel width σ and regularisation parameter λ are optimised using five-fold cross validation. In other words, after initially shuffling the datasets, they are split into five folds, where the first 4 (80%) are used for testing, and the last (20%) for validation. The best hyperparameters are those with a minimal MAE on the test set across the five folds. Details of the hyperparameters tested and the eventual optimal values are given in the supplementary material.

For all datasets except the Hydroform-22-TS set, learning curves are generated using standard ten-fold cross validation. The last fold (10%) of the data is held out as a test set. Of the 90% available for train, increasing fractions of the data are included to generate the learning curves. The MAEs on the test set are averaged over the ten folds. Learning curves for the Hydroform-22-TS set are generated in a different way, owing to its multi-modal nature (see figure 2). The data is split into ten folds as before, where the last fold is always held out as a test set. Rather than including the first 200, 400, etc points for training, the most *diverse* 200, 400, etc points are included at a time. Farthest point sampling (FPS) is used greedily to determine the most diverse points in the training set. The out-of-sample MAEs are again averaged over the ten folds.

3. Results and discussion

3.1. From molecular to reaction representations

In order to maximise transferability, we benchmark the selected reaction representations (see section 2.2) on the four datasets covering a range of reaction types, properties and molecular sizes. The learning curves in figure 3 report the performance of one molecular representation per category (BoB, SLATM and SOAP) but the remaining ones (CM and FCHL19) can be found in the supplementary material.

With the kernel models used herein, thermodynamic properties can be learned to higher accuracy than kinetic properties as illustrated by the lowest prediction errors obtained for the SN2-20 set. This result is expected, as the reaction energy (thermodynamic) depends on the isolated reactant and product structures and not on the reaction barrier. In contrast, other test sets with energy barriers as a target (e.g. Proparg-21-TS, Hydroform-22-TS) are more difficult to learn, as they depend on transition state structures [31]. As noted by previous authors [29, 30], the large errors observed for the GDB7-20-TS dataset arise from the inherent challenging nature of this set caused by the significant spread in the target property (0–200 kcal mol^{-1}). While pre-training combined with deep neural network architectures can offer improvements for reaction tasks [30], such a comparison is beyond the scope of this work dedicated to physics-based reaction representations.

In the following analysis, we identify three key design principles for an effective reaction representation: first, the difference between products and reactants should be meaningful (i.e. the representation should be additive). Second, emphasis should be placed on the pairwise (two-body) interactions. Third, distinct pairwise interactions should be distinguished. Each of these ingredients and their motivation are discussed in the next sections.

3.1.1. Representations of meaningful differences

In figure 3, it is evident that representations including both reactants and products (i.e. $\mathbf{X}_{rp}$ and $\mathbf{X}_d$) systematically outperform the equivalent representations based on the reactants or products only (i.e. $\mathbf{X}_r$ and $\mathbf{X}_p$). For most datasets, the slope of $\mathbf{X}_{rp}$ and $\mathbf{X}_d$ models are steeper, leading to an improved performance over $\mathbf{X}_r$ and $\mathbf{X}_p$. For the Proparg-21-TS set, it is instead the offset of $\mathbf{X}_d$ that improves the predictions. Previous efforts in the literature suggested that a single reactant might be sufficient to learn reaction properties [56, 67]. This was motivated for example by Hammond's postulate [56, 68], where the transition state (and therefore reaction barrier) should closely resemble either the reactant or product. In general, a description of both sides of the chemical equation is fundamental, as neither the reactant nor the product consistently resemble the transition state. Additionally, in a given reaction dataset, the same set of reactants may have multiple products. A representation based on reactants alone therefore leads to a direct violation of the representation uniqueness and injectivity. Therefore a transferable reaction representation must describe all of the participating molecules in the reaction.

As mentioned in section 2.2.2, some representations, like CM and BoB, use an internal sorting procedure which renders the notion of difference between representations meaningless. Representations based on additive potentials (FCHL and SLATM) or densities (SOAP) allow for a suitable notion of difference. Figure 3 illustrates that difference-based representations $\mathbf{X}_d$ outperform those that concatenate reactants and products $\mathbf{X}_{rp}$ in the prediction of reaction properties. The additivity criterion is well-justified from physical laws, as most reaction properties are algebraically additive. For instance, reaction energies are defined as:

$$\Delta E_r = \sum_{\text{products}} E_{\text{products}} - \sum_{\text{reactants}} E_{\text{reactants}}. \tag{7}$$

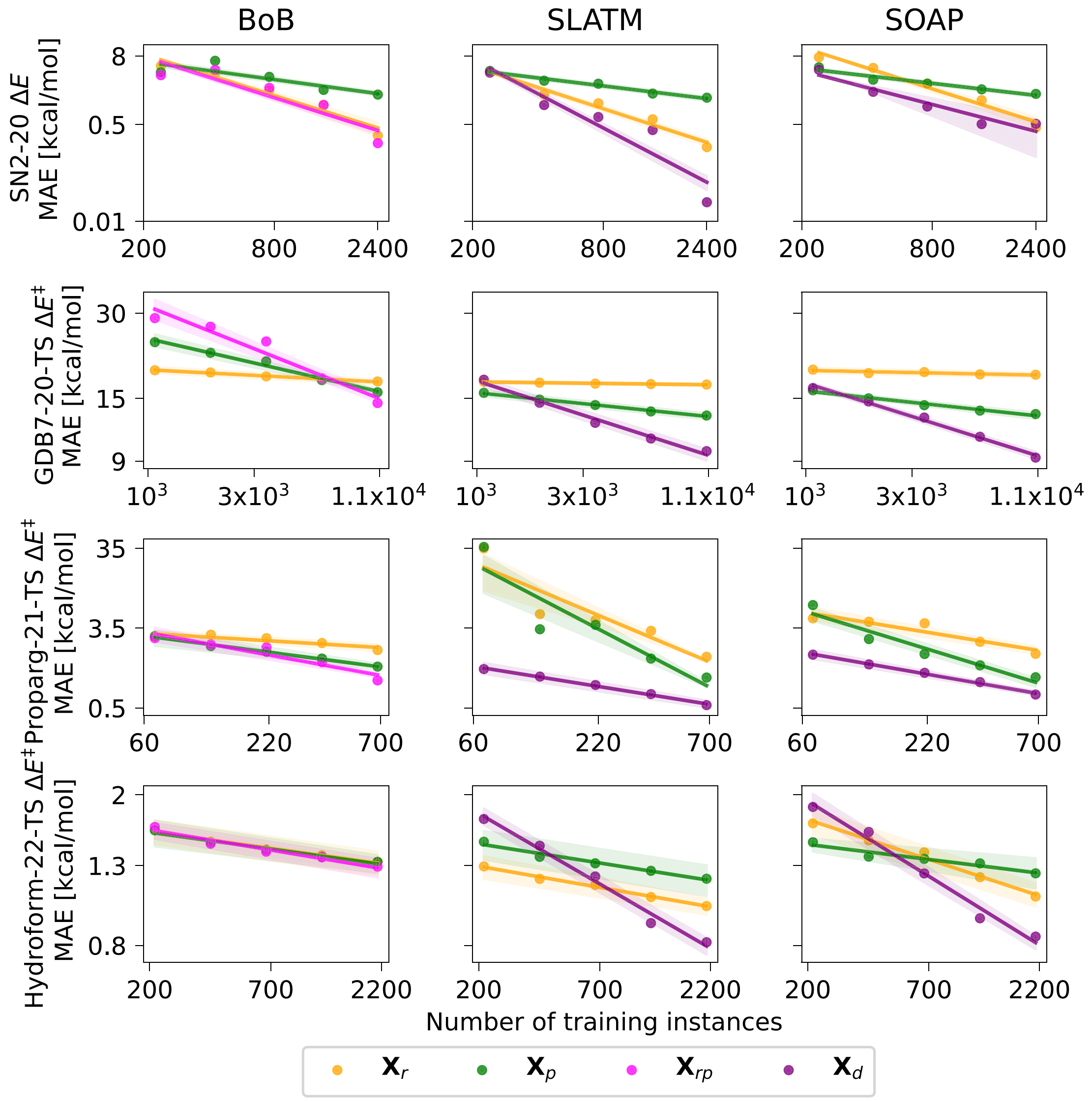

Figure 3. Learning curves are shown for reaction representations constructed from molecular representations of reactants only $\mathbf{X}_r$, products only $\mathbf{X}_p$, concatenations $\mathbf{X}_{rp}$ or a difference $\mathbf{X}_d$.

If a representation is additive, then we can write similarly:

$$\mathbf{X}_d = \sum_{\text{products}} \mathbf{X}_{\text{products}} - \sum_{\text{reactants}} \mathbf{X}_{\text{reactants}}. \tag{8}$$

Assuming that molecular representations contain sufficient information to regress molecular energies, then the algebraic sum of the molecular fingerprints should also correlate well with the algebraic sum of the energies. Reaction barriers are not an explicit algebraic difference like reaction energies. However, transition state structures often can be reasonably approximated as a (weighted) interpolation between reactant and product structures, analogous to interpolation in the Nudged Elastic Band [69] method. This allows $\mathbf{X}_d$ to resemble $\mathbf{X}_{TS}$:

$$\mathbf{X}_d \approx \mathbf{X}_{TS} = a\mathbf{X}_p - b\mathbf{X}_r. \tag{9}$$

In the simplest case, $a = b = 0.5$. In previous work [31], it was found that adapting $a \neq b \neq 0.5$ did not improve the performance of the reaction representation. Taking the simplest interpolation allows for a consistent definition of $\mathbf{X}_d$ which performs well for thermodynamic and kinetic property prediction. Additionally, $\mathbf{X}_d$ has the same dimensions as the single-molecule representation equivalent $\mathbf{X}$, whereas $\mathbf{X}_{rp}$ doubles the number of features. $\mathbf{X}_d$ thereby enhances the predictive performance of reaction properties in the same number of features as single-molecule features.

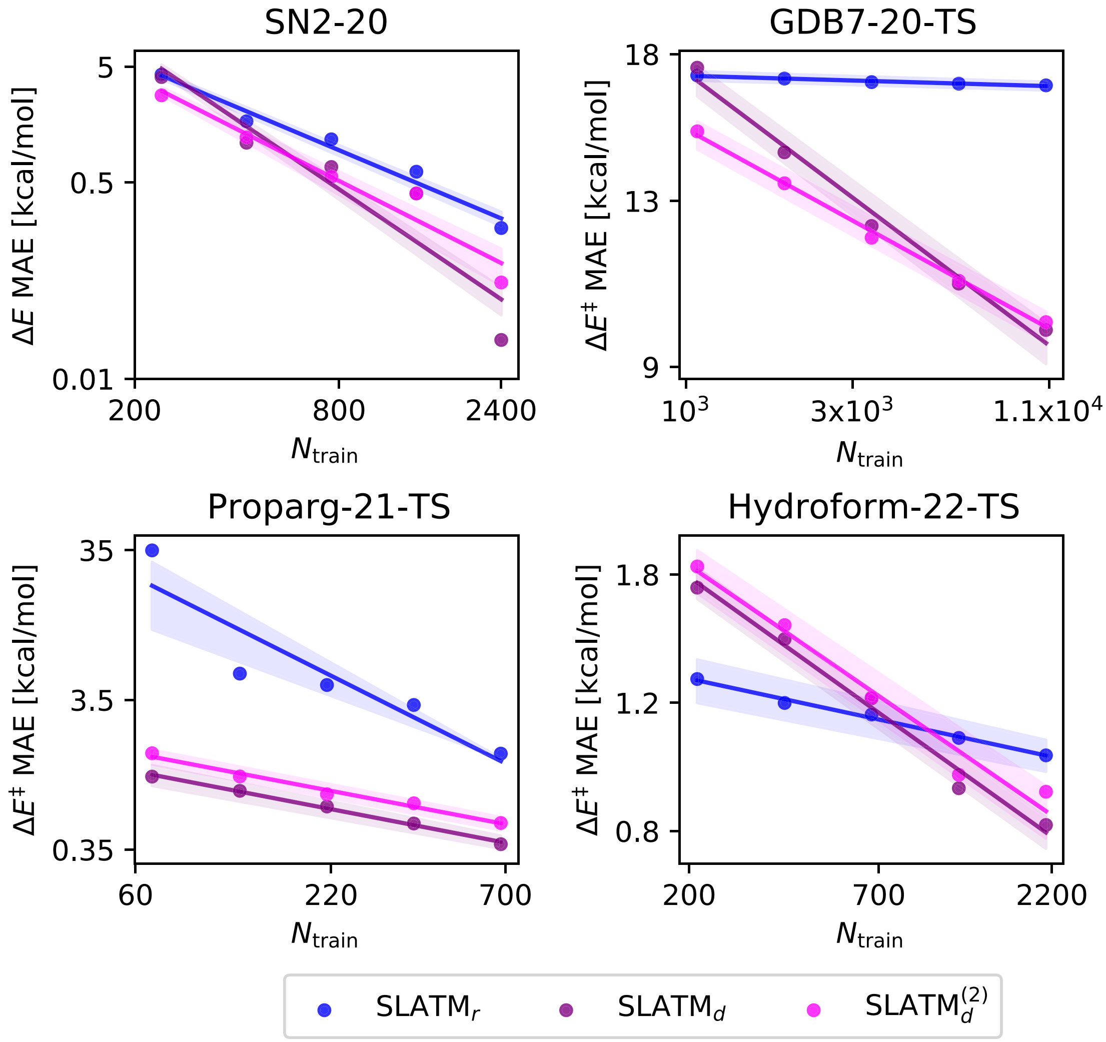

Figure 4. Learning curves are shown for full SLATM_d and the two-body only version $\text{SLATM}_d^{(2)}$. SLATM_r is included for reference.

3.1.2. Importance of two-body features

While representations incorporating higher-order terms (SLATM, SOAP) outperform those with two-body interactions only (BoB) in figure 3, it is known that the two-body interactions are critical, and that higher order interactions only modestly improve upon the initial necessary terms [3]. For chemical reactions in particular, the predominant interactions are bonds breaking and forming. We hypothesised that only two-body interactions were necessary for good predictive performance of reaction properties. To this aim, we tested the two-body terms of SLATM ($\text{SLATM}_d^{(2)}$) and compared it to full SLATM_d. Figure 4 illustrates that $\text{SLATM}_d^{(2)}$ indeed performs almost equivalently to SLATM_d across the four datasets. The eventual MAE of $\text{SLATM}_d^{(2)}$ and SLATM_d models is very close. For the smaller molecule datasets (SN2-20 and GDB7-20-TS), the slope of $\text{SLATM}_d^{(2)}$ is a little more shallow, but not significantly so. We note that SLATM places a higher weight on short-range interactions by describing pairwise interactions using the London potential $1/R^6$ [70] which diminishes the description of longer-range interactions, which might be needed to allow $\text{SLATM}_d^{(2)}$ to match SLATM_d in all cases. Removing or reducing this higher weight on shorter-range interactions might allow for a two-body only representation to be sufficiently accurate.

3.1.3. Separation of distinct two-body features

The two-body terms of SLATM_d are separated into pairwise bags of element types (C–C, C–N, C–O, etc). This separation allows for the isolation of the description of important bonds that are involved in a reaction. In figure 5, we leverage the conceptual simplicity of an S_N2 reaction to look at SLATM_r, SLATM_p and SLATM_d two-body features. A C–F bond is broken and C–H bond is formed, while all other bonds remain unchanged during the reaction. The pairwise interaction bags are organised into a fixed order, such that the same bond type (e.g. C–H) is at the same position in the reactant and product representation vector. The

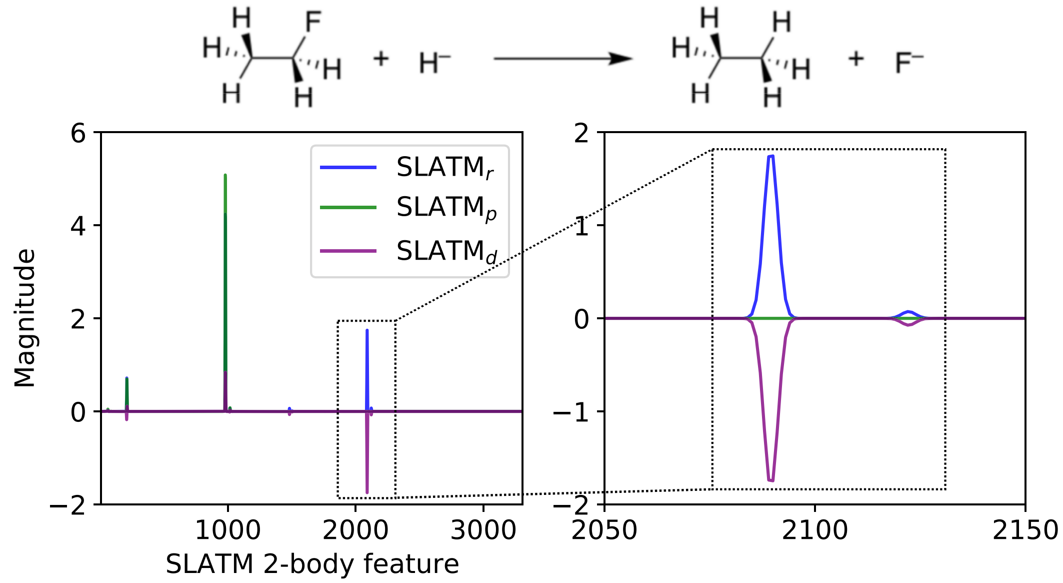

Figure 5. Two-body features of SLATM$_r$, SLATM$_p$ and SLATM$_d$ for a simple S$_N$2 reaction.

ultimate difference captured in SLATM$_d$ then highlights the breaking of the C–F bond and creation of a C–H bond, with a large amplitude in these bags. Further moderate structural rearrangements in the reactant and product are highlighted in H–H, C–C, and H–F feature bags.

The pairwise bagging approach provides clearly separated and interpretable features, but has relatively inefficient computational scaling at $O(n^2)$, where n is the number of unique elements in the dataset. We present the B^2R^2 representation first with the same scaling, and later propose alternative approaches: at $O(n)$, still separating two-body features, and at $O(1)$, reducing feature separation and slightly suffering in performance accordingly.

3.2. The bond-based reaction representation

SLATM$_d$ is a robust reaction representation as a consequence of three key ingredients: (i) a meaningful difference, (ii) an emphasis on two-body interactions and (iii) separation of the relevant two-body features. These ingredients enable SLATM$_d$ to capture and amplify changes in individual bonding environments, which are the ultimate drivers of reaction properties. Relying upon these concepts, we present the Bond-Based Reaction Representation (B^2R^2), a dedicated reaction representation built on the notion of difference in pairwise interactions between reactants and products. There are three variations depending on the bagging strategy: (i) the canonical variation, with pairwise bags ($O(n^2)$); (ii) a linear variation ($O(n)$) and (iii) a constant-size variation ($O(1)$). In all cases, the B^2R^2 is significantly smaller than SLATM or other higher-order potential based representations which scale as $O(n^3)$.

3.2.1. Canonical B^2R^2

The canonical variant of B^2R^2 employs the same bagging strategy as SLATM: that is, by pairwise element types. To describe an interaction between a pair of atoms I and J (in a reactant or product molecule), with nuclear charges Z_I and Z_J respectively, only the distance between the atoms R_{IJ} is needed, since the element types are encoded in the bag. Here, we use simple Gaussian functions to encode pairwise interactions, which we found to offer good predictive accuracy without the need for physically-informed potential terms (as is typical in QML representations [8, 9, 71, 72]). As outlined in section 3.1, it is the additivity of the functions employed that is responsible for their performance. We choose Gaussian functions centred on the bond between two atoms ($\mu = R_{IJ}/2$), with a standard deviation that we found to perform well on average across the datasets tested ($\sigma = R_{IJ}/8$).

For each bag, all $p \in P$ bond distances R_p in the products and $r \in R$ bond distances R_r in the reactants are collected and used to construct a difference:

$$B^2R^2_{\mathrm{bag}} = \sqrt{\frac{32}{\pi}} \left[\sum_p^P \frac{1}{R_p} \exp\left(-\frac{32(x - R_p/2)^2}{R_p^2} \right) - \sum_r^R \frac{1}{R_r} \exp\left(-\frac{32(x - R_r/2)^2}{R_r^2} \right) \right]. \qquad (10)$$

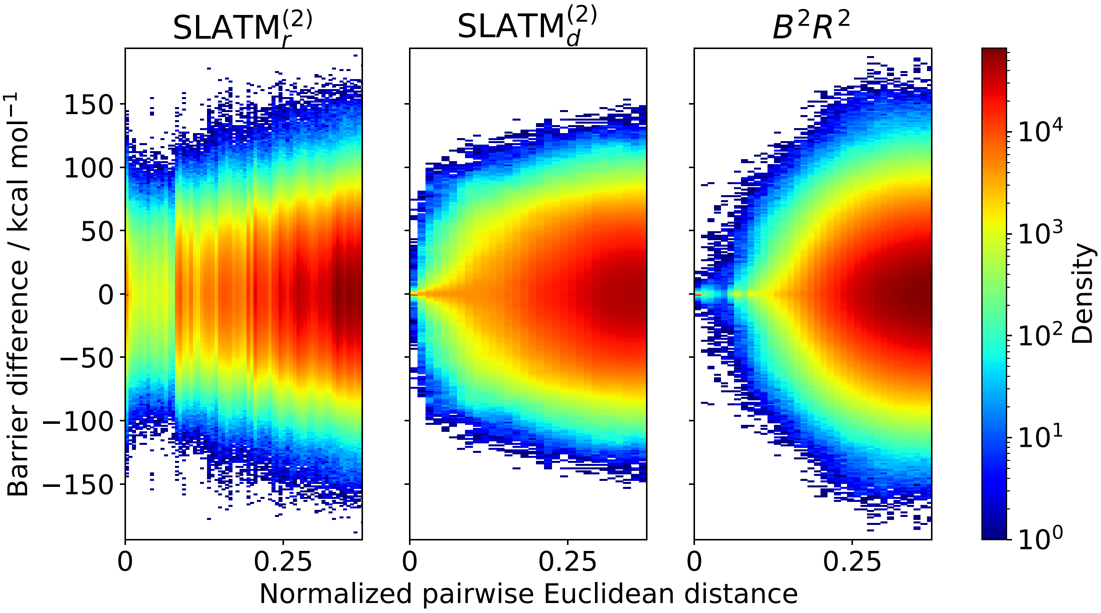

Figure 6. Dissimilarity plots, i.e. difference in target values (barriers) vs the normalised Euclidean distance are shown for SLATM$_r^{(2)}$, SLATM$_d^{(2)}$ and B^2R^2 on the GDB7-20-TS dataset.

To emphasize the relevant bond distances, only $R_{IJ} < R_{\text{cut}}$ are included in the representation. R_{cut} is optimised for each dataset (see supplementary material), but typically lies between 3 and 5 Å. This suggests that most information needed to predict reaction properties is local, i.e. within the range of a few bond lengths. Depending on the dataset, in some cases longer-range interactions are needed to describe the structural rearrangements associated with bond-breaking and -making. Unlike in SLATM, the shorter-range interactions are not weighted. Instead, R_{cut} accounts for the nature of bonding interactions per dataset. For example, if longer-range interactions dominate, a longer R_{cut} would be optimal.

Figure 6 illustrates B^2R^2's desirable properties based on diagnostics introduced in our previous work [31]. Dissimilarity plots were introduced to determine whether a representation correlates with the target property. These plots map the pairwise Euclidean distance between representations against the pairwise difference between target values. A suitable representation should recognise that a small distance between representations should correspond to a small difference in the target property. Plots are constructed for SLATM$_r^{(2)}$, SLATM$_d^{(2)}$ and B^2R^2 on the GDB7-20-TS dataset (plots for other sets are given in the supplementary material). Both SLATM$_d^{(2)}$ and B^2R^2 exhibit ideal behaviour, whereas SLATM$_r^{(2)}$ results in a noisy dissimilarity plot which actually increases the distance between representations around zero barrier difference.

Correspondingly, in figure 7, except for minor performance differences, SLATM$_d^{(2)}$ and B^2R^2 perform similarly well, while SLATM$_r^{(2)}$ fails to accurately predict the target property. Except for the SN2-20 set, the B^2R^2 curves have the same slope as SLATM$_d$, or even steeper (Proparg-21-TS). While B^2R^2 does not offer an overall improvement in performance vs. SLATM$_d$, it does encapsulate the same critical information in far fewer features (*vide infra*).

3.2.2. Linear- and constant-size versions

While pairwise bags are the canonical choice, and they allow for interpretable features, they also scale as $O(n^2)$. While less than SLATM's $O(n^3)$, other bagging strategies reduce the scaling.

To reduce the scaling to $O(n)$, $B^2R^2_l$ uses bags constructed using elements rather than pairwise elements. In other words, bags contain information about Z_I only rather than about Z_I and Z_J. As a consequence, information about Z_J should be included in the representation. To this aim, we use a skew-normal distribution rather than a Gaussian distribution to describe pairwise interaction terms:

$$B^2R^2_{l,\text{bag}} = 16\left[\sum_p^P \frac{Z_J}{R_p}\phi\left(\frac{x-R_p/2}{R_p/8}\right)\Phi\left(Z_J\frac{x-R_p/2}{R_p/8}\right) - \sum_r^R \frac{Z_J}{R_r}\phi\left(\frac{x-R_r/2}{R_r/8}\right)\Phi\left(Z_J\frac{x-R_r/2}{R_r/8}\right)\right] \quad (11)$$

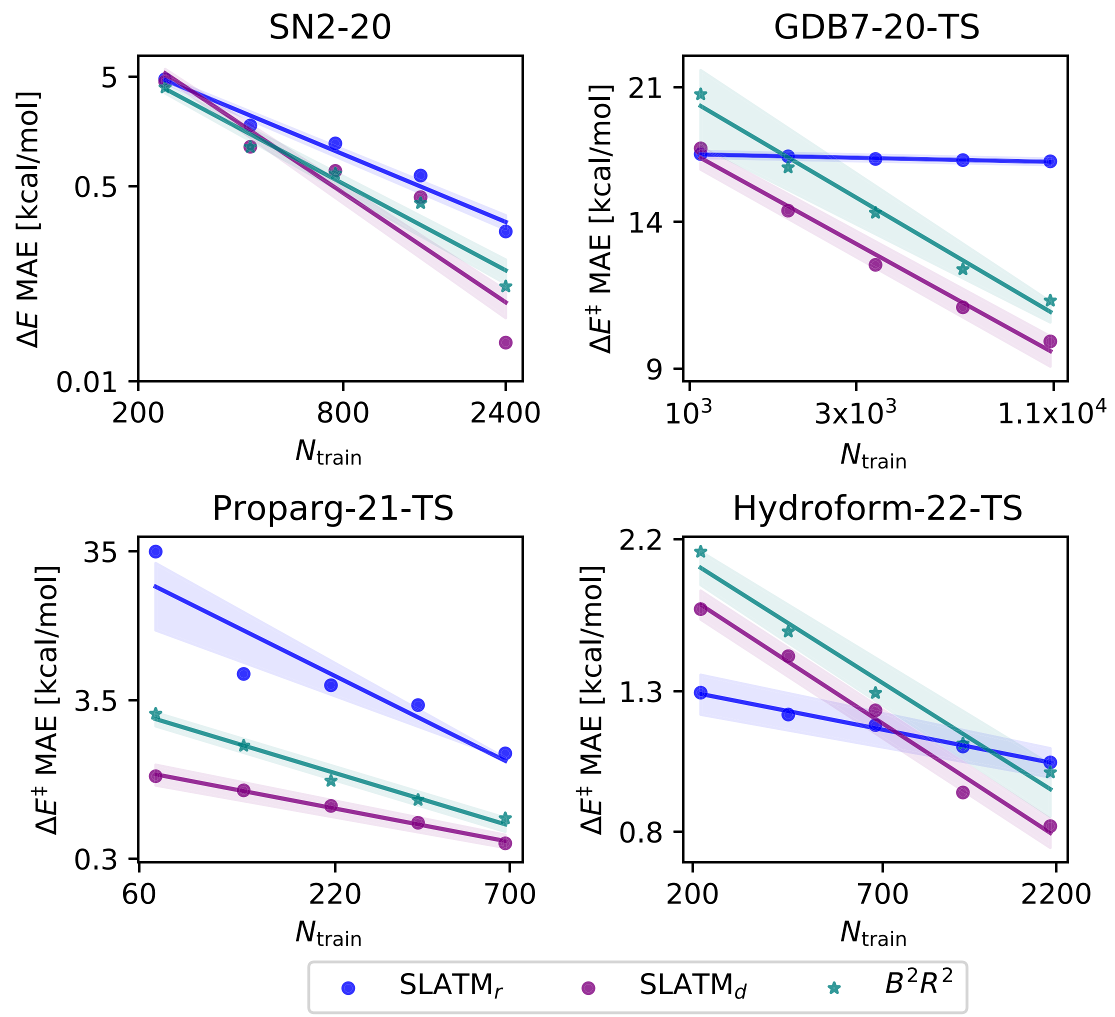

Figure 7. Learning curves are presented for our reaction representation B^2R^2. SLATM_r and SLATM_d are included for reference.

where $\phi(x)$ is the standard normal probability density function:

$$\phi(x) = \frac{1}{\sqrt{2\pi}} \exp -\frac{x^2}{2} \tag{12}$$

and $\Phi(x)$ is the cumulative distribution function:

$$\Phi(x) = \frac{1}{2}\left[1 + \mathrm{erf}\left(\frac{x}{\sqrt{2}}\right)\right]. \tag{13}$$

The degree of skewness is modulated by the second nuclear charge Z_J, which also amplifies the magnitude of the function. A similar form is used for the third no bags $B^2R_n^2$ representation, where interaction terms are no longer separated by pairs of elements (as in B^2R^2) or elements (as in $B^2R_l^2$), instead collected in the same fixed-size vector. Since there is no bagging to isolate Z_I, the magnitude is modulated by Z_I rather than Z_J as in $B^2R_l^2$:

$$B^2R_n^2 = 16\left[\sum_p^P \frac{Z_I}{R_p}\phi\left(\frac{x - R_p/2}{R_p/8}\right)\Phi\left(Z_J\frac{x - R_p/2}{R_p/8}\right) - \sum_r^R \frac{Z_I}{R_r}\phi\left(\frac{x - R_r/2}{R_r/8}\right)\Phi\left(Z_J\frac{x - R_r/2}{R_r/8}\right)\right]. \tag{14}$$

3.2.3. Evaluation and performance

The features of $B^2R_n^2$ correspond to equation (14) evaluated for each grid point equally spaced between 0 Å and R_{cut}. The same grid-spacing of 0.03 Å is used in all B^2R^2 variations, such that the length of $B^2R_n^2$ is determined only by the cut-off R_{cut}. Similarly, for B^2R^2 and $B^2R_l^2$, the features of each bag are evaluated using equations (10) and (11) respectively across an equally spaced grid. Their size is determined by both R_{cut} and

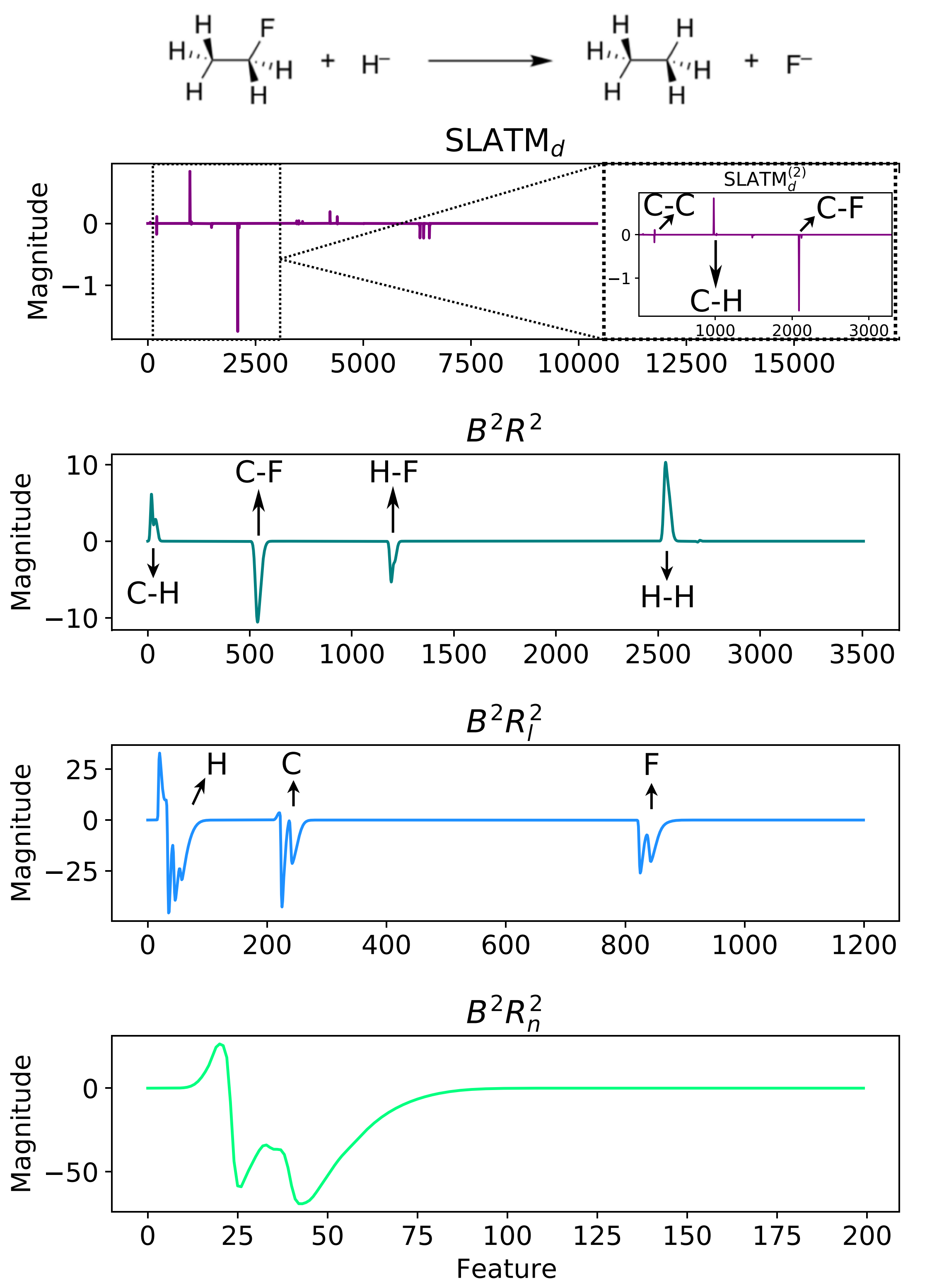

Figure 8. The features of the three variations of B^2R^2 (pairwise B^2R^2, linear $B^2R_l^2$ and constant-size $B^2R_n^2$) are shown for a simple S_N2 reaction. SLATM$_d$ is included for reference.

the number of unique element types which determine the number of bags. For each dataset and variation of B^2R^2, R_{cut} is optimised on a grid. Optimal R_{cut} values are provided in the supplementary material.

Features of the three B^2R^2 variants are illustrated for an S_N2 reaction example in figure 8. Akin to SLATM$_d^{(2)}$, B^2R^2 emphasises bonds breaking and forming in pairwise bags. $B^2R_l^2$ encodes the same information, but collects it in element bags, where now the relevant features are in the H, C and F bags. Finally, $B^2R_n^2$ no longer separates the features into bags, and simply collects all pairwise interactions into the

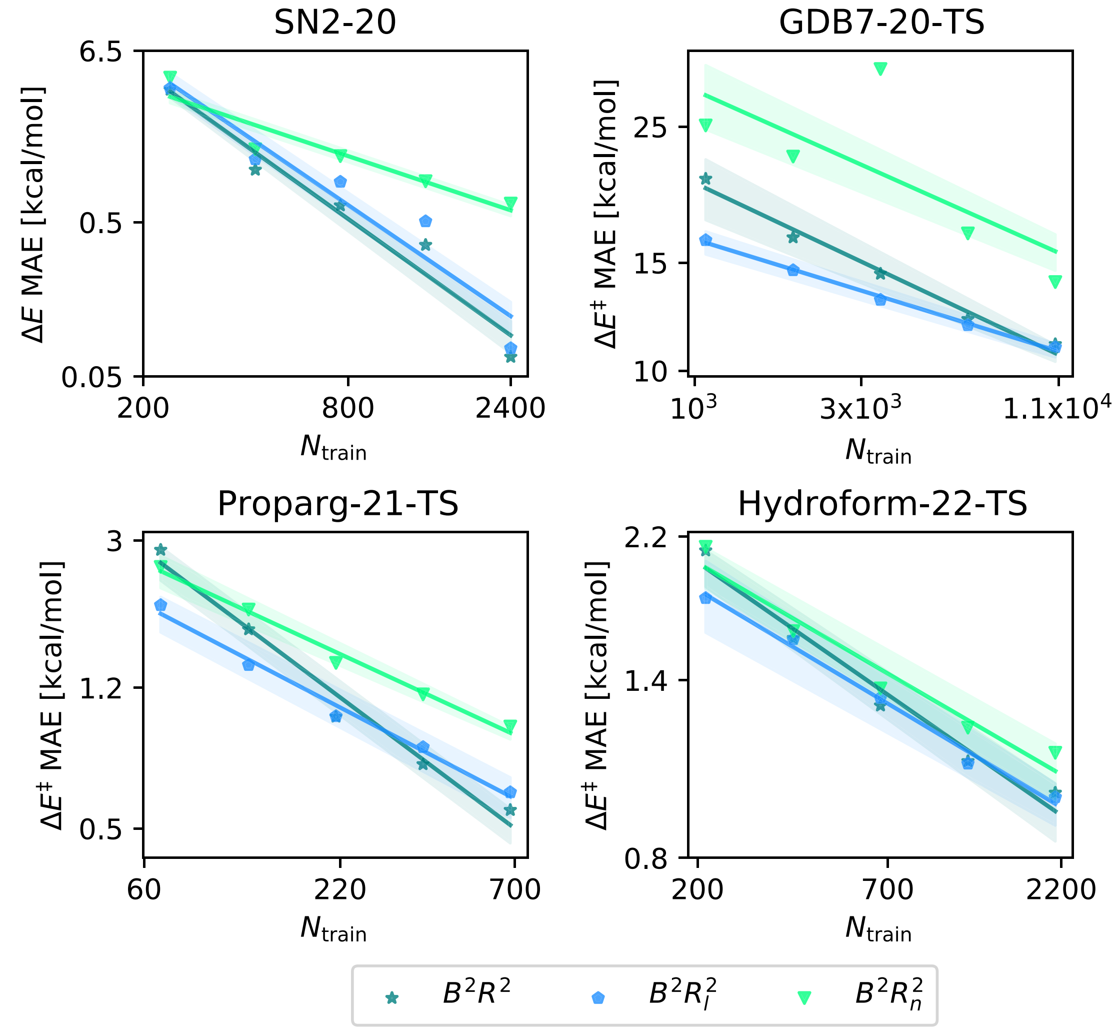

Figure 9. Learning curves are presented for the three variations of B^2R^2: pairwise B^2R^2, linear $B^2R^2_l$ and constant-size $B^2R^2_n$.

same feature vector. Moving between variations subsequently reduces the representation size, all of which are significantly smaller than the baseline SLATM$_d$.

Figure 9 compares the predictive accuracy of the three B^2R^2 variants. Interestingly, $B^2R^2_l$ achieves nearly identical performance to B^2R^2. Except for the GDB7-20-TS and Proparg-21-TS sets, the slopes are also equally steep. This suggests that incorporating information about Z_J in the functional form of the interaction terms, and bagging only by Z_I, is as effective as incorporating information about both Z_I and Z_J. A similar strategy is employed in the design of FCHL19 [10], which uses element bags rather than pairwise bags, and typically achieves similar performance to SLATM in the prediction of molecular properties. FCHL19 does not construct a single function for each pairwise interaction, but rather, in the spirit of ACSF [73], a set of radial basis functions for each unique element type. $B^2R^2_l$ maintains the single function concept of SLATM while employing an intelligent bagging strategy similar to that of FCHL19. The $B^2R^2_l$ results thus demonstrates the relevance of this approach for reaction properties.

Finally, $B^2R^2_n$ (variant with no bags) achieves surprisingly good predictive capabilities. As illustrated for an S_N2 example in figure 8, features that are otherwise separated into four (B^2R^2) or three ($B^2R^2_l$) distinct bags now overlap. Nevertheless, the overall reduction in performance is not drastic, especially for the Hydroform-22-TS and Proparg-21-TS sets. Both sets correspond to larger molecules containing a more diverse set of chemical elements than those in SN2-20 and GDB7-20-TS. As a consequence of the diversity in chemical elements, there are a larger range in bond lengths, likely preventing significant overlap in the same features and allowing for the effects of different pairwise interactions to be distinguished. For GDB7-20-TS, $B^2R^2_n$ exhibits unusual behaviour, with increased out-of-sample MAE for intermediate training set sizes. Since this is the most challenging dataset, $B^2R^2_n$ is not the most suitable choice here. In any case, such oscillations at intermediate training set sizes can likely be corrected with FPS learning curves as done for the Hydroform-22-TS set.

In practice, all three variations of B^2R^2 are available but we recommend the users to default to $B^2R_l^2$ for a compromise in representation size and predictive capability. Overall, the B^2R^2 series emphasize the necessary attributes for an effective reaction representation and constitutes a first step towards exploring more sophisticated functional forms. We note that the B^2R^2 representations might not perform well in all cases. For example, they do not naturally encode symmetry of reaction energies: a model trained on a dataset of forward reactions with their corresponding reaction energies will not naturally predict the same reaction energy with opposing sign for a dataset of backward reactions. Additionally, while a trained model tested on reactions consisting of identical reactants and products should recognise that the reaction energy should be zero, it will not necessarily recognise that the barrier should not be. Previous authors [29, 30] addressed such issues by including both forward and backward reactions in the training set. However, it might be feasible to encode such symmetries in the representation directly.

4. Conclusion

We systematically explore the construction of reaction representations from existing molecular ones, and benchmark their performance on four illustrative datasets of chemical reaction properties. One of these, the Hydroform-22-TS (2350 hydroformylation reaction barriers), is introduced as part of this work. Key design principles for a high-performing, transferable reaction representation are extracted: (i) a meaningful notion of difference between products and reactants (i.e. additivity); (ii) an emphasis on bonding (pairwise) interactions and (iii) effective separation of the description of these bonding interactions. We use these principles to design the Bond-Based Reaction Representation (B^2R^2) and illustrate its competitive performance. Finally, strategies are proposed to manipulate the size of the representation while maintaining excellent predictive capabilities. We expect these findings will expand QML from a molecule-focused discipline to a reaction-focused one.

Data availability statement

The data that support the findings of this study are openly available at the following URL/DOI: https://doi.org/10.5281/zenodo.6627913.

Acknowledgments

The authors are grateful to the EPFL for financial support. P V G, M D W and C C acknowledge the National Centre of Competence in Research (NCCR) 'Sustainable chemical process through catalysis (Catalysis)' of the Swiss National Science Foundation (SNSF, Grant No. 180544) for financial support. A F and C C acknowledge the NCCR 'Materials' Revolution: Computational Design and Discovery of Novel Materials (MARVEL)' for financial support. The authors thank Rubén Laplaza, Raimon Fabregat and Philippe Schwaller for helpful discussions.

ORCID iDs

Puck van Gerwen https://orcid.org/0000-0002-7992-5529
Alberto Fabrizio https://orcid.org/0000-0002-4440-3149
Matthew D Wodrich https://orcid.org/0000-0002-6006-671X
Clemence Corminboeuf https://orcid.org/0000-0001-7993-2879

References

[1] von Lilienfeld O A 2018 Quantum machine learning in chemical compound space *Angew. Chem., Int. Ed.* **57** 4164–9

[2] Huang B and von Lilienfeld O A 2021 *Ab initio* machine learning in chemical compound space *Chem. Rev.* **121** 10001–36

[3] Musil F, Grisafi A, Bartok A P, Ortner C, Csanyi G and Ceriotti M 2021 Physics-inspired structural representations for molecules and materials *Chem. Rev.* **121** 9759–815

[4] Langer M F, Goessmann A and Rupp M 2022 Representations of molecules and materials for interpolation of quantum-mechanical simulations via machine learning *npj Comput. Mater.* **8** 41

[5] Kulik H *et al* 2022 Roadmap on machine learning in electronic structure *Electron. Struct.* **4** 023004

[6] Deringer V L, Bartók A P, Bernstein N, Wilkins D M, Ceriotti M and Csányi G 2021 Gaussian process regression for materials and molecules *Chem. Rev.* **121** 10073–141

[7] Aspuru-Guzik A, Lindh R and Reiher M 2018 The matter simulation (r)evolution *ACS Cent. Sci.* **4** 144–52

[8] Rupp M, Tkatchenko A, Müller K-R and von Lilienfeld O A 2012 Fast and accurate modeling of molecular atomization energies with machine learning *Phys. Rev. Lett.* **108** 058301

[9] Rupp M, Ramakrishnan R and von Lilienfeld O A 2015 Machine learning for quantum mechanical properties of atoms in molecules *J. Phys. Chem. Lett.* **6** 3309–13

[10] Christensen A S, Bratholm L A, Faber F A and von Lilienfeld O A 2020 FCHL revisited: faster and more accurate quantum machine learning *J. Chem. Phys.* **152** 044107

[11] Li Z, Kermode J R and De Vita A 2015 Molecular dynamics with on-the-fly machine learning of quantum-mechanical forces *Phys. Rev. Lett.* **114** 096405

[12] Chmiela S, Tkatchenko A, Sauceda H E, Poltavsky I, Schütt K T and Müller K-R 2017 Machine learning of accurate energy-conserving molecular force fields *Sci. Adv.* **3** e1603015

[13] Chmiela S, Sauceda H E, Müller K-R and Tkatchenko A 2018 Towards exact molecular dynamics simulations with machine-learned force fields *Nat. Commun.* **9** 3887

[14] Bartók A P, Payne M C, Kondor R and Csányi G 2010 Gaussian approximation potentials: the accuracy of quantum mechanics, without the electrons *Phys. Rev. Lett.* **104** 136403

[15] Behler J 2017 First principles neural network potentials for reactive simulations of large molecular and condensed systems *Angew. Chem., Int. Ed.* **56** 12828–40

[16] Smith J S, Nebgen B, Lubbers N, Isayev O and Roitberg A E 2018 Less is more: sampling chemical space with active learning *J. Chem. Phys.* **148** 241733

[17] Westermayr J and Marquetand P 2020 Machine learning for electronically excited states of molecules *Chem. Rev.* **121** 9873–926

[18] Grisafi A, Wilkins D M, Csányi G and Ceriotti M 2018 Symmetry-adapted machine learning for tensorial properties of atomistic systems *Phys. Rev. Lett.* **120** 036002

[19] Wilkins D M, Grisafi A, Yang Y, Lao K U, DiStasio R A and Ceriotti M 2019 Accurate molecular polarizabilities with coupled cluster theory and machine learning *Proc. Natl Acad. Sci.* **116** 3401–6

[20] Brockherde F, Vogt L, Li L, Tuckerman M E, Burke K and Müller K-R 2017 Bypassing the Kohn–Sham equations with machine learning *Nat. Commun.* **8** 1–10

[21] Grisafi A, Fabrizio A, Meyer B, Wilkins D M, Corminboeuf C and Ceriotti M 2018 Transferable machine-learning model of the electron density *ACS Cent. Sci.* **5** 57–64

[22] Weinreich J, Browning N J and von Lilienfeld O A 2021 Machine learning of free energies in chemical compound space using ensemble representations: reaching experimental uncertainty for solvation *J. Chem. Phys.* **154** 134113

[23] Sandfort F, Strieth-Kalthoff F, Kühnemund M, Beecks C and Glorius F 2020 A structure-based platform for predicting chemical reactivity *Chem* **6** 1379–90

[24] Probst D, Schwaller P and Reymond J-L 2022 Reaction classification and yield prediction using the differential reaction fingerprint DRFP *Digit. Discovery* **1** 91–97

[25] Ahneman D T, Estrada J G, Lin S, Dreher S D and Doyle A G 2018 Predicting reaction performance in C–N cross-coupling using machine learning *Science* **360** 186–90

[26] Zahrt A F, Henle J J, Rose B T, Wang Y, Darrow W T and Denmark S E 2019 Prediction of higher-selectivity catalysts by computer-driven workflow and machine learning *Science* **363** eaau5631

[27] Chuang K V and Keiser M J 2018 Comment on "Predicting reaction performance in C–N cross-coupling using machine learning" *Science* **362** eaat8603

[28] Granda J M, Donina L, Dragone V, Long D-L and Cronin L 2018 Controlling an organic synthesis robot with machine learning to search for new reactivity *Nature* **559** 377–81

[29] Grambow C A, Pattanaik L and Green W H 2020 Deep learning of activation energies *J. Phys. Chem. Lett.* **11** 2992–7

[30] Heid E and Green W H 2022 Machine learning of reaction properties via learned representations of the condensed graph of reaction *J. Chem. Inf. Model.* **62** 2101–10

[31] Gallarati S, Fabregat R, Laplaza R, Bhattacharjee S, Wodrich M D and Corminboeuf C 2021 Reaction-based machine learning representations for predicting the enantioselectivity of organocatalysts *Chem. Sci.* **12** 6879–89

[32] von Rudorff G F, Heinen S N, Bragato M and von Lilienfeld A O 2020 Thousands of reactants and transition states for competing E2 and S_N2 reactions *Mach. Learn.: Sci. Technol.* **1** 045026

[33] Frisch M J, Head-Gordon M and Pople J A 1990 A direct MP2 gradient method *Chem. Phys. Lett.* **166** 275–80

[34] Krishnan R, Binkley J S, Seeger R and Pople J A 1980 Self-consistent molecular orbital methods. XX. A basis set for correlated wave functions *J. Chem. Phys.* **72** 650–4

[35] McLean A D and Chandler G S 1980 Contracted Gaussian basis sets for molecular calculations. I. Second row atoms, $z = 11$–18 *J. Chem. Phys.* **72** 5639–48

[36] Frisch M J, Pople J A and Binkley J S 1984 Self-consistent molecular orbital methods 25. Supplementary functions for Gaussian basis sets *J. Chem. Phys.* **80** 3265–9

[37] Neese F 2012 The ORCA program system *Wiley Interdiscip. Rev.-Comput. Mol. Sci.* **2** 73–78

[38] Neese F 2018 Software update: the ORCA program system, version 4.0 *Wiley Interdiscip. Rev.-Comput. Mol. Sci.* **8** e1327

[39] Grambow C A, Pattanaik L and Green W H 2020 Reactants, products and transition states of elementary reactions based on quantum chemistry *Sci. Data* **7** 137

[40] Blum L C and Reymond J-L 2009 970 million druglike small molecules for virtual screening in the chemical universe database GDB-13 *J. Am. Chem. Soc.* **131** 8732–3

[41] Reymond J-L 2015 The chemical space project *Acc. Chem. Res.* **48** 722–30

[42] Ramakrishnan R, Dral P O, Rupp M and von Lilienfeld O A 2014 Quantum chemistry structures and properties of 134 kilo molecules *Sci. data* **1** 1–7

[43] Doney A C, Rooks B J, Lu T and Wheeler S E 2016 Design of organocatalysts for asymmetric propargylations through computational screening *ACS Catal.* **6** 7948–55

[44] Wodrich M D, Fabrizio A, Meyer B and Corminboeuf C 2020 Data-powered augmented volcano plots for homogeneous catalysis *Chem. Sci.* **11** 12070–80

[45] Guan Y, Ingman V M, Rooks B J and Wheeler S E 2018 AARON: an automated reaction optimizer for new catalysts *J. Chem. Theory Comput.* **14** 5249–61

[46] Ingman V M, Schaefer A J, Andreola L R and Wheeler S E 2021 QChASM: quantum chemistry automation and structure manipulation *Wiley Interdiscip. Rev.-Comput. Mol. Sci.* **11** e1510

[47] Perdew J P, Burke K and Ernzerhof M 1996 Generalized gradient approximation made simple *Phys. Rev. Lett.* **77** 3865–8

[48] Adamo C and Barone V 1999 Toward reliable density functional methods without adjustable parameters: the PBE0 model *J. Chem. Phys.* **110** 6158–70

[49] Grimme S, Antony J, Ehrlich S and Krieg H 2010 A consistent and accurate *ab initio* parametrization of density functional dispersion correction (DFT-D) for the 94 elements H-Pu *J. Chem. Phys.* **132** 154104

[50] Grimme S, Ehrlich S and Goerigk L 2011 Effect of the damping function in dispersion corrected density functional theory *J. Comput. Chem.* **32** 1456–65

[51] Weigend F and Ahlrichs R 2005 Balanced basis sets of split valence, triple zeta valence and quadruple zeta valence quality for H to Rn: design and assessment of accuracy *Phys. Chem. Chem. Phys.* **7** 3297–305

[52] Marenich A V, Cramer C J and Truhlar D G 2009 Universal solvation model based on solute electron density and on a continuum model of the solvent defined by the bulk dielectric constant and atomic surface tensions *J. Phys. Chem.* B **113** 6378–96

[53] Frisch M J *et al* 2016 *Gaussian 16, Revision C.01* (Wallingford, CT: Gaussian Inc.)

[54] Grimme S 2012 Supramolecular binding thermodynamics by dispersion-corrected density functional theory *Chem. Eur. J.* **18** 9955–64

[55] Luchini G, Alegre-Requena J V, Funes-Ardoiz I and Paton R S 2020 GoodVibes: automated thermochemistry for heterogeneous computational chemistry data *F1000Research* **9** 291

[56] Heinen S, von Rudorff G F and von Lilienfeld O A 2021 Toward the design of chemical reactions: machine learning barriers of competing mechanisms in reactant space *J. Chem. Phys.* **155** 064105

[57] Hansen K, Biegler F, Ramakrishnan R, Pronobis W, von Lilienfeld O A, Müller K-R and Tkatchenko A 2015 Machine learning predictions of molecular properties: accurate many-body potentials and nonlocality in chemical space *J. Phys. Chem. Lett.* **6** 2326–31

[58] Huang B and von Lilienfeld O A 2020 Quantum machine learning using atom-in-molecule-based fragments selected on the fly *Nat. Chem.* **12** 945–51

[59] Faber F A, Christensen A S, Huang B and von Lilienfeld O A 2018 Alchemical and structural distribution based representation for universal quantum machine learning *J. Chem. Phys.* **148** 241717

[60] Christensen A S, Bratholm L A, Faber F A and von Lilienfeld O A 2020 FCHL revisited: faster and more accurate quantum machine learning *J. Chem. Phys.* **152** 044107

[61] Bartók A P, Kondor R and Csányi G 2013 On representing chemical environments *Phys. Rev.* B **87** 184115

[62] De S, Bartók A P, Csányi G and Ceriotti M 2016 Comparing molecules and solids across structural and alchemical space *Phys. Chem. Chem. Phys.* **18** 13754–69

[63] Himanen L, Jäger M O J, Morooka E V, Federici Canova F, Ranawat Y S, Gao D Z, Rinke P and Foster A S 2020 DScribe: library of descriptors for machine learning in materials science *Comput. Phys. Commun.* **247** 106949

[64] Christensen A S, Faber F, Huang B, Bratholm L, Tkatchenko A, Müller K-R and von Lilienfeld O A 2017 QML: a Python toolkit for quantum machine learning (available at: https://github.com/qmlcode/qml)

[65] von Lilienfeld O A and Burke K 2020 Retrospective on a decade of machine learning for chemical discovery *Nat. Commun.* **11** 4895

[66] Faber F A, Hutchison L, Huang B, Gilmer J, Schoenholz S S, Dahl G E, Vinyals O, Kearnes S, Riley P F and von Lilienfeld O A 2017 Prediction errors of molecular machine learning models lower than hybrid DFT error *J. Chem. Theory Comput.* **13** 5255–64

[67] Friederich P, dos Passos Gomes G, Bin R D, Aspuru-Guzik A and Balcells D 2020 Machine learning dihydrogen activation in the chemical space surrounding Vaska's complex *Chem. Sci.* **11** 4584–601

[68] Hammond G S 1955 A correlation of reaction rates *J. Am. Chem. Soc.* **77** 334–8

[69] Henkelman G, Uberuaga B P and Jónsson H 2000 A climbing image nudged elastic band method for finding saddle points and minimum energy paths *J. Chem. Phys.* **113** 9901–4

[70] Huang B and von Lilienfeld O A 2020 Quantum machine learning using atom-in-molecule-based fragments selected on the fly *Nat. Chem.* **12** 945–51

[71] Huang B and von Lilienfeld O A 2016 Communication: understanding molecular representations in machine learning: the role of uniqueness and target similarity *J. Chem. Phys.* **145** 161102

[72] Bartók A P, Kondor R and Csányi G 2013 On representing chemical environments *Phys. Rev.* B **87** 184115–21

[73] Behler J and Parrinello M 2007 Generalized neural-network representation of high-dimensional potential-energy surfaces *Phys. Rev. Lett.* **98** 146401–5