Digital Discovery

ROYAL SOCIETY OF CHEMISTRY

PAPER

View Article Online
View Journal | View Issue

Cite this: *Digital Discovery*, 2024, **3**, 932

Benchmarking machine-readable vectors of chemical reactions on computed activation barriers†

Puck van Gerwen, [ab] Ksenia R. Briling, [a] Yannick Calvino Alonso, [a] Malte Franke[a] and Clemence Corminboeuf *[ab]

Received 6th September 2023
Accepted 28th February 2024

DOI: 10.1039/d3dd00175j

rsc.li/digitaldiscovery

In recent years, there has been a surge of interest in predicting computed activation barriers, to enable the acceleration of the automated exploration of reaction networks. Consequently, various predictive approaches have emerged, ranging from graph-based models to methods based on the three-dimensional structure of reactants and products. In tandem, many representations have been developed to predict experimental targets, which may hold promise for barrier prediction as well. Here, we bring together all of these efforts and benchmark various methods (Morgan fingerprints, the DRFP, the CGR representation-based Chemprop, SLATM$_d$, $B^2R_l^2$, EquiReact and language model BERT + RXNFP) for the prediction of computed activation barriers on three diverse datasets.

1 Introduction

The activation barrier is a fundamental quantity required to understand elementary reaction steps, allowing the estimation of reaction rates and determining dominant reaction pathways.[1–5] Obtaining an accurate Transition State (TS) structure (and therefore, its energy) remains a major computational bottleneck in reaction exploration tasks.[2,4,6–10] Machine learning of activation barriers offers a cheaper alternative than direct computation. Consequently, there has been a recent flurry of works aiming at accurately predicting them.[11–29] These models have featurized reactions using the 2D graph of reactants and products,[13–16] Physical Organic (PO) descriptors derived from computations on reactant and product molecules,[17–20] or 3D structure of reactants and/or products.[21–29]

Of the 2D-graph-based models, the frontrunner is the "Condensed Graph of Reaction" (CGR)[13,14] used to represent reactions in the Chemprop[30,31] model, which is constructed by exploiting atom-mapped reaction SMILES such that a node in the Graph Neural Network (GNN) describes an atomic centre that is transformed during a reaction. Note that atom-mapping remains a challenge in digital chemistry[32,33] despite the progress made over decades of effort.[34–44] State-of-the-art atom-mapping tools either enumerate a subset of known chemical transformations and try to identify them in the reaction of interest,[45] or "template-free" models attempt to extract chemical transformation rules from data.[46] Both approaches are restricted to commonly-occurring organic chemistry: the former requiring the enumeration of reaction rules, and the latter trained on organic chemistry from patent data.[47] Both codes pose practical problems, too, being closed-source. In many cases, it is still preferable to atom-map manually, as was done by Heid *et al.*[13] for specific reaction classes. Atom-mapping by hand requires knowledge of the reaction mechanism for every step in a multi-step process, which is not available for most new chemistry. Given a dataset is correctly atom-mapped however, the CGR representation in Chemprop (hereafter referred to as Chemprop) is a promising approach to encode reactions.[13,14,31] Chemprop has been used to predict computed reaction barriers[48–51] as well as computed reaction energies,[52] reaction rate constants,[53] experimental activation energies,[54] yields[55] and reaction classes.[56,57]

"Physical Organic" (PO) descriptors are based on the properties of molecules involved in the reaction, typically using quantum-chemical computations.[18,54,58–67] The computed properties can be broadly divided into steric and electronic parameters.[54,58,63,64,68] Steric properties include the molecular volume and surface area or Sterimol parameters,[69,70] or in the case of ligands, buried volume[63] or Tolman cone angles.[71] Electronic descriptors include frontier molecular orbital energies, reactivity indices from conceptual DFT,[72] natural bond orbital (NBO)-derived descriptors, atomic charges, and NMR chemical shifts.[64] There are also conformation-specific descriptors such as the Average Steric Occupancy (ASO)[60] and Molecular Field Analysis (MFA)-based descriptors.[73] Typically, PO descriptors that are relevant for a specific reaction are chosen by an expert.

[a]*Laboratory for Computational Molecular Design, Institute of Chemical Sciences and Engineering, École Polytechnique Fédérale de Lausanne, 1015 Lausanne, Switzerland. E-mail: clemence.corminboeuf@epfl.ch*

[b]*National Center for Competence in Research-Catalysis (NCCR-Catalysis), École Polytechnique Fédérale de Lausanne, 1015 Lausanne, Switzerland*

† Electronic supplementary information (ESI) available. See DOI: https://doi.org/10.1039/d3dd00175j

However, databases of descriptors are being pre-computed and made publicly available,[74] and automated workflows developed,[75–77] that alleviate the need of an expert to construct these representations.[65] So-called QM-augmented GNNs[77–79] combine atom-mapped graph-based models (specifically, Weisfeiler–Lehman Networks (WLNs)[80]) and PO descriptors. QM-GNNs[78,79] as well as PO descriptors combined with simpler (often linear) models[17,18,20,81] have been used to predict activation barriers. PO-based models have also been used to predict experimental targets such as yield[58,59] and e.e.[60,62,82,83]

3D-structure-based models can be broadly separated into two categories: (i) those that predict the TS structure, by virtue of Generative Adversarial Networks (GANs),[24] GNNs,[23] Reinforcement Learning (RL)[22] or diffusion models,[21,84] or (ii) those that directly predict the activation barrier,[13,14,25,27–29] which will be our focus here. As illustrated in Fig. 1, such 3D-structure based reaction representations can be thought of as "thermochemistry-inspired", using an interpolation between reactants' and products' geometries as a proxy[85] for the reactants and TS, that allows for a mapping to the activation barrier. Examples are SLATM$_d$,[28,29] constructed from molecular representations[86] of reactants and products and B^2R^2,[29] a dedicated reaction representation. Molecular variants of these representations[86–91] are often referred to as "physics-based",[92–94] taking as input molecules' atom types and coordinates (and in some cases charge and spin information[95,96]) thereby mimicking the role of the Hamiltonian in the Schrödinger equation to solve for molecular properties.

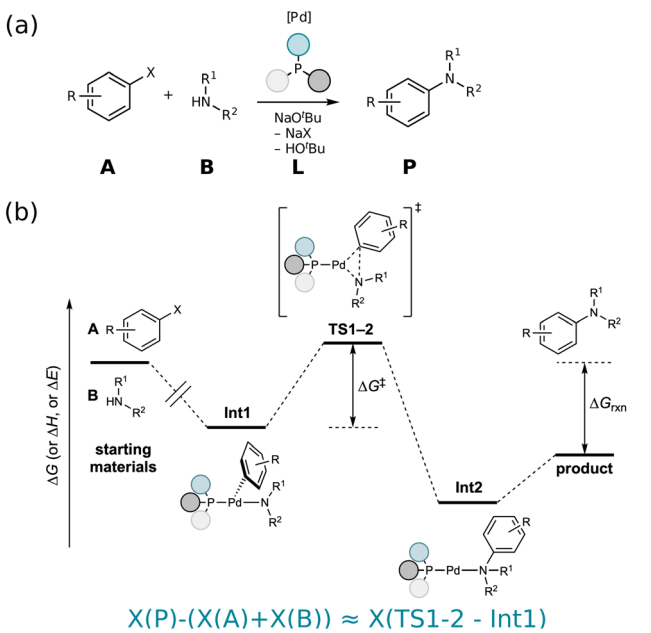

Fig. 1 (a) An example Buchwald–Hartwig amination is illustrated with starting materials A and B, catalyzed by a Pd-based catalyst with phosphene ligands L resulting in product P. (b) Thermochemistry-inspired representations aim to predict the activation barrier $\Delta G^{\ddagger}$, which is a function of the TS1-2 and Int1 energies, using starting materials and products (or intermediates Int1 and Int2, if this mechanistic information is available).

In a similar spirit, physics-based deep learning models[97–106] learn their own representations from the structural input data (*i.e.* nuclear charges and coordinates) and encode known physical priors, such as symmetries, into the network architecture. While these have been demonstrated to obtain impressive out-of-sample accuracies for molecular properties on countless occasions,[97–106] they have only recently been introduced to predict reaction properties.[14,107] EquiReact,[107] an Equivariant Neural Network (ENN) that uses 3D information from reactants and products, as well as optionally atom-mapping information, has been shown to exhibit competitive performance on datasets of reaction barriers.

In tandem to the work on activation barrier prediction, there has been a broader effort to predict experimental targets,[11,58,60,108–111] including reaction class,[56] synthetic routes[46,112–115] and experimental outcomes (yield,[58,59,109,110] activation energy[54] and enantioselectivity e.e.[60,62]). The workhorse reaction representation in this domain is a combination of Morgan FingerPrints (MFPs)[116] of reaction components.[117–119] MFPs identify the presence of circular substructures in a 2D description of molecules. Alternatively, reactions can be represented as strings in the form of reaction SMILES. Powerful transformer models[120,121] developed for language translation can then be applied to text-based descriptions of reactions for classification or regression tasks.[46,56,109,112] Such models were originally explored in chemistry with synthesis planning tasks in mind.[46,112] Since then, the RXNFP[56] has been used to predict reaction classes[56] and yields[109] as well. The Differential Reaction FingerPrint (DRFP), which takes a symmetric difference between substructures identified in the reactants and products to generate a fixed-size vector per reaction, has been illustrated to be competitive with the RXNFP for related yield and reaction class prediction tasks.[122]

Here, we benchmark different reaction representations, whether previously used to predict activation barriers (the graph-based Chemprop,[13,31] physics-based SLATM$_d$ and $B^2R_l^2$,[29] EquiReact[107]) or not (MFPs,[116] DRFP,[122] and RXNFP[56,109]) for the prediction of activation barriers across several datasets: from the general-scope GDB7-22-TS[123] to a single-reaction class dataset Cyclo-23-TS[51] to a specific dataset, the Proparg-21-TS.[28,124] In the process, we discover what information needs to be captured in the reaction representation for accurate activation barrier prediction.

2 Computational details

2.1 Representations and ML models

We study a variety of reaction representations: the 2D-structure based MFP[116] and DRFP,[122] the 2D-graph based Chemprop,[13,31] RXNFP[56] trained using the BERT language model,[120] 3D-structure based thermochemistry-inspired representations SLATM$_d$ and $B^2R_l^2$,[29] and ENN EquiReact.[107] Note that we exclude PO descriptors because of the computational cost associated with generating these representations for the larger databases studied in this work.

2.1.1 MFP and DRFP. The input to these models is the SMILES strings of reactants and products, or the reaction

SMILES. A difference of Morgan FingerPrints of reactants and products (MFP) was generated using RDKit version 2023.3.3.[125] The Differential Reaction FingerPrint (DRFP) was generated using the drfp version 0.3.6.[122] A fingerprint size of 1024 was used throughout. Fingerprints are combined with random forest (RF) models as implemented in sklearn[126] version 1.3.1, as these models are naturally suited to the binary features in MFP and DRFP fingerprints. Correspondingly, MFP/DRFP have often been combined with RF or gradient boosting[127] models in previous publications for best results.[13,31,58,122,128]

2.1.2 Chemprop. The Condensed Graph of Reaction (CGR)[13] is built from atom-mapped SMILES strings of reactants and products, which is then passed through the directed message-passing neural network Chemprop[30] (version 1.6.1, using RDKit[125] version 2023.9.4).

In order to assess the sensitivity of the trained models to the quality of the atom-mapping, we tested three versions of each Chemprop model: (i) with "true" atom-mapping ("Chemprop True"), as provided by the authors of the datasets (typically obtained using the transition state structure, or using known reaction rules); (ii) with "automatic" atom-mapping, performed by the open-source tool RXNMapper[129] ("Chemprop RXNMapper"); and (iii) with no atom-mapping, which removes the atom-mapping indices from reactants and products ("Chemprop None"). The latter option evaluates the efficacy of the graph-based models without atom-mapping information. We note that this is different to how the "no-mapping" model for Chemprop was run for a previous publication,[130] where the mappings of product atoms were randomly shuffled with respect to reactant atoms. In the case of no maps, Chemprop interprets the reactants' and products' graphs as separate entities. For an unmapped reaction A + B → C, it extends the reactants' and products' graphs with "ghost" graphs shadowing the missing counterparts producing the disconnected reaction graph $(A.B.0^{C}) \rightarrow (0^{A}.0^{B}.C)$. The condensed graph, being (with default settings) a difference between products and reactants, reads (−A.−B.C). This leads to a ∼1 kcal mol^{-1} improvement compared to "no mapping" results with random maps.[130]

Where available (for the GDB7-22-TS and Proparg-21-TS sets with "True" maps), we used explicit hydrogen atoms in the Chemprop model. Otherwise, the Hs are implicit.

We also note that the version of RXNMapper used here (0.3.0) runs successfully on all datasets studied, while the previous version (0.2.9) failed on some datasets in a previous publication.[130]

2.1.3 Language models. Language models are built from reaction SMILES. Input data is augmented using SMILES randomization:[131,132] first, SMILES strings of reactants and products are de-canonicalized. Then, atoms of each SMILES string are renumbered by rotation of their index. For each renumbering, a grammatically correct SMILES is constructed. Duplicate SMILES are removed after the randomization procedure. Using a randomization factor of 10, this effectively multiplies the training and test set sizes by 10 for each dataset. Since most of our reactions consist of a single reactant and product, we did not employ molecule permutations, also illustrated to be an effective data augmentation strategy.[132]

We use a BERT model[120] pre-trained on a reaction MLM task from the rxnfp library.[133] We fine-tune the model on the training data. 10× data augmentation was used (see the ESI† for a comparison of models with and without data augmentation).

2.1.4 Thermochemistry-inspired representations. In Fig. 1a, an example Buchwald–Hartwig amination is illustrated with starting materials **A** and **B**, catalyzed by a Pd-based catalyst with phosphene ligands **L** resulting in product **P**. The reaction mechanism is shown in Fig. 1b. While the activation barrier $\Delta G^{\ddagger}$ can be obtained from the energies of the Transition State (TS1-2) and the preceding state (Int1), determining the structure of the TS is computationally expensive. Likewise, an ML model to predict $\Delta G^{\ddagger}$ would be most accurate if the geometry of the TS was used to construct the representation. However, if the geometry is known, the energy is known, making the ML model redundant. Thermochemistry-inspired representations, that instead use 3D structures of reactants and products, have therefore been developed.[28,29]

Depending on the mechanistic information available, "reactants" can be either starting materials or the intermediate preceding the relevant TS. Similarly, "products" may be an intermediate following the TS or the final product. For uncatalyzed reactions, using starting materials and products only is sufficient. For catalyzed reactions however, both the catalyst/ligand and substrates must be encoded. Representations built from intermediates preceding and following the TS have the advantage of naturally encoding the structure of the catalyst.

A previous benchmarking study[29] identified the best-performing thermochemistry-inspired representation as the difference in SLATM vectors[86] of reactants and products ($SLATM_d$). The molecular SLATM representation is built from increasing orders of potential terms that describe interactions between atoms in a molecule.[86] The one-, two- and three-body terms are concatenated to form the eventual molecular vector. The B^2R^2 family of representations[29] are constructed in a similar way, except using different potential functions and being truncated at two-body terms. The $B^2R_l^2$ representation employed here combines the appropriate features into element-wise "bags" depending on the identity of the element I in each pairwise interaction between atoms I, J. $SLATM_d$ instead bags pairwise interactions into pairwise bags, as well as three-body interactions into bags corresponding to the identity of the three involved elements.

Molecular SLATM vectors were generated using the qml python package[134] before being combined to form the reaction version $SLATM_d$. The $B^2R_l^2$ representation is generated from the github repository,[135] using default parameters. These representations are combined with Kernel Ridge Regression (KRR) models, as has been the standard for the molecular representations since their initial development.[89,128,136–141] So-called "physics-based" representations are typically high dimensional, continuous, and used in a low-data regime. Kernel methods are then well-suited to these, allowing for meaningful interpretation of the similarity kernel, finding trends in high dimensions and with little data.[92–94]

2.1.5 EquiReact. EquiReact[107] builds on the thermochemistry-inspired representations, taking the same

structures as input to the model. However, the representation is learned end-to-end as part of the training process. The model consists of independent equivariant channels for reactant and product molecules, followed by different possible combination modes to obtain a latent reaction representation (which can take into account atom-mapping information, mimic atom-maps with cross-attention, or use simple arithmetic operations like the difference between reactants' and products' representations). The latent representation is provided as input to an MLP to predict the reaction barrier. Here, we take the model either using atom mapping or not resulting in the best performance in each case (GDB7-22-TS with atom-maps, the other models without). The set of hyperparameters for these models is given in the ESI.†

As in the original work,[107] we run EquiReact without explicit Hs, as it has been shown that there is no consistent improvement in performance when including Hs, and the models become significantly more expensive to train.

2.2 Data splits, hyperparameters and performance metrics

All datasets are split into 10 random 80% train/10% validation/10% test splits. For all models, performance is reported as mean absolute error (MAE) on the test set, averaged over the 10 folds.

In the case of the EquiReact model, we use the previously published hyperparameters,[107] which correspond to those optimised on the first data split for each of the three datasets. These are repeated in the ESI† of this paper.

For all other models, hyperparameters are tuned for the first data split and then used for all other splits. The space of hyperparameters tested, as well as the optimal parameters obtained, can be found in the ESI.† In the case of a large number of parameter combinations (the RF models and Chemprop), the parameters were optimized using Bayesian optimization with hyperopt version 0.2.7.[142–144] For the RF models, we optimized the maximum depth of the decision trees, the number of decision trees, the maximum number of features used to split each tree, the minimum number samples per split, minimum number of samples per leaf, and whether to bootstrap the models. A maximum of 100 combinations of hyperparameters were tested for each model.

For Chemprop, the hyperparameter search using hyperopt was modified from the chemprop codebase,[145] version 1.6.1, to evaluate the hyperparameters on the first cross-validation fold only (the default behaviour finding the best hyperparameters over k folds). The modified code can be found in forked version[146] of the original repository. As per the default in the original codebase, the hyperparameter search optimizes the hidden size, depth (number of message passing iterations), dropout probability and number of feed-forward layers after message passing. For the hyperparameter search, we train the models for 100 epochs (50 for the GDB7-22-TS set), rather than the 300 epochs used to train and test the model performance. A maximum of 100 (50 for the GDB7-22-TS set) combinations of hyperparameters were tested for each model.

For the kernel models with relatively few hyperparameters, a grid search was used to optimize the kernel width and regularization parameter. For RXNFP, following the suggestions in the documentation,[147] we optimize the learning rate and dropout probability parameters using a grid search. We use a batch size of 32 and train for 10 epochs.

2.3 Datasets of reaction barriers

In order to compare reaction representations built from either reaction SMILES or three-dimensional structure, we include two recently published datasets of activation barriers that provide both input formats: the GDB7-22-TS[123] and Cyclo-23-TS.[51] Several models[31,79,107] have already been tested on these sets, which allows for an interesting broader comparison across different reaction fingerprints. We also include the Proparg-21-TS set of reaction barriers,[28,124] which provides only three-dimensional structure as the input format. In order to allow comparison to other methods we convert these to reaction SMILES (*vide supra*).

The three datasets are diverse in their respective challenges. Fig. 2 illustrates the spread in the barrier ($\Delta G^{\ddagger}$ or $\Delta E^{\ddagger}$) for the sets, highlighting their difference. Each is described in detail below.

2.3.1 GDB7-22-TS. The GDB7-22-TS dataset[123] consists of close to 12 000 diverse un-catalyzed organic reactions automatically constructed from the GDB7 dataset[148–150] using the growing string method[151] along with corresponding energy barriers ($\Delta E^{\ddagger}$) computed at the CCSD(T)-F12a/cc-pVDZ-F12//ωB97X-D3/def2-TZVP level. These molecules have a maximum of 7 heavy atoms and a maximum of 23 atoms. The input structures to the ML models are optimized reactants (starting materials) and products. This is an updated version of the GDB7-20-TS set[48] used in previous works.[15,29] The dataset is chemically diverse, spanning several reaction classes, reflected in the large span of the target property in Fig. 2.

Correspondingly, no "direct" model (without pre-training on lower levels of theory)[13–15] has predictive mean absolute errors (MAEs) of less than 4.1 kcal mol^{-1},[31] (reported error on a single 80/10/10% split). Errors of 4–5 kcal mol^{-1} are in the realm of

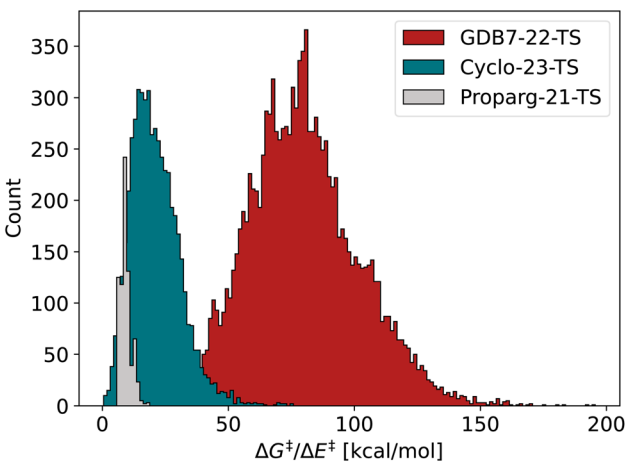

Fig. 2 Distribution of barriers $\Delta G^{\ddagger}$ or $\Delta E^{\ddagger}$ for the three datasets GDB7-22-TS, Cyclo-23-TS and Proparg-21-TS.

DFT errors with respect to the CCSD(T) data,[123] making these predictions as useful as DFT.

2.3.2 Cyclo-23-TS. The Cyclo-23-TS dataset[51] encompasses 5269 reaction profiles for un-catalyzed [3 + 2] cycloaddition reactions with activation free energy barriers ($\Delta G^{\ddagger}$) computed at the B3LYP-D3(BJ)/def2-TZVP//B3LYP-D3(BJ)/def2-SVP level. These molecules have a maximum of 50 heavy atoms and a maximum of 94 atoms. The input structures to the ML models are optimized reactants (starting materials) and products. Since the dataset focuses on a single reaction class, the spread in the target property is narrower than for the GDB7-22-TS, illustrated in Fig. 2. The best published MAE is 2.3 kcal mol^{-1},[107] (reported error averaged over 10 random 90/5/5% splits).

2.3.3 Proparg-21-TS. The Proparg-21-TS dataset[28,124] contains 754 structures of intermediates before and after the enantioselective transition state of the propargylation of benzaldehyde, with activation energies ($\Delta E^{\ddagger}$) computed at the B97D/TZV(2p,2d) level. These molecules have a maximum of 52 heavy atoms and a maximum of 89 atoms. The input structures to the ML models are optimized intermediates preceding and following the TS (see Fig. S1†). This dataset is separate from the other two: while it reports energy barriers, these are then converted into enantioselectivity (e.e.) values using the competing barriers of (*R*)- and (*S*)-enantiomers of the product. Thus the focus is not on different reaction classes and transformations, but rather on a stereochemical level, given that stereoisomers of intermediates/TSs are present in the dataset. Correspondingly, there is a narrower spread in the target value. In addition, the dataset is smaller than the other two, providing a more challenging test case for deep learning models. The best reported MAE is 0.27 kcal mol^{-1},[107] (reported error averaged over 10 random 90/5/5% splits). As the relationship between the barrier and computed selectivity is exponential, a low error in $\Delta E^{\ddagger}$ is essential.

Unlike the other datasets which have both xyz files containing Cartesian coordinates of reactants and products and (atom-mapped) SMILES strings, this dataset contains only xyz files. xyz were converted to SMILES using the xyz2mol[152] function of cell2mol.[153] We note that the original xyz2mol[152] fails to convert these structures to SMILES, since the encoded chemical rules do not include the elements present in the Proparg-21-TS dataset. cell2mol[153] extends the program to inorganic chemistry and is able to convert all but one structure to SMILES (noted in the ESI†). On inspection however the resulting SMILES strings are unreasonable, breaking the aromaticity in the cycles.

To address this problem, we generated an alternative set of SMILES strings. The ligands in the intermediates' structures were constructed from a library of fragments.[28] This allowed for the generation of chemically-meaning SMILES using simple combinatorial rules. We also extend these SMILES strings to partially encode the relevant stereochemistry. Since we noticed no difference in performance of the SMILES-based models depending on the quality of the SMILES strings, we use the lower-quality (generated from cell2mol) and put the results associated with the higher-quality variations, as well as more details as to their construction, in the ESI.†

In order to atom-map the SMILES strings, we modified cell2mol's xyz2mol to keep atom indices from xyz. Since each reaction consists of only one reactant and one product, whose atom order in the xyz files is preserved, the reaction SMILES are easily correctly atom-mapped.

2.3.4 Geometries. The aforementioned datasets provide geometries optimized using DFT. A set of geometries at GFN2-*x*TB[154] level is taken from ref. 107 for all datasets to compare the performance of the 3D-structure based models with lower quality geometries. Note that for a handful of reactions in the GDB7-22-TS (171) and Cyclo-23-TS (60) sets, at least one of the molecules in the reaction did not converge and therefore are excluded from the geometry quality tests (Fig. 4).

3 Results and discussion

Fig. 3 illustrates the performance of various models to predict barriers of 3 diverse datasets: (a) the GDB7-22-TS, (b) the Cyclo-23-TS and (c) the Proparg-21-TS. The three datasets pose different challenges: the GDB7-22-TS set has the largest chemical diversity (and therefore spread in its target property), resulting in a higher overall predicted MAE as well as a larger difference in the performance of various models. This dataset provides a challenging test case for ML models for barrier prediction.

The Proparg-21-TS set provides a different sort of challenge: for a fixed set of starting materials, catalysts with different stereochemistries lead to either the (*R*)- or (*S*)-enantiomeric products. While we constructed modified SMILES strings that enumerate different octahedral arrangements of the ligands (*i.e.*, stereochemistry), they yield the same results as the SMILES strings (see ESI†). Only the 3D-structure-based models, capturing the stereochemistry of the intermediates before and after the enantiodetermining TS, are injective and therefore effective representations. The resulting MAEs are over 20×

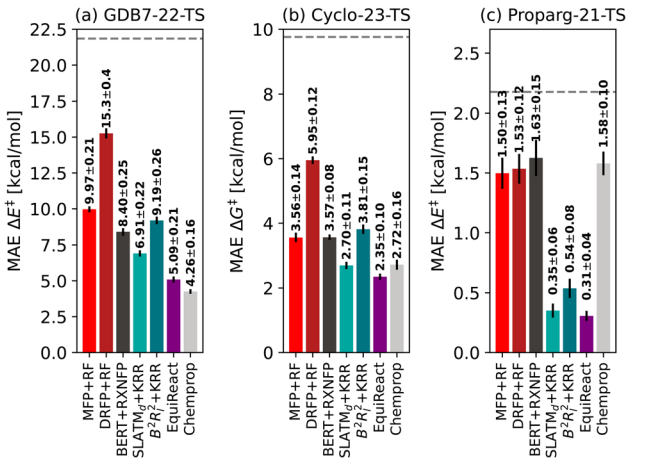

Fig. 3 Mean absolute errors of different fingerprints based on 2D structure (MFP + RF, DRFP + RF), language model BERT + RXNFP, 3D-structure based models (SLATM$_d$ + KRR, $B^2R_l^2$ + KRR, EquiReact) and 2D graph-based model Chemprop on three different datasets of reaction barriers GDB7-22-TS, Cyclo-23-TS and Proparg-21-TS. The standard deviation of each dataset is given as a dashed line.

lower than those of the 2D-structure based representations. The Proparg-21-TS set serves as an important reminder of the diversity of chemical reactions beyond changes in connectivity in reactants and products, by including changes in stereochemistry. Only recently are datasets emerging that explore variations beyond connectivity changes, such as conformations in the recent work of Zhao *et al.*[50]

The Cyclo-23-TS dataset, with a fixed reaction class and no diversity in 3D structure, is the simplest of the three, though distinctions can still be seen between different models, and interestingly, the hierarchy of models for the GDB7-22-TS set are mostly maintained in the Cyclo-23-TS. The following subsections discuss the relative performance of the different models as interesting case studies for ML models for activation barrier prediction.

3.1 Models based on three-dimensional geometries

van Gerwen *et al.*[29] previously compared the performance of various physics-based molecular fingerprints[86,87,89–91] adapted for the prediction of reaction properties on four different datasets (the GDB7-20-TS,[48] Hydroform-22-TS,[29] SN2-20,[29,49] and Proparg-21-TS[29,124]). The authors found that the SLATM$_d$ representation resulted in the lowest MAE across the datasets studied. They introduced a related reaction fingerprint, the $B^2R_l^2$, based on similar design principles, but at a compact size, resulting in a small increase in error compared to SLATM$_d$. Here, we observe a bigger gap between the performance of the two representations. While $B^2R_l^2$ + KRR still produces reasonable errors, it becomes less competitive in comparison to other ML models studied here. In line with previously published results,[107] we find that the more sophisticated architecture in EquiReact (including an end-to-end learned representation and incorporating equivariance for molecular components) allows for improved accuracy compared to the fingerprint models.

In an out-of-sample setting, optimizing geometries at a high level of theory (here, ωB97X-D3/def2-TZVP level for GDB7-22-TS, B3LYP-D3(BJ)/def2-SVP for Cyclo-23-TS, B97D/TZV(2p,2d) for the Proparg-21-TS) is computationally demanding. Fig. 4 illustrates the resultant MAE with training and predicting on cheaper GFN2-*x*TB geometries. Except for EquiReact on the GDB7-22-TS set, all models suffer from a deterioration of predicted MAE when moving from DFT to *x*TB geometries. The effect is most pronounced for the Proparg-21-TS set, likely because GFN2-*x*TB is not parameterised on 5- or 6-coordinated silicon systems.[154] The thermochemistry-inspired representations combined with kernel models are more sensitive than EquiReact to the geometry quality, as discussed in ref. 107. SLATM$_d$ + KRR is also more sensitive than $B^2R_l^2$ to the geometry quality, since B^2R^2 uses only distances whereas SLATM also uses angles between atoms.

Comparing the models with GFN2-*x*TB geometries to other methods tested, Chemprop and BERT + RXNFP become more attractive options (for Chemprop this relies on atom-mapping quality however, *vide infra*). 3D-structure-based models using xTB geometries still outperform 2D-structure based methods for the Proparg-21-TS set, however, due to the inability of SMILES strings to capture stereochemistry. Methods based on 3D structure remain the only viable option for accurate predictions of datasets with stereochemical diversity.

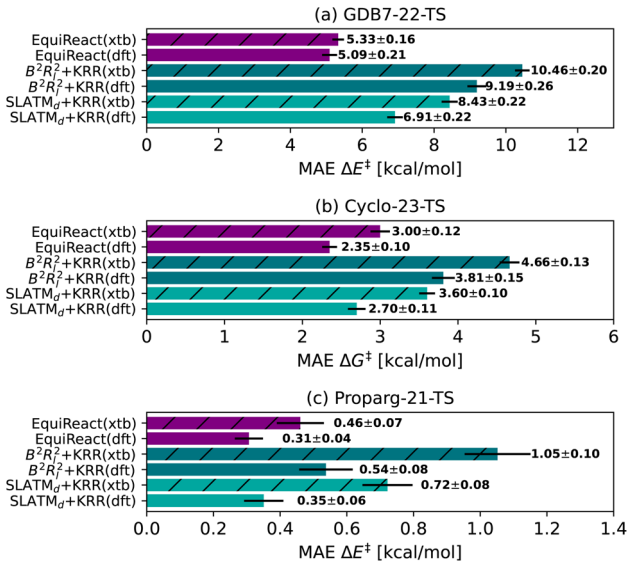

Fig. 4 Mean Absolute Error (MAE) for models based on 3D geometry: SLATM$_d$ + KRR, $B^2R_l^2$ + KRR and EquiReact using GFN2-*x*TB[154] (xtb) or DFT (dft) levels of theory (ωB97X-D3/def2-TZVP level for GDB7-22-TS, B3LYP-D3(BJ)/def2-SVP for Cyclo-23-TS, B97D/TZV(2p,2d) for the Proparg-21-TS). Note that EquiReact uses a smaller subset of geometries due to technical reasons.[107] This results in a larger data loss from DFT to *x*TB geometries.

3.2 Graph-based models and their reliance on atom-mapping

The graph-based Chemprop model gives excellent predictions of reaction barriers for the GDB7-22-TS and Cyclo-23-TS datasets. Our final Chemprop MAEs are slightly higher than those published by Heid *et al.*,[31] because we optimized hyperparameters only on the first fold. For the Cyclo-23-TS dataset, we note that many models (Chemprop, SLATM$_d$ + KRR, EquiReact) are more accurate than the QM-GNN model published by Stuyver *et al.*[79] following the publication of the dataset[51] with a published MAE of 2.96 kcal mol^{-1} (reported error on a single 90/5/5% split, for a non-ensembled model. This is the closest setting to ours published by the authors, otherwise using ensembled models).

In line with Spiekermann *et al.*,[14] we were surprised to see that the Chemprop model based on 2D graphs of reactants and products could compete with models with 3D information. However, the graph-based models do encode additional information compared to all others considered here: they rely on atom-mapped SMILES as input. The atom-mapped reaction SMILES are used to construct a single graph for a reaction. Then, each node in the graph describes a transformation taking part at each of the atoms involved in the reaction. This is valuable information that resembles a reaction mechanism. It is likely the Chemprop model is more effective than the WLN-type[78–80] at exploiting the atom-mapping information, as the WLN-type does not explicitly create node features that encode

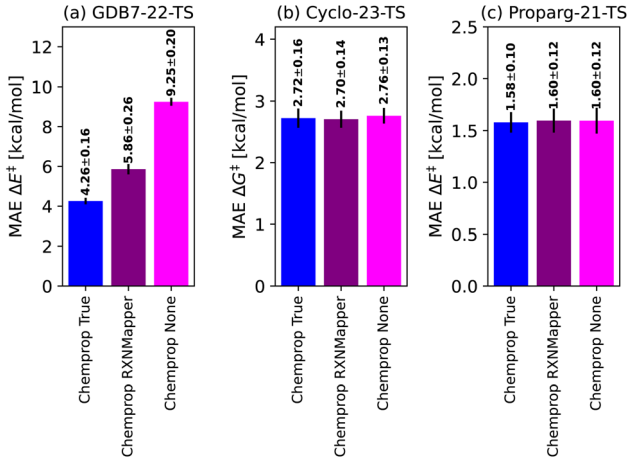

Fig. 5 Comparison of the Chemprop model performance depending on the atom-mapping quality: by-hand ("True"), using automatic tools[129] ("RXNMapper") or no maps ("None").

transformations from reactants to products like Chemprop. If the Chemprop model were enhanced with quantum-chemical features, we might see further performance improvements. In any case, high quality atom mapping is an unrealistic prerequisite for most reaction prediction scenarios (*vide supra*).

To place Chemprop's performance on fairer footing, we compared its out-of-sample MAE when using high-quality atom-mapping (done by hand or using closed-source software,[155] "Chemprop True") to atom-mapping performed by open-source software ("Chemprop RXNMapper") to the naive graph model without atom-mapping information ("Chemprop None"). Results are illustrated in Fig. 5. The GDB7-22-TS dataset contains exotic chemistry, having been generated using automated PES exploration. This illustrates a realistic test case where RXNMapper might be needed, as mapping by hand or using heuristic rules would be a challenge. Due to the presence of unseen chemistries, RXNMapper struggles to correctly map all the reactions, and correspondingly Chemprop RXNMapper reduces in accuracy compared to Chemprop True, illustrated in Fig. 5a. Entirely removing atom-mapping information results in a poor Chemprop None model.

The RXNMapper is trained on patent data, likely including cycloaddition reactions, therefore predicting the correct maps and Chemprop RXNMapper achieving the same results as Chemprop True for the Cyclo-23-TS dataset (Fig. 5b). For these reactions, atom mapping does not seem critical to good performance, as Chemprop None obtains similar performance to Chemprop True/RXNMapper. Since these reactions consist of a fixed reaction class, atom-mapping likely does not provide new information to the models. Finally, as illustrated in Fig. 5c for the Proparg-21-TS set, all the available Chemprop versions perform poorly, due to graph-based models' inability to distinguish stereochemical variations in the dataset.

These results illustrate that graph-based models are not necessarily best-performing because of the nature of the GNN architecture, but rather because of their dependence on atom-mapping, an additional input that may be challenging to come by for realistic reaction settings. For the GDB7-22-TS dataset, all models without atom-mapping information result in relative high predictive errors (>5.9 kcal mol^{-1}), highlighting that more work needs to be done on such challenging sets before ML models can reliably replace computation of activation barriers.

3.3 Text-based models

Probst *et al.*[122] had observed that simple 2D fingerprints (*i.e.*, the DRFP) perform as well as the deep-learned RXNFP[133] for the tasks of reaction classification on the USPTO 1k TPL dataset[56] and reaction yield prediction on high-throughput datasets of Buchwald–Hartwig cross-coupling reactions[58] and Suzuki–Miyaura reactions,[156] as well as from patents.[47,109] Jorner and co-workers[54] previously observed that BERT + RXNFP fingerprints outperformed MFPs in the prediction of experimental activation energies (reaction rates). Now comparing MFPs and DRFPs to deep learned representations built from SMILES for the prediction of computed reaction barriers for the first time, the BERT + RXNFP model outperforms the simpler representations (as well as other sophisticated representations/models). The BERT + RXNFP performs better or equivalently to $B^2R_l^2$, within standard deviations, on the GDB7-22-TS and Cyclo-23-TS datasets. Thus, it is feasible to obtain good predictive accuracy of reaction barriers using only (un-mapped) reaction SMILES as input. Since RXNMapper results from an attention head of a related transformer model,[129] it is possible that a similar unsupervised atom-mapping-like task was performed in its training stages.

These results suggest that a new generation of ML models might be able to achieve accurate predictions on reaction property prediction tasks with less information. It could be interesting to investigate whether a tokenization involving atom-maps could further improve these models.

4 Timings and representation sizes

Table 1 gives the training and inference times for a subset of 750 points (80/10/10 split) for each model and dataset combination, as well as a description of the scaling of the dimensionality of the representation with the number of unique elements.

All fingerprint-based models (MFP + RF, DRFP + RF, $B^2R_l^2$ + KRR, SLATM$_d$ + KRR) were run on a CPU: an Apple Macbook Pro 2022 with an Apple M2 chip (8 CPU cores, 3.5 GHz). All neural networks (EquiReact, Chemprop, BERT + RXNFP) were run on a GPU-enabled cluster: specifically, an Intel Xeon-Gold based cluster (20 CPU cores, 2.1 GHz) with an NVIDIA V100 PCIe 32 GB GPU. A single CPU is used for all jobs.

Of the 2D fingerprints, MFP is more efficient to train than DRFP by several orders of magnitude. Similarly for SLATM$_d$ *vs.* $B^2R_l^2$. In both cases this is due to the former representations being implemented in low-level languages: MFP in C and SLATM$_d$ in Fortran. B^2R^2 is simpler than SLATM$_d$, using only two-body terms (as indicated in the scaling of the dimensionality of the representations), and therefore could run faster than SLATM$_d$. A more efficient implementation of B^2R^2 will soon be

Table 1 Training times (including generation of fingerprints/representations) and inference times, for a total dataset size of 750 (split into 80%/10%/10%), trained with previously established optimal hyperparameters. Timings reported for a single random fold. Fingerprint models are run on a CPU, neural networks on a GPU. The representation (rep.) size describes the scaling of the representation size with the number of unique elements n

Model	Dataset	Train time (s)	Inference time (s)	Rep. size
DRFP + RF	GDB7-22-TS	4.560	0.005	$O(1)$
	Cyclo-23-TS	7.590	0.008	$O(1)$
	Proparg-21-TS	10.597	0.001	$O(1)$
MFP + RF	GDB7-22-TS	0.855	0.010	$O(1)$
	Cyclo-23-TS	0.672	0.007	$O(1)$
	Proparg-21-TS	0.654	0.008	$O(1)$
$B^2R_l^2$ + KRR	GDB7-22-TS	9.5082	0.0004	$O(n)$
	Cyclo-23-TS	41.543	0.001	$O(n)$
	Proparg-21-TS	125.3702	0.0006	$O(n)$
SLATM$_d$ + KRR	GDB7-22-TS	1.1902	0.0007	$O(n^3)$
	Cyclo-23-TS	7.2619	0.0007	$O(n^3)$
	Proparg-21-TS	13.548	0.0009	$O(n^3)$
EquiReact	GDB7-22-TS	794.265	0.141	$O(1)$
	Cyclo-23-TS	1171.083	0.538	$O(1)$
	Proparg-21-TS	3735.803	0.207	$O(1)$
Chemprop	GDB7-22-TS	131.331	0.102	$O(1)$
	Cyclo-23-TS	181.070	0.160	$O(1)$
	Proparg-21-TS	507.706	0.197	$O(1)$
BERT + RXNFP	GDB7-22-TS	82.681	2.675	$O(1)$
	Cyclo-23-TS	85.931	2.915	$O(1)$
	Proparg-21-TS	88.639	3.385	$O(1)$

released as part of the Q-stack python package.[157] Only SLATM$_d$ and B^2R^2 suffer from an increasing representation size with the number of unique elements in the dataset. All other methods fix the size of the representation *a priori*. SLATM$_d$'s cubic scaling can pose memory errors if working with datasets of diverse molecules.

Molecule sizes increase from GDB7-22-TS (up to 7 heavy atoms) to Cyclo-23-TS (up to 50 heavy atoms) to Proparg-21-TS (up to 52 heavy atoms). While the Cyclo-23-TS and Proparg-21-TS have a similar maximum molecule size, all of the molecules in the Proparg-21-TS set are large, whereas the Cyclo-23-TS set also contains small molecules. The train times therefore increase accordingly from GDB7-22-TS to Cyclo-23-TS to Proparg-21-TS for all methods except the MFP + RF. Since the MFP and BERT + RXNFP operate on SMILES strings, while the SMILES lengths do increase with molecule size, the effect is not as dramatic as the graph-based methods or atom-in-molecule based methods. The DRFP also operates on SMILES strings, but creates a set of sub-structures, where these sets increase in size with increased lengths of SMILES strings.

All deep learning models are more costly to train and use for inference than the fingerprint models. EquiReact, which uses tensor operations, is the most expensive to run. BERT + RXNFP is the most expensive to use for inference due to the data augmentation pre-processing step, but is the cheapest of the deep learning models to train.

In summary, if users have limited compute time and/or no GPU, the fingerprint-based models are the best choice. Based on current implementations, cheaper fingerprint-based models are the MFP + RF and SLATM$_d$ + KRR respectively for 2D and 3D. SLATM$_d$ can pose memory problems due to its cubic scaling of representation size with maximum number of elements. BERT + RXNFP is the cheapest deep learning model.

5 Recommendations for activation barrier prediction

Following the analysis on the three datasets studied here, we give our general recommendations as to which currently available ML methods might be best suited in different scenarios.

- In the case of **chemically diverse datasets**, such as the GDB7-22-TS,[123] its precursor the GDB7-20-TS,[48] and the more recent RGD1 dataset[50] (in our terminology would be the RGD1-23-TS), where reactions are already mapped or can be readily mapped by RXNMapper,[129] **Chemprop**[13] offers a promising model, as it can better distinguish between different reaction classes in the dataset.

- In the case of **geometry-sensitive datasets**, such as the Proparg-21-TS,[28,124] a subset of the RGD1 dataset, and any datasets of enantioselectivity data, models based on **3D geometry**,[29] **especially EquiReact**[107] offers the best performance. The accuracy of the final model might depends on the **quality of the three-dimensional geometries** of reactants and products provided, depending on how well GFN2-*x*TB captures the systems studied.

- In the case of **minimal input information**: no atom-mapping, nor three-dimensional geometries, **only SMILES strings of reactants and products**, language model **BERT + RXNFP**[56] offers the best performance. This might be practical for more challenging reactions, where atom-mapping with

open-source tools remains a challenge (due to their difference from the data used for pre-training RXNMapper) and estimating three-dimensional geometries is difficult with cheap methods.

- If **resources are limited**, particularly if users do not have access to a GPU, the fingerprint models are the best choice. Based on current implementations, the cheapest models are the **MFP + RF**[116] and **SLATM$_d$ + KRR**[29] for 2D and 3D models respectively. If users are working with diverse datasets, $B^2R_l^2$ has a more favourable scaling of representation size with number of elements. **BERT + RXNFP** is the cheapest deep learning model to train (though more expensive at inference time than the other models).

6 Conclusion

With the surge in interest in both the dedicated prediction of activation barriers as well as measured performance metrics (yield, enantioselectivity, *etc.*) of chemical reactions, various approaches have emerged to featurize chemical reactions for use in ML models. We compared a diverse set of fingerprints (the MFP and DRFP built from 2D structure, the RXNFP deriving from a pre-trained BERT model on reaction SMILES, the 2D graph-based Chemprop, and 3D-structure based SLATM$_d$, $B^2R_l^2$ and EquiReact) on three different datasets of reaction barriers, from the chemically diverse GDB7-22-TS to the fixed reaction class Cyclo-23-TS to the stereochemistry-sensitive Proparg-21-TS. We find that 3D-structure based models are needed particularly for configuration–sensitive reaction properties. The graph-based Chemprop model exhibits excellent performance in the absence of stereochemical diversity, but this is contingent on high-quality atom-mapped reaction SMILES. Language-based models offer the convenience of only using unmapped reaction SMILES as input. These results suggest the way forward for a new generation of ML models for chemical reactions.

Data availability

The code to reproduce all results can be found at **https://github.com/lcmd-epfl/benchmark-barrier-learning**. This includes scripts to run the ML models, to parse the results and to generate the plots in the paper. The outputs of the models are saved in the github repository, such that the results can be easily parsed to re-generate the results in the paper. A detailed description can be found in the README of the repository. Three datasets are studied in this work: the Proparg-21-TS, Cyclo-23-TS and GDB7-22-TS. While all sets were previously published and made available open-source, we made some modifications to these sets and made these versions available in two places: in the same github repository mentioned above, as well as on zenodo at the following link: **https://zenodo.org/record/8309465**. This record includes the datasets studied as well as the saved ML models.

Author contributions

The project was conceptualized by P. v. G. and C. C. M. F. and Y. C. A. contributed to preliminary experiments initiating this work. Data was curated by P. v. G. and K. R. B. Data was analyzed by P. v. G., K. R. B. and Y. C. A. The original draft was written by P. v. G. with reviews and edits from all authors. C. C. provided supervision throughout and is acknowledged for acquiring funding.

Conflicts of interest

There are no conflicts to declare.

Acknowledgements

The authors would like to thank Oliver Schilter for debugging the code for the language models as well as for helpful discussions. Simone Gallarati and Ruben Laplaza are thanked for feedback on the text. P. v. G. and C. C. acknowledge the National Centre of Competence in Research (NCCR) "Sustainable chemical process through catalysis (Catalysis)", grant number 180544, of the Swiss National Science Foundation (SNSF), for financial support. K. R. B., Y. C. A. and C. C. acknowledge the National Centre of Competence in Research (NCCR) "Materials' Revolution: Computational Design and Discovery of Novel Materials (MARVEL)", grant number 205602, of the Swiss National Science Foundation. K. R. B. was supported by the European Research Council (grant number 817977). Y. C. A. thanks the EPFL for financial support.

References

1 F. D. Qian Peng and R. S. Paton, *Chem. Soc. Rev.*, 2016, **22**, 6093–6107.
2 O. Engkvist, P.-O. Norrby, N. Selmi, Y.-h. Lam, Z. Peng, E. C. Sherer, W. Amberg, T. Erhard and L. A. Smyth, *Drug Discovery Today*, 2018, **23**, 1203–1218.
3 A. L. Dewyer, A. J. Argüelles and P. M. Zimmerman, *Wiley Interdiscip. Rev.: Comput. Mol. Sci.*, 2018, **8**, e1354.
4 D. G. Truhlar, B. C. Garrett and S. J. Klippenstein, *J. Phys. Chem.*, 1996, **100**, 12771–12800.
5 F. Battin-Leclerc, E. Blurock, R. Bounaceur, R. Fournet, P.-A. Glaude, O. Herbinet, B. Sirjean and V. Warth, *Chem. Soc. Rev.*, 2011, **40**, 4762–4782.
6 J. P. Unsleber and M. Reiher, *Annu. Rev. Phys. Chem.*, 2020, **71**, 121–142.
7 Y.-h. Lam, M. N. Grayson, M. C. Holland, A. Simon and K. Houk, *Acc. Chem. Res.*, 2016, **49**, 750–762.
8 P. Zimmerman, *J. Chem. Theory Comput.*, 2013, **9**, 3043–3050.
9 S. Maeda, K. Ohno and K. Morokuma, *Phys. Chem. Chem. Phys.*, 2013, **15**, 3683–3701.
10 Y. Guan, V. M. Ingman, B. J. Rooks and S. E. Wheeler, *J. Chem. Theory Comput.*, 2018, **14**, 5249–5261.
11 P. Schwaller, A. C. Vaucher, R. Laplaza, C. Bunne, A. Krause, C. Corminboeuf and T. Laino, *Wiley Interdiscip. Rev.: Comput. Mol. Sci.*, 2022, **12**, e1604.
12 T. Lewis-Atwell, P. A. Townsend and M. N. Grayson, *Wiley Interdiscip. Rev.: Comput. Mol. Sci.*, 2022, **12**, e1593.

13 E. Heid and W. H. Green, *J. Chem. Inf. Model.*, 2022, **62**, 2101–2110.
14 K. A. Spiekermann, L. Pattanaik and W. H. Green, *J. Phys. Chem. A*, 2022, **126**, 3976–3986.
15 C. A. Grambow, L. Pattanaik and W. H. Green, *J. Phys. Chem. Lett.*, 2020, **11**, 2992–2997.
16 D. Lemm, G. F. von Rudorff and O. A. von Lilienfeld, *Nat. Commun.*, 2021, **12**, 4468.
17 A. R. Singh, B. A. Rohr, J. A. Gauthier and J. K. Norskov, *Catal. Lett.*, 2019, **149**, 2347–2354.
18 S. Choi, Y. Kim, J. W. Kim, Z. Kim and W. Y. Kim, *Chem.–Eur. J.*, 2018, **24**, 12354–12358.
19 I. Migliaro and T. R. Cundari, *J. Chem. Inf. Model.*, 2020, **60**, 4958–4966.
20 E. H. E. Farrar and M. N. Grayson, *Chem. Sci.*, 2022, **13**, 7594–7603.
21 C. Duan, Y. Du, H. Jia and H. J. Kulik, Accurate transition state generation with an object-aware equivariant elementary reaction diffusion model, *Nat. Comput. Sci.*, 2023, **3**, 1045–1055.
22 J. Zhang, Y.-K. Lei, Z. Zhang, X. Han, M. Li, L. Yang, Y. I. Yang and Y. Q. Gao, *Phys. Chem. Chem. Phys.*, 2021, **23**, 6888–6895.
23 L. Pattanaik, J. B. Ingraham, C. A. Grambow and W. H. Green, *Phys. Chem. Chem. Phys.*, 2020, **22**, 23618–23626.
24 M. Z. Makoś, N. Verma, E. C. Larson, M. Freindorf and E. Kraka, *J. Chem. Phys.*, 2021, **155**, 024116.
25 B. Savoie, Q. Zhao, D. Anstine and O. Isayev, *Chem. Sci.*, 2023, **14**, 13392–13401.
26 P. Friederich, G. dos Passos Gomes, R. D. Bin, A. Aspuru-Guzik and D. Balcells, *Chem. Sci.*, 2020, **11**, 4584–4601.
27 S. Heinen, G. F. von Rudorff and O. A. von Lilienfeld, *J. Chem. Phys.*, 2021, **155**, 064105.
28 S. Gallarati, R. Fabregat, R. Laplaza, S. Bhattacharjee, M. D. Wodrich and C. Corminboeuf, *Chem. Sci.*, 2021, **12**, 6879–6889.
29 P. van Gerwen, A. Fabrizio, M. Wodrich and C. Corminboeuf, *Mach. Learn.: Sci. Technol.*, 2022, **3**, 045005.
30 K. Yang, K. Swanson, W. Jin, C. Coley, P. Eiden, H. Gao, A. Guzman-Perez, T. Hopper, B. Kelley, M. Mathea, *et al.*, *J. Chem. Inf. Model.*, 2019, **59**, 3370–3388.
31 E. Heid, K. P. Greenman, Y. Chung, S.-C. Li, D. E. Graff, F. H. Vermeire, H. Wu, W. H. Green and C. J. McGill, *J. Chem. Inf. Model.*, 2024, **64**, 9–17.
32 W. L. Chen, D. Z. Chen and K. T. Taylor, *Wiley Interdiscip. Rev.: Comput. Mol. Sci.*, 2013, **3**, 560–593.
33 G. A. Preciat Gonzalez, L. R. El Assal, A. Noronha, I. Thiele, H. S. Haraldsdóttir and R. M. Fleming, *J. Cheminf.*, 2017, **9**, 1–15.
34 M. F. Lynch and P. Willett, *J. Chem. Inf. Comput. Sci.*, 1978, **18**, 154–159.
35 J. J. McGregor and P. Willett, *J. Chem. Inf. Comput. Sci.*, 1981, **21**, 137–140.
36 T. E. Moock, J. G. Nourse, D. Grier and W. D. Hounshell, *Chemical structures: the international language of chemistry*, Springer, 1988, pp. 303–313.
37 K. Funatsu, T. Endo, N. Kotera and S.-I. Sasaki, *Tetrahedron Comput. Methodol.*, 1988, **1**, 53–69.
38 R. Körner and J. Apostolakis, *J. Chem. Inf. Model.*, 2008, **48**, 1181–1189.
39 J. Apostolakis, O. Sacher, R. Körner and J. Gasteiger, *J. Chem. Inf. Model.*, 2008, **48**, 1190–1198.
40 C. Jochum, J. Gasteiger and I. Ugi, *Angew. Chem., Int. Ed.*, 1980, **19**, 495–505.
41 T. Akutsu, *Proceedings of the seventh annual international conference on research in computational molecular biology*, 2003, pp. 1–8.
42 J. D. Crabtree and D. P. Mehta, *ACM J. Exp. Algorithmics*, 2009, **13**, 1–15.
43 E. L. First, C. E. Gounaris and C. A. Floudas, *J. Chem. Inf. Model.*, 2012, **52**, 84–92.
44 M. Latendresse, J. P. Malerich, M. Travers and P. D. Karp, *J. Chem. Inf. Model.*, 2012, **52**, 2970–2982.
45 W. Jaworski, S. Szymkuć, B. Mikulak-Klucznik, K. Piecuch, T. Klucznik, M. Kaźmierowski, J. Rydzewski, A. Gambin and B. A. Grzybowski, *Nat. Commun.*, 2019, **10**, 1434.
46 P. Schwaller, B. Hoover, J.-L. Reymond, H. Strobelt and T. Laino, *Sci. Adv.*, 2021, **7**, eabe4166.
47 D. M. Lowe, PhD thesis, University of Cambridge, 2012.
48 C. Grambow, L. Pattanaik and W. Green, *Sci. Data*, 2020, **7**, 137.
49 G. von Rudorff, S. Heinen, M. Bragato and O. A. von Lilienfeld, *Mach. Learn.: Sci. Technol.*, 2020, **1**, 045026.
50 Q. Zhao, S. M. Vaddadi, M. Woulfe, L. A. Ogunfowora, S. S. Garimella, O. Isayev and B. M. Savoie, *Sci. Data*, 2023, **10**, 145.
51 T. Stuyver, K. Jorner and C. W. Coley, *Sci. Data*, 2023, **10**, 66.
52 S. Stocker, G. Csányi, K. Reuter and J. T. Margraf, *Nat. Commun.*, 2020, **11**, 5505.
53 P. L. Bhoorasingh, B. L. Slakman, F. Seyedzadeh Khanshan, J. Y. Cain and R. H. West, *J. Phys. Chem. A*, 2017, **121**, 6896–6904.
54 K. Jorner, T. Brinck, P.-O. Norrby and D. Buttar, *Chem. Sci.*, 2021, **12**, 1163–1175.
55 H. Huang, C. Pandya, C. Liu and J. D. Farelli, *Proc. Natl. Acad. Sci. U.S.A.*, 2015, **112**, E1974–E1983.
56 P. Schwaller, D. Probst, A. Vaucher, V. Nair, D. Kreutter, T. Laino and J.-L. Reymond, *Nat. Mach. Intell.*, 2021, **3**, 1–9.
57 Nextmove Software, *Pistachio*, 2022, **https://www.nextmovesoftware.com/pistachio.html**.
58 D. T. Ahneman, J. G. Estrada, S. Lin, S. D. Dreher and A. G. Doyle, *Science*, 2018, **360**, 186–190.
59 A. M. Żurański, J. I. Martinez Alvarado, B. J. Shields and A. G. Doyle, *Acc. Chem. Res.*, 2021, **54**, 1856–1865.
60 A. F. Zahrt, J. J. Henle, B. T. Rose, Y. Wang, W. T. Darrow and S. E. Denmark, *Science*, 2019, **363**, eaau5631.
61 A. R. Singh, B. A. Rohr, J. A. Gauthier and J. K. Nørskov, *Catal. Lett.*, 2019, **149**, 2347–2354.
62 J. P. Reid and M. S. Sigman, *Nature*, 2019, **571**, 343–348.

63 T. Gensch, G. dos Passos Gomes, P. Friederich, E. Peters, T. Gaudin, R. Pollice, K. Jorner, A. Nigam, M. Lindner-D'Addario, M. S. Sigman, *et al.*, *J. Am. Chem. Soc.*, 2022, **144**, 1205–1217.
64 C. B. Santiago, J.-Y. Guo and M. S. Sigman, *Chem. Sci.*, 2018, **9**, 2398–2412.
65 K. Jorner, *Chimia*, 2023, **77**, 22.
66 L. C. Gallegos, G. Luchini, P. C. St. John, S. Kim and R. S. Paton, *Acc. Chem. Res.*, 2021, **54**, 827–836.
67 W. L. Williams, L. Zeng, T. Gensch, M. S. Sigman, A. G. Doyle and E. V. Anslyn, *ACS Cent. Sci.*, 2021, **7**, 1622–1637.
68 D. J. Durand and N. Fey, *Chem. Rev.*, 2019, **119**, 6561–6594.
69 A. Verloop, *Pesticide Chemistry: Human Welfare and Environment*, Elsevier, 1983, pp. 339–344.
70 A. V. Brethomé, S. P. Fletcher and R. S. Paton, *ACS Catal.*, 2019, **9**, 2313–2323.
71 C. A. Tolman, *Chem. Rev.*, 1977, **77**, 313–348.
72 A. R. Jupp, T. C. Johnstone and D. W. Stephan, *Inorg. Chem.*, 2018, **57**, 14764–14771.
73 S. Yamaguchi, *Org. Biomol. Chem.*, 2022, **20**, 6057–6071.
74 T. Gensch, G. dos Passos Gomes, P. Friederich, E. Peters, T. Gaudin, R. Pollice, K. Jorner, A. Nigam, M. Lindner-D'Addario, M. S. Sigman and A. Aspuru-Guzik, *J. Am. Chem. Soc.*, 2022, **144**, 1205–1217.
75 A. M. Żurański, J. Y. Wang, B. J. Shields and A. G. Doyle, *React. Chem. Eng.*, 2022, **7**, 1276–1284.
76 S. Sowndarya S. V., J. N. Law, C. E. Tripp, D. Duplyakin, E. Skordilis, D. Biagioni, R. S. Paton and P. C. S. John, *Nat. Mach. Intell.*, 2022, **4**, 720–730.
77 Y. Guan, C. W. Coley, H. Wu, D. Ranasinghe, E. Heid, T. J. Struble, L. Pattanaik, W. H. Green and K. F. Jensen, *Chem. Sci.*, 2021, **12**, 2198–2208.
78 T. Stuyver and C. W. Coley, *J. Chem. Phys.*, 2022, **156**, 084104.
79 T. Stuyver and C. W. Coley, *Chem.–Eur. J.*, 2023, **29**, e202300387.
80 C. W. Coley, W. Jin, L. Rogers, T. F. Jamison, T. S. Jaakkola, W. H. Green, R. Barzilay and K. F. Jensen, *Chem. Sci.*, 2019, **10**, 370–377.
81 I. Migliaro and T. R. Cundari, *J. Chem. Inf. Model.*, 2020, **60**, 4958–4966.
82 A. Schoepfer, R. Laplaza, M. Wodrich, J. Waser and C. Corminboeuf, *ChemRxiv*, preprint, 2023, chemrxiv-2023-pknnt.
83 L.-C. Xu, J. Frey, X. Hou, S.-Q. Zhang, Y.-Y. Li, J. C. A. Oliveira, S.-W. Li, L. Ackermann and X. Hong, *Nat. Synth.*, 2023, **2**, 321–330.
84 S. Kim, J. Woo and W. Y. Kim, *Nat. Commun.*, 2023, **15**, 341.
85 D. Sheppard, R. Terrell and G. Henkelman, *J. Chem. Phys.*, 2008, **128**, 134106.
86 B. Huang and O. A. von Lilienfeld, *Nat. Chem.*, 2020, **12**, 945–951.
87 A. S. Christensen, L. A. Bratholm, F. A. Faber and O. A. von Lilienfeld, *J. Chem. Phys.*, 2020, **152**, 044107.
88 F. A. Faber, A. S. Christensen, B. Huang and O. A. von Lilienfeld, *J. Chem. Phys.*, 2018, **148**, 241717.
89 M. Rupp, A. Tkatchenko, K.-R. Müller and O. A. von Lilienfeld, *Phys. Rev. Lett.*, 2012, **108**, 058301.
90 K. Hansen, F. Biegler, R. Ramakrishnan, W. Pronobis, O. A. von Lilienfeld, K.-R. Müller and A. Tkatchenko, *J. Phys. Chem. Lett.*, 2015, **6**, 2326–2331.
91 A. P. Bartók, R. Kondor and G. Csányi, *Phys. Rev. B: Condens. Matter Mater. Phys.*, 2013, **87**, 184115.
92 B. Huang and O. A. von Lilienfeld, *Chem. Rev.*, 2021, **121**, 10001–10036.
93 F. Musil, A. Grisafi, A. P. Bartók, C. Ortner, G. Csányi and M. Ceriotti, *Chem. Rev.*, 2021, **121**, 9759–9815.
94 M. F. Langer, A. Goessmann and M. Rupp, *npj Comput. Mater.*, 2022, **8**, 41.
95 A. Fabrizio, K. R. Briling and C. Corminboeuf, *Digital Discovery*, 2022, **1**, 286–294.
96 S. Llenga and G. Gryn'ova, *J. Chem. Phys.*, 2023, **158**, 214116.
97 B. Anderson, T. S. Hy and R. Kondor, *Adv. Neural Inf. Process. Syst.*, 2019, **32**, 14537–14546.
98 J. Gasteiger, F. Becker and S. Günnemann, *arXiv*, 2021, preprint, arXiv:2106.08903, DOI: **10.48550/ARXIV.2106.08903**.
99 K. Schütt, O. Unke and M. Gastegger, *Proceedings of the 38th International Conference on Machine Learning*, 2021, pp. 9377–9388.
100 O. T. Unke and M. Meuwly, *J. Chem. Theory Comput.*, 2019, **15**, 3678–3693.
101 B. K. Miller, M. Geiger, T. E. Smidt and F. Noé, *arXiv*, 2020, preprint, arXiv:2008.08461, DOI: **10.48550/ARXIV.2008.08461**.
102 J. Brandstetter, R. Hesselink, E. van der Pol, E. Bekkers and M. Welling, *arXiv*, 2021, preprint, arXiv:2110.02905, DOI: **10.48550/ARXIV.2110.02905**.
103 K. T. Schütt, O. T. Unke and M. Gastegger, *arXiv*, 2021, preprint, arXiv:2102.03150, DOI: **10.48550/ARXIV.2102.03150**.
104 J. Gasteiger, S. Giri, J. T. Margraf and S. Günnemann, *arXiv*, 2022, preprint, arXiv:2011.14115, DOI: **10.48550/ARXIV.2011.14115**.
105 V. G. Satorras, E. Hoogeboom and M. Welling, *Proceedings of the 38th International Conference on Machine Learning*, 2021, pp. 9323–9332.
106 N. Thomas, T. Smidt, S. Kearnes, L. Yang, L. Li, K. Kohlhoff and P. Riley, *arXiv*, 2018, preprint, arXiv:1802.08219, DOI: **10.48550/ARXIV.1802.08219**.
107 P. van Gerwen, K. R. Briling, C. Bunne, V. R. Somnath, R. Laplaza, A. Krause and C. Corminboeuf, *arXiv*, 2023, preprint, arXiv:2312.08307, DOI: **10.48550/ARXIV.2312.08307**.
108 O. Engkvist, P.-O. Norrby, N. Selmi, Y.-h. Lam, Z. Peng, E. C. Sherer, W. Amberg, T. Erhard and L. A. Smyth, *Drug Discovery Today*, 2018, **23**, 1203–1218.
109 P. Schwaller, A. C. Vaucher, T. Laino and J.-L. Reymond, *Mach. Learn.: Sci. Technol.*, 2021, **2**, 015016.
110 A. L. Haywood, J. Redshaw, M. W. D. Hanson-Heine, A. Taylor, A. Brown, A. M. Mason, T. Gärtner and J. D. Hirst, *J. Chem. Inf. Model.*, 2022, **62**, 2077–2092.
111 N. Ree, A. H. Göller and J. H. Jensen, *J. Cheminf.*, 2021, **13**, 1–9.
112 P. Schwaller, T. Laino, T. Gaudin, P. Bolgar, C. A. Hunter, C. Bekas and A. A. Lee, *ACS Cent. Sci.*, 2019, **5**, 1572–1583.

113 J. N. Wei, D. Duvenaud and A. Aspuru-Guzik, *ACS Cent. Sci.*, 2016, **2**, 725–732.
114 C. W. Coley, R. Barzilay, T. S. Jaakkola, W. H. Green and K. F. Jensen, *ACS Cent. Sci.*, 2017, **3**, 434–443.
115 M. H. S. Segler and M. P. Waller, *Chem.–Eur. J.*, 2017, **23**, 5966–5971.
116 D. Rogers and M. Hahn, *J. Chem. Inf. Model.*, 2010, **50**, 742–754.
117 L. Pattanaik and C. W. Coley, *Chem*, 2020, **6**, 1204–1207.
118 N. Schneider, D. M. Lowe, R. A. Sayle and G. A. Landrum, *J. Chem. Inf. Model.*, 2015, **55**, 39–53.
119 F. Sandfort, F. Strieth-Kalthoff, M. Kühnemund, C. Beecks and F. Glorius, *Chem*, 2020, **6**, 1379–1390.
120 J. Devlin, M.-W. Chang, K. Lee and K. Toutanova, *arXiv*, 2018, preprint, arXiv:1810.04805, DOI: **10.48550/ARXIV.1810.04805**.
121 A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser and I. Polosukhin, *Adv. Neural Inf. Process. Syst.*, 2017, **30**, 5998–6008.
122 D. Probst, P. Schwaller and J.-L. Reymond, *Digital Discovery*, 2022, **1**, 91–97.
123 K. Spiekermann, L. Pattanaik and W. H. Green, *Sci. Data*, 2022, **9**, 417.
124 A. C. Doney, B. J. Rooks, T. Lu and S. E. Wheeler, *ACS Catal.*, 2016, **6**, 7948–7955.
125 RDKit, *Open-source cheminformatics*, 2023, **https://www.rdkit.org**.
126 F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot and E. Duchesnay, *J. Mach. Learn. Res.*, 2011, **12**, 2825–2830.
127 T. Chen and C. Guestrin, *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, New York, NY, USA, 2016, pp. 785–794.
128 F. A. Faber, L. Hutchison, B. Huang, J. Gilmer, S. S. Schoenholz, G. E. Dahl, O. Vinyals, S. Kearnes, P. F. Riley and O. A. von Lilienfeld, *J. Chem. Theory Comput.*, 2017, **13**, 5255–5264.
129 P. Schwaller, B. Hoover, J.-L. Reymond, H. Strobelt and T. Laino, *Sci. Adv.*, 2021, **7**, eabe4166.
130 P. van Gerwen, M. D. Wodrich, R. Laplaza and C. Corminboeuf, *Mach. Learn.: Sci. Technol.*, 2023, **4**, 048002.
131 G. Lambard and E. Gracheva, *Mach. Learn.: Sci. Technol.*, 2020, **1**, 025004.
132 P. Schwaller, A. C. Vaucher, T. Laino and J.-L. Reymond, *ChemRxiv*, preprint, 2020, chemrxiv.13286741.
133 RXN for Chemistry team/University of Bern, *RXNFP - chemical reaction fingerprints*, 2021, **https://rxn4chemistry.github.io/rxnfp/**.
134 A. S. Christensen, F. Faber, B. Huang, L. Bratholm, A. Tkatchenko, K.-R. Müller and O. A. von Lilienfeld, *QML: A Python Toolkit for Quantum Machine Learning*, 2017, **https://github.com/qmlcode/qml**.
135 P. van Gerwen, A. Fabrizio, M. Wodrich and C. Corminboeuf, *b2r2-reaction-rep*, 2022, **https://github.com/lcmd-epfl/b2r2-reaction-rep**.
136 A. P. Bartók, R. Kondor and G. Csányi, *Phys. Rev. B: Condens. Matter Mater. Phys.*, 2013, **87**, 184115.
137 K. Hansen, F. Biegler, R. Ramakrishnan, W. Pronobis, O. A. von Lilienfeld, K.-R. Muller and A. Tkatchenko, *J. Phys. Chem. Lett.*, 2015, **6**, 2326–2331.
138 H. Huo and M. Rupp, *arXiv*, 2017, preprint, arXiv:1704.06439, DOI: **10.48550/ARXIV.1704.06439**.
139 A. Grisafi, J. Nigam and M. Ceriotti, *Chem. Sci.*, 2021, **12**, 2078–2090.
140 R. Drautz, *Phys. Rev. B*, 2019, **99**, 014104.
141 J. Nigam, S. Pozdnyakov and M. Ceriotti, *J. Chem. Phys.*, 2020, **153**, 121101.
142 J. Bergstra, D. Yamins and D. Cox, *Proceedings of the 30th International Conference on Machine Learning*, 2013, pp. 115–123.
143 J. Bergstra, B. Komer, C. Eliasmith, D. Yamins and D. D. Cox, *Comput. Sci. Discovery*, 2015, **8**, 014008.
144 J. Bergstra, *hyperopt*, 2024, **https://github.com/hyperopt/hyperopt**.
145 K. Swanson, K. Yang, W. Jin, L. Hirschfeld and A. Tam, *Chemprop*, 2022, **https://github.com/chemprop/chemprop**.
146 P. van Gerwen, *Chemprop*, 2024, **https://github.com/puckvg/chemprop**.
147 RXN for Chemistry team/University of Bern, *Predicting Chemical Reaction Yields*, 2022, **https://rxn4chemistry.github.io/rxn_yields/**.
148 L. C. Blum and J.-L. Reymond, *J. Am. Chem. Soc.*, 2009, **131**, 8732–8733.
149 J.-L. Reymond, *Acc. Chem. Res.*, 2015, **48**, 722–730.
150 R. Ramakrishnan, P. O. Dral, M. Rupp and O. A. von Lilienfeld, *Sci. Data*, 2014, **1**, 1–7.
151 P. M. Zimmerman, *J. Comput. Chem.*, 2015, **36**, 601–611.
152 Y. Kim and W. Kim, *Bull. Korean Chem. Soc.*, 2015, **36**, 1769–1777.
153 S. Vela, R. Laplaza, Y. Cho and C. Corminboeuf, *npj Comput. Mater.*, 2022, **8**, 188.
154 C. Bannwarth, S. Ehlert and S. Grimme, *J. Chem. Theory Comput.*, 2019, **15**, 1652–1671.
155 W. Jaworski, S. Szymkuć, B. Mikulak-Klucznik, K. Piecuch, T. Klucznik, M. Kaźmierowski, J. Rydzewski, A. Gambin and B. A. Grzybowski, *Nat. Commun.*, 2019, **10**, 1434.
156 D. Perera, J. W. Tucker, S. Brahmbhatt, C. J. Helal, A. Chong, W. Farrell, P. Richardson and N. W. Sach, *Science*, 2018, **359**, 429–434.
157 K. R. Briling, A. Fabrizio, Y. Calvino Alonso, R. Laplaza, P. van Gerwen, O. Hernandez Cuellar and L. Marsh, *Q-stack*, 2023, **https://github.com/lcmd-epfl/Q-stack**.