Chemical Science

ROYAL SOCIETY OF CHEMISTRY

EDGE ARTICLE

View Article Online
View Journal

Cite this: DOI: 10.1039/d2sc02925a

All publication charges for this article have been paid for by the Royal Society of Chemistry

Machine learning and semi-empirical calculations: a synergistic approach to rapid, accurate, and mechanism-based reaction barrier prediction†

Elliot H. E. Farrar and Matthew N. Grayson *

Modern QM modelling methods, such as DFT, have provided detailed mechanistic insights into countless reactions. However, their computational cost inhibits their ability to rapidly screen large numbers of substrates and catalysts in reaction discovery. For a C–C bond forming nitro-Michael addition, we introduce a synergistic semi-empirical quantum mechanical (SQM) and machine learning (ML) approach that allows the prediction of DFT-quality reaction barriers in minutes, even on a standard laptop using widely available modelling software. Mean absolute errors (MAEs) are obtained that are below the accepted chemical accuracy threshold of 1 kcal mol^{-1} and substantially better than SQM methods without ML correction (5.71 kcal mol^{-1}). Predictive power is shown to hold when the ML models are applied to an unseen set of compounds from the toxicology literature. Mechanistic insight is also achieved *via* the generation of full SQM transition state (TS) structures which are found to be very good approximations for the DFT-level geometries, revealing important steric interactions in some TSs. This combination of speed, accuracy, and mechanistic insight is unprecedented; current ML barrier models compromise on at least one of these important criteria.

Received 25th May 2022
Accepted 8th June 2022

DOI: 10.1039/d2sc02925a

rsc.li/chemical-science

Introduction

In the last thirty years, the ease-of-use, power, and accessibility of highly sophisticated computational chemistry techniques has increased substantially.[1,2] These methods allow us to obtain detailed understandings of reaction mechanisms *via* exploration of their potential energy surface (PES) to identify reactant, product, and transition state (TS) structures. Typically, the most telling insights come from analysis of competing TS geometries, from which key steric and electronic effects can be identified, and their respective reaction barriers. This allows us to rationalise the mechanisms and selectivities of a huge range of chemical reactions.[3–6] In turn, this enables the rational design of new reactions and catalysts,[7–9] and reduces the need for experimental trial-and-error approaches.

Although numerous toolkits have been developed to automate the location and subsequent optimisation of TS structures,[10–12] the cost of these methods remains limited by the level of molecular modelling method used in geometry optimisation and energy calculations. Despite some reported shortcomings,[13–15] for example when TSs contain ion pairs,[16] density functional theory (DFT)[17,18] is one of the most widely used quantum mechanical (QM) reaction modelling techniques and has been involved in the successful modelling of countless reactions.[2,19] However, a trade-off between speed and accuracy must be made; typical DFT calculations take on the order of hours to days, but this can be further exacerbated by the exact combination of functional, basis set and solvation model used, the number of atoms and complexity of the system in question, and the necessity, in most cases, to optimise many distinct chemical species and multiple conformations of each to draw practical conclusions.

In terms of calculation time, molecular mechanics (MM) typically improves on DFT by around six orders of magnitude.[20] Accordingly, several force field and force field-cost methods have been developed that allow the approximation of various thermochemical properties.[21–31] However, many do not provide geometric information, or produce less accurate pictures of the TS, for example by treating them as minima. Furthermore, force field methods are often parameterised on niche training domains with expensive DFT calculations or experimental data or require complex and lengthy parameterisation procedures to be effective, limiting their transferability and making their implementation more difficult.

In contrast, most semi-empirical quantum mechanical (SQM) methods are extensively parameterised, and thus widely applicable to many areas of chemistry.[32–35] Additionally, many are embedded in widely available software packages, such as Gaussian,[36] GAMESS,[37] Spartan,[38] MOPAC,[39] and ORCA,[40] and

Department of Chemistry, University of Bath, Claverton Down, Bath, BA2 7AY, UK. E-mail: M.N.Grayson@bath.ac.uk

† Electronic supplementary information (ESI) available: Additional details on dataset generation, feature extraction, machine learning, and model evaluation; complete list of all metrics, features, and hyperparameters for all computed models. See https://doi.org/10.1039/d2sc02925a

thus are straightforward to implement. Like force fields, SQM methods are several orders of magnitude faster than DFT, on the scale of seconds to minutes per calculation.[20] However, this comes at the expense of some chemical accuracy as parts of the more expensive QM calculations are replaced with empirical parameters tuned *via* experiment or full QM.[41] Despite some reported disagreements with more accurate *ab initio* methods, for example in the assignment of Diels–Alder reactions as either step-wise or concerted,[42] SQM methods have been shown to produce reliable geometries for TSs of many reactions, including nucleophilic substitutions, isomerisations, alkene epoxidations, metal-catalysed oxidations, and even some cycloadditions.[43–45] However, they require expensive, high-accuracy DFT single point energy (SPE) corrections to give reliable barriers,[43] and thus SQM data alone is not of a sufficient quality for accurate reaction modelling.

Machine learning (ML) is an increasingly prevalent tool in the field of chemistry, used to identify patterns in chemical datasets.[46,47] Once a relationship is identified, the target property can be predicted from the features (inputs) of the model. Previously, ML has been used for the prediction of reaction rates and barriers derived from both experiment[48–53] and high-level reference calculations.[54–63] However, many models have errors significantly above the accepted threshold for chemical accuracy of 1 kcal mol^{-1},[64,65] and offer little or no mechanistic insight, for example using molecular fingerprints or graphical representations of molecules. A few examples do make use of TSs in their predictions but use DFT to generate them, making the prediction process very time-consuming.[48,66] Therefore, no current ML barrier model offers the combination of fast and accurate predictions with mechanistic insight derived from TSs. In recent years, several studies have used ML to bridge the gap between SQM and high-level QM, allowing prediction of various ground state thermochemical properties.[67–71] We believe this combined SQM/ML approach could be used to improve current standards in the prediction of reaction barriers. Thus, we proposed to learn the relationship between simple, interpretable, and readily available SQM-derived molecular and atomic features and target DFT barriers, and hence afford DFT-quality barriers with the calculation speed of SQM. By calculating TSs using SQM, rapid mechanistic insight would also be available.

Herein, we use a synergistic SQM/ML approach to predict DFT-quality free energy activation barriers for a diverse class of C–C bond forming nitro-Michael additions (Fig. 1). Michael additions are one of the most efficient and prevalent methods for formation of C–C bonds in organic and biosynthesis,[72] finding important applications in asymmetric catalysis[5,73–75] and several natural product syntheses.[72,76–79] Among the many classes of these versatile reactions, the nitro-Michael addition is one of the most useful;[80–84] insertion of the nitro group into the organic framework *via* Michael addition enables a variety of synthetically important stereoselective reactions,[85] and the resulting nitro compounds can be the precursor for an assortment of highly useful chemical functionalities, including pyrrolidines, lactones, aminocarbonyls, and aminoalkanes.[86]

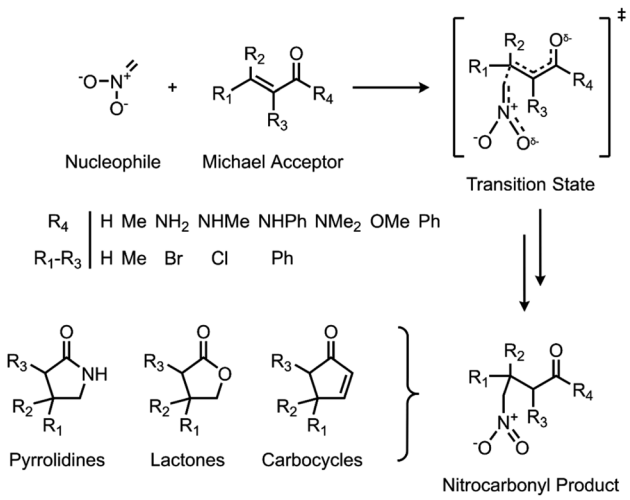

Fig. 1 C–C bond forming nitro-Michael additions used to generate the ML dataset.

Methodology

Reactant and TS geometries for 1000 unique Michael addition reactions were built using Schrödinger's R-Group enumeration[87] to vary four positions of a generic α,β-unsaturated carbonyl Michael acceptor (MA) core with common organic fragments across synthesis,[79] toxicology,[88,89] and covalent drug design (Fig. 1).[90] A set of 37 additional reactions were built using Michael acceptors (aldehydes, ketones, and esters) from the toxicology literature to be used for external validation (Fig. S2†).[91] Full details of dataset generation are provided in the ESI, section 1.†

All structures were conformationally searched using Schrödinger's MacroModel (version 12.7)[87,92] with the OPLS3e force field[93] before optimising the lowest energy conformation of each with AM1,[32] PM6,[35] and ωB97X-D/def2-TZVP[94,95] using Gaussian16 (Revision A.03).[36] Additionally, each MA was optimised with UFF for a comparison of SQM with classical force field methods.[96] To incorporate the effect of solvent, SPE corrections were performed with the same method as the optimisation and the integral equation formalism of the polarisable continuum model (IEFPCM)[97] with toluene. Temperature (298.15 K) and concentration-corrected (1 mol l^{-1}) quasi-harmonic free energies were calculated with GoodVibes[98] and used to calculate reaction barriers (AM1 barrier range: 7.83–42.38 kcal mol^{-1}; PM6 barrier range: 2.54–42.01 kcal mol^{-1}; DFT barrier range: 3.17–39.35 kcal mol^{-1}). Full computational details are provided in the ESI, section 1.†

A variety of simple and interpretable molecular and atomic physical organic chemical features were extracted for each MA and TS at each level of theory (Table S4 and Fig. S7†). Prior to fitting, all features were standardised and processed to deal with collinear and zero-variance features before dividing them into several distinct feature subsets per level of theory (Table S5†); MA features, TS features, and combined MA and TS features (including the reaction barrier) denoted by "All". Full details of feature extraction are provided in the ESI, section 2.†

The enumerated dataset was randomly split into an 80% train set (800 reactions) and 20% test set (200 reactions) and the former used to train each feature subset to predict the DFT free energy reaction barrier. Seven scikit-learn[99] regression algorithms were used for training: ridge regression (Ridge), *k*-nearest neighbour regression (NNR), random forest regression (RFR), gradient boosting regression (GBR), support vector regression (SVR), kernel ridge regression (KRR), and Gaussian process regression (GPR). Feature selection and hyperparameter tuning were employed using scikit-learn[99] and mlxtend[100] within the 80% train set to prevent overfitting of the feature subsets to the regression models and optimise the models parameters, respectively.[101] 5-Fold cross validation (CV) was performed within the train set to generate mean absolute errors (MAEs). To assess the individual model performances, external validation was performed using the unseen 20% test set to generate MAEs with standard errors. To further assess the generalisability of the models, MAE scores with standard errors, which are comparable across different sample sizes, were also calculated for the set of 37 unseen reactions from literature. However, among these 37 reactions, two (E5 and E7) contain alcohol groups within their R-groups that allow intra- and intermolecular hydrogen bonding to take place within their respective MA and TS geometries (Fig. S3†). As no such structures were present in the train set, the generated ML models cannot reasonably be expected to learn to account for hydrogen bonding. Indeed, for all models and feature subsets, reactions E5 and E7 were found to exhibit disproportionately worse predictions compared to the MAE of the other 35 structures; for example, absolute errors of 4.26 and 4.74 kcal mol^{-1} were obtained for E5 and E7, respectively, with GPR (AM1 All), compared to an MAE of 0.92 kcal mol^{-1} over the other 35 reactions. Thus, herein, the final literature set was defined as only the 35 structures other than E5 and E7. Full details of the ML process and analyses are provided in the ESI, sections 3 and 4.†

Results and discussion

The test MAEs and standard errors (from external validation with the 20% test set) for each model and feature subset are provided in Fig. 2. Full metrics, features, and hyperparameters for each model are provided in the ESI, section 5.† In general, train (5-fold CV) MAEs closely match the test MAEs, indicating that no significant overfitting takes place in the models.

Impressive results are achieved by all models, with test MAEs quenched below 2 kcal mol^{-1} for all feature subsets and each regressor. However, the performances of the kernel-based models (SVR, KRR, and GPR) are the most remarkable, with each producing MAEs below the accepted threshold for chemical accuracy of 1 kcal mol^{-1}.[64,65] Indeed, examples of SVR,[102,103] KRR,[67,104,105] and GPR[48,106] are prevalent in chemistry. These algorithms employ the Kernel trick, mapping the original input features into a higher dimensional feature space in a computationally efficient way and thus allowing the generation of highly complex models at a relatively low cost. In contrast, NNR is a conceptually simple method, and so the complex relationship between the input features and targets cannot be captured as easily. For the remaining models, the data presented does not appear to be of the most suitable form to deliver optimal performance. For example, ridge regression is a linear algorithm that relies on the features having strong linear correlations with the target barriers, which is often not the case in chemical systems;[46] indeed, only three features across all levels of theory (the AM1 barrier, and the percent buried volume (PBV) of the nucleophile carbon for AM1 and PM6) have linear correlations (Pearson's *r*) with the DFT barrier that are above 0.7, often quoted as the threshold for collinearity.[107] Similarly, the decision-tree-based models (RFR and GBR) generally lend themselves to the prediction of discrete data, such as reaction selectivities, rather than continuous variables such as reaction barriers.[62,66] However, despite their apparent unsuitability, each of these models still produces MAEs approaching chemical accuracy, demonstrating the overall success of our SQM to DFT ML approach.

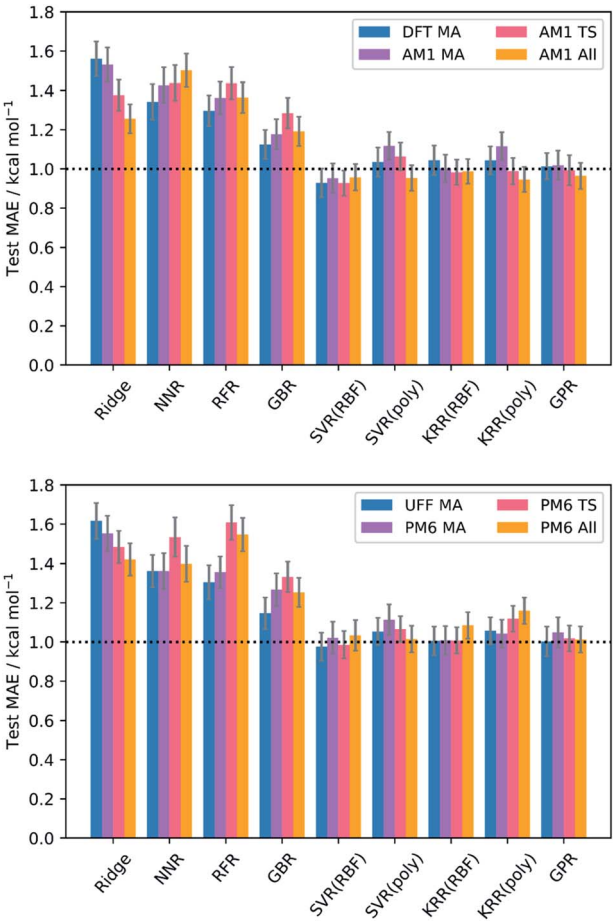

Fig. 2 Test MAEs and standard errors (20% test set) for each model and feature subset; RBF = radial basis function kernel; poly = polynomial kernel. GPR used the Matern kernel.

Evaluation of the average performances of the SVR, KRR, and GPR models (Table 1) reveals that all feature subsets produce excellent train and test MAEs. In fact, the predictive power of the MA-only and TS-only feature subsets were generally found to be comparable with the combined subsets. This is important

Table 1 Average MAEs of all SVR, KRR, and GPR models

	MAE/kcal mol^{-1}		
Feature subset	Train	Test	Literature
UFF MA	0.98	1.02	1.68
AM1 MA	0.99	1.04	1.44
AM1 TS	0.94	0.99	1.17
AM1 All	0.91	0.96	1.03
PM6 MA	0.98	1.05	1.43
PM6 TS	0.99	1.04	1.42
PM6 All	0.98	1.06	1.28
DFT MA	0.95	1.01	1.65

because conformationally searching and obtaining converged structures for reactant states is a more trivial task than for TSs, typically requiring less user input and computational expense. This gives ML users the opportunity to use purely MA-derived features and thus trade off a small amount of accuracy for what could be a substantial amount of time when learning or predicting over enough reactions. However, mechanistic insight from TSs would not be available in such an approach. Additionally, both the MA-only and TS-only feature subsets were found to perform relatively poorly for prediction of the literature structures, with TS features performing slightly better but still worse than their respective train and test metrics. This indicates that TS features are more important than MA features with respect to a model's ability to generalise. However, only when both MA and TS features are combined with the reaction barrier from the SQM method do literature predictions begin to approach the accuracy of the train and test metrics. By inclusion of the reaction barrier, these combined feature subsets represent a version of the Δ-ML approach, in which models are trained to learn the difference between the SQM and DFT barriers, rather than predicting the DFT barrier directly.[67] Indeed, Δ-ML has previously been found to perform well for out-of-sample predictions.[58] Therefore, the use of feature subsets without both MA and TS information, as well as the reaction barrier, is generally not recommended.

Comparing performances across the different levels of theory, the classical UFF method is found to perform similarly to the AM1, PM6, and DFT MA-only subsets for the train and test sets. However, with only MA information available, extension of UFF to the literature set is poor and mechanistic insight from TSs is not available. In addition to these same drawbacks, the DFT MA subset also suffers from the relatively high cost of the DFT calculations; on a 16-core node, the average DFT MA calculation took over an hour, compared to 5, 14, and 32 seconds with UFF, AM1, and PM6, respectively. Furthermore, DFT calculations scale poorly as the size of the chemical structures become larger. Thus, the use of either UFF or DFT MA features are not generally recommended.

Conceptually, AM1 and PM6 are very similar methods; both are based on the neglect of diatomic differential overlap (NDDO) formalism and share several fundamental approximations, although PM6 is more extensively parameterised and makes several improvements to the core–core potentials.[35] Accordingly, the initial MAE between the PM6 and DFT barriers (4.17 kcal mol^{-1}) was found to be slightly better than between the AM1 and DFT barriers (5.71 kcal mol^{-1}). Nevertheless, AM1 was found to perform marginally better after ML, particularly when making predictions on the literature dataset; overall, the combined AM1 MA and TS feature subset (AM1 All) performed the best across the train, test, and literature sets.

Overall, the best model obtained is *via* GPR using the combined AM1 MA and TS feature subset with 101 features (GPR (AM1 All)), yielding train, test, and literature MAEs of 0.93, 0.96 ± 0.07, and 0.92 ± 0.18 kcal mol^{-1}, respectively. Indeed, GPR was found to be the best model in several other studies for the prediction of thermochemical properties,[48,106] however, using the same feature subset, both KRR(RBF) and KRR(poly) also produced models with all MAEs below 1 kcal mol^{-1}. By plotting the AM1 and GPR (AM1 All)-predicted barriers against the DFT barriers, the improvement gained over the untrained model *via* our ML approach with respect to both the test set and literature set predictions can be visualised (Fig. 3).

To validate that the success of the GPR model is genuine and not due to any fortuitous test–train splitting, we performed an extensive double CV approach by retraining the model at five additional random test–train splittings (Fig. 4).[108] These

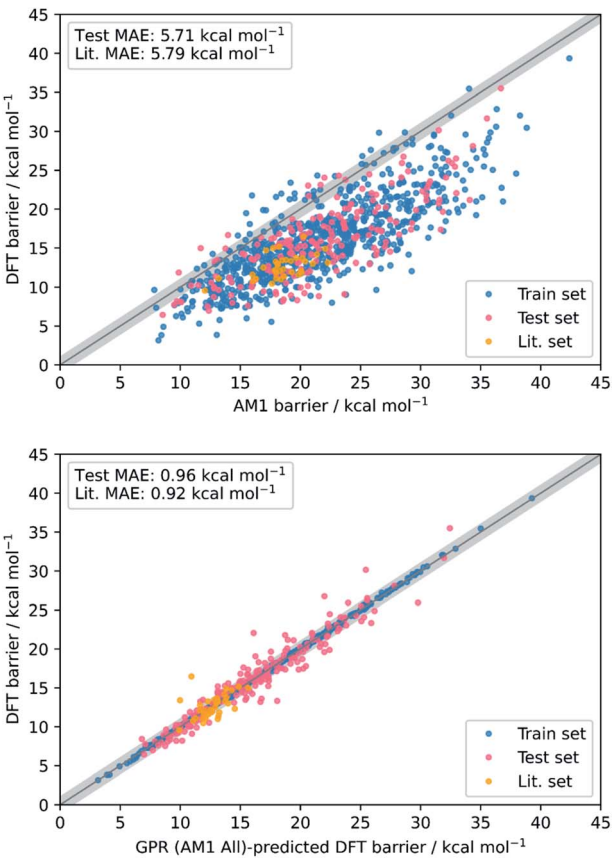

Fig. 3 Scatter plots showing the accuracy of DFT barrier prediction using only the AM1 barrier (top) compared to ML with GPR (AM1 All) (bottom), both with respect to the identity line (grey bands correspond to ±1 kcal mol^{-1}).

produced average train, test, and literature MAEs of 0.94, 0.89 ± 0.06, and 0.98 ± 0.18 kcal mol^{-1}, respectively, in line with the original metrics. Additionally, learning curves for the model (Fig. 5) flatten out as the number of training points is increased towards the maximum of 800, indicating that the size of the train set is acceptable and does not substantially limit the accuracy of the model or its ability to make predictions. Finally, the train and test scores tend to be very similar, indicating that no significant overfitting takes place in the model at any point. Learning curves for the SVR and KRR models with the AM1 All feature subset were found to display the same trends (see ESI, section 6†).

To ascertain which features make the largest contributions to determining the DFT barrier, as well as the overall generalisability of the models, their permutation feature importances were analysed. This process, in a typical ML study, can provide useful chemical insights into the mechanism of the reaction, depending on the interpretability of the highest-ranking features. For the GPR (AM1 All) model, the AM1 barrier was found to rank very highly in feature importance (Fig. 6).

The next most important feature, and the most important for the literature set, is the Mulliken charge of the carbonyl oxygen. The conjugated π-system of α,β-unsaturated carbonyls is soft and highly polarisable, whilst oxygen is a highly electronegative atom.[109] Thus, the charge of the oxygen is a good indicator of the total electron density in the π-system of each MA. Accordingly, as the charge of the oxygen becomes more negative, interaction between the MA and negatively charged nucleophile becomes

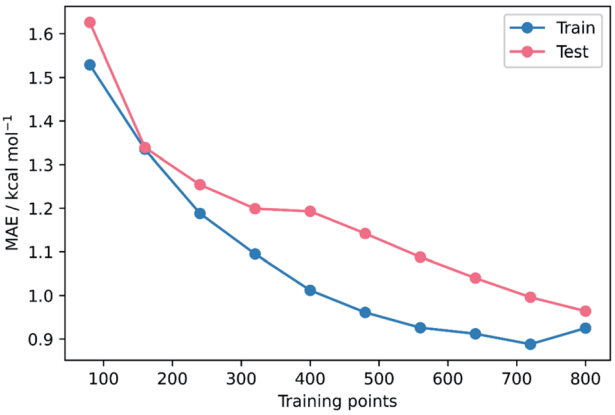

Fig. 5 MAE learning curves for GPR (AM1 All).

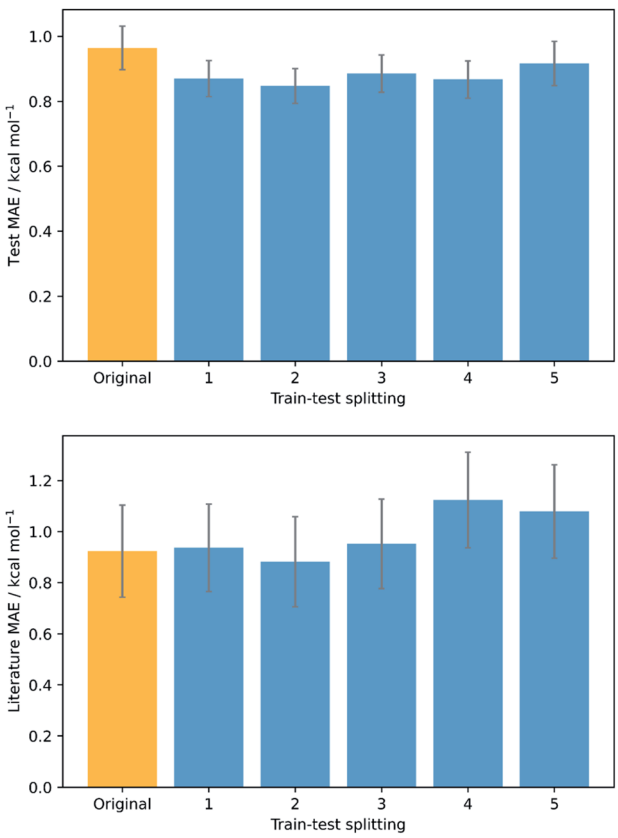

Fig. 4 Double cross validation; test (top) and literature (bottom) MAE and standard errors for GPR (AM1 All) at the original test–train splitting (yellow) and five additional splittings (blue).

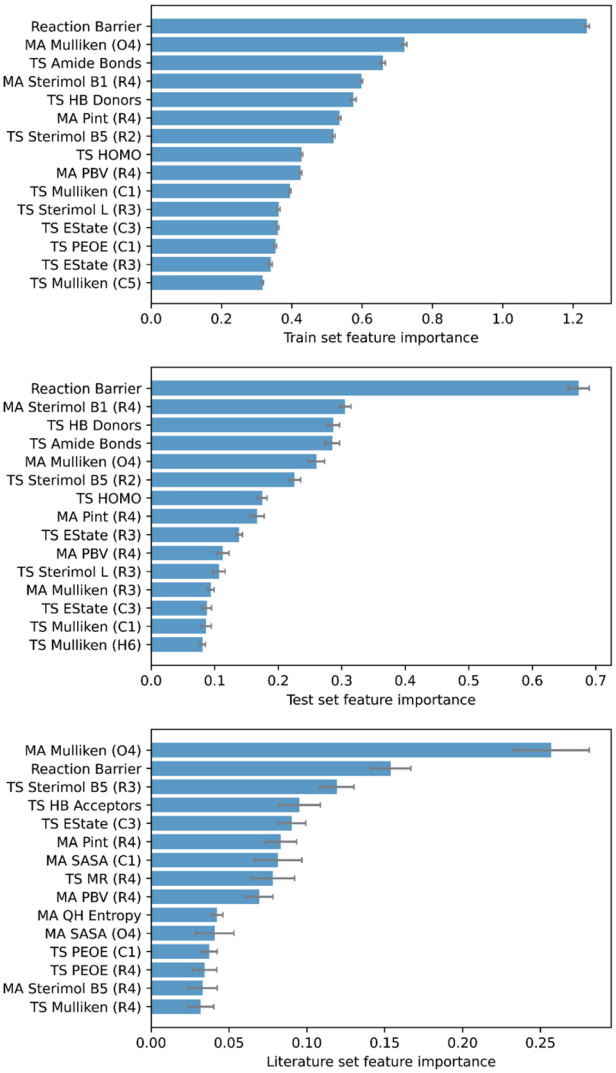

Fig. 6 Top 15 train, test, and literature set permutation features importances for GPR (AM1 All).

more difficult, and the reaction barrier increases. In addition, several other features encoding electrostatic information for specific atoms (largely the R_1–R_4 substituents and core Michael acceptor atoms, C_1–O_4) are important, including orbital electronegativities (PEOE), dispersion descriptors (P_{int}), and electrotopological characteristics (EState). Overall, these features account for a significant proportion of the electrostatic component of the nitro-Michael addition reaction.

Finally, several steric features, including sterimol parameters, PBVs, and solvent accessible surface areas (SASA), comprise a substantial proportion of the remaining high-importance features. This is in line with previous findings that a significant extent of steric control exists in the rate of Michael addition reactions with glutathione.[89,110] For example, steric parameters in the vicinity of the α,β-unsaturated carbonyl describe the steric accessibility of the β-carbon, where the nucleophile needs to be for 1,4-Michael addition to occur, and thus correlate very strongly with the reaction barrier.

The top-performing features of GPR models with subsets derived from PM6 calculations, in addition to each of the SVR and KRR models with AM1 features, were of a similar nature to GPR (AM1 All) (see ESI, section 7†), validating the importance of the features discussed. However, whilst chemical insights can be drawn from these features, allowing a subsequent understanding of their impact on the reaction mechanism, one of the unique benefits of our ML approach is that full SQM TS geometries are produced as part of the feature generation process. In turn, this allows mechanistic insights into the reaction to be obtained *via* visualisation of the geometries, without the need to analyse individual feature importances. But how accurate are these SQM geometries, and can they be used as approximations for the DFT TS geometries?

To test this, we calculated the root-mean-squared deviation of atomic positions (RMSD)[111] and difference in bond-forming distance between each of the TS geometries at the AM1 and DFT levels of theory (Fig. 7). In the context of molecular docking, an RMSD below 2 Å is considered successful when comparing the conformations of organic ligands to their protein-bound conformation.[112] The average RMSD for our enumerated dataset of 1000 structures was calculated at 0.75 Å, with 99.2% of TSs falling below the 2 Å threshold. The bond-forming distance is on average 0.04 Å larger in the DFT structure than in the AM1 structure, with 96.4% of TSs having an absolute difference in bond-forming distance below 0.3 Å. Fig. 8 depicts the AM1 and DFT geometries of the TS with the RMSD closest to 0.75 Å, and hence represents the approximate average deviation that would be expected between an AM1 and DFT geometry in our dataset. Close inspection of the structures reveals that the major origin of deviation results from the angle of approach of the nucleophile and the orientation of R-groups, rather than any changes to the core structure of the MA. Accordingly, removing the nucleophile or the R-groups from each structure and recalculating the RMSDs drops the averages to 0.6 Å and 0.45 Å, respectively, whilst removing both (leaving only the core enone functionality) drops it to 0.14 Å (Fig. S20–22†). In fact, even when the AM1 MA geometries were compared to the DFT TS geometries with the nucleophile removed, an

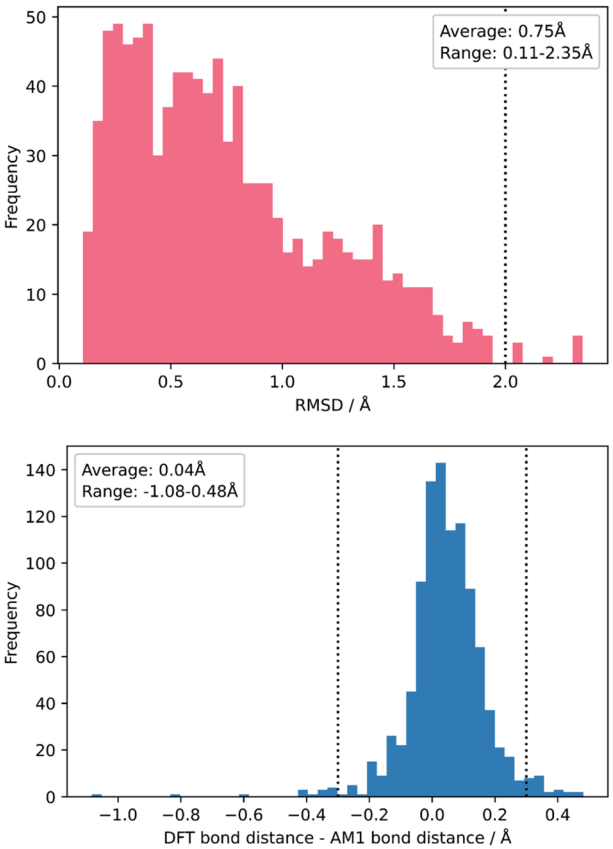

Fig. 7 Distribution of RMSDs (top) and bond distance differences (bottom) between each TS in the enumerated dataset at its AM1 and DFT geometries.

average RMSD of 1.35 Å was calculated, substantially below the 2 Å threshold (Fig. S23†). In all cases, similar distributions and average values were also calculated for the literature structures and at each level of theory (see ESI, section 8†); notably, UFF and DFT MA geometries were similarly comparable with DFT-derived TS geometries (nucleophile removed), whilst PM6 was found be slightly worse at predicting DFT TS geometries than AM1, with a larger average RMSD of 0.87 Å.

The inclusion of solvent *via* SPE corrections results in highly flexible models that allow different solvents to be incorporated without reoptimising every structure. However, to examine the impact of solvent, AM1, PM6, and DFT TS optimisations with the IEFPCM solvent model were performed on the literature set.

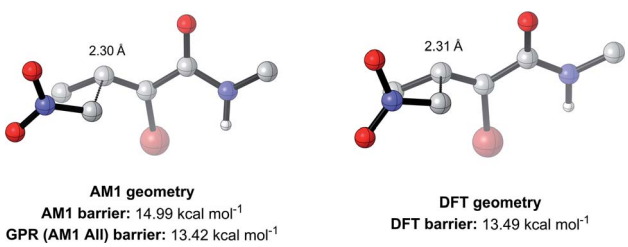

Fig. 8 AM1 and DFT geometries of the enumerated TS with the RMSD closest to the average of 0.75 Å (hydrogens omitted for clarity).

We then superimposed each pair of gas and solvent optimised TSs and calculated average RMSDs of 0.13 Å for AM1, 0.30 Å for PM6, and 0.16 Å for DFT (Fig. S27 and S28†). These low RMSDs indicate that, for the reaction investigated in our work, the use of gas phase optimised structures is a valid approximation.

Overall, these analyses indicate that SQM geometries of TSs, and even MAs to some extent, can be considered good substitutes for the full DFT level geometries, on average, allowing accurate mechanistic analysis of the TSs simply by their visualisation. Importantly, no other ML barrier model that we are aware of offers this level of mechanistic insight without the need for time-consuming DFT calculations. For example, analysis of one TS structure at the DFT level reveals how severe C–C steric interactions between the nucleophile and the R_1 and R_2 groups of the MA destabilise the structure, and these interactions are also captured by the AM1 geometry (Fig. 9). Whilst insights into the electronics and sterics of the reaction were revealed by analysis of the most important predictive features, direct visualisation of the geometries and interactions improves upon this approach by identifying which groups are directly involved in this steric clash and revealing more about their exact nature. Such information helps to validate the predictions made by ML and can guide rational reaction design.

For the nitro-Michael addition reaction investigated here, the full DFT level of theory took approximately 7 hours on a 16-core node to obtain each of the 1000 enumerated MA and TS structures, compared to only 51 seconds for the respective AM1 calculations (Table S3†). On a 1-core laptop, this corresponds to over 100 hours of calculation for DFT per structure, compared to less than 15 minutes for AM1. Thus, the ease and efficiency of our SQM/ML approach is demonstrated; with a prebuilt ML model, the user simply calculates a baseline barrier for a reaction (in seconds, using widely available SQM or force field approaches), applies a correction *via* ML (in seconds), and accurate DFT-level barriers and geometries can be obtained. Although users may be concerned that the deficiencies of a particular baseline method, for example poor applicability to a particular chemical domain (see introduction), may inhibit a model's ability to make accurate predictions, our results demonstrate that feature subsets derived from both classical force field and SQM methods all lead to comparable predictive performances. Thus, the user may simply select an appropriate baseline method for the reaction in question and, by the same principles described above, accurate predictions should be possible. For a more detailed account of the applicability of SQM methods in modelling organic chemistry, we recommend a review by Thiel.[113]

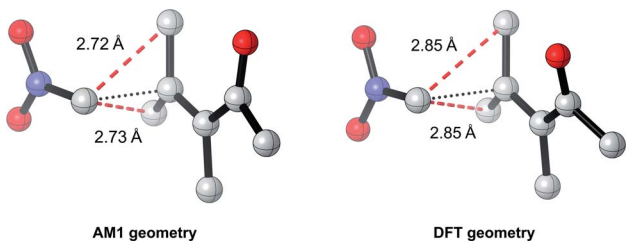

Fig. 9 AM1 and DFT geometries of an example TS revealing steric interactions between the nucleophile and the R_1 (Me) and R_2 (Me) groups of the MA; C–C distances within 90% of the sum of the van der Waals radii (3.4 Å) are indicated by dotted red lines.

Overall, the combination of highly accurate SQM geometries and ML-derived energies represents a significantly cheaper way to obtain very good approximations of DFT-level geometries and energies. In turn, this enables the rapid prediction of reaction barriers and delivers mechanistic insight for this essential class of nitro-Michael additions, which could lead to much faster screening of these kinds of reactions, and thus much more efficient design of new synthetic methodology.

Conclusion

We have combined ML and SQM calculations to achieve the fast and accurate prediction of DFT-quality free energy activation barriers using widely available computational techniques. Using a variety of regression algorithms with simple and highly interpretable features, MAEs below the accepted chemical accuracy threshold of 1 kcal mol^{-1} were achieved with a calculation time of seconds, even when making predictions on an unseen set of compounds from the toxicology literature. Evaluation of the most predictive features provided clear insights into important aspects of the nitro-Michael addition reaction mechanism. However, SQM geometries of TSs, and to some extent MAs, were found to be very good approximations to the full DFT TS geometries and thus offer mechanistic insight with no additional work required. Combination of these SQM geometries with highly accurate ML-derived energies allows the prediction of barriers and the screening of reactions at DFT level, without the need for time-consuming DFT calculations. No current ML barrier models offer our combination of speed, accuracy, and mechanistic insight. The generalised nature of the study means the ML approach can be highly customised, for example by choosing from various regression algorithms, features, and molecular modelling methods. We believe that the same principles could also be applied to achieve the rapid prediction of reaction barriers and mechanisms for other important classes of chemical reactions, paving the way for more efficient drug discovery and rational reaction design.

Data availability

Gaussian16 output files for all computed structures are openly available in Dataset for "Machine learning and semi-empirical calculations: a synergistic approach to rapid, accurate, and mechanism-based reaction barrier prediction" in the University of Bath Research Data Archive at **https://doi.org/10.15125/BATH-01092**.

Author contributions

This manuscript was written through contributions from all authors.

Conflicts of interest

There are no conflicts to declare.

Acknowledgements

This research made use of the Balena High Performance Computing (HPC) Service at the University of Bath. The authors thank the EPSRC (studentship to E. H. E. F., grant number EP/R513155/1) and the University of Bath for funding. Dr Florian Roessler and Dr Natalie Fey are thanked for their help and discussions.

Notes and references

1 Y. H. Lam, M. N. Grayson, M. C. Holland, A. Simon and K. N. Houk, *Acc. Chem. Res.*, 2016, **49**, 750–762.
2 K. N. Houk, M. N. Paddon-Row, N. G. Rondan, Y. D. Wu, F. K. Brown, D. C. Spellmeyer, J. T. Metz, Y. Li and R. J. Loncharich, *Science*, 1986, **231**, 1108–1117.
3 E. H. E. Farrar and M. N. Grayson, *J. Org. Chem.*, 2020, **85**, 15449–15456.
4 M. N. Grayson, S. C. Pellegrinet and J. M. Goodman, *J. Am. Chem. Soc.*, 2012, **134**, 2716–2722.
5 M. N. Grayson, *J. Org. Chem.*, 2017, **82**, 4396–4401.
6 P. H.-Y. Cheong, C. Y. Legault, J. M. Um, N. Çelebi-Ölçüm and K. N. Houk, *Chem. Rev.*, 2011, **111**, 5042–5137.
7 J. K. Nørskov, F. Abild-Pedersen, F. Studt and T. Bligaard, *Proc. Natl. Acad. Sci. U. S. A.*, 2011, **108**, 937–943.
8 I. Konstantinov, S. Ewart, H. Brown, C. Eddy, J. Mendenhall and S. Munjal, *Mol. Syst. Des. Eng.*, 2018, **3**, 228–242.
9 X. Du, X. Gao, W. Hu, J. Yu, Z. Luo and K. Cen, *J. Phys. Chem. C*, 2014, **118**, 13617–13622.
10 Y. Guan, V. M. Ingman, B. J. Rooks and S. E. Wheeler, *J. Chem. Theory Comput.*, 2018, **14**, 5249–5261.
11 L. D. Jacobson, A. D. Bochevarov, M. A. Watson, T. F. Hughes, D. Rinaldo, S. Ehrlich, T. B. Steinbrecher, S. Vaitheeswaran, D. M. Philipp, M. D. Halls and R. A. Friesner, *J. Chem. Theory Comput.*, 2017, **13**, 5780–5797.
12 T. A. Young, J. J. Silcock, A. J. Sterling and F. Duarte, *Angew. Chem., Int. Ed.*, 2021, **60**, 4266–4274.
13 R. E. Plata and D. A. Singleton, *J. Am. Chem. Soc.*, 2015, **137**, 3811–3826.
14 M. Linder and T. Brinck, *Phys. Chem. Chem. Phys.*, 2013, **15**, 5108–5114.
15 L. Goerigk, A. Hansen, C. Bauer, S. Ehrlich, A. Najibi and S. Grimme, *Phys. Chem. Chem. Phys.*, 2017, **19**, 32184–32215.
16 Y. P. Chin and E. H. Krenske, *J. Org. Chem.*, 2022, **87**, 1710–1722.
17 P. Hohenberg and W. Kohn, *Phys. Rev.*, 1964, **136**, B864–B871.
18 W. Kohn and L. J. Sham, *Phys. Rev.*, 1965, **140**, A1133–A1138.
19 A. D. Becke, *J. Chem. Phys.*, 2014, **140**(18), 18A301.
20 Q. Cui and M. Elstner, *Phys. Chem. Chem. Phys.*, 2014, **16**, 14368–14377.
21 A. Warshel and R. M. Weiss, *J. Am. Chem. Soc.*, 1980, **102**, 6218–6226.
22 P. Grochowski, B. Lesyng, P. Bała and J. A. McCammon, *Int. J. Quantum Chem.*, 1996, **60**, 1143–1164.
23 A. C. T. Van Duin, S. Dasgupta, F. Lorant and W. A. Goddard, *J. Phys. Chem. A*, 2001, **105**, 9396–9409.
24 F. Jensen, *J. Am. Chem. Soc.*, 1992, **114**, 1596–1603.
25 J. E. Eksterowicz and K. N. Houk, *Chem. Rev.*, 1993, **93**, 2439–2461.
26 F. Jensen, *J. Chem. Phys.*, 2003, **119**, 8804–8808.
27 Á. Madarász, D. Berta and R. S. Paton, *J. Chem. Theory Comput.*, 2016, **12**, 1833–1844.
28 E. Hansen, A. R. Rosales, B. Tutkowski, P. O. Norrby and O. Wiest, *Acc. Chem. Res.*, 2016, **49**, 996–1005.
29 A. R. Rosales, T. R. Quinn, J. Wahlers, A. Tomberg, X. Zhang, P. Helquist, O. Wiest and P. O. Norrby, *Chem. Commun.*, 2018, **54**, 8294–8311.
30 J. S. Smith, O. Isayev and A. E. Roitberg, *Chem. Sci.*, 2017, **8**, 3192–3203.
31 N. Weill, C. R. Corbeil, J. W. De Schutter and N. Moitessier, *J. Comput. Chem.*, 2011, **32**, 2878–2889.
32 M. J. S. Dewar, E. G. Zoebisch, E. F. Healy and J. J. P. Stewart, *J. Am. Chem. Soc.*, 1985, **107**, 3902–3909.
33 G. B. Rocha, R. O. Freire, A. M. Simas and J. J. P. Stewart, *J. Comput. Chem.*, 2006, **27**, 1101–1111.
34 J. J. P. Stewart, *J. Mol. Model.*, 2013, **19**, 1–32.
35 J. J. P. Stewart, *J. Mol. Model.*, 2007, **13**, 1173–1213.
36 M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, G. Scalmani, V. Barone, B. Mennucci, G. A. Petersson, H. Nakatsuji, M. Caricato, X. Li, H. P. Hratchian, A. F. Izmaylov, J. Bloino, J. Zheng, J. L. Sonnenberg, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, T. Vreven, J. A. Montgomery, J. E. Peralta, F. Ogliaro, M. Bearpark, J. J. Heyd, E. Brothers, K. N. Kudin, V. N. Staroverov, R. Kobayashi, J. Normand, K. Raghavachari, J. C. A. Rendell, S. Burant, S. Iyengar, J. Tomasi, M. Cossi, N. Rega, J. M. Millam, M. Klene, J. E. Knox, J. B. Cross, V. Bakken, C. Adamo, J. Jaramillo, R. Gomperts, R. E. Stratmann, O. Yazyev, A. J. Austin, R. Cammi, C. Pomelli, J. W. Ochterski, R. L. Martin, K. Morokuma, V. G. Zakrzewski, G. A. Voth, P. Salvador, J. J. Dannenberg, S. Dapprich, A. D. Daniels, O. Farkas, J. B. Foresman, J. V. Ortiz, J. Cioslowski and D. J. Fox, *Gaussian16, Revision A.03*, 2016.
37 G. M. J. Barca, C. Bertoni, L. Carrington, D. Datta, N. De Silva, J. E. Deustua, D. G. Fedorov, J. R. Gour, A. O. Gunina, E. Guidez, T. Harville, S. Irle, J. Ivanic, K. Kowalski, S. S. Leang, H. Li, W. Li, J. J. Lutz, I. Magoulas, J. Mato, V. Mironov, H. Nakata, B. Q. Pham, P. Piecuch, D. Poole, S. R. Pruitt, A. P. Rendell, L. B. Roskop, K. Ruedenberg, T. Sattasathuchana, M. W. Schmidt, J. Shen, L. Slipchenko, M. Sosonkina, V. Sundriyal, A. Tiwari, J. L. Galvez Vallejo, B. Westheimer, M. Włoch, P. Xu, F. Zahariev and M. S. Gordon, *J. Chem. Phys.*, 2020, **152**, 154102.
38 *Spartan*, Wavefunction, Inc., 18401, Von Karman Avenue, Suite 370, Irvine CA 92612 USA.

39 J. J. P. Stewart, *J. Comput.-Aided Mol. Des.*, 1990, **4**, 1–105.
40 F. Neese, F. Wennmohs, U. Becker and C. Riplinger, *J. Chem. Phys.*, 2020, **152**, 224108.
41 A. S. Christensen, T. Kubař, Q. Cui and M. Elstner, *Chem. Rev.*, 2016, **116**, 5301–5337.
42 P. Caramella, K. N. Houk and L. N. Domelsmith, *J. Am. Chem. Soc.*, 1977, **99**, 4511–4514.
43 M. Gruden, L. Andjeklović, A. K. Jissy, S. Stepanović, M. Zlatar, Q. Cui and M. Elstner, *J. Comput. Chem.*, 2017, **38**, 2171–2185.
44 M. H. Rasmussen and J. H. Jensen, *PeerJ Phys. Chem.*, 2020, **2**, e15.
45 T. Saito and Y. Takano, *Bull. Chem. Soc. Jpn.*, 2018, **91**, 1377–1389.
46 F. Strieth-Kalthoff, F. Sandfort, M. H. S. Segler and F. Glorius, *Chem. Soc. Rev.*, 2020, **49**, 6154–6168.
47 C. B. Santiago, J. Y. Guo and M. S. Sigman, *Chem. Sci.*, 2018, **9**, 2398–2412.
48 K. Jorner, T. Brinck, P. O. Norrby and D. Buttar, *Chem. Sci.*, 2021, **12**, 1163–1175.
49 S. Choi, Y. Kim, J. W. Kim, Z. Kim and W. Y. Kim, *Chem.–Eur. J.*, 2018, **24**, 12354–12358.
50 F. Palazzesi, M. R. Hermann, M. A. Grundl, A. Pautsch, D. Seeliger, C. S. Tautermann and A. Weber, *J. Chem. Inf. Model.*, 2020, **60**, 2915–2923.
51 E. Heid and W. H. Green, *J. Chem. Inf. Model.*, 2022, **62**, 2101–2110.
52 T. Stuyver and C. W. Coley, *J. Chem. Phys.*, 2022, **156**, 084104.
53 M. Glavatskikh, T. Madzhidov, D. Horvath, R. Nugmanov, T. Gimadiev, D. Malakhova, G. Marcou and A. Varnek, *Mol. Inf.*, 2019, **38**, 1–8.
54 A. R. Singh, B. A. Rohr, J. A. Gauthier and J. K. Nørskov, *Catal. Lett.*, 2019, **149**, 2347–2354.
55 K. Mikami, *Polymer*, 2020, **203**, 122738.
56 C. A. Grambow, L. Pattanaik and W. H. Green, *J. Phys. Chem. Lett.*, 2020, **11**, 2992–2997.
57 S. Wannakao, N. Artrith, J. Limtrakul and A. M. Kolpak, *J. Phys. Chem. C*, 2017, **121**, 20306–20314.
58 M. Bragato, G. F. Von Rudorff and O. A. Von Lilienfeld, *Chem. Sci.*, 2020, **11**, 11859–11868.
59 K. Takahashi and I. Miyazato, *J. Comput. Chem.*, 2018, **39**, 2405–2408.
60 S. Heinen, G. F. von Rudorff and O. A. von Lilienfeld, *J. Chem. Phys.*, 2021, **155**, 064105.
61 I. Migliaro and T. R. Cundari, *J. Chem. Inf. Model.*, 2020, **60**, 4958–4966.
62 X. Li, S. Q. Zhang, L. C. Xu and X. Hong, *Angew. Chem., Int. Ed.*, 2020, **59**, 13253–13259.
63 T. Bučko, M. Gešvandtnerová and D. Rocca, *J. Chem. Theory Comput.*, 2020, **16**, 6049–6060.
64 K. N. Houk and F. Liu, *Acc. Chem. Res.*, 2017, **50**, 539–543.
65 K. A. Peterson, D. Feller and D. A. Dixon, *Theor. Chem. Acc.*, 2012, **131**, 1–20.
66 S. M. Maley, D. H. Kwon, N. Rollins, J. C. Stanley, O. L. Sydora, S. M. Bischof and D. H. Ess, *Chem. Sci.*, 2020, **11**, 9665–9674.
67 R. Ramakrishnan, P. O. Dral, M. Rupp and O. A. Von Lilienfeld, *J. Chem. Theory Comput.*, 2015, **11**, 2087–2096.
68 P. Zhang, L. Shen and W. Yang, *J. Phys. Chem. B*, 2019, **123**, 901–908.
69 P. Zheng, R. Zubatyuk, W. Wu, O. Isayev and P. O. Dral, *Nat. Commun.*, 2021, **12**, 1–13.
70 M. Stöhr, L. Medrano Sandonas and A. Tkatchenko, *J. Phys. Chem. Lett.*, 2020, **11**, 6835–6843.
71 S. Wengert, G. Csányi, K. Reuter and J. T. Margraf, *Chem. Sci.*, 2021, **12**, 4536–4546.
72 A. Miyanaga, *Nat. Prod. Rep.*, 2019, **36**, 531–547.
73 T. Das, S. Mohapatra, N. P. Mishra, S. Nayak and B. P. Raiguru, *ChemistrySelect*, 2021, **6**, 3745–3781.
74 M. Heravi, P. Hajiabbasi and H. Hamidi, *Curr. Org. Chem.*, 2014, **18**, 489–511.
75 S. J. Connon, *Chem. Commun.*, 2008, 2499.
76 L. R. Zhong and Z. J. Yao, *Sci. China Chem.*, 2016, **59**, 1079–1087.
77 C. Hui, F. Pu and J. Xu, *Chem.–Eur. J.*, 2017, **23**, 4023–4036.
78 D. Ni, Y. Wei and D. Ma, *Angew. Chem., Int. Ed.*, 2018, **57**, 10207–10211.
79 Q. Gu and S. L. You, *Chem. Sci.*, 2011, **2**, 1519–1522.
80 Y. Wang, P. Li, X. Liang, T. Y. Zhang and J. Ye, *Chem. Commun.*, 2008, 1232.
81 L. Hojabri, A. Hartikka, F. M. Moghaddam and P. I. Arvidsson, *Adv. Synth. Catal.*, 2007, **349**, 740–748.
82 H. Gotoh, D. Okamura, H. Lshikawa and Y. Hayashl, *Org. Lett.*, 2009, **11**, 4056–4059.
83 W. Yang and D. M. Du, *Org. Lett.*, 2010, **12**, 5450–5453.
84 B. Vakulya, S. Varga, A. Csámpai and T. Soós, *Org. Lett.*, 2005, **7**, 1967–1969.
85 A. Y. Sukhorukov, A. A. Sukhanova and S. G. Zlotin, *Tetrahedron*, 2016, **72**, 6191–6281.
86 R. Ballini, G. Bosica, D. Fiorini, A. Palmieri and M. Petrini, *Chem. Rev.*, 2005, **105**, 933–971.
87 *Schrödinger Release 2020-1*, Maestro, Schrödinger, LLC, New York, NY, 2020.
88 A. Pérez-Garrido, A. M. Helguera, F. G. Rodríguez and M. N. D. S. Cordeiro, *Dent. Mater.*, 2010, **26**, 397–415.
89 J. A. H. Schwöbel, D. Wondrousch, Y. K. Koleva, J. C. Madden, M. T. D. Cronin and G. Schüürmann, *Chem. Res. Toxicol.*, 2010, **23**, 1576–1585.
90 P. A. Jackson, J. C. Widen, D. A. Harki and K. M. Brummond, *J. Med. Chem.*, 2017, **60**, 839–885.
91 A. Böhme, A. Laqua and G. Schüürmann, *Chem. Res. Toxicol.*, 2016, **29**, 952–962.
92 F. Mohamadi, N. G. J. Richards, W. C. Guida, R. Liskamp, M. Lipton, C. Caufield, G. Chang, T. Hendrickson and W. C. Still, *J. Comput. Chem.*, 1990, **11**, 440–467.
93 K. Roos, C. Wu, W. Damm, M. Reboul, J. M. Stevenson, C. Lu, M. K. Dahlgren, S. Mondal, W. Chen, L. Wang, R. Abel, R. A. Friesner and E. D. Harder, *J. Chem. Theory Comput.*, 2019, **15**, 1863–1874.
94 J.-D. Chai and M. Head-Gordon, *Phys. Chem. Chem. Phys.*, 2008, **10**, 6615.
95 F. Weigend and R. Ahlrichs, *Phys. Chem. Chem. Phys.*, 2005, **7**, 3297.

96 A. K. Rappé, C. J. Casewit, K. S. Colwell, W. A. Goddard and W. M. Skiff, *J. Am. Chem. Soc.*, 1992, **114**, 10024–10035.
97 B. Mennucci, R. Cammi and J. Tomasi, *J. Chem. Phys.*, 1998, **109**, 2798–2807.
98 G. Luchini, J. V. Alegre-Requena, I. Funes-Ardoiz and R. S. Paton, *F1000Research*, 2020, **9**, 291.
99 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 É. Duchesnay, *J. Mach. Learn. Res.*, 2011, **12**, 2825–2830.
100 S. Raschka, *J. Open Source Software*, 2018, **3**, 638.
101 D. M. Hawkins, *J. Chem. Inf. Comput. Sci.*, 2004, **44**, 1–12.
102 M. Fernandez, N. R. Trefiak and T. K. Woo, *J. Phys. Chem. C*, 2013, **117**(27), 14095–14105.
103 R. M. Balabin and E. I. Lomakina, *Phys. Chem. Chem. Phys.*, 2011, **13**, 11710–11718.
104 B. Meyer, B. Sawatlon, S. Heinen, O. A. Von Lilienfeld and C. Corminboeuf, *Chem. Sci.*, 2018, **9**, 7069–7077.
105 J. P. Janet and H. J. Kulik, *J. Phys. Chem. A*, 2017, **121**, 8939–8954.
106 P. Friederich, G. Dos Passos Gomes, R. De Bin, A. Aspuru-Guzik and D. Balcells, *Chem. Sci.*, 2020, **11**, 4584–4601.
107 C. F. Dormann, J. Elith, S. Bacher, C. Buchmann, G. Carl, G. Carré, J. R. G. Marquéz, B. Gruber, B. Lafourcade, P. J. Leitão, T. Münkemüller, C. McClean, P. E. Osborne, B. Reineking, B. Schröder, A. K. Skidmore, D. Zurell and S. Lautenbach, *Ecography*, 2013, **36**, 27–46.
108 I. Tsamardinos, A. Rakhshani and V. Lagani, *Int. J. Artif. Intell. Tool.*, 2015, **24**, 1540023.
109 R. M. LoPachin, T. Gavin, A. DeCaprio and D. S. Barber, *Chem. Res. Toxicol.*, 2012, **25**, 239–251.
110 J. A. H. Schwöbel, J. C. Madden and M. T. D. Cronin, *SAR QSAR Environ. Res.*, 2010, **21**, 693–710.
111 D. L. Theobald, *Acta Crystallogr., Sect. A: Found. Crystallogr.*, 2005, **61**, 478–480.
112 H. Gohlke, M. Hendlich and G. Klebe, *J. Mol. Biol.*, 2000, **295**, 337–356.
113 W. Thiel, *Wiley Interdiscip. Rev.: Comput. Mol. Sci.*, 2014, **4**, 145–157.