Chemical Science

ROYAL SOCIETY OF CHEMISTRY

EDGE ARTICLE

View Article Online
View Journal | View Issue

Cite this: *Chem. Sci.*, 2021, **12**, 1163

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

Machine learning meets mechanistic modelling for accurate prediction of experimental activation energies†

Kjell Jorner, [a] Tore Brinck, [b] Per-Ola Norrby [c] and David Buttar *[a]

Accurate prediction of chemical reactions in solution is challenging for current state-of-the-art approaches based on transition state modelling with density functional theory. Models based on machine learning have emerged as a promising alternative to address these problems, but these models currently lack the precision to give crucial information on the magnitude of barrier heights, influence of solvents and catalysts and extent of regio- and chemoselectivity. Here, we construct hybrid models which combine the traditional transition state modelling and machine learning to accurately predict reaction barriers. We train a Gaussian Process Regression model to reproduce high-quality experimental kinetic data for the nucleophilic aromatic substitution reaction and use it to predict barriers with a mean absolute error of 0.77 kcal mol^{-1} for an external test set. The model was further validated on regio- and chemoselectivity prediction on patent reaction data and achieved a competitive top-1 accuracy of 86%, despite not being trained explicitly for this task. Importantly, the model gives error bars for its predictions that can be used for risk assessment by the end user. Hybrid models emerge as the preferred alternative for accurate reaction prediction in the very common low-data situation where only 100–150 rate constants are available for a reaction class. With recent advances in deep learning for quickly predicting barriers and transition state geometries from density functional theory, we envision that hybrid models will soon become a standard alternative to complement current machine learning approaches based on ground-state physical organic descriptors or structural information such as molecular graphs or fingerprints.

Received 4th September 2020
Accepted 2nd November 2020

DOI: 10.1039/d0sc04896h

rsc.li/chemical-science

Introduction

Accurate prediction of chemical reactions is an important goal both in academic and industrial research.[1–3] Recently, machine learning approaches have had tremendous success in quantitative prediction of reaction yields based on data from high-throughput experimentation[4,5] and enantioselectivities based on carefully selected universal training sets.[6] At the same time, traditional quantitative structure–reactivity relationship (QSRR) methods based on linear regression have seen a renaissance with interpretable, holistic models that can generalize across reaction types.[7] In parallel with these developments of quantitative prediction methods, deep learning models trained on reaction databases containing millions of patent and literature data have made quick qualitative yes/no feasibility prediction routine for almost any reaction type.[8]

In the pharmaceutical industry, prediction tools have great potential to accelerate synthesis of prospective drugs (Fig. 1a).[9] Quick prediction is essential in the discovery phase, especially within the context of automation and rapid synthesis of a multitude of candidates for initial activity screening.[3,10,11] In these circumstances, a simple yes/no as provided by classification models is usually sufficient. More accurate prediction is necessary in the later drug development process, where the synthesis route and formulation of one or a few promising drug candidates is optimized. Here, regression models that give the reaction activation energy can be used to predict both absolute reactivity and selectivity (Fig. 1b). Prediction of absolute reactivity can be used to assess feasibility under process-relevant conditions, while prediction of selectivity is key to reducing purification steps. Predictive tools therefore hold great promise for accelerating route and process development, ultimately delivering medicines to patients both faster and at lower costs.

The current workhorse for computational studies of organic reactions is density functional theory (DFT, Fig. 2a). Since rising to prominence in the early 90s, DFT has enjoyed extraordinary success in rationalizing reactivity and selectivity across the reaction spectrum by modelling the full reaction mechanism.[12]

[a]*Early Chemical Development, Pharmaceutical Sciences, R&D, AstraZeneca, Macclesfield, UK. E-mail: david.buttar@astrazeneca.com*

[b]*Applied Physical Chemistry, Department of Chemistry, CBH, KTH Royal Institute of Technology, Stockholm, Sweden*

[c]*Data Science & Modelling, Pharmaceutical Sciences, R&D, AstraZeneca, Gothenburg, Sweden*

† Electronic supplementary information (ESI) available: Detailed computational methods, full description of machine learning workflow and results. Description of the experimental and computed datasets. See DOI: 10.1039/d0sc04896h

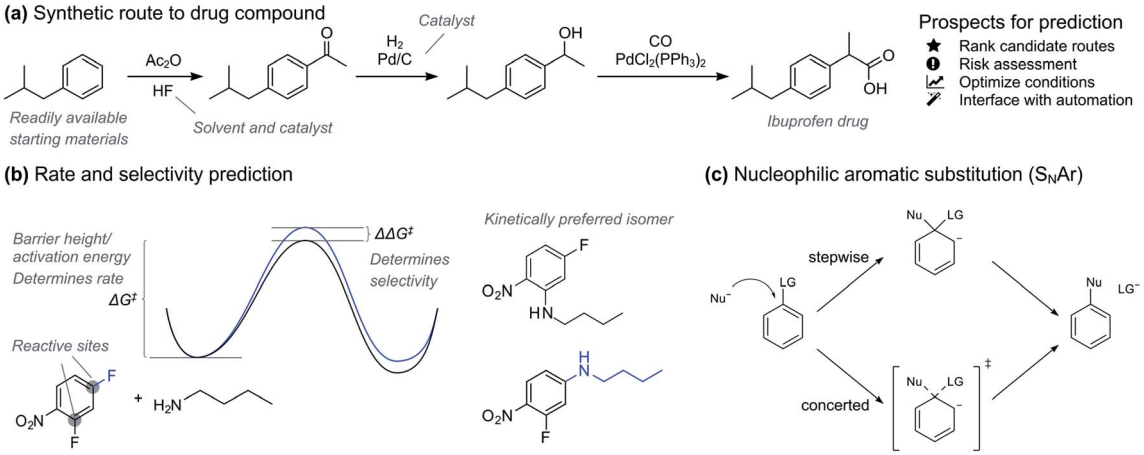

Fig. 1 (a) Example of synthetic route to a drug compound. Prospects for AI-assisted route design. (b) Accurate prediction of reaction barriers gives both rate and selectivity. (c) The nucleophilic aromatic substitution (S_NAr) reaction.

The success of DFT can be traced in part due to a fortuitous cancellation of errors, which makes it particularly suited for properties such as enantioselectivity, which depends on the relative energies of two structurally very similar transition states (TSs). However, this cancellation of errors does not generally extend to the prediction of the absolute magnitude of reactions barriers (activation free energies, $\Delta G^{\ddagger}$). In particular, DFT struggles with one very important class of reactions: ionic reactions in solution. Plata and Singleton even suggested that computed mechanisms of this type can be so flawed that they are "not even wrong".[13] Similarly, Maseras and co-workers only achieved agreement with experiment for the simple condensation of an imine and an aldehyde in water by introducing an *ad hoc* correction factor, even when using more accurate methods than DFT.[14] These results point to the fact that the largest error in the DFT simulations is often due to the poor performance of the solvation model.

Machine learning represents a potential solution to the problems of DFT. Based on reaction data in different solvents, machine learning models could in principle learn to compensate for both the deficiencies in the DFT energies and the solvation model. Accurate QSRR machine learning models (Fig. 2b) for reaction rates or barriers have been constructed for, *e.g.*, cycloaddition,[15,16] S_N2 substitution,[17] and E2 elimination.[18] While these models are highly encouraging, they treat reactions that occur in a single mechanistic step and they are based on an amount of kinetic data (>500 samples) that is only available for very few reaction classes. Another promising line of research uses machine learning to predict DFT barrier heights and then use these barrier heights to predict experimental outcomes.[19] A recent study from Hong and co-workers used the ratio of predicted DFT barriers to predict regioselectivity in radical C–H functionalization reactions.[20] While these models can show good performance, the predicted barriers still suffer from the

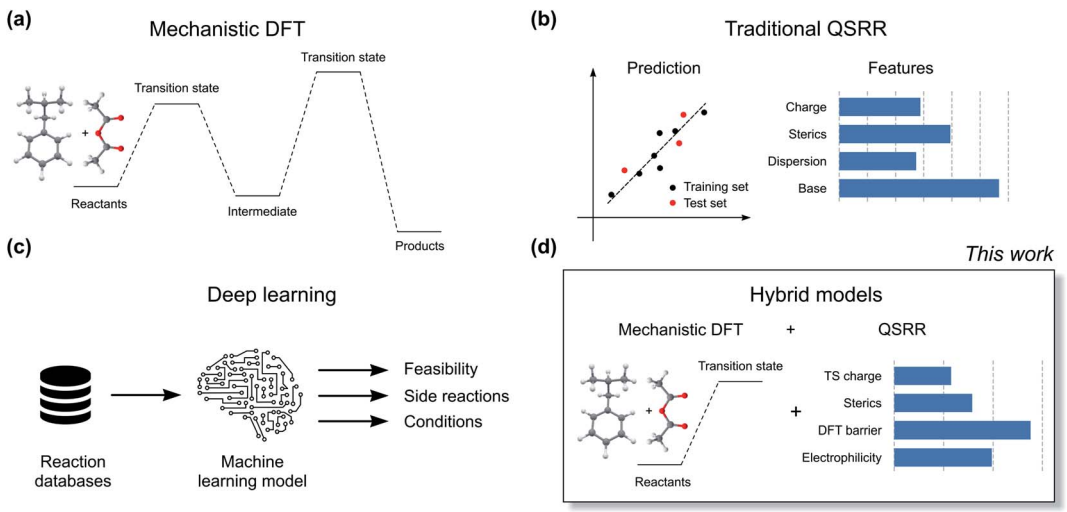

Fig. 2 Different types of quantitative reaction prediction approaches. Mechanistic DFT (a) and QSRR (b) are the current gold standard methods. Deep learning models (c) are emerging as an alternative. Hybrid models (d) combine mechanistic DFT modelling with traditional QSRR.

shortcomings of the underlying DFT method and solvation model. We therefore believe that for models to be broadly applicable in guiding experiments, they should be trained to reproduce experimental rather than computed barrier heights.

Based on the recent success of machine learning for modelling reaction barriers, we wondered if we could combine the traditional mechanistic modelling using DFT with machine learning in a hybrid method (Fig. 2d). Machine learning would here be used to correct for the deficiencies in the mechanistic modelling. Hybrid models could potentially reach useful chemical accuracy (error below 1 kcal mol^{-1})[21,22] with fewer training data than QSRR models, be able to treat more complicated multi-step reactions, and naturally incorporate the effect of catalysts directly in the DFT calculations. Mechanistic models are also chemically understandable and the results can be presented to the chemist with both a view of the computed mechanism and a value for the associated barrier. As a prototype application for a hybrid model, we study the nucleophilic aromatic substitution (S_NAr) reaction (Fig. 1c), one of the most important reactions in chemistry in general and the pharmaceutical chemistry in particular. The S_NAr reaction comprises 9% of all reaction carried out in pharma,[23] and features heavily in commercial routes to block-buster drugs.[24,25] It has recently seen renewed academic interest concerning whether it occurs through a stepwise or a concerted mechanism.[26] We show that hybrid models for the S_NAr reaction reach chemical accuracy with *ca.* 100–200 reactions in the training set, while traditional QSRR models based on quantum-chemical features seem to need at least 200 data points. Models based on purely structural information such as reaction fingerprints need data in the range of 350–400 samples. If these results hold also for other reaction classes, we envision a hierarchy of predictive models depending on how much data is available. Here, transfer learning might ultimately represent the best of both worlds. Models pre-trained on a very large number of DFT-calculated barriers[27] can be retrained on a much smaller amount of high-quality experimental data to achieve chemical accuracy for a wide range of reaction classes.

Results and discussion

First, we describe how the S_NAr reaction dataset was collected and analysed. We then describe the featurization of the reactions in terms of ground state and TS features. Machine learning models are then built and validated. Finally, we use the model for regio- and chemoselectivity prediction on patent reaction data, a task the model was not explicitly trained for.

Reaction dataset

We collected 449 rate constants for S_NAr reactions from the literature and ran 443 (98.7%) successfully through the modelling procedure. Of these 443 reactions, 336 corresponded to unique sets of reactants and products, of which 274 were performed under only one set of conditions (temperature and solvent) while 62 were performed under at least two sets of conditions. Activation energies were obtained from the rate constants *via* the Eyring equation at the reaction temperature, and were in the range 12.5–42.4 kcal mol^{-1} with a mean of 21.3 kcal mol^{-1} (Fig. 3a). The dataset is diverse, with nitrogen, oxygen and sulphur nucleophiles (Fig. 3b) and oxygen, halogen and nitrogen leaving groups (Fig. 3c), although the combinations of nucleophilic atom and leaving atoms is unevenly populated (Fig. 3d). The two most common nucleophiles are piperidine (96 entries) and methoxide (49 entries), while the most common substrates are dinitroarenes (Fig. 3e). A principal components analysis of the full feature space (*vide infra*) reveals

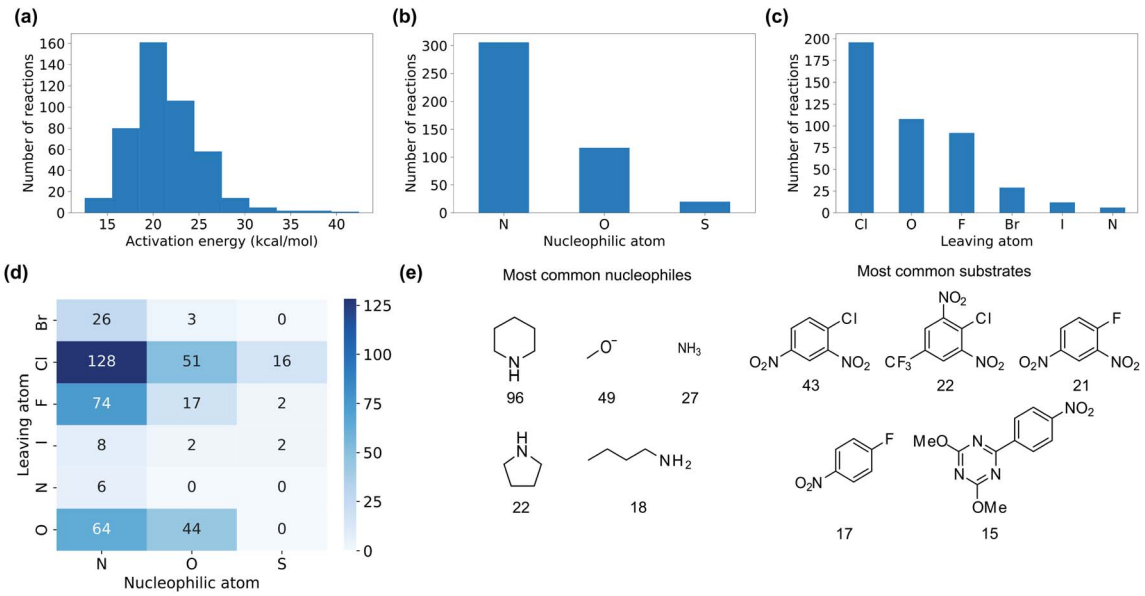

Fig. 3 Distribution of (a) activation energies (b) nucleophilic atoms and (c) leaving atoms in the dataset. (d) Number of reactions in the training set for combinations of nucleophilic atom and leaving atom. (e) Frequency of the five most common nucleophiles and substrates in the training set.

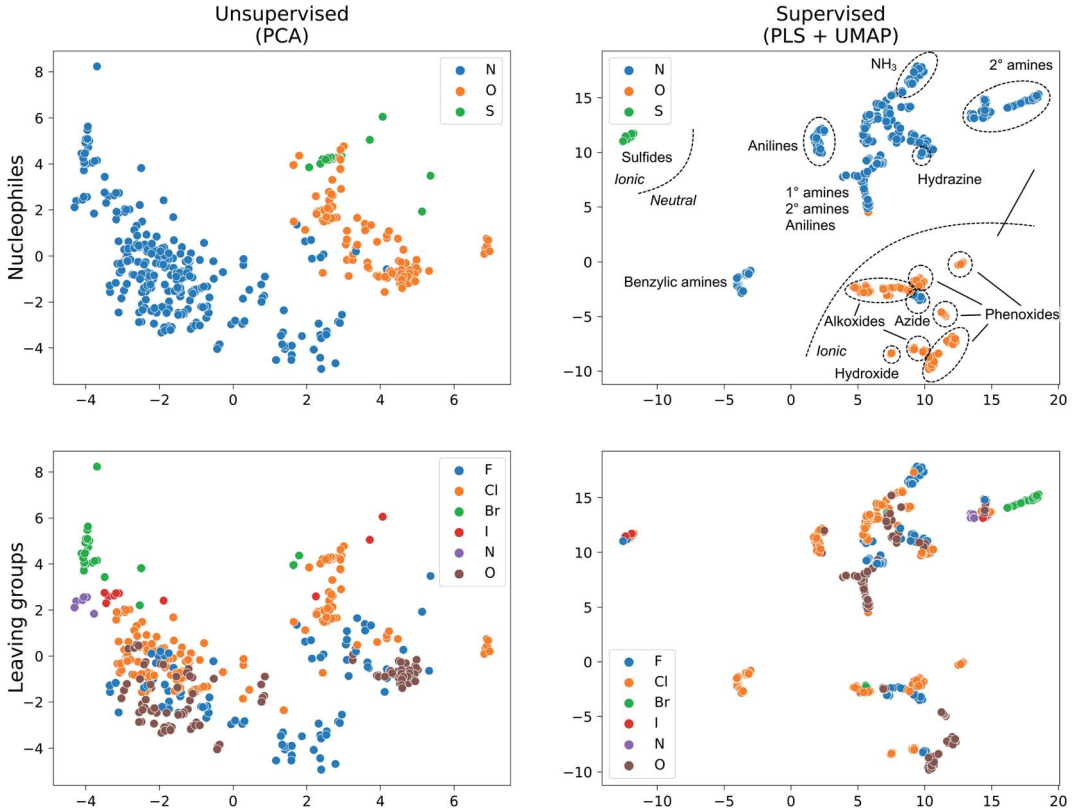

Fig. 4 Dataset visualization with unsupervised (PCA) and supervised (PLS + UMAP) dimensionality reduction for the X_{full} feature set.

a clear separation of reactions with respect to different nucleophile and leaving group atom types (Fig. 4). Supervised dimensionality reduction with partial least squares (PLS) to 5 dimensions, followed by unsupervised dimensionality reduction with the uniform manifold approximation and projection (UMAP) method[28] yielded separated clusters with clear chemical interpretation (Fig. 4).

Reaction feature generation

To calculate the reaction features automatically, we constructed a workflow called predict-S_NAr. The workflow takes a reaction SMILES representation as input, deduces the reactive atoms and computes reactants, transition states and products with a combination of semi-empirical (SE) methods and DFT (Fig. 5a). Both concerted and stepwise mechanisms are treated and all structures are conformationally sampled, making use of the lowest-energy conformer (Fig. 5b). In the case of anionic nucleophiles, a mixed explicit/implicit solvation model was employed to reduce the errors of the computed barriers. The workflow also automatically calculates the quantum mechanical reaction features (Fig. 5c). When reactions corresponded to substitution on several electrophilic sites in a substrate, each reaction was treated with a separate calculation. Initially, we attempted to calculate the transition states with SE methods, but this was unsuccessful. Anionic nucleophiles are artificially destabilized by the lack of diffuse basis functions in the SE method, and the resulting potential energy surface is therefore highly distorted. We therefore used a more robust combination of SE and DFT as outlined in the ESI, Section 2.3.† There are still some limitations of our S_NAr workflow that could be improved in future work, such as treating counter-ions and acid and base catalysis.

For the hybrid model, we needed features for both the ground state molecules and the rate-determining transition state. We opted for physical organic chemistry features which would be chemically understandable and transferable to other reactions.[29] We selected features associated with nucleophilicity, electrophilicity, sterics, dispersion and bonding as well as features describing the solvent. As "hard" descriptors of nucleophilicity and electrophilicity, we used the surface average of the electrostatic potential ($\bar{V}_s$) of the nucleophilic or electrophilic atom.[30,31] ("Descriptor" and "feature" are here used as synonyms.) As "soft" descriptors we used the atomic surface minimum of the average local ionization energy ($I_{s,min}$),[32] as well as the local electron attachment energy ($E_{s,min}$), which has been shown to correlate well with S_NAr reactivity.[33,34] From conceptual DFT, we used the global electrophilicity descriptor ω and nucleophilicity descriptor N, as well as the corresponding local nucleophilicity and electrophilicity descriptors ℓ_N and ℓ_ω.[35] These electronic features were complemented with atomic charges (q) from the DDEC6 scheme[36] and the electrostatic potential at the nuclei (V_N) of the reactive atoms.[37] In terms of sterics and dispersion, we use the ratio of solvent accessible surface area ($SASA_r$)[38] of the reactive atoms and the universal

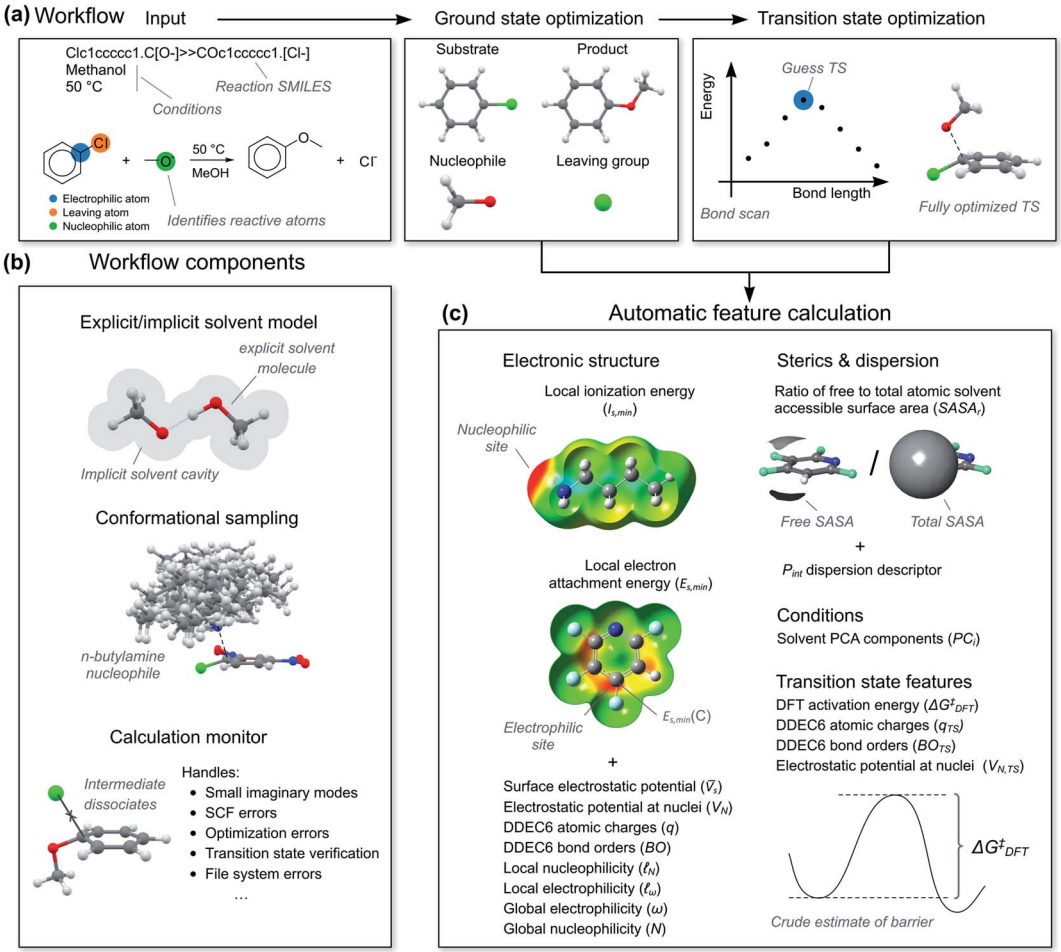

Fig. 5 (a) Automatic workflow for calculation of reaction mechanism and features. (b) Important components of the workflow. (c) Features calculated for ground and transition states. Conformational sampling done with GFN2-xTB using the CREST tool, geometries refined with ωB97X-D/6-31+G(d) with SMD solvation. Final single points energies with ωB97X-D/6-311+G(d,p). Electronic structure features calculated with B3LYP/6-31+G(d).

quantitative dispersion descriptor P_{int}.[39] For bonding, we used the DDEC6 bond orders (BO) of the carbon-nucleophile and carbon-leaving group bonds.[40] Solvents were described using the first five principal components (PC_1–PC_5) in the solvent database by Diorazio and co-workers.[41] The most important TS feature was the DFT-calculated activation free energy ($\Delta G^{\ddagger}_{\text{DFT}}$), *i.e.*, a "crude estimation" of the experimental target.[42] We also added V_N, DDEC6 charges and bond orders at the TS geometry. We decided to not include the reaction temperature as one of the features, as reactions with higher barriers tend to be run at higher temperatures as they are otherwise too slow, and therefore correlate unduly with the target (see ESI, Section 5.6†). Atomic features are denoted with C (central), N (nucleophilic) or L (leaving) in parenthesis, *e.g.*, q(C) for the atomic charge of the central atom. Features at the TS geometry are indicated with a subscript "TS", *e.g.*, $q(C)_{TS}$.

We chose to investigate three main feature sets: (1) $\boldsymbol{X}_{\text{full}}$, containing all the features (34 in total) for maximum predictive accuracy, (2) $\boldsymbol{X}_{\text{noTS}}$ without any information from the TS, to assess whether hybrid models are indeed more accurate, and (3) $\boldsymbol{X}_{\text{small}}$, which represents a minimal set of 12 features that can be interpreted more easily, with $\Delta G^{\ddagger}_{\text{DFT}}$ as the only TS feature (see ESI† for complete list). We also made two versions of $\boldsymbol{X}_{\text{noTS}}$, excluding either surface electronic descriptors ($\boldsymbol{X}_{\text{trad}}$, missing $\bar{V}_s$, $I_{s,\text{min}}$, $E_{s,\text{min}}$ and P_{int}) or traditional features ($\boldsymbol{X}_{\text{surf}}$, missing ω, N, ℓ_ω, ℓ_N, V_N, q, and BO). As a comparison to the physical organic features, we investigated four feature sets that only make use of the 2D structural information of the molecule: the Condensed Graph of Reaction (CGR) with the *In Silico* Design and Data Analysis (ISIDA) descriptors[43] ($\boldsymbol{X}_{\text{ISIDA,atom}}$ and $\boldsymbol{X}_{\text{ISIDA,seq}}$), the Morgan reaction fingerprints[44] as implemented in the RDKit ($\boldsymbol{X}_{\text{Morgan}}$),[45] as well as the deep learning reaction fingerprints from Reymond and co-workers ($\boldsymbol{X}_{\text{BERT}}$).[46] These structural features can be calculated almost instantaneously and are useful for fast prediction. We added solvent information to the structural features by concatenating PC_1–PC_5.

Choosing the best machine learning model

We split the data randomly into 80% used for model selection (training set) and 20% used to validate the final model (test set).

To compare the performance of a series of machine learning models on the training set, we used bootstrap bias-corrected cross-validation (BBC-CV) with 10 folds.[47] BBC-CV is an economical alternative to nested cross-validation to avoid overfitting in the model selection process and also gives an estimate of the bias for choosing the model that performs best on the training set. We measure performance with the squared correlation coefficient (R^2), the mean absolute error (MAE) and the root mean squared error (RMSE). We focus on the MAE as its scale is directly comparable to the prediction error. Error bars are given in terms of one standard error of the mean.

The results of the model validation (Fig. 6a and Table S5†) show a clear progression from simpler models such as linear regression (LR) with a MAE of 1.20 kcal mol^{-1}, to intermediate methods such as random forests (RF) at 0.98 kcal mol^{-1}, to more advanced Support Vector Regression (SVR) and Gaussian Process Regression (GPR) models at 0.80 kcal mol^{-1}. Most importantly, the best methods are well below chemical accuracy (1 kcal mol^{-1}). In comparison, the raw DFT barriers $\Delta G^{\ddagger}_{\mathrm{DFT}}$ show a high MAE of 2.93 kcal mol^{-1} and have the same predictive value as just guessing the mean of the training dataset (Fig. 6b). Compensating for systematic errors in the DFT energies by linear correction helps, but still has an unacceptable MAE of 1.74 kcal mol^{-1}. Interestingly, simpler linear methods such as the Automatic Relevance Determination (ARD) can achieve the same performance as the non-linear RF when polynomial and interaction features of second order (PF2) are used to capture non-linear effects. The overall best method considering MAE is GPR with the Matern 3/2 kernel ($\mathrm{GPR_{M3/2}}$). This result is very gratifying as GPRs are resistant to overfitting, do hyper-parameter tuning internally and produce error bars that can be used for risk assessment. We therefore selected $\mathrm{GPR_{M3/2}}$ as our final method and also used it to make comparisons between different feature sets. In the BBC-CV evaluation, it had an R^2 of 0.87, an MAE of 0.80 kcal mol^{-1}, and an RMSE of 1.41 kcal mol^{-1}. Importantly, the BBC-CV method indicates a very low bias of only 0.02 kcal mol^{-1} for MAE in choosing $\mathrm{GPR_{M3/2}}$ as the best method, so overfitting in the model selection can be expected to be small. Indeed, $\mathrm{GPR_{M3/2}}$ shows excellent performance on the external test set, with a R^2 of 0.93, a MAE of 0.77 kcal mol^{-1} and an RMSE of 1.01 kcal mol^{-1} (Fig. 7). The prediction intervals measuring the uncertainty of the individual prediction have a coverage of 99% for the test set, showing that the model can also accurately assess how reliable its predictions are.

Performance of different feature sets

Now, are hybrid models using TS features better than the traditional QSRR models based on just ground-state features? The validation results indicate that hybrid models built with the full feature set $\boldsymbol{X}_{\mathrm{full}}$ perform the best, but not significantly better than models built on $\boldsymbol{X}_{\mathrm{noTS}}$ without TS features (Fig. 6c). Also

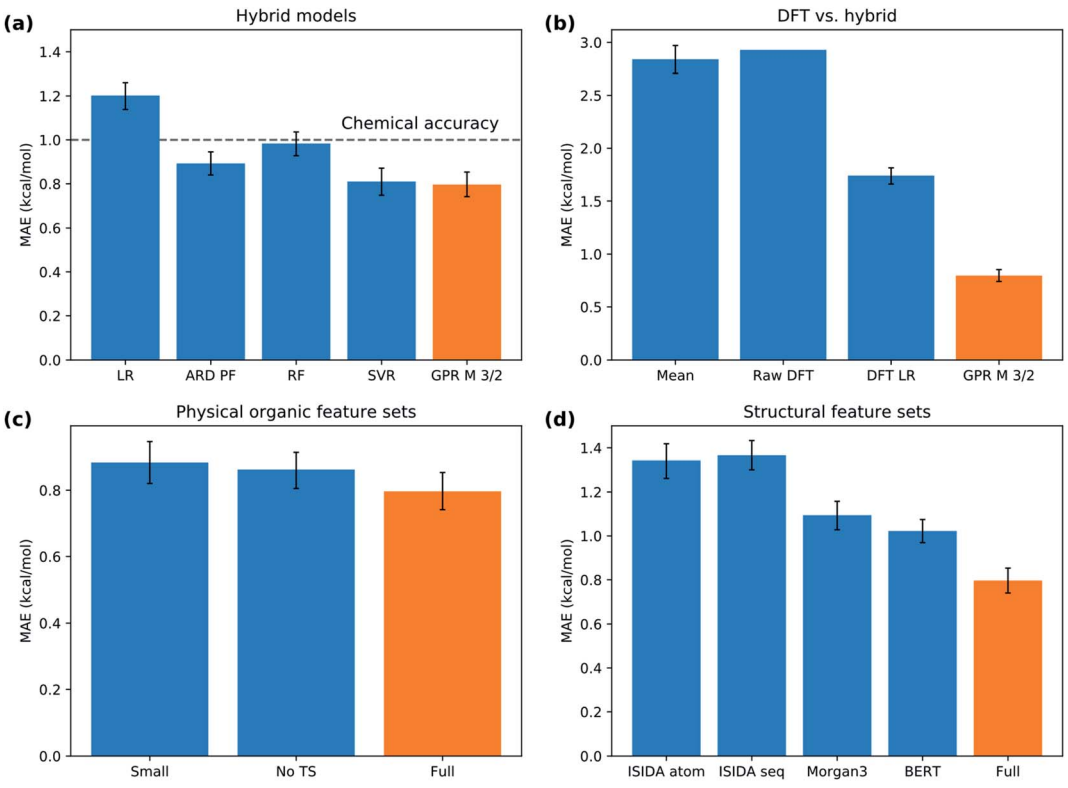

Fig. 6 Model performance. (a) Selection of hybrid models of increasing complexity. (b) Performance of DFT *versus* hybrid model. (c) Comparison of physical organic feature sets. (d) Comparison of structural feature sets. Error bar corresponds to one standard error. Abbreviations explained in the text. The orange bar corresponds to the same model in all subplots, $\mathrm{GPR_{M3/2}}$ with the $\boldsymbol{X}_{\mathrm{full}}$ feature set. A comparison of all the models on the same scale is given in Fig. S6.†

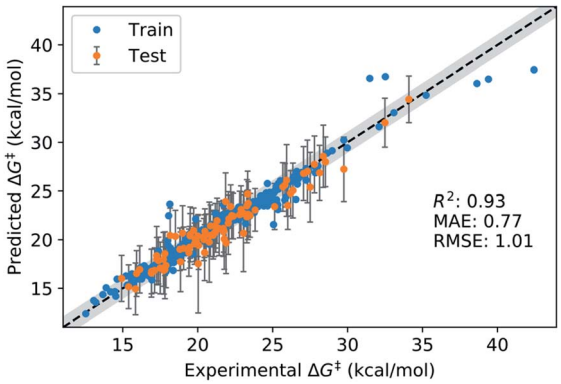

Fig. 7 Performance of final $GPR_{M3/2}$ model on the external test set. Coverage for test set: 99%. The grey band corresponds to ±1 kcal mol^{-1} with respect to the identity line.

the model built on the small expert-chosen feature set $\boldsymbol{X}_{small}$ shows similar performance. To investigate the matter more deeply, we calculated the learning curves of $GPR_{M3/2}$ using the different features sets (Fig. 8a) on the full dataset. We see that all models indeed do perform similarly with the amount of training data used for the model selection (318 reactions), but that the hybrid models based on $\boldsymbol{X}_{full}$ and $\boldsymbol{X}_{small}$ seem to have an advantage below 150 samples. Indeed, the learning curve for predicting the $\Delta G^{\ddagger}_{DFT}$ based on $\boldsymbol{X}_{noTS}$ also starts levelling off after *ca.* 150 samples (Fig. 8b). This indicates that the model is able to implicitly learn $\Delta G^{\ddagger}_{DFT}$ from the ground state features given sufficient data. It seems that there is still a residual advantage using the full hybrid model even with larger dataset sizes, although this advantage becomes smaller and smaller. For larger datasets it would therefore make sense to apply the "one-standard-error" rule and use the less complex model based on $\boldsymbol{X}_{noTS}$ which is easier to implement. This rule states that a simpler model could be chosen in place of a better-scoring and more complex one if the score of the simpler model is within one standard error of the more complex model.[48,49] But with fewer datapoints, it would instead make sense to use $\boldsymbol{X}_{full}$ for maximum performance. Models built on ground state features lacking either surface features ($\boldsymbol{X}_{trad}$) or lacking more traditional electronic descriptors ($\boldsymbol{X}_{surf}$) showed worse performance (MAE: 1.00 and 1.10 kcal mol^{-1}, respectively) than when both were included as in $\boldsymbol{X}_{noTS}$ (MAE: 0.86 kcal mol^{-1}). Therefore, both should be included for maximum performance and seem to capture different aspects of reactivity. It will be interesting to see if these trends with regard to dataset size hold up also for other reaction classes.

How good can the model get given even more data? Any machine learning model is limited by the intrinsic noise of the underlying training data, given in our case by the experimental error of the kinetic measurement. In the dataset, there are four reactions reported with the same solvent and temperature but in different labs or on different occasions. Differences between the activation energies are 0.1, 0.1, 0.5 and 1.6 kcal mol^{-1}. The larger difference is probably an outlier, and we estimate the experimental error is on the order 0.1–0.5 kcal mol^{-1}. In comparison, the interlaboratory error for a data set of S_N2 reactions was estimated to *ca.* 0.7 kcal mol^{-1}.[17] It is thus reasonable to believe that the current model with a MAE of 0.77 kcal mol^{-1} on the external test set is getting close to the performance that can be achieved given the quality of the underlying data. Gathering more data is therefore not expected to significantly improve the accuracy of the average prediction, but may widen the applicability domain by covering a broader range of structures (*vide infra*) and reduce the number of outlier predictions.

Models based on structural information

Given the time needed to develop both traditional and hybrid QSRR models, an attractive option is using features derived from just chemical connectivity. We investigated this option using the CGR/ISIDA approach with either atom-centred ($\boldsymbol{X}_{ISIDA,atom}$) or sequential ($\boldsymbol{X}_{ISIDA,seq}$) fragment features, as well as reaction difference fingerprints of the Morgan type with a radius of three atoms ($\boldsymbol{X}_{Morgan3}$). The results with $GPR_{M3/2}$ show good performance using $\boldsymbol{X}_{Morgan3}$, almost reaching

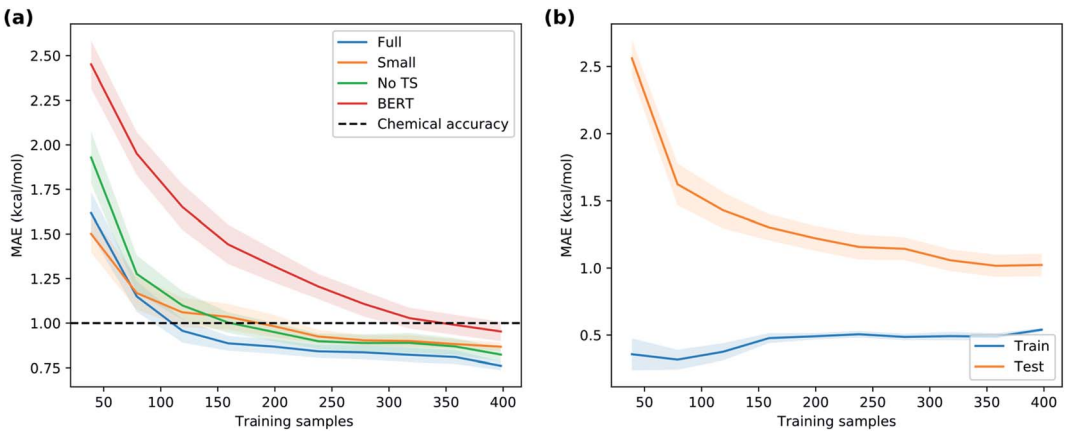

Fig. 8 (a) Learning curve giving the mean absolute error as a function of number of reactions in the training set. (b) Learning curve to predict the DFT activation energies using the ground state features. Shaded regions correspond to one standard error.

chemical accuracy with a MAE of 1.09 kcal mol^{-1} (Fig. 6d). ISIDA sequence and atom features perform worse (Fig. 6d). They also require an accurate atom-mapping of the reaction, and we found that automatic atom mapping failed for 50 (11%) of the reactions studied. The methods above rely on expert-crafted algorithms to generate fingerprints or feature vectors. In recent years, deep learning has emerged as a method for creating such representations from the data itself, *i.e.*, representation learning. The recent reaction fingerprint from Reymond and co-workers is one such example, where the fingerprint is learned by a BERT deep learning model that is pre-trained in an unsupervised manner on reaction SMILES from patent data.[46] Gratifyingly, the model built on the $\boldsymbol{X}_{\mathrm{BERT}}$ feature set performs on par or slightly better than $\boldsymbol{X}_{\mathrm{Morgan3}}$ with an MAE of 1.03 kcal mol^{-1} (Fig. 6d). The big advantage with the BERT fingerprint is that it does not need atom mapping and can potentially be used with noisy reaction SMILES with no clear separation between reactants, reagents, catalysts and solvents. We also used BERT fingerprints specially tuned for reaction classification to construct reaction maps, which show clear separation between different types of nucleophiles and electrophiles in the dataset (Fig. S7†). The learning curve shows that the model based on $\boldsymbol{X}_{\mathrm{BERT}}$ is more data-hungry than the models based on physical organic features, and requires *ca.* 350–400 data points to reach chemical accuracy (Fig. 8a). The radius of three for the Morgan fingerprint was chosen to incorporate long-range effects of electron-donating and electron-withdrawing groups in the *para* position of the aromatic ring (Fig. 9b). Plotting the MAE as a function of the fingerprint radius shows a clear minimum at a radius of three (Fig. 9a). With this radius, all relevant reaction information seems to be captured, and increasing the radius further probably just adds noise through bit clashes in the fingerprint generation. Encouraged by the promising results with the structural data, we wondered if a combined feature set with both the physical organic features in $\boldsymbol{X}_{\mathrm{full}}$ and the structural features in $\boldsymbol{X}_{\mathrm{BERT}}$ would perform even better. However, the model built on the combined feature set performs on par with the model built on only $\boldsymbol{X}_{\mathrm{full}}$ (Table S5†).

In summary, models based on reaction fingerprints are an attractive alternative when a sizeable dataset of at least 350 reactions are available as they are easy to develop and make very fast predictions.

Interpretability

There has a been a push in the machine learning community in recent years to not only predict accurately but also to understand the factors behind the prediction.[50] Models are often interpreted in terms of their feature importances, *i.e.*, how much a certain feature contributes to the prediction. Feature importances can be obtained directly from multivariate linear regression models as the regression coefficients and have been used extensively to give insight on reaction mechanisms based on ground-state features.[51] A number of modern techniques can obtain feature importances for any machine learning technique, including SHAP values[52] and permutation importances,[53] potentially allowing the simple interpretation of linear models to be combined with the higher accuracy of more modern non-linear methods.

Although feature importances can be easily calculated, they are not always easily interpretable. In particular, correlation between features poses severe problems. This problem is present for our feature set, as shown by the Spearman rank correlation matrix and the variance inflation factors (VIFs)[54] of $\boldsymbol{X}_{\mathrm{full}}$ (Fig. S8†). Therefore, special care has to be taken when calculating the feature importances, and the final interpretation of them will not be straightforward. To get around this technical problem of multicollinearity, we clustered the features based on their Spearman rank correlations, keeping only one of the features in each cluster (Fig. 10a). The stability of the feature ranking was further analysed by using ten bootstrap samples of the data.

First, we looked at how important $\Delta G^{\ddagger}_{\mathrm{DFT}}$ is in our hybrid model. It turns out that it is consistently ranked as the most important feature across all bootstrap samples (Fig. 10b). The second-most important feature is the soft electrophilicity feature $E_{\mathrm{s,min}}$(C). To see more clearly which are the most important ground-state features, we analysed a model built on $\boldsymbol{X}_{\mathrm{noTS}}$ (Fig. 10b). The most important feature is again the soft electrophilicity descriptor $E_{\mathrm{s,min}}$(C). Other important features are the soft nucleophilicity descriptor $I_{\mathrm{s,min}}$(N) and the feature cluster of hard electrophilicity represented by $\bar{V}_{\mathrm{s}}$(C). The global nucleophilicity descriptor N and the electrophile–nucleophile bond strength through BO(C–N) are also ranked consistently high.

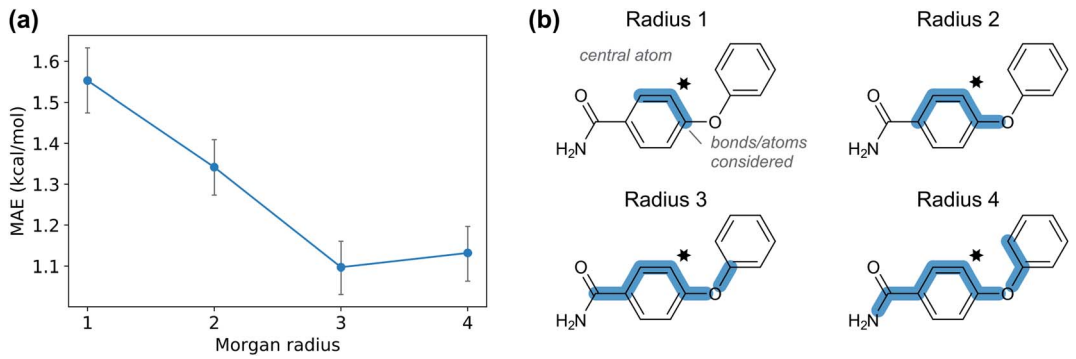

Fig. 9 (a) Mean absolute error for $\mathrm{GPR}_{\mathrm{M3/2}}$ as a function of Morgan fingerprint radius. (b) Atoms and bonds considered by Morgan fingerprints of different radius. A radius of radius three or more is needed to capture the effect from groups in the *para* position of the aromatic ring.

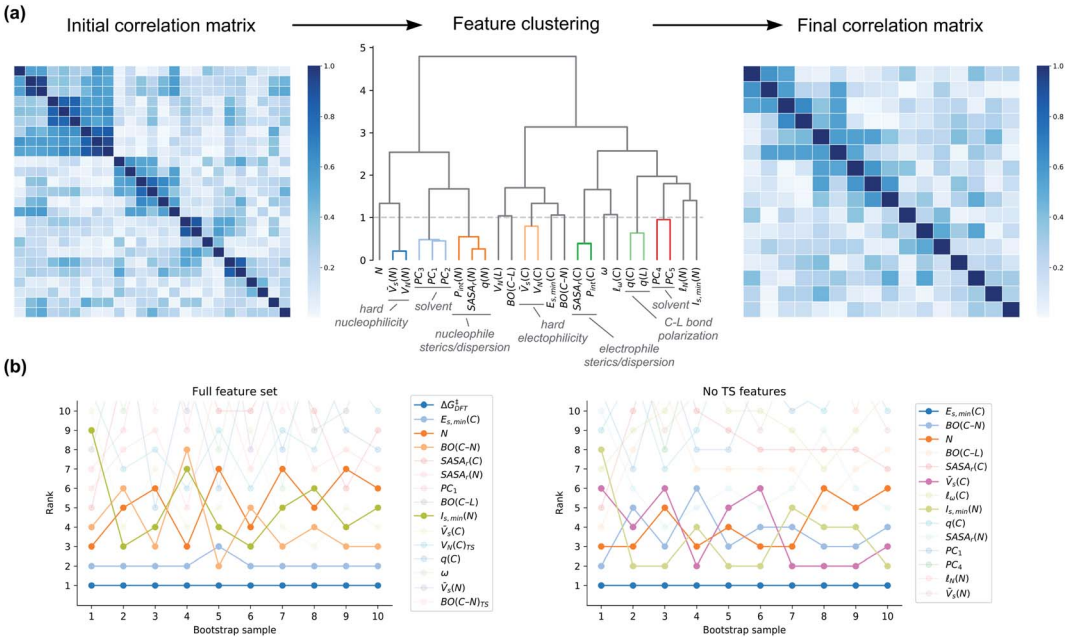

Fig. 10 (a) Clustering of correlated features based on Spearman rank correlation. (b) Bootstrapped feature ranking for clustered X_{full} and X_{noTS}. For identity of feature clusters, see the ESI.†

In summary, the most important feature for the hybrid models is, as expected, the DFT-computed activation free energy $\Delta G^{\ddagger}_{DFT}$. Models trained without the TS features give insight into the features of substrates and nucleophiles that govern reactivity. Here, the most important features are the electrophilicity of the central carbon atom of the substrate, followed by those related to nucleophilicity. Steric and solvent features are less important. There are a number of plausible reasons why solvent features have lower importance. Firstly, the effect of the solvent is already incorporated in the $\Delta G^{\ddagger}_{DFT}$ through the use of both implicit and explicit solvation (for anions of second and third row elements of the periodic table). Secondly, the implicit solvent influences the values of the quantum-mechanically derived features. For the $E_{s,min}$ descriptor this effect has been shown to be substantial.[34,55] Thirdly, the solvent correction can likely be learnt implicitly also from the other features for problematic nucleophiles such as those with anionic oxygen. This aspect is also connected to the fourth factor, that solvent variation is low for the anionic nucleophiles, where for example reactions with oxygen nucleophiles are only carried out in water or methanol (Fig. S25†). There is therefore limited data for the model to learn from the solvent features for some nucleophile classes. Collection of more balanced data will be key to improved models. Steric features may become more important for other types of substrates as the current data set doesn't include very sterically crowded substrates or nucleophiles.

Applicability domain

The applicability domain (AD) is a central concept to any QSRR model used for regulatory purposes according to the OECD guidelines.[56] The AD is broadly defined by the OECD as "the response and chemical structure space in which the model makes predictions with a given reliability". Although there have been many attempts to define the AD for prediction of molecular properties, there has been little work for reaction prediction models. Here, we will follow a practical approach to (1) define a set of strict rules for when the model shouldn't be applied based on the reaction type and the identity of the reactive atoms, (2) identify potentially problematic structural motifs from visualization of outliers, and (3) assess whether the uncertainty provided by the $GPR_{M3/2}$ model can be used for risk assessment.

First, the model should only be applied to S_NAr reactions. Second, the model can only be used with confidence for those reactive atom types that are part of the training set (Fig. 3b). One example of reactions falling outside the applicability domain according to these rules are those with anionic carbon nucleophiles. A provisional analysis of the outliers (residual of >2 kcal mol^{-1}) for the training set identifies many reactions involving methoxide nucleophile (Fig. 11). This tendency can be partly understood based on the poor performance of implicit solvation models for such small, anionic nucleophiles which is probably not corrected fully by our model. Additionally, some of these reactions with methoxide and unactivated substrates are very slow and have been run at high temperature and the rate constants have been determined through extrapolation techniques, leading to a larger error in the corresponding rates. Outliers from the test set involve azide and secondary amine nucleophiles. However, secondary amines are also the most common type of nucleophile in the dataset and are therefore expected to contribute to some outliers. For a complete list of all outliers, see the ESI.†

We also investigated whether the prediction uncertainty given by the $GPR_{M3/2}$ model could be trusted. For predicting

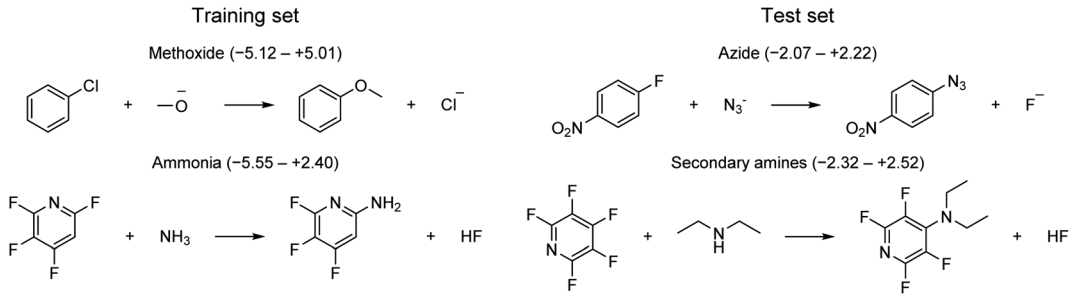

Fig. 11 Example of outliers for training and test set. Ranges in parenthesis correspond to signed errors in kcal mol^{-1}.

molecular properties, similarity to the training set is usually used to assess whether a prediction should be trusted (in the applicability domain) or not (outside the domain). The similarity is measured by distance metrics based on molecular fingerprints.[57] For reactions, difference fingerprints have been shown to differentiate between reaction classes, but their suitability for defining an applicability domain within one reaction class is not clear.[58] In the absence of a clear similarity metric, we used the Distance to Model in X-space (DModX) of a PLS model with two components (dimensions) to compare to GPR$_{M3/2}$. PLS is a type of linear method that uses dimensionality reduction and that is used widely in chemometrics. DModX is based on distance in the latent space used by the PLS model and is an established metric for defining the applicability domain.[59] We compared the performance of DModX to the standard deviation (std) of the GPR$_{M3/2}$ predictions using integral accuracy averaging curves (Fig. 12). The integral accuracy averaging curve is a standard tool for evaluating uncertainty metrics, where the predicted values are ordered from most to least reliable according to the uncertainty metric.[60] The MAE is then plotted as a function of the portion of included data. A good uncertainty metric should show a curve with an upward slope from left to right, as including more points with larger uncertainty should lead to a larger error. As can be seen from the plot, both DModX and GPR$_{M3/2}$ std are decent measures of uncertainty. As the GPR$_{M3/2}$ std performs best, the prediction intervals (as shown in Fig. 7) can be used directly as a measure of how reliable the prediction is.

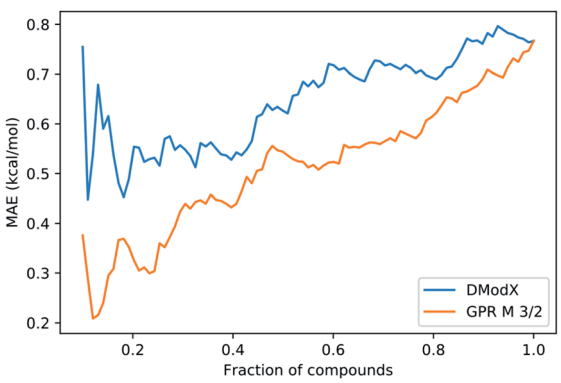

Fig. 12 Integral accuracy averaging curve. Predicted values are ordered according to an uncertainty measure from most certain to least certain. MAE is calculated from the predictions of the GPR$_{M3/2}$ model.

As there are 62 reactions which occur in the dataset with more than one reaction condition (different temperature or solvent), we investigated the performance of leaving these reactions out altogether in the modelling. With this leave-one-reaction-out validation approach, we observed a MAE of 1.00 kcal mol^{-1} for GPR$_{M3/2}$ (compared to 0.80 kcal mol^{-1} from normal cross-validation). In comparison, a model trained on $\boldsymbol{X}_{BERT}$ decreased from 1.03 to 1.11 kcal mol^{-1}. We also tested leave-one-electrophile-out, giving a MAE of 1.20 kcal mol^{-1} and leave-one-nucleophile-out, giving a of MAE 0.68 kcal mol^{-1}. These results indicate that the model is able to predict outside its immediate chemical space with good accuracy, and not only interpolate.

Taken together, the applicability domain of our model is defined strictly in terms of the type of reaction (S_NAr) and the types of reactive atoms in the training set. The outlier analysis identified that extra care should be taken when interpreting the results of reactions at high temperature and with certain nucleophile classes. The width of the prediction interval from the GPR$_{M3/2}$ model is a useful measure of the uncertainty of the prediction.

Validation on selectivity data

To assess whether the model can be used not only for predicting rates, but also selectivities, we compiled a dataset of reactions with potential regio- or chemoselectivity issues from patent data.[61] A total of 4365 S_NAr reactions were considered, of which 1214 had different reactive sites on the same aromatic ring, while only one product was recorded experimentally. A reactive site was considered as a ring carbon with a halogen substituent (F, Cl, Br and I) that could be substituted with S_NAr. Regioselectivity means distinguishing between reactive sites with the same type of halogen, while for chemoselectivity the halogens are different. Out of these 1214, we selected the 100 with lowest product molecular weight for a preliminary evaluation. As reaction solvent or temperature was not available, we used acetonitrile for neutral reactions and methanol for ionic reactions and a reaction temperature of 25 °C. It is possible that neglecting conditions information in this way reduces the accuracy of our model. Each possible reaction leading to different isomers was modelled by a separate predict-S_NAr

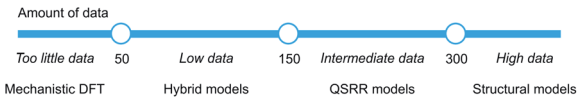

Fig. 13 Models for quantitative rate prediction for different data regimes based on the S_NAr dataset.

calculation (209 in total), and the predicted major isomer was taken as the one corresponding to the lowest $\Delta G^{\ddagger}$. As another testament to the robustness of the workflow, only 3 of 209 TS calculations failed (1.4%), leading finally to results for 97 reactions with selectivity issues. Of these, 66 involved regioselectivity and 31 chemoselectivity. The results show that $GPR_{M3/2}$ was able to predict the correct site in 86% of the cases (top-1 accuracy), with 85% for regioselectivity and 87% for chemoselectivity. In comparison, predictions based on only the DFT activation energies give a comparable top-1 accuracy of 87%, with 91% for regioselectivity and 77% for chemoselectivity. It is notable that the $GPR_{M3/2}$ model shows a much better score for chemoselectivity (87%) than DFT (77%). For regioselectivity prediction, where the same element is being substituted, DFT probably works well due to error cancellation. That the hybrid ML model performs better than DFT for chemoselectivity clearly shows that it has learnt to compensate for the systematic errors in the DFT calculations. In fact, it seems that the ML model is balancing the gain in chemoselectivity prediction with a loss in regioselectivity (85%) compared to DFT (91%), leading to a comparable accuracy on both tasks. More training data is expected to alleviate this loss in regioselectivity accuracy. Also, inclusion of the actual solvents and temperatures could also increase accuracy. Further validation work could extract the conditions from the original patents, or use a dataset that already contains this information.

Overall, it is remarkable that the hybrid model $GPR_{M3/2}$ performs so well (86%) for selectivity as it was not explicitly trained on this task. In comparison, for electrophilic aromatic substitution (S_EAr), single-task deep learning models trained for selectivity prediction of bromination, chlorination, nitration and sulfonylation achieved top-1 accuracies of 50–87%.[62] For bromination, the RegioSQM model based on semiempirical energies of the regioisomeric intermediate σ-complexes achieved 80%, compared to 85% for the neural network just mentioned. In light of these models, the 86% top-1 accuracy obtained with $GPR_{M3/2}$ for S_NAr looks very competitive. Likely, hybrid models explicitly trained for selectivity prediction can perform even better. Another approach would be to use transfer learning to repurpose deep learning models for barrier prediction to selectivity prediction.

Conclusions and outlook

We have created hybrid mechanistic/machine learning models for the S_NAr reaction which incorporate TS features in addition to the traditional physical organic features of reactants and products. The chosen Gaussian Process Regression model achieved a mean absolute error of only 0.77 kcal mol^{-1} on the external test set, well below the targeted chemical accuracy of 1 kcal mol^{-1}. Furthermore, the model achieves a top-1 accuracy for regio- and chemoselectivity prediction on patent reaction data of 86%, without being explicitly trained for this task. Finally, the model comes with a clear applicability domain specification and prediction error bars that enables the end user to make a contextualized risk assessment depending on what accuracy is required.

By studying models built on reduced sets of the physical organic features, as well as reaction fingerprints, we identified separate data regimes for modelling the S_NAr reaction (Fig. 13). In the range 0–50 samples, it is questionable whether accurate and generalizable machine learning models can be constructed.[63] Instead, we suggest that traditional mechanistic modelling with DFT should be used, with appropriate consideration of its weaknesses. With 50–150 samples, hybrid models are likely the most accurate choice, and should be used if time and resources for their development is available. In the range 150–300 samples, traditional QSRR models based on physical organic features reach similar accuracy as hybrid models, while models based on purely structural information become competitive with over 300 samples. It will be very interesting to see if these numbers generalize to other reaction classes. For choosing features for QSRR models, it is notable that electronic reactivity features from DFT were consistently ranked high in the feature importances, and we saw that the inclusion of surface features makes the models significantly better.

Our workflow can handle the mechanistic spectrum of concerted and stepwise S_NAr reactions, and we are currently working on extending it to handle the influence of general or specific acid and base catalysts as well as treating related reaction classes. This general model is a significant improvement in generality compared to previous work, which modelled selectivity of S_NAr reactions in terms of the relative stability of the σ addition complexes and therefore was only applicable to reactions with step-wise mechanisms.[64–66] Although promising, we believe that widespread use of hybrid models is currently held back by difficulties in computing transition states in an effective and reliable way. We envision that this problem will be solved in the near future by deep learning approaches that can predict both TS geometries[67] and DFT-computed barriers[27] based on large, publicly available datasets.[68,69] In the end, machine learning for reaction prediction needs to reproduce experiment, and transfer learning will probably be key to utilizing small high-quality kinetic datasets together with large amounts of computationally generated data. Regardless of their construction, accurate reaction prediction models will be key components of accelerated route design, reaction optimization and process design enabling the delivery of medicines to patients faster and with reduced costs.

Conflicts of interest

There are no conflicts to declare.

Acknowledgements

Stefan Grimme is kindly acknowledged for permission to use the *xtb* code and for technical advice regarding its use. Ramil

Nugmanov is acknowledged for advice on calculation of the ISIDA descriptors. Philippe Schwaller is acknowledged for advice on using the BERT reaction fingerprint. Tore Brinck acknowledges support from the Swedish Research Council (VR). Kjell Jorner is a fellow of the AstraZeneca post doc programme.

References

1 E. N. Muratov, J. Bajorath, R. P. Sheridan, I. V. Tetko, D. Filimonov, V. Poroikov, T. I. Oprea, I. I. Baskin, A. Varnek, A. Roitberg, O. Isayev, S. Curtalolo, D. Fourches, Y. Cohen, A. Aspuru-Guzik, D. A. Winkler, D. Agrafiotis, A. Cherkasov and A. Tropsha, *Chem. Soc. Rev.*, 2020, **49**, 3525–3564.
2 C. W. Coley, N. S. Eyke and K. F. Jensen, *Angew. Chem., Int. Ed.*, DOI: 10.1002/anie.201909989.
3 O. Engkvist, P.-O. Norrby, N. Selmi, Y. Lam, Z. Peng, E. C. Sherer, W. Amberg, T. Erhard and L. A. Smyth, *Drug Discovery Today*, 2018, **23**, 1203–1218.
4 D. T. Ahneman, J. G. Estrada, S. Lin, S. D. Dreher and A. G. Doyle, *Science*, 2018, **360**, 186–190.
5 F. Sandfort, F. Strieth-Kalthoff, M. Kühnemund, C. Beecks and F. Glorius, *Chem*, 2020, **6**, 1379–1390.
6 A. F. Zahrt, J. J. Henle, B. T. Rose, Y. Wang, W. T. Darrow and S. E. Denmark, *Science*, 2019, **363**, eaau5631.
7 J. P. Reid and M. S. Sigman, *Nature*, 2019, **571**, 343–348.
8 P. Schwaller, T. Laino, T. Gaudin, P. Bolgar, C. A. Hunter, C. Bekas and A. A. Lee, *ACS Cent. Sci.*, 2019, **5**, 1572–1583.
9 T. Klucznik, B. Mikulak-Klucznik, M. P. McCormack, H. Lima, S. Szymkuć, M. Bhowmick, K. Molga, Y. Zhou, L. Rickershauser, E. P. Gajewska, A. Toutchkine, P. Dittwald, M. P. Startek, G. J. Kirkovits, R. Roszak, A. Adamski, B. Sieredzińska, M. Mrksich, S. L. J. Trice and B. A. Grzybowski, *Chem*, 2018, **4**, 522–532.
10 A. Tomberg, M. J. Johansson and P.-O. Norrby, *J. Org. Chem.*, 2019, **84**, 4695–4703.
11 A. Tomberg, M. É. Muratore, M. J. Johansson, I. Terstiege, C. Sköld and P.-O. Norrby, *iScience*, 2019, **20**, 373–391.
12 P. Vogel and K. N. Houk, *Organic Chemistry: Theory, Reactivity and Mechanisms in Modern Synthesis*, Wiley, 2019.
13 R. E. Plata and D. A. Singleton, *J. Am. Chem. Soc.*, 2015, **137**, 3811–3826.
14 R. Pérez-Soto, M. Besora and F. Maseras, *Org. Lett.*, 2020, **22**, 2873–2877.
15 J. M. J. M. Ravasco and J. A. S. Coelho, *J. Am. Chem. Soc.*, 2020, **142**, 4235–4241.
16 M. Glavatskikh, T. Madzhidov, D. Horvath, R. Nugmanov, T. Gimadiev, D. Malakhova, G. Marcou and A. Varnek, *Mol. Inf.*, 2019, **38**, 1800077.
17 T. Gimadiev, T. Madzhidov, I. Tetko, R. Nugmanov, I. Casciuc, O. Klimchuk, A. Bodrov, P. Polishchuk, I. Antipin and A. Varnek, *Mol. Inf.*, 2019, **38**, 1800104.
18 T. I. Madzhidov, A. V. Bodrov, T. R. Gimadiev, R. I. Nugmanov, I. S. Antipin and A. A. Varnek, *J. Struct. Chem.*, 2015, **56**, 1227–1234.
19 P. Friederich, G. dos Passos Gomes, R. De Bin, A. Aspuru-Guzik and D. Balcells, *Chem. Sci.*, 2020, **11**, 4584–4601.
20 X. Li, S.-Q. Zhang, L.-C. Xu and X. Hong, *Angew. Chem., Int. Ed.*, 2020, **59**, 13253.
21 K. N. Houk and F. Liu, *Acc. Chem. Res.*, 2017, **50**, 539–543.
22 K. A. Peterson, D. Feller and D. A. Dixon, *Theor. Chem. Acc.*, 2012, **131**, 1079.
23 J. Boström, D. G. Brown, R. J. Young and G. M. Keserü, *Nat. Rev. Drug Discovery*, 2018, **17**, 709–727.
24 M. R. V. Finlay, M. Anderton, S. Ashton, P. Ballard, P. A. Bethel, M. R. Box, R. H. Bradbury, S. J. Brown, S. Butterworth, A. Campbell, C. Chorley, N. Colclough, D. A. E. Cross, G. S. Currie, M. Grist, L. Hassall, G. B. Hill, D. James, M. James, P. Kemmitt, T. Klinowska, G. Lamont, S. G. Lamont, N. Martin, H. L. McFarland, M. J. Mellor, J. P. Orme, D. Perkins, P. Perkins, G. Richmond, P. Smith, R. A. Ward, M. J. Waring, D. Whittaker, S. Wells and G. L. Wrigley, *J. Med. Chem.*, 2014, **57**, 8249–8267.
25 M. Baumann and I. R. Baxendale, *Beilstein J. Org. Chem.*, 2013, **9**, 2265–2319.
26 E. E. Kwan, Y. Zeng, H. A. Besser and E. N. Jacobsen, *Nat. Chem.*, 2018, **10**, 917–923.
27 C. A. Grambow, L. Pattanaik and W. H. Green, *J. Phys. Chem. Lett.*, 2020, **11**, 2992–2997.
28 L. McInnes, J. Healy and J. Melville, arXiv:1802.03426.
29 W. Beker, E. P. Gajewska, T. Badowski and B. A. Grzybowski, *Angew. Chem., Int. Ed.*, 2019, **58**, 4515–4519.
30 J. S. Murray and P. Politzer, *Wiley Interdiscip. Rev.: Comput. Mol. Sci.*, 2011, **1**, 153–163.
31 T. Brinck, P. Jin, Y. Ma, J. S. Murray and P. Politzer, *J. Mol. Model.*, 2003, **9**, 77–83.
32 P. Sjoberg, J. S. Murray, T. Brinck and P. Politzer, *Can. J. Chem.*, 1990, **68**, 1440–1443.
33 T. Brinck, P. Carlqvist and J. H. Stenlid, *J. Phys. Chem. A*, 2016, **120**, 10023–10032.
34 J. H. Stenlid and T. Brinck, *J. Org. Chem.*, 2017, **82**, 3072–3083.
35 J. Oller, P. Pérez, P. W. Ayers and E. Vöhringer-Martinez, *Int. J. Quantum Chem.*, 2018, **118**, e25706.
36 T. A. Manz and N. Gabaldon Limas, *RSC Adv.*, 2016, **6**, 47771–47801.
37 B. Galabov, S. Ilieva, G. Koleva, W. D. Allen, H. F. Schaefer III and P. v. R. Schleyer, *Wiley Interdiscip. Rev.: Comput. Mol. Sci.*, 2013, **3**, 37–55.
38 B. Lee and F. M. Richards, *J. Mol. Biol.*, 1971, **55**, 379–IN4.
39 R. Pollice and P. Chen, *Angew. Chem., Int. Ed.*, 2019, **58**, 9758–9769.
40 T. A. Manz, *RSC Adv.*, 2017, **7**, 45552–45581.
41 L. J. Diorazio, D. R. J. Hose and N. K. Adlington, *Org. Process Res. Dev.*, 2016, **20**, 760–773.
42 Y. Zhang and C. Ling, *npj Comput. Mater.*, 2018, **4**, 25.
43 A. Varnek, D. Fourches, F. Hoonakker and V. P. Solov'ev, *J. Comput.-Aided Mol. Des.*, 2005, **19**, 693–703.
44 D. Rogers and M. Hahn, *J. Chem. Inf. Model.*, 2010, **50**, 742–754.
45 *RDKit: Open-source cheminformatics*, http://www.rdkit.org.
46 P. Schwaller, D. Probst, A. C. Vaucher, V. H. Nair, T. Laino and J.-L. Reymond, *ChemRxiv*, DOI: 10.26434/chemrxiv.9897365.v2.

47 I. Tsamardinos, E. Greasidou and G. Borboudakis, *Mach. Learn.*, 2018, **107**, 1895–1922.
48 T. Hastie, R. Tibshirani and J. Friedman, *The elements of statistical learning: data mining, inference, and prediction*, Springer, New York, 2nd edn, 2009.
49 M. Kuhn and K. Johnson, *Applied predictive modeling*, Springer, New York, 2013.
50 C. Molnar, *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*, 2019.
51 M. S. Sigman, K. C. Harper, E. N. Bess and A. Milo, *Acc. Chem. Res.*, 2016, **49**, 1292–1301.
52 S. M. Lundberg and S.-I. Lee, in *Advances in Neural Information Processing Systems 30*, ed. I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan and R. Garnett, Curran Associates, Inc., 2017, pp. 4765–4774.
53 L. Breiman, *Mach. Learn.*, 2001, **45**, 5–32.
54 G. James, D. Witten, T. Hastie and R. Tibshirani, *An introduction to statistical learning: with applications in R*, Springer, New York, 2013.
55 T. Brinck and J. H. Stenlid, *Adv. Theory Simul.*, 2019, **2**, 1800149.
56 OECD, 2014.
57 T. Hanser, C. Barber, J. F. Marchaland and S. Werner, *SAR QSAR Environ. Res.*, 2016, **27**, 865–881.
58 N. Schneider, D. M. Lowe, R. A. Sayle and G. A. Landrum, *J. Chem. Inf. Model.*, 2015, **55**, 39–53.
59 E. Lennart, J. Joanna, P. Worth Andrew, T. D. Cronin Mark, M. McDowell Robert and G. Paola, *Environ. Health Perspect.*, 2003, **111**, 1361–1375.
60 M. Mathea, W. Klingspohn and K. Baumann, *Mol. Inf.*, 2016, **35**, 160–180.
61 N. Schneider, N. Stiefl and G. A. Landrum, *J. Chem. Inf. Model.*, 2016, **56**, 2336–2346.
62 T. J. Struble, C. W. Coley and K. F. Jensen, *React. Chem. Eng.*, 2020, **5**, 896–902.
63 Y. S. Abu-Mostafa, M. Magdon-Ismail and H. T. Lin, *Learning from Data: A Short Course*, 2012, http://AMLBook.com.
64 M. Liljenberg, T. Brinck, B. Herschend, T. Rein, G. Rockwell and M. Svensson, *Tetrahedron Lett.*, 2011, **52**, 3150–3153.
65 M. Liljenberg, T. Brinck, B. Herschend, T. Rein, S. Tomasi and M. Svensson, *J. Org. Chem.*, 2012, **77**, 3262–3269.
66 M. Liljenberg, T. Brinck, T. Rein and M. Svensson, *Beilstein J. Org. Chem.*, 2013, **9**, 791–799.
67 L. Pattanaik, J. B. Ingraham, C. A. Grambow and W. H. Green, *Phys. Chem. Chem. Phys.*, 2020, **22**(41), 23618–23626.
68 C. A. Grambow, L. Pattanaik and W. H. Green, *Sci. Data*, 2020, **7**, 137.
69 G. F. von Rudorff, S. N. Heinen, M. Bragato and O. A. von Lilienfeld, *Mach. Learn.: Sci. Technol.*, 2020, **1**, 045026.