Accurate prediction of binding energies for two-dimensional catalytic materials using machine learning

Julia Melisande Fischer,[a] Michelle Hunter,[b] Marlies Hankel,[b] Debra J. Searles,[b, c] Amanda J. Parker,[a] and Amanda S. Barnard*[d]

The binding energy of small molecules on two-dimensional (2D) single atom catalysts influences their reaction efficiency and suitability for different applications. In this study, the binding energy on single metal atoms to N-doped graphene defects was predicted using random forest regression based on approximately 1700 previously generated density functional theory simulations of catalytic reactions. Three different structural feature groups containing hundreds of individual structural features were created and used to characterise the active sites. This approach was found to be accurate and reliable using either fully relaxed output structures or pre-simulation input structures, with coefficients of determination of $R^2 = 0.952$ and $R^2 = 0.865$, respectively. The ability to predict optimal 2D-catalysts before undertaking expensive quantum chemical calculations is an attractive basis for future research, and could be extended to other 2D-materials.

1. Introduction

Heterogeneous catalysis often takes place on scarce and expensive metals. Single atom catalysts (SACs) have attracted significant interest because they catalyse reactions using single metal atoms instead of a metal surface, which can reduce the amount of metal required by at least three orders of magnitude.[1] This has the potential to reduce the cost of the catalyst as well as reducing the quantity of heavy metals required. In early studies single Pt atoms were dispersed on a ceramic substrate (FeO_x),[2] and found to be 2–3 times more active than usually used Au nanoparticle on FeO_x for CO oxidation. This discovery led to testing SACs for other reactions such as hydrogen evolution reaction (HER) or water-gas shift reaction (WGS).[3,4]

Along with ceramics, defective graphene surfaces have been shown to provide an ideal substrate to stabilise single metal atoms and serve as active catalysts (see Figure 1a),[5,6] and recently shown to be engineerable based on the adsorption of different surface groups.[7] Pairs of SACs separated by different degrees on graphene-based materials have also been shown to be effective (see Figure 1b–1f). For example, in a combined experimental and computational study on the oxygen reduction reaction (ORR), a synergistic effect between two single metal atoms on a nitrogen/carbon-based catalyst was proposed.[8] It was shown that paired single cobalt and platinum atoms have a mass activity for ORR that is 267 times higher than commercial platinum on carbon. The calculations were performed on a N-doped graphene (GR) defect, and further studies showed that a different N-doped GR defect with Co and Pt metal atoms could also catalyse HER.[9] Potentially, with the appropriate N-doped GR defects and two metal atoms, an even more efficient catalyst for different reactions could be created.[10] The strength of interaction between the metal atoms and the reactants and intermediates is critical to their performance, and is known to be correlated to the local environment of the metal centre.[11,12] This includes the geometric coordination of the surface to the metal as well as the electronic environment. To make informed decisions about where to start and what materials to pursue, it would be desirable to predict which of the potential metal pairs have the right strength of adsorption for the relevant adsorbates *a priori*.

Finding optimal metal pairs and the right 2D-support environment requires multiple experiments with a high level of precision and control. Scientists have to firstly confirm the thermochemical stability of the system, and secondly identify different configurations of atoms and molecules adsorbed on the surfaces.[13] Given the large number of possible metals and GR-defect combinations, even a high-throughput approach can become prohibitively expensive. In the case of the paired single metal catalysts, the possible combinations of metals comprising the catalyst increases by orders of magnitude. The problem is even greater when moving to other 2D-materials. This type of exploration is challenging for conventional synthesis and characterisation workflows, but is ideally suited to computational materials design where the composition and environment can be controlled explicitly, in a more cost effective manner.[14,15] Computationally, the most likely adsorption site

[a] Dr. J. Melisande Fischer, Dr. A. J. Parker
Data61 CSIRO, Docklands, 3008, Australia
[b] M. Hunter, Dr. M. Hankel, Prof. Dr. D. J. Searles
Centre for Theoretical and Computational Molecular Science,
Australian Institute for Bioengineering and Nanotechnology,
The University of Queensland, Brisbane 4072, Australia
[c] Prof. Dr. D. J. Searles
School of Chemistry and Molecular Biosciences,
The University of Queensland,
Brisbane 4072, Australia
[d] Prof. Dr. A. S. Barnard
Research School of Computer Science,
Australian National University,
Acton 2601, Australia
E-mail: amanda.s.barnard@anu.edu.au

This publication is part of a Special Collection on "Data Science in Catalysis". Please check the ChemCatChem homepage for more articles in the collection.

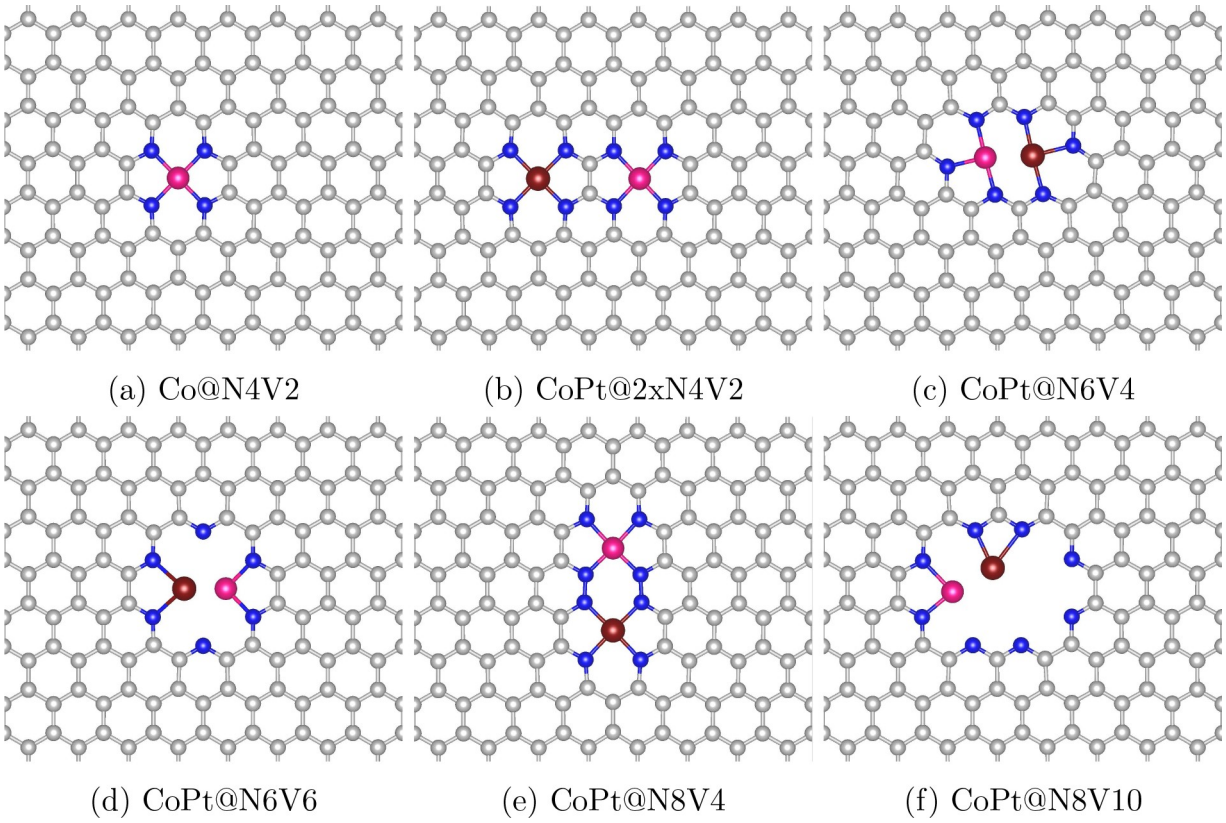

Figure 1. Typical configurations of the six different pores supporting single metal atoms or paired single metal atoms that were considered in this study. Atoms are coloured in grey (carbon), blue (nitrogen), pink (cobalt) and brown (platinum).

and configuration is determined by the strongest binding energy from multiple calculations. Detailed quantum chemical calculations are capable of determining the binding energy of the specific species and defect configurations, but an exhaustive search can rapidly become too computationally intensive at this level of sophistication.[16,17] Moreover, at this stage, it is not clear which information from these calculations is most effective in generating future predictions of binding energies.

Machine learning is a useful tool for scanning over a large number of possible new materials and it does not require all these decisions to be made at the outset.[18–20] Machine learning is capable of identifying patterns in complicated multivariate data, and inferring relationships between structural features such as the type of GR defect, type of metal, and a functional property such as the binding energy.[19,21–24] It has been successfully used to predict important properties of graphene in the past.[25–27] In cases such as these, the success and usefulness of machine learning models is heavily influenced by the choice of features used to describe the material. Feature extraction and engineering is the process of converting data into non-redundant derived values suitable for machine learning, which ideally can also assist in the interpretation of the data.[29] Some features may be scientifically intuitive, such as molecular features including bond lengths or angles,[30] or topological features such as relative distances,[31] while others such as molecular fingerprints are not.[32] Topological features from quantum mechanical energy optimisation calculations are extracted from three-dimensional atomic coordinates. They are typically one-dimensional quantities (e.g. distance or angle values), or two-dimensional quantities such as distribution functions.[33] SACs on defective (continuous) graphene present a particularly interesting case for feature extraction, being something in between a purely molecular and a periodic solid system, characterised by both different adsorption sites and different adsorbates.

Binding energy prediction can be achieved with a variety of different algorithms[34–37] including neural networks (e.g. deep learning), tree-based models (e.g. gradient boosting trees, extra tree or random forest) and kernel methods (e.g. support vector machine (SVM)). The best algorithm for materials prediction depends on the data set and the research objective,[38,24] however, tree-bases algorithms have the advantage of being able to process different types of data (categorical or on a continues scale) and generate a list of important features, making them flexible and interpretable. Among the available tree-based models random forests have been shown to perform well with a high dimensional feature spaces,[39] as is typical of low dimensional SACs.

In this study, we apply machine learning to predict the binding energy of molecules on metal atoms stabilised with doped, defective graphene, based on an large number of electronic structure simulations and an extensive search of the

feature space. Considerable focus has been given to feature extraction and engineering to determine which types of features are most suitable to describe this hybrid system, and how applicable they will be to other 2D-materials. As a tacit desire of the research community is to use machine learning to inform research planning and experimental design, we also take this opportunity to determine whether reliable predictions can be made based on inputs used to generate the data alone, rather than the outputs from costly calculations. We used random forest (RF) and support vector machine (SVM), as both methods are well suited to small data sets with a large feature space.[38] We find that both methods are capable of predicting the binding energy from input (unrelaxed) configurations with an accuracy of $R^2 = 0.865$ for RF and 0.866 for SVM. This could be particularly useful to assist in the pre-selection of a few specific structures out of thousands of possibilities, and reduce the number of simulations required by orders of magnitude. Our newly defined features giving rise to these results can similarly be applied in other machine learning predictions of any structure/property relationship.

2. Methods

2.1. Data set collection

The data set was comprised of 1694 single structure density functional theory (DFT) optimisation calculations on rectangular sheets of graphene-based material with an area 20×17 Å. Each of the systems contained different nitrogen doped pores to stabilise single or pairs of metal atoms (Fe, Pt, Co and Ni). Here metal atoms are part of the surface. In this set, there are 96 surface structures without adsorbates and 1587 structures with adsorbates.

All of the original optimisation calculations are performed using the same density functional and computational hyper-parameters (such as super-cell volume, mass of carbon, and number of k-points), as described in previous publications.[8–10] Details to the calculations are in the Supporting Information.

The structures have been separated into five groups depending on the number of carbon vacancies and the number of nitrogen dopants, as illustrated in Figure 1. The notation is NxVy for x number of nitrogen and for y number of vacancies. As we can see, the active sites are either a double pore (2xN4V2, Figure 1b) or pores large enough to stabilise two metal atoms (N6V4, N6V6, N8V4, and N8V10, Figures 1c–1f). In the cases of N6 V4 and N8 V4, all possible combinations of the four metals (Pt, Co, Ni and Fe) were captured, and a subset of combinations for 2xN4V2, N6V6 and N8V10, as summarised in Table 1. The adsorbed species were H*, O*, OH* and OOH*. While this is not an exhaustive ensemble of possible configurations, there is sufficient structural and chemical diversity to provide considerable insights into 2D-catalyst design.

The surface structures were formed by removal and replacement of C atoms with N atoms in graphene. The pores were then filled with different metals and energy optimised with DFT. The input structures for the binding energy calculation are partly arbitrarily selected and partly automatically generated based on previous optimised structures with other metals. The arbitrary structures are molecules in either their DFT gas-phase geometry or previous optimised geometries on a similar surface on different surface sites and orientations. The automatically generated structures were all generated from the same set of adsorption site and geometries for molecules.

Table 2 shows the partitioning of the data by adsorbed species, where O* and H* include any number of oxygen and hydrogen atoms adsorbed in a structure, eg. one or two H*. In 2/3 of the hydrogen calculations there was a single hydrogen adsorbed, and in 1/3 there were 2 H-atoms on the surface. In four systems, a coverage of up to seven hydrogen were tested. As we can see from Table 1 and Table 2, the data set shows significant frequency imbalances for both the pore type and adsorbate species. For all models the data set was stratified with regards to these two features during cross-validation.

2.2. Machine learning algorithms

Regression was undertaken with random forest (RF) regressors that average over an ensemble of decision trees trained on the data.[40] Although they are more computationally expensive and difficult to implement than other methods, they generally lower the risk of over-fitting, are more accurate when dealing with high variance features and require limited hyper-parameters (requiring only the number of trees and depth, which is optional). Overall, the RF cost is still very small and only takes minutes compared to the hours and days of quantum simulations. Accuracy of the regression models was assessed by maximising the coefficient of determination (R^2), with under-fitting (bias) and over-fitting (variance) evaluated through comparisons of the R^2 from the cross-validation of the 80/20 train/test split. The split values were chosen as the dataset is small compared to the type of datasets these algorithms were intended for. Different splitting values (e.g. 75/25) gave slightly lower scores in pre-tests, but the learning curve (Figure S1 in Supporting Information) shows that the accuracy and generalisable of the prediction is largely unaffected with a training set of at least 1200 sample instances.

Random forests are collections of multiple decision trees. For a single tree at each node, the data is split into two groups (e.g. with a binary separation of a feature such as ‚Fe'; iron (1)

Table 1. Partitioning of the data set by pore type.

Type of Surface	N6V4	N8V4	2x(N4V2)	N6V6	N8V10
No. of samples	847	554	42	66	78

Table 2. Partitioning of the data set by adsorbate species.

Type	H*	OH*	OOH*	O*
No. of samples	585	423	332	247

or no iron (0)). Splitting based on feature values is repeated at each node until a leaf is reached. Each leaf predicts a narrow range of the binding energy, or the exact binding energy is the tree is deep. Features are selected by the mean reduction in tree impurity, which is the best estimation of the binding energy in this case.[41] The results from the trees are averaged to predict a single result. An advantage of RF is that features can be ordered depending on how many of the trees require the feature to make a decision, providing interpretable feature importance profiles.

To compare the results of RF we have also used support vector regression (SVR). Based on statistical learning, SVR is a type of support vector machine (SVM) that is known to generalise well on unseen data. SVR is characterized by the use of kernels, sparse solutions, and Vapnik-Chervonenkis control of the margin and the number of support vectors, resulting in an effective tool in real-value function estimation. With an appropriate kernel function, complex problems can be easily addressed. One of the main advantages of SVR is that its computational complexity does not depend on the dimensionality of the input space, but SVR does required more refinement than RF during implementation (hyper-parameter optimisation). Details of the SVR results in this study are provided in the Supporting Information.

K-fold cross-validation (CV) is a re-sampling procedure that provides an important measure of how well each model and hyper-parameter set generalises to unseen (validation) data. During CV the training set is split into a number of (*k*) folds, with a single fold reserved as a 'hold-out' fold for validation, while the model/hyper-parameter set is trained on the remaining (*k*-1) folds. A test of the performance of the model on the hold-out set provides single CV score, analogous to a series of smaller test/train splits. This process is repeated *k* times, with each fold contributing a validation score when assigned as the 'hold-out'. The mean of the *k* validation scores provides the CV score, while the standard deviation in the individual validation test scores offers a measure of confidence in the consistency of the model hyper-parameter set. In this study, we have defined CV uncertainty as 5 standard deviations in the CV set. For SVM the hyper-parameters were selected with grid-search and CV. These are: C = 100, cache_size = 200, coef0 = 0.0, degree = 3, epsilon = 0.1, gamma = 0.001, kernel = 'rbf'. For RF, there is no restriction on tree depth and 5,000 trees were used to train the model.

3. Results

Using the data set described in the Methods section, we began by collecting a list of conventional molecular features that define the scope of the study. This includes the total number of atoms of each element; the number of metal atoms; the distance between the two metals (0 represents only one atom in the system); the number of surface atoms; the number of valence electrons on the separated metals; the number of vacancies in the GR sheet; the adsorbate-type, and the type of GR pore. All of these features represent *inputs* to the original electronic structure simulation that remain unchanged following structural relaxation, but their importance in determining an optimal 2D-SAC are largely unknown.

The feature list was expanded to include other *output* features characterising the interatomic network that changes as a result of the DFT structure optimisation, each describing the system in different ways. We include periodic information to capture the structure of the graphene-based material (radial distribution) as well as molecular information to capture the surface chemistry (bond lengths and angles), and statistical information to capture the overall structural heterogeneity. For 2D-catalyst materials each of these three feature groups alone could describe the system uniquely. This can be understood from the fact that each feature group uses the 3D-atomic coordinates in a different way and every sample instance can be distinguished from every other sample instance.

The topological features were extracted from the interatomic configurations contained in the input or output structure files. These three structural features groups were then used with the unrelaxed structures to predict the binding energy of the adsorbed molecules purely from the *input* before the density functional theory (DFT) calculations of the adsorbate. For these 1694 sample instances, 96 surface structures and the adsorbed molecules (O*, OOH*, OH* and H*) were optimised with DFT.

3.1. Feature generation

The new topological features X are sorted into three different topological feature groups. These are bond lengths and angles X_{ba}, statistical features X_{stat} and partial radial distances X_{PRD}.

$$X_{ba} \cup X_{stat} \cup X_{PRD} = X \tag{1}$$

Firstly, considering the connectivity as a primary descriptor, inter-atomic bond lengths and angles were calculated for all atoms participating in the pore or adsorbate. The carbon atoms participating in the pore were designated within a distance of 2.1 Å from the centrosymmetric point between the two adjacent N atoms. Distances between all atom tes (H, O, N, C, Fe, Pt, Co and Ni) were calculated for each atom pair and all distances below 2.1 Å were classified as bonds, after testing different cutoffs (see Supporting Information). Bond angles were calculated for all atoms with two or more bonds.

The specific species (bond length and angle) at each site were recorded to generate a feature list for C–N, Pt–O–H, etc. This naming convention is consistent for all sample structures in the data set, with values that are specific to a given structure. The metals were also treated collectively as 'M' in addition to the specific metals (Pt, Co etc). These generalised bond lengths and angles (i.e. M–N, M–O–H, etc) were captured to maximise the number of features X_{ba} that can be defined. All the bond lengths and angle feature values x_{ba} were normalised by the mean of all equivalent types of bond lengths or angles in all samples:

$$x_{ba} = \frac{x - \mu}{\mu}, \quad x_{ba} \in X_{ba} \tag{2}$$

where x is the distance or angle between specific atoms, and μ is the ensemble average, defined as

$$\mu = \frac{1}{N_x}\sum_{i=1}^{N_x} x_i \tag{3}$$

with N_x being the total number of samples with a non-zero value of x. In cases where x is not defined (such as a Pt–N bond in structures without Pt) the feature was set to -1 to allow for discrimination and omission during subsequent machine learning. Failure to tag these cases in this way results in an inconveniently high number of NaNs that cannot be identified and eliminated.

Secondly, in recognition that many experimental characterisation methods measure averaged values over entire samples, the mean $\bar{x}$ and standard deviation x_{std} of all bond lengths and angles x_i for each individual sample, partitioned by atom type, were also calculated. These mean features $\bar{x}$ were also normalised. We separated C–C bonds from the whole system (Std/Mean CC bond length) from C–C bonds roundabout the pore (Std/Mean of bond C C). N_{ba} is the total number of a specific bond length or angle in a single sample.

$$\bar{x} = \frac{1}{N_{ba}}\sum_{i=1}^{N_{ba}} x_i, \quad \bar{x} \in X_{stat} \tag{4}$$

$$x_{std} = \frac{1}{N_{ba}}\sum_{i=1}^{N_{ba}} (x_i - \bar{x}), \quad x_{std} \in X_{stat} \tag{5}$$

Thirdly, considering distribution functions as primary descriptors, a series of discrete distances and angles were compiled for a specific number of atoms; referred to as the partial radial distance (PRD), and partial angular radial distribution (PARD). In this context *partial* reflects the fact that only the a limited number of atoms are included, depending on the substrate structure. As mentioned above, the centre of the pore was defined as the centrosymmetric point between the N dopants, which is insensitive to the exact position of the metal, buckling or other distortions affecting 2D-materials which may not be supported by a metal substrate, and is consistent for pores with one or two metal atoms.

PARD was defined as the corresponding angle from the first N-atom (*N*1) to the centre of the pore to the atom in the neighbour list. Including more neighbours in the PRD or PARD is superfluous, as it only captures more carbon atoms from the GR surrounding the pore but does not change the description of the active site; a smaller number of neighbours risks missing some of the critical atoms participating in the active site itself. In each case we have numbered the feature in order of occurrence (with respect to the original list of atoms in the structural file), though any other ordering or naming convention would also suffice. Overall this process resulted in 36 distances and 35 angles.

$$\vec{r}_c = \frac{1}{N_N}\sum_{i=1}^{N_N} \vec{r}_i \tag{6}$$

$$x_j = \left|\vec{r}_j - \vec{r}_c\right|, \quad \{n_j | n_j \le n_{36}\} \subseteq X_{PRD} \tag{7}$$

$$\theta_j = \arccos\left(\frac{(\vec{r}_c - \vec{r}_{N1}) \cdot (\vec{r}_{nj} - \vec{r}_c)}{|(\vec{r}_c - \vec{r}_{N1})|\,|(\vec{r}_{nj} - \vec{r}_c)|}\right), \quad \{\theta_j | n_j \le n_{36}\} \subseteq X_{PRD} \tag{8}$$

Here, N_N is the number of nitrogen atoms in the sample with the coordinates $\vec{r}_i$, which are used to calculate the vector to the centre of the pore $\vec{r}_c$. From all calculated distances x_j, between the $\vec{r}_c$ and all atoms in the system $\vec{r}_j$ only the 36 shortest distances and their corresponding angle θ_j are used for PRD features X_{PRD}. This method of choosing a specific number of atoms close to an adsorption centre (instead of all atoms in an area or in the whole structure) gives the advantage of creating a fixed number of features for systems where the total number of atoms can vary.

In total, this feature extraction process provided 1031 features for output and 1218 for input structures: 17 generalised, 688 (819 input) bond lengths and angles, 269 (311 input) statistical averages and standard deviations, and 71 PRD and PARD. The number of features for input and output is different, as these are automatically generated from the structures with a cutoff of 2.1 Å for bond lengths. After energy optimisation fewer atom pairs are below that value, which means fewer features are generated. The advantage of this method is that it can be implemented for any number of atoms and any 2D-system.

3.2. Dimension reduction

With so many features, it is almost certain that many will be highly correlated, but it will not be immediately obvious which ones. To identify and eliminate strongly correlated features we used a correlation matrix and automatically retained all features below 90% correlation (see tables S1-S2 in Supporting Information). Failure to eliminate strongly correlated features results in a variance error (overfitting). The 459 (output) and 583 (input) features retained for subsequent analysis and the 635 eliminated features, for completeness, are listed in the Supporting Information in tables S3–S6.

3.3. Regression

3.3.1. Prediction from output structures

Using the retained topological features we investigated their relationship with the target property label, the binding energy

(output from the original electronic structure simulations), and generated feature importance profiles to predict how the binding energy of adsorbates to graphene SACs may be tuned in practice. We have used random forest regression with 5,000 estimators (due to the large number of features) and no maximum tree depth (which means the nodes are expanded until all leaves are pure or until all leaves contain fewer than 2 samples). This model gave a training score of $R^2 = 0.990$, and a testing score of $R^2 = 0.952$, as shown in Figure 2. The cross-validation score is 0.936 when using purely energy optimised structures to generate features, here referred to as *output data*. These results were confirmed using support vector regression which gave a training score of $R^2 = 0.994$, and a testing score of 0.965 (the details of which are provided in the Supporting Information, Figure S6).

Structures with a residual over an absolute value of 1 eV are referred to as high variance structures (HVS). When using output data we found only seven HVS; three lower and four higher than the calculated (true) value. There is no clear trend in these predicted HVS values in terms of a specific structure, but six of the HVS were with hydrogen.

The 20 most important features for predicting the binding energy are shown in Figure 3 on a log scale. In this list five features are statistical (one mean and four standard deviations), ten are based on connectivity (three bond lengths and seven angles), and four are from pair distributions (three PRD and one PARD). Eight of the top 20 involve the generalised metal atom (M), but only one with the specific metal species (which contrasts with SVR, shown in Figure S5 of the Supporting Information). Half of these top 20 features are angular. This result indicates that three-body features contain more information than two-body features, and suggests that the higher level of specificity provided by the individual metals is not necessary for most SACs with RF. The binding energy can be confidently predicted using averaged measures of the pore structure of the type collected from spectroscopic instruments, and metal-dependent trends are either already in the general distances or in the noise. This is counter-intuitive but very useful, as it supports the possibility of directly using outputs from standard characterisation instruments on a large scale. As would be intuitively expected, the input feature 'Adsorbate-type' is also highly influential.

To better explain some of the features, three important features (one from each category) were plotted in Figure 4. In each plot the distribution of sample instances *vs* the binding energy is shown, with a separation of the different adsorbate species. The left plot shows the standard deviation of the H–O distance. The standard deviation is zero for all H, O and OH instances as well as for OOH, where the second H–O distance is larger than 2.1 Å. When only focusing on the yellow (OOH) markers, it is clear that the highest binding energies have a larger second OH distance than the 2.1 Å and hence 'std of bond H_O' = 0. Chemically, the stronger OOH bonds to the surface the longer the O–O bond, which increases the distance of the surface bonded O to the H, and decreases the direct OH bond length. As a result the standard deviation of the H–O bond increases.

In Figure 4, the middle plot shows the distance of the first oxygen-metal distance normalised by the mean of all metal oxygen distances *vs* the binding energy. The value −1 indicates systems with H adsorbed or without an oxygen metal distance below 2.1 Å; i.e. for a few systems, oxygen atoms were placed on the carbon or nitrogen atoms of the pore. For all other systems, the O_M1 instances are approximately 0.

In Figure 4, the right plot is the feature PRD8 *vs* the binding energy. Here the distances are from the middle of the pore to the 8th atom in the system of the 36 closest atoms, which are between 2.4 to 4.0 Å depending in the pore structure and independent of the adsorbate type.

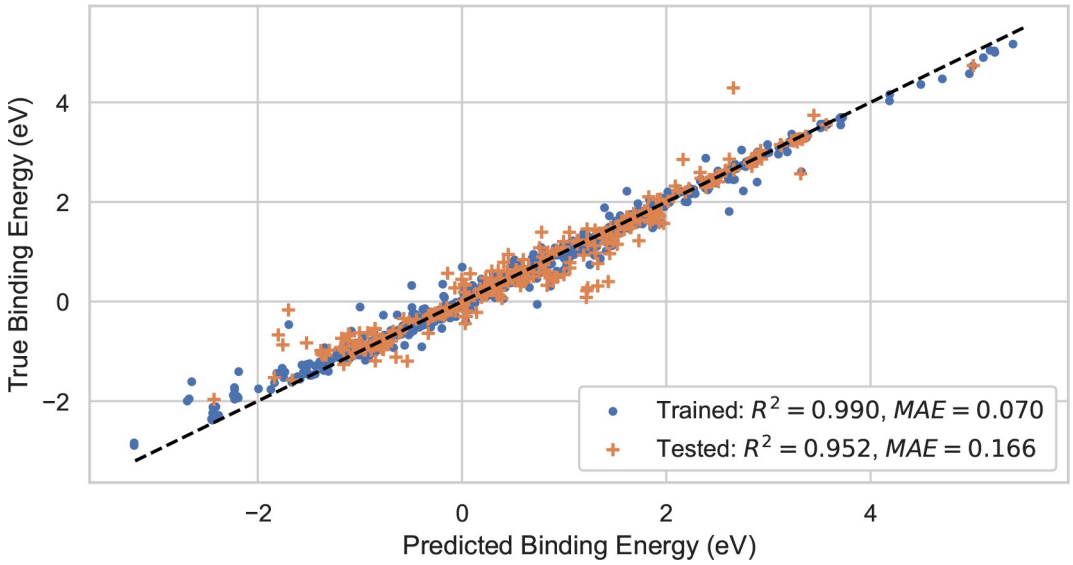

Figure 2. Scatter plot predicting the binding energy on N-doped porous graphene paired single atoms metal catalysts using random forest regression. The training set is shown in yellow and the test set in green, providing training and testing scores of $R^2 = 0.990$ and $R^2 = 0.952$, respectively.

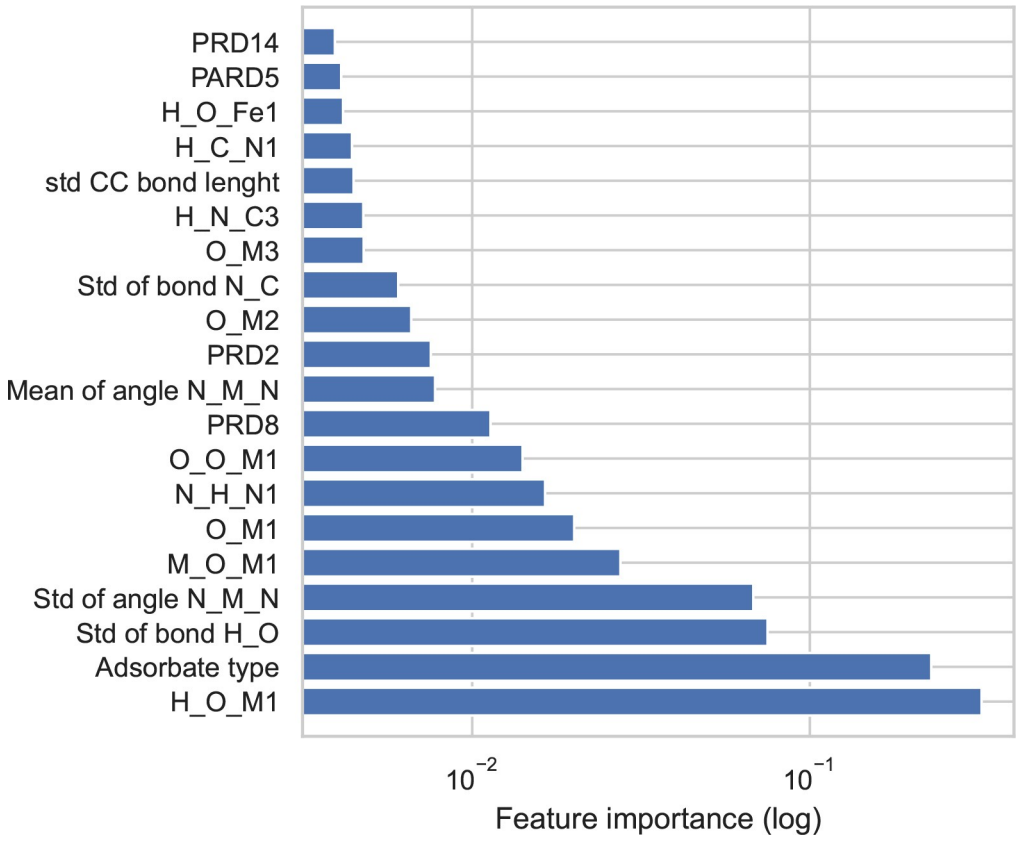

Figure 3. Feature importance profiles on a log scale showing the top 20 retained features for predicting the binding energy on porous graphene single atoms metal catalysts using random forest regression.

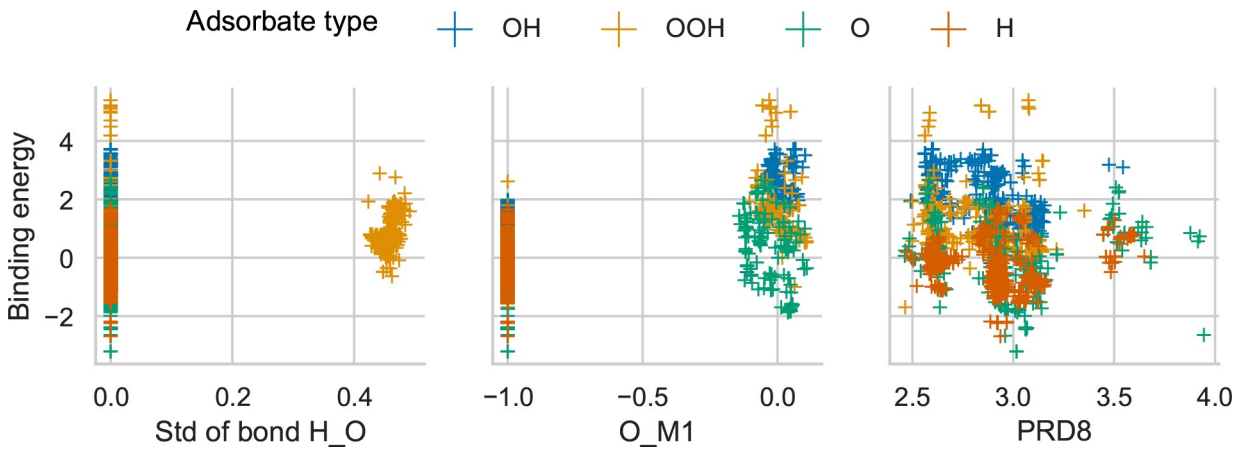

Figure 4. The distribution of three output features (from left to right: Standard deviation of the H_O distance, first mean normalised O–M distance, and 8th atom from the partial radial distance, respectively) *vs* the binding energy. The sample instances are coloured by their adsorbate type: blue – OH, yellow – OOH, green – O and red – H.

3.3.2. Prediction from input structures

While models predicting the patterns in observations is instructive, predicting possible outcomes based on input parameters alone is highly desirable.

By repeating the feature extraction process on the *input* structures before the electronic structure relaxation we developed values for the 1218 features, which were reduced to 583 by eliminating strongly correlated features using a correlation matrix with 90% threshold (see Supporting Information). We then repeated the machine learning using the same hyper-parameters to aid comparison. We find that the *input data* was remarkably capable of predicting the (output) binding energy with a testing accuracy of $R^2 = 0.834$ with 5-fold cross-validation

and a CV score of $R^2 = 0.865$. These results are shown in Figure 5. This high score evidences similar performance to the RF model trained on hundreds of features that required thousands of CPU hours to generate, and is confirmed using SVR to predict the binding energy with $R^2 = 0.866$ (see Supporting Information, Figure S8).

When using only the input structures, the 20 most important features, shown in Figure 6, differ from the order

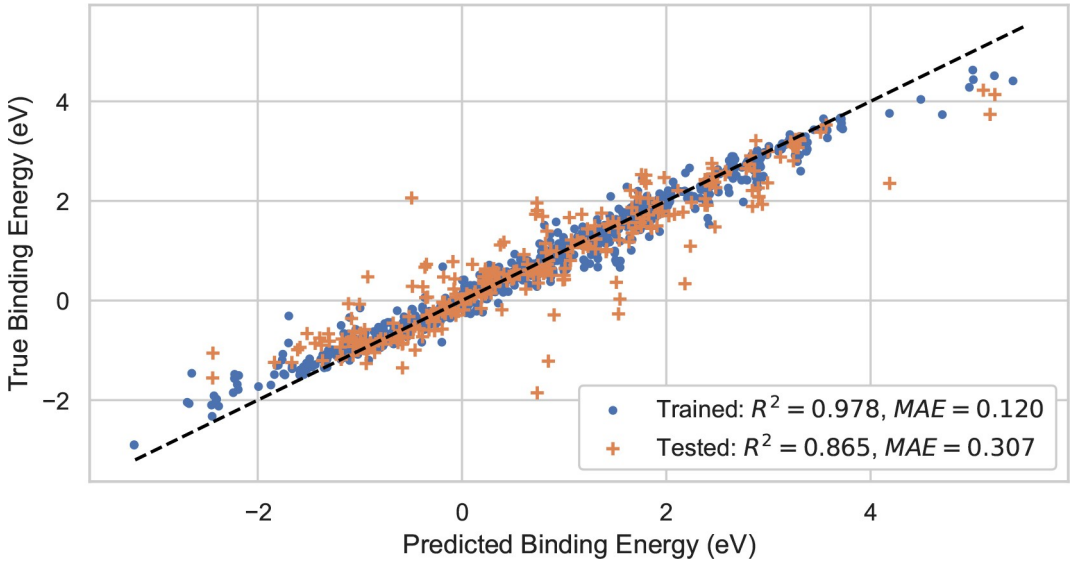

Figure 5. Scatter plot predicting the binding energy on porous graphene single atoms metal catalysts using random forest regression and only *input* features. The training set is shown in yellow and the test set in green, providing training and testing scores of $R^2 = 0.978$ and $R^2 = 0.865$, respectively.

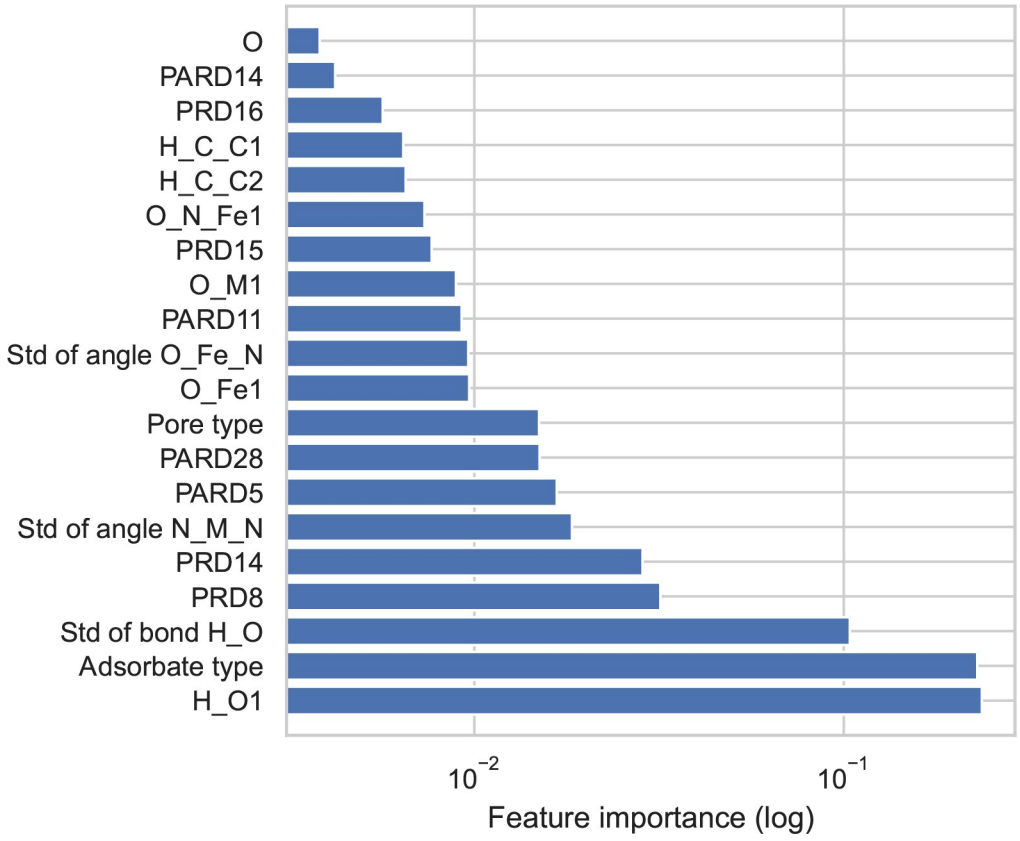

Figure 6. Feature importance profiles on a log scale showing the top 20 retained features for predicting the binding energy based on the *input* structures of porous graphene single atoms metal catalysts using random forest regression.

obtained using the outputs of electronic structure relaxations. The top 20 input features include eight distributions, three statistical features, six based on connectivity and three general features defining the input space. We see a decreased reliance on bond lengths and angles and an increased importance of distributional features, but a similar dependence on three-body features. The 'Adsorbate type', the 'Pore type' and 'O' (the number of oxygen atoms) are notable additions to the list. From all the metals included in this only Fe appears on the list more often than the generalised metal 'M'.

In the case of the input feature set there are 21 HVS (12 over-predicted and nine under-predicted HVS), all but one of which were from the arbitrarily produced input structures. There is no clear trend regarding adsorbate species but nine of these HVS have the same pore structure (N6V6), three are from the double pore 2x(N4V2), and two are from N8V10. This might be attributed to the fact that these pores are under-represented in the data set. When we restrict the study to the two most dominant pore structures (N8V4, N6V4) a higher accuracy of $R^2 = 0.900$ can be achieved (see Figure S3 in the Supporting Information). However, including these under-represented pores does reduce the number of HVS in the most important catalytic pores (N8V4, N6V4); improving the accuracy where it matters most. Increasing the diversity of the GR-defect reduces the over-fitting to the two main systems.

In Figure 7, the distribution of three input structure features *vs* the binding energy are shown and categorised by the adsorbate species. In the left plot, the standard deviation of the H–O distance is shown. Compared to Figure 4 we can see that the values are spread over a wider range, but most a close to 0.43, which is attributable to how the dataset was created. The first calculations were randomly distributed and the intra molecular distances could vary considerably. Based on these results later calculations used automatically produced input structures, so the yellow markers are aligned at a certain value.

The middle plot of Figure 7 shows the mean normalised metal oxygen distance. There are two main difference to the output structure distribution shown in Figure 4. Firstly, at a value of −1, there are more OH instances for input structures than for output structures. This is because initially more molecules were placed more than 2.1 Å away from the metal atom and their position changed during the DFT relaxation. Secondly, as we saw for the 'Std of bond H_O' feature, the range of values is wider for input features.

For the right plot in Figure 7, the PRD8 feature distribution for input structures is shown. Here we see an alignment of the markers at the same PRD8 value and different binding energy (e.g. at 3.9 Å). With the value 2.884707 Å, there are 32 occurrences. This highlights that the values of the PRD features depend on the SAC system or adsorption site, resulting in a distribution between 2.4 to 4 Å.

Further analysis was undertaken on the 17 features which are not structure-related (see Supporting Information table S3), with the exception of the distance between the two metals in the input structure. Here $R^2 = 0.704$ with a MAE of 0.526 are attained with 5-fold cross-validation. This drop in accuracy was expected as no adsorption site is defined, nor any information about the 2-D surface provided, and serves to highlight the importance of combining domain knowledge with interpretable machine learning models, rather than relying on scores alone.

To determine if this approach can reduce the number of calculations required to describe SACs, we also tested the HER predicted binding energy. Here the binding strength of hydrogen should be in a range of −0.15 to −0.33 eV to be more efficient than Pt/C.[42] In our test set there were 119 systems with hydrogen, 19 of which were predicted to be in this range, and 11 of which were DFT calculated to be in the range. With such a small dataset it is unclear if these instances include the strongest binding sites, but the fact that the method could select 19 from 119 calculations is a testament to the potential for reducing the number of overall calculations required to identify SACs for HER.

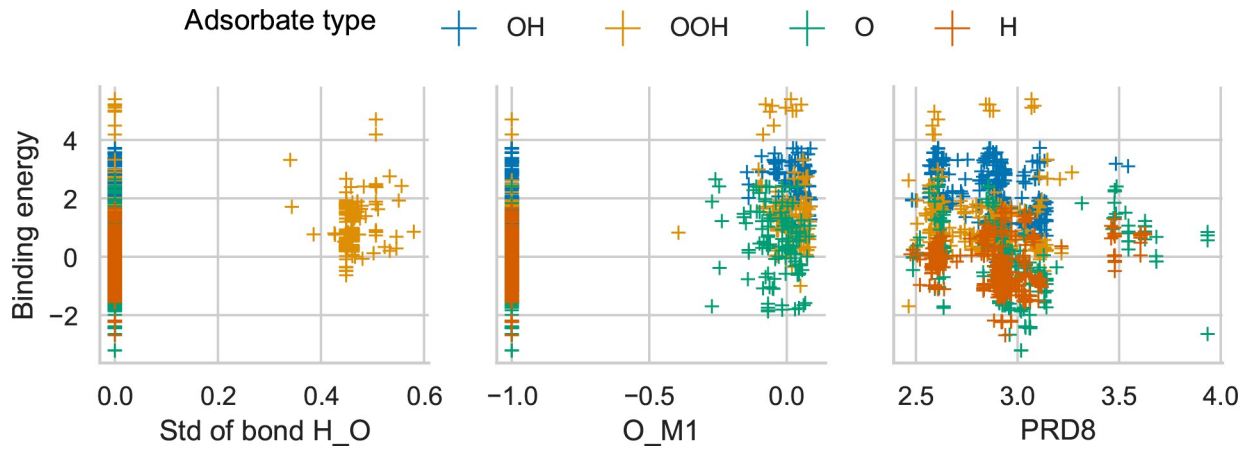

Figure 7. The distribution of three input structure features (from left to right: Standard deviation of the H–O distance, first mean normalised O–M distance and 8th atom from the partial radial distance) *vs* the binding energy. The sample instances are coloured by their adsorbate type: blue – OH, yellow – OOH, green – O and red – H.

4. Discussion

In this study we have used regression to predict the binding energy for a range of different adsorbates on graphene SACs. We show that producing a large feature space with three different feature groups can predict the binding energy accurately from output structures ($R^2 = 0.952$), or from input structures with an accuracy of $R^2 = 0.865$. The input data set consisted of 1587 unique samples and no high performance computing, except for relaxing the 96 surface structures with DFT. The features are automatically produced from the overall input structures and are applicable to various types of reactions, GR-defects and 2D-structures. The feature extraction methods used herein are equally applicable to classifications where a range of binding energies is predefined depending on the desired energy. However, in the present work the adsorbates have different ranges with considerable overlap, making classification less useful to this study.

As eluded to above, there were imbalances in the distribution of samples; some pore structures were under-represented. One of three under-represented structures were visibly deficient and produced a high number of HVSs. This indicates that while the calculation of different adsorbate species and systems can be used to increase the sample size for the prediction of a species or pore structure, this does not always mean that minority structures will be predicted well. In our case we estimate that at least 78 samples of a specific group have to be included to reduce the number of HVSs. The number of HVSs can also be reduced by automatically generating input structures from previously quantum mechanically relaxed surfaces; saving time, number of overall DFT calculations and improving machine learning accuracy.

The distribution features were also found to be particularly important for predicting the binding energy. In the Supporting Information we show the prediction with a higher and lower number of PRD features (26, 46 and 56 PRD) (see table S2) to further investigate this trend. We found no noticeable difference between any of these numbers and the accuracy of the trained models, either for the input or output data sets. For all systems 36 PRD includes all pore atoms and atoms adsorbed and was therefore considered the optimal balance between information and computational efficiency in this case. For other systems, where a ring or set of (under-coordinated) atoms around the adsorption site are assigned as the reference, fewer atoms in the PRD or PARD distributions may be sufficient. If the adsorption site does not include a defect other selection processes will be needed, but we recommend restricting the range of the number of redundant features where the values do not change. The same applies to the inclusion of bond lengths and angles.

Comparing the importance of the top 20 features from regression on output and input data sets, one of the most important features is the adsorbate type (encoded *via* classification with a number from 1 to 4), followed by angular three-body and interatomic two-body interactions. This means that if a specific angle exists the energy can be better estimated with the angle value than with the bond distances alone. We used three different ways to describe the pore structurally, all of which are represented in the top 20 features from both input and output data sets.

In the top 20 features from output data set the non-specific metal 'M' appears in eight features while a specific metal, 'Fe' only appears once. This is not the case for input structures, where more 'Fe' than 'M' are of relevance. This is reflected in a previous publication, as the binding energies observed over the Fe containing surface changed significantly, while with the other metals the potential energy profile seemed very flat. This indicates that all other metals (Pt, Co, Ni) can be treated equally when purely regarding topological features (as opposed to chemical or electronic features), but Fe is clearly a special case worth of more focused research. Undoubtedly other 2D-materials would benefit from similar investigations.

Another point worth mentioning is that, as mentioned above, the RF features are selected and split by a node in two parts by a certain value. This could mean that a structural feature is selected on its existence rather than by a distance (or angle) value being larger or smaller a certain average. For Fe and the general metal 'M' this could mean that when using the output structures the distances between metals and adsorbed species are more diverse and can be well separated in RF-leaves by the general distance to the metal 'M'. On the other hand, when using only input structures, metal and adsorbate distances are less diverse as a lot of structures are automatic prepared. In this case the algorithm may only check if a bond (e.g. O_Fe1) exists with a value larger than −1. In a further node (e.g. the 'Std of angle N_M_N') this would give more information of the specific surface and further improve the binding energy prediction, as this angle was optimised in the surface structure calculation. It has been previously reported that the environment of the metal atoms is correlated to the binding energy which further emphasises that the geometric description of the accurate DFT optimised surface with the approximated adsorption site of adsorbate are important descriptors for machine learning studies of catalysts.

Overall this study shows that for small material data sets, a large feature space with diverse and automated structural feature descriptions can be useful in planning future workflows and focusing calculations on the materials that matter most. For example, to explore HER one would automatically create multiple potential adsorption sites or multiple different potential active sites, then estimate the binding energies with our RF model. These estimated energies could then be screened and a limited subset of structures calculated with quantum chemical simulations. These results could then be added to the data set and the models retrained until self-consistency is achieved. Running the machine learning algorithm to estimate binding energies only takes minutes on a laptop, while each calculation takes hours or days on multi-core high performance supercomputers. Once the optimum binding energy is found the specific systems can become the target of more detailed analysis.

5. Conclusion

Three different descriptors were created to predict the binding energy of small molecules on N-doped GR defects, consisting of hundreds of general, molecular and statistical features. These include discrete partial radial distribution of distances and angles from the center of the defect, bond lengths and angles below 2.1 Å, and the mean and standard deviation of these bond lengths and angles. Using random forest regression these features were shown to accurately predict the binding energy of various molecules on metals using output structures relaxed from quantum chemical simulations with $R^2 = 0.952$, and unrelaxed input structures requiring no costly simulations with $R^2 = 0.865$. These results were confirmed with support vector regression, and features from all three groups were confirmed as important in determining the results.

With this method, the expensive computational testing of different adsorbates configurations and sites for SAC can be reduced. This could save resources and researchers' time by finding the right material morphology and composition. As this method predicts most structures binding energy within 1 eV and many materials can be excluded because they are beyond that. However, that is not sufficient to determine a good catalyst where the differences of 0.1 eV is significant for the performance. Nevertheless, it is a inexpensive method for screening a wide range of 2D-materials followed by more accurate calculations. Particular for the ORR, three binding energies are necessary to predict the performance (reduce the overpotential). Here ruling out materials because one species is predicted to be too strongly or weakly bond, saves the calculation of three adsorbed species.

The ability to estimate the binding energy and structural feature importance from input structures, whether automatically to sample a configuration space or manually based on domain knowledge or intuition, allows machine learning to be used as a research planning tool, in addition to providing analysis. We found that generating structures based on domain knowledge or intuition increases the number of high variance structure (where models perform poorly) but increases the diversity of the training data and increases the overall accuracy of well-represented materials. Ideally, a homogeneous input set would be used, where domain knowledge is incorporated and captured during automated structure generation, so all structures are equally created.

While we currently exclude other property labels such as the magnetic moment or charges on specific atoms, the features extracted herein are applicable to other structure/property relationships, and will form the basis for future work. The same principle applies to other 2D-materials, defects and catalytic reactions.

Acknowledgements

We acknowledge access to computational resources at the NCI National Facility through the National Computational Merit Allocation Scheme supported by the Australian Government. This work was also supported by resources provided by the Pawsey Supercomputing Centre with funding from the Australian Government and the Government of Western Australia. We also acknowledge support from the Queensland Cyber Infrastructure Foundation (QCIF) and the University of Queensland Research Computing Centre.

Conflict of Interest

The authors declare no conflict of interest.

Keywords: molecular topology · feature engineering · single atom catalysts · random forest regression · graphene defects

[1] X.-F. Yang, A. Wang, B. Qiao, J. Li, J. Liu, T. Zhang, *Acc. Chem. Res.* **2013**, *46*, 1740–1748.
[2] B. Qiao, A. Wang, X. Yang, L. F. Allard, Z. Jiang, Y. Cui, J. Liu, J. Li, T. Zhang, *Nat. Chem.* **2011**, *3*, 634–64.
[3] J. Lin, A. Wang, B. Qiao, X. Liu, X. Yang, X. Wang, J. Liang, J. Li, J. Liu, T. Zhang, *J. Am. Chem. Soc.* **2013**, *135*, 15314–15317.
[4] N. Cheng, S. Stambula, D. Wang, M. Norouzi Banis, J. Liu, A. Riese, B. Xiao, R. Li, T. Kong Sham, L. M. Liu, G. A. Botton, X. Sun, *Nat. Commun.* **2016**, *7*, 1.
[5] S. Liang, C. Hao, Y. Shi, *Chem. CatChem* **2015**, *7*, 2559–2567.
[6] A. Wang, J. Li, T. Zhang, *Nat. Rev. Chem.* **2018**, *2*, 65–81.
[7] B. Motevalli, B. Sun, A. S. Barnard, *J. Phys. Chem. C* **2020**, *124*, 7404–7413.
[8] L. Zhang, J. Melisande, T. A. Fischer, Y. Jia, X. Yan, W. Xu, X. Wang, J. Chen, Dongjiang Yang, H. Liu, L. Zhuang, M. Hankel, D. J. Searles, K. Huang, S. Feng, C. L. Brown, Xiangdong Yao, *J. Am. Chem. Soc.* **2018**, *140*, 10757–10763.
[9] M. A. Hunter, J. Fischer, M. Hankel, Q. Yuan, D. J. Searles, *J. Chem. Inf. Model.* **2019**, *59*, 2242–2247.
[10] M. A. Hunter, J. Fischer, Q. Yuan, M. Hankel, D. J. Searles, *ACS Catal.* **2019**, *9*, 7660–7667.
[11] H. Xu, D. Cheng, D. Cao, X. C. Zeng, *Nat. Catal.* **2018**, *1*, 339–348.
[12] G. Gao, S. Bottle, A. Du, *Catal. Sci. Technol.* **2018**, *8*, 996–1001.
[13] H. Zhang, G. Liu, L. Shi, J. Ye, *Adv. Energy Mater.* **2018**, *8*, 1–24.
[14] B. Hammer, J. K. Nørskov, *Adv. Catal.* **2000**, *45*, 71–129.
[15] J. Hafner, C. Wolverton, G. Ceder, *MRS Bull.* **2006**, *31*, 659–668.
[16] A. Jain, G. Hautier, C. J. Moore, S. P. Ong, C. C. Fischer, T. Mueller, K. A. Persson, G. Ceder, *Comput. Mater. Sci.* **2011**, *50*, 2295–2310.
[17] T. Heine, *Front. Mater.* **2014**, *1*.
[18] A. Jain, Y. Shin, K. A. Persson, *Nat. Rev. Mater.* **2016**, *1*, 15004.
[19] B. R. Goldsmith, J. Esterhuizen, J.-X. Liu, C. J. Bartel, C. Sutton, *AIChE J.* **2018**, *64*, 2311–2323.
[20] Y. Liu, T. Zhao, W. Ju, S. Shi, *J. Materiomics* **2017**, *3*, 159–177.
[21] Z. Li, S. Wang, W. S. Chin, L. E. Achenie, H. Xin, *J. Mater. Chem. A* **2017**, *5*, 24131–24138.
[22] T. S. Choksi, L. T. Roling, V. Streibel, F. Abild-Pedersen, *J. Phys. Chem. Lett.* **2019**, *10*, 1852–1859.
[23] R. A. Hoyt, M. M. Montemore, I. Fampiou, W. Chen, G. Tritsaris, E. Kaxiras, *J. Chem. Inf. Model.* **2019**, *59*, 1357–1365.
[24] P. Schlexer Lamoureux, K. T. Winther, J. A. Garrido Torres, V. Streibel, M. Zhao, M. Bajdich, F. Abild-Pedersen, T. Bligaard, *ChemCatChem* **2019**, *11*3581–36.
[25] M. Fernandez, H. Shi, A. S. Barnard, *Carbon* **2016**, *103*, 142–150.
[26] M. Fernandez, J. I. Abreu, H. Shi, A. S. Barnard, *ACS Comb. Sci.* **2016**, *18*, 661–664.
[27] H. Yang, Z. Zhang, J. Zhang, X. C. Zeng, *Nanoscale* **2018**, *10*, 19092–19099.
[28] M. Fernandez, H. Shi, A. S. Barnard, *J. Chem. Inf. Model.* **2015**, *55*, 2500–2506.
[29] T. Cox, B. Motevalli, G. Opletal, A. S. Barnard, *Adv. Theory Simul.* **2020**, *2*, 1900190.
[30] K. Hansen, F. Biegler, R. Ramakrishnan, W. Pronobis, O. Anatole von Lilienfeld, K.-R. Müller, A. Tkatchenko, *J. Phys. Chem. Lett.* **2015**, *6*, 2326–2331.

[31] B. Motevalli, A. J. Parker, B. Sun, A. S. Barnard, *Nano Futures* **2019**, *3*, 045001.
[32] D. C. Elton, Z. Boukouvalas, M. S. Butrico, M. D. Fuge, P. W. Chung, *Sci. Rep.* **2018**, *8*, 9059.
[33] K. T. Schütt, H. Glawe, F. Brockherde, A. Sanna, K. R. Müller, E. K. U. Gross, *Phys. Rev. B* **2014**, *89*, 205118.
[34] Z. W. Ulissi, M. T. Tang, J. Xiao, X. Liu, D. A. Torelli, M. Karamad, K. Cummins, C. Hahn, N. S. Lewis, T. F. Jaramillo, K. Chan, J. K. Nørskov, *ACS Catal.* **2017**, *7*,6600–660.
[35] T. Nhat Nguyen, T. T. Phuong Nhat, K. Takimoto, A. Thakur, S. Nishimura, Junya Ohyama, I. Miyazato, L. Takahashi, J. Fujima, K. Takahashi, T. Taniike, *ACS Catal.* **2020**, *10*, 921–932.
[36] T. Toyao, K. Suzuki, S. Kikuchi, S. Takakusagi, K. Ichi Shimizu, I. Takigawa, *J. Phys. Chem. C* **2018**, *122*, 8315–8326.
[37] H. Zhao, X. Zhang, L. Ji, H. Hu, Q. Li, *Corros. Sci.* **2014**, *83*, 261–271.
[38] A. S. Barnard, B. Motevalli, A. J. Parker, J. M. Fischer, C. A. Feigl, G. Opletal, *Nanoscale* **2019**, *11*, 19190–1920.
[39] P. Geurts, D. Ernst, L. Wehenkel, *Mach. Learn.* 2006, 63, 42.
[40] L. Breiman, R. Forests, *Mach. Learn.* **2001**, *45*, 5–32.
[41] U. Grömping, *Am. Stat.* **2009**, *63*, 308–319.
[42] J. K. Nørskov, T. Bligaard, A. Logadottir, J. R. Kitchin, J. G. Chen, S. Pandelov, U. Stimming, *J. Electrochem. Soc.* **2005**, *152*, J2.

Manuscript received: March 28, 2020
Revised manuscript received: May 26, 2020
Accepted manuscript online: June 2, 2020
Version of record online: September 6, 2020