NJC

ROYAL SOCIETY OF CHEMISTRY

PAPER

View Article Online
View Journal | View Issue

Cite this: *New J. Chem.*, 2020, **44**, 8545

Understanding the ML black box with simple descriptors to predict cluster–adsorbate interaction energy†

Sheena Agarwal,[ab] Shweta Mehta[ab] and Kavita Joshi *[ab]

Density functional theory (DFT) is currently one of the most accurate and yet practical theories used to gain insight into the properties of materials. Although successful, the computational cost required is still the main hurdle even today. In recent years, there has been a trend of combining DFT with Machine Learning (ML) to reduce the computational cost without compromising accuracy. Finding the right set of descriptors that are simple to understand in terms of giving insights about the problem at hand, lies at the heart of any ML problem. In this work, we demonstrate the use of nearest neighbor (NN) distances as descriptors to predict the interaction energy between the cluster and an adsorbate. The model is trained over a size range of 5 to 75 atom clusters. When the training and testing is carried out on mutually exclusive cluster sizes, the mean absolute error (MAE) in predicting the interaction energy is ~ 0.24 eV. MAE reduces to 0.1 eV when testing and training sets include information from the complete range. Furthermore, when the same set of descriptors are tested over individual sizes, the MAE further reduces to ~0.05 eV. We bring out the correlation between dispersion in the nearest neighbor distances and variation in MAE for individual sizes. Our detailed and extensive DFT calculations provide a rationale as to why nearest neighbor distances work so well. Finally, we also demonstrate the transferability of the ML model by applying the same recipe of descriptors to systems of different elements like (Na_{10}), bimetallic systems (Al_6Ga_6, Li_4Sn_6, and $Au_{40}Cu_{40}$) and also different adsorbates (N_2, O_2, and CO).

Received 6th February 2020,
Accepted 28th April 2020

DOI: 10.1039/d0nj00633e

rsc.li/njc

Introduction

Artificial intelligence (AI) has unleashed a new era leading to a paradigm shift in the fields of science and engineering. And hence, it is rightly referred to as both the ''fourth paradigm of science''[1] and the ''fourth industrial revolution''.[2] The increasing application that it is finding in the field of science and particularly the chemical domain was unimaginable just a few decades ago. Machine Learning, a subset of AI is growing rapidly and gaining popularity amongst the scientific community. With the advent of ML assisted by open access materials property databases[3–5] new doors have opened up. ML is finding application to a variety of problems like rapid materials discovery, energy storage, catalysis, hydrogen storage, and many more.[6–15] Chemistry, or specifically quantum chemistry, is yet another field in which we see rapid progresses being made with the nexus of ML and first principles quantum calculations.[16–25]

The key to unlock the power that ML holds, is in finding the right set of descriptors for any model. Better the description of the problem at hand, better the learning and predictive capacity of any ML algorithm. And hence, the search of accurate descriptors is still an active area of research.[26–29] The developed set of features then find varied applications like finding similarity between two structures,[30,31] finding the structure–activity relation for various systems,[32–34] screening the chemical space to discover novel materials of desired properties[7,8] or even predict properties for a given material.[9–12] In a study by Hansen *et al.*, they outlined a number of established machine learning techniques and investigated the influence of the molecular representation on the ML method's performance. The best methods achieve prediction errors of 3 kcal mol^{-1} (0.13 eV) for the atomization energies of a wide variety of molecules.[35] An issue that arises during this is the interpretability of the chosen set of descriptors. If the descriptors that work excellently for any ML algorithm do not translate in terms of the scientific insights of the problem, then its transferability becomes restricted.[36] It limits the use of ML models to a computational black-box.

In this work, we aim at dealing with this question on gaining insights about the problem at hand by choosing the right set of

[a] *Physical and Materials Chemistry Division, CSIR-National Chemical Laboratory, Dr. Homi Bhabha Road, Pashan, Pune-411008, India. E-mail: k.joshi@ncl.res.in, kavita.p.joshi@gmail.com*

[b] *Academy of Scientific and Innovative Research (AcSIR), India*

† Electronic supplementary information (ESI) available. See DOI: 10.1039/d0nj00633e

descriptors to describe the system. We train our ML model on various recipes for correct description of the system and understand systematically as to why a particular set of descriptors perform well. We study the interaction of an H atom as an adsorbate with Al_n clusters in the size range of 5–75 (n = 5 to 20 and 25, 36, 47, 55, 67, 74, 75). Owing to the high surface area and hence enhanced catalytic activity, clusters have always been of immense interest for catalysis.[37–44] Also in this size regime where every atom counts, addition or removal of just one atom dramatically changes its properties.[40,41,45–56] Thus, it becomes very difficult to highlight general trends in this size range. And hence, to model the interaction of clusters with an incoming adsorbate, all possible adsorption sites for all the clusters must be scanned, which in turn leads to a prohibitively large number of DFT calculations. To overcome this problem, we have used data driven algorithms of ML to predict the site specific interaction energies for Al_n clusters. Resorting to ML for discovering correlations between geometric structure and catalytic activities[57,58] of metal surfaces as well as nanoparticles[59] is becoming more common. Recently, an ML scheme was proposed to understand the catalytic activities based on local atomic configurations and applied to study direct NO decomposition on RhAu alloy nanoparticles.[60] A local structural similarity kernel known as a smooth overlap of atomic positions (SOAP) was used to find similarities between two geometries based on structural descriptors. Gasper *et al.* used the gradient-boosting algorithm, for the prediction of CO adsorption energies on Pt clusters.[61] They built predictive models of site-specific adsorbate binding on realistic, low-symmetry nanostructures, with MAE $\sim$ 0.1 eV (with respect to DFT). Descriptors used during the training of the ML model in this study comprised of d-band center energy, s and p band center energies, Bader charges, generalized coordination number, *etc.* With the intention of capturing the cluster adsorbate interaction, we developed a method of simple descriptors that throws light on the cluster chemistry. The use of descriptors like the Coulomb Matrix (CM),[62] Smooth Overlap of Atomic Positions (SOAP),[63] Atom Centered Symmetry Function (ACSF) and so on is seen in problems of finding structural similarities. These functions, while they describe the structure to a very reasonable extent, are limited to local environments.

In the present work, we propose nearest neighbor distances as descriptors to predict the interaction energy between a cluster and an adsorbate. Although the model is trained on Al clusters in the present work, it is demonstrated to work for any homogeneous (Na_{10}) as well as heterogeneous clusters (Al_6Ga_6, Li_4Sn_6, $Au_{40}Cu_{40}$). Furthermore, we also demonstrate a one to one correlation between dispersion in the NN distances and variation of MAE for individual clusters. Finally, the rationale as to why NN distances work so well for predicting the interaction energy is brought out by our DFT calculations.

Computational details

We have computed the interaction energy of various atoms like H, N and molecules like N_2, O_2, and CO with Al_n clusters in the size range of 5–75 (n = 5 to 20 and 25, 36, 47, 55, 67, 75). Also, the interaction of a H atom with a cluster of another element like Na_{10} and bimetallic clusters like Al_6Ga_6, Li_4Sn_6, and $Au_{40}Cu_{40}$ were computed. All these resulted into about 18 000 single point calculations. The adsorbates were placed at the on-top position of all the surface sites (*i.e.*, surface atoms) for the selected clusters in this size range. The GS geometries for all the clusters were taken from previously reported work.[64–67] As shown in Fig. 1, the adsorbate was kept along the outward radial vector from the center of mass of the cluster to the surface atom. The distance of the adsorbate was varied between 1.30 Å to 3.00 Å from the surface site. All the calculations were carried out within the Kohn–Sham formulation of DFT. Projector Augmented Wave potential[68,69] was used, with Perdew–Burke–Ehrzenhof (PBE)[70] approximation for the exchange–correlation and generalized gradient[71] approximation, as implemented in planewave, pseudopotential based code, VASP.[72–74] Cubic simulation cell, with the image in each direction separated by at least 15 Å of vacuum, was used. An energy convergence criteria of 10^{-4} eV was used for SCF calculations. Interaction energy between the cluster and adsorbate was calculated using the formula:

$$E_{\text{I.E.}} = E_{\text{system}} - (E_{\text{cluster}} + E_{\text{adsorbate}})$$

where E_{system} is the energy of the cluster + adsorbate system, E_{cluster} is the energy of the bare cluster and $E_{\text{adsorbate}}$ is the energy of the bare adsorbate. The same value of vacuum was used for all the clusters so all the parameters are consistent throughout for all the sizes to avoid any shift in energies.

Data collected from the DFT calculations was then used to train an ML model. We used the Gradient Boosting Regression

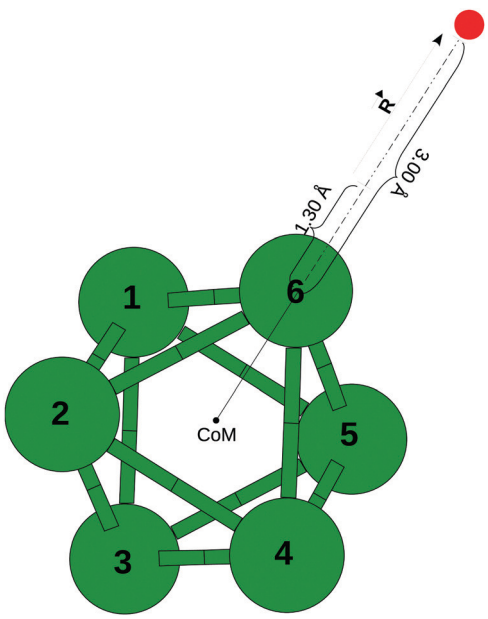

Fig. 1 Schematic showing the radial vector along which the adsorbate was placed. Al_6 cluster is shown with a center of mass marked at the center. The adsorbate is placed along the radial vector pointing outwards from the center of mass to the surface atom. Distance between the adsorbate and surface atom is varied from 1.30 Å to 3.00 Å.

(GBR) algorithm as implemented in the scikit-learn python package.[75] The GBR was selected after comparing it against five other regression algorithms *viz.* Linear Regression, Ridge Regression, LASSO, Artificial Neural Networks, and Stochastic Gradient Descent (SGD). GBR is a regression technique that uses decision tree based classifiers as weak learners. The tree-based models finally develop into a sequence of trees with the right choice of descriptors positioned correctly at different nodes. The relative importance of the descriptors that we arrive at finally is the model that the algorithm has developed to further test the unseen data. This model basically predicts the interaction energy for newer data points with the proposed descriptor ranking at the end of training. Architecture of the GBR is explained in detail in the ML section of the ESI.† Furthermore, we used the mean squared error function as our loss function (*i.e.*, the objective function to be optimized). Tree-based algorithms need to be tuned for the best values of various hyper parameters. An exhaustive grid search was carried out to find the best parameter values of an estimator. The tuned hyper parameters were: n_estimators, max_depth, min_samples_split, loss, learning_rate. Different train-test splits *viz.* 60–40, 75–25, and 80–20 were tested. Finally, the train and test datasets were split in 75–25 percentage. To avoid bias, random shuffling of the complete data set before the train-test split were performed. This exercise was repeated for 50 randomly generated seeds for shuffling. 5-Fold cross validation was performed to test the accuracy of the model. MAE was used as the scoring parameter during cross validation. Multiple checks like plotting the validation curve and learning curves (see Fig. 2) were used to ensure that the model did not overfit the data. We can see from the Learning curve that both training score and cross validation score reach a point of stability and hence do not represent over fitting. However, after 400 points, there is a minimal decrease in the train score while the cross validation score continues to saturate. And hence, in general, 400 data points would be sufficient to train this model. Finally, feature ranking was plotted to understand the relative importance for a chosen set of descriptors, which as explained later translates in terms of physical understanding of the problem. MAE for the test set is reported throughout the work to understand the ML results in comparison with DFT.

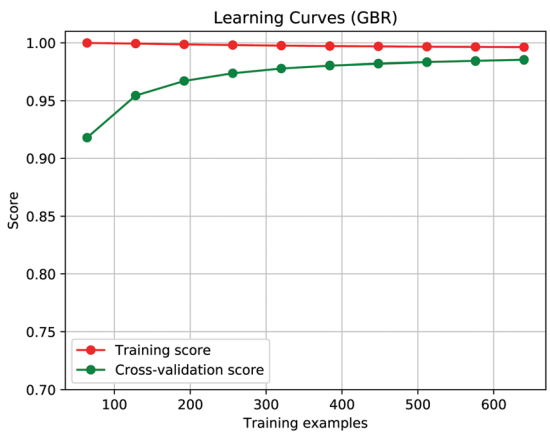

Fig. 2 Learning curve plotted with training score and cross validation score as model verification against over fitting.

Results and discussion

Choosing the right set of descriptors lies at the heart of any Machine Learning problem. Structural properties like size of the cluster, structural arrangement of atoms within a cluster, orientation of approaching adsorbate, *etc.* play a role in governing the interaction of the cluster with an incoming adsorbate. The set of descriptors used were based on the atomic arrangement of atoms within the cluster and properties of the cluster as a whole. Descriptors like the number of surface atoms (SA), number of unique adsorption sites on the surface (UAS), and coordination number (CN) would vary with changing cluster sizes. SA gives information about the overall shape of the cluster. A cluster which is elongated would have a greater fraction of atoms on the surface than one which is spherical. The number of unique adsorption sites on the surface would provide information about the symmetry of the cluster. A symmetrical cluster like Al_{13} has only 2 inequivalent sites on the surface whereas that of Al_{12} has 7 inequivalent sites on the surface. Thus UAS brings out the symmetry within the cluster. On the other hand, coordination number is a site specific descriptor. Similarly, descriptors like the nearest neighbor distances would provide information regarding the specific adsorption site. We used the distances of the first 5 nearest neighbors from the adsorption site as descriptors (nn1 to nn5) while training ML model for all 23 cluster sizes. The size specific descriptors like size, SA, and UAS provide information about the cluster whereas the nearest neighbor distances provide information about the specific site within the cluster. These descriptors were the ones that did not require any expensive calculations but were calculated from the geometry of the clusters. The ML model was applied to Al_n clusters for combinations of these descriptors. In Fig. 3, the feature ranking for all the geometric descriptors discussed above is shown. To avoid any kind of bias, the model was tested for 50 random test-train splits. Each time the feature ranking was noted. Values of 50 trials were divided in 4 quartiles centered at the median. The error bars represent the variation in the feature ranking values of each feature over 50 trials in the 1st and 4th quartile. We see that nn1 (the distance between the adsorbate and nearest atom in the cluster) has the highest importance followed by another adsorption site specific descriptor CN. The remaining size specific descriptors like the UAS, size and SA rank higher than the farther neighbor distances. Hence, we see a mixture of both site as well as size specific descriptors ranking highly. For the 23 clusters in this size range, it can be seen that using the geometric set of descriptors, the model performed well with MAE $\sim$ 0.10 eV (Fig. 4(a)). To examine the effect of changing atomic environment on the interaction energy, we further used only interatomic distances that essentially capture the site specific variations in atomic arrangements.

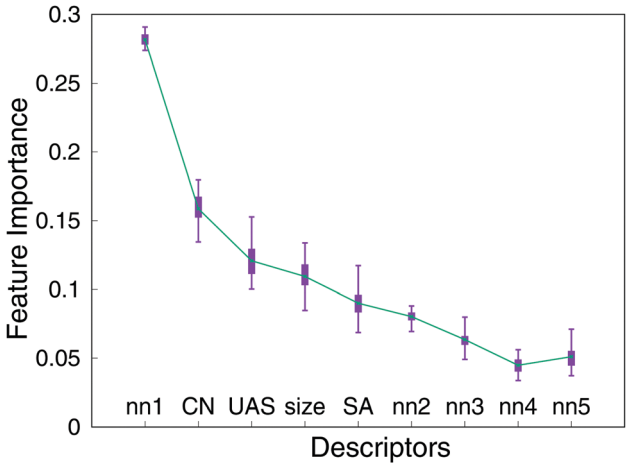

Fig. 3 Feature ranking plots for 23 clusters averaged over 50 trials. It is interesting to note that the first two features (nn1 and CN) are site specific features.

The ML model was trained on only the nearest neighbor (nn1 to nn5) distances as descriptors along with size. We report that though the accuracy of the model goes down to MAE $\sim$ 0.12 eV, the increase in errors is not substantial (Fig. 4(b)). And hence, our model does work to predict the site specific interaction between adsorbate and atoms for clusters over a size range with an MAE of 0.10 eV. Now, to test the model for size specific interaction, the model was further trained on data points for cluster sizes from 5 to 20 except for a few on which it was then tested. Every time model performed well giving out an MAE in the range of $\sim$0.23–0.26 eV. The same exercise when repeated with only the nearest neighbor distances as descriptors along with size, giving an error in the range of 0.24–0.28 eV. This result is particularly important as it demonstrates that computation of interaction energies with DFT could be completely bypassed with reasonable accuracy. All these experiments also point at the fascinating possibility of unfolding questions related to size or even site specific cluster chemistry by understanding the atomic arrangement of atoms. And hence, to understand the role that NN distances are playing we perform studies on individual cluster sizes with varying distance information. Cases of both small and large cluster sizes are shown in Tables 1 and 2, respectively. The reported cases for larger clusters are a balanced mixture of ordered and disordered clusters.

A trend of reducing prediction errors with increasing system representation was seen for all smaller ($N < 20$) clusters that we have studied. For any surface site of an N atom cluster, there will be $N - 1$ distances as descriptors. The model was trained each time by gradually including more descriptors *i.e.* distances. In Table 1 we list the variation in MAE as a function of increasing number of descriptors for smaller cluster sizes. The variation in MAE is correlated with interatomic distances from the H atom. We will discuss this further by closely analyzing the specific case of Al_{13}. In Fig. 5(a) we have plotted MAE as a function of the number of descriptors (number of nn distances between the adsorbate and atoms within the cluster) used to fit the model for predicting interaction energies for Al_{13} on the $y1$ axis. We have also simultaneously plotted the nearest neighbor distribution for all surface sites of the Al_{13} cluster on the $y2$ axis. We observe that all 12 surface atoms of the Al_{13} cluster follow only two distance distributions, indicating that there are only two types of unique atoms (in terms of neighbor distance environment) in this cluster. The difference between these two types of atoms in their nearest neighbor distribution is picked up in the ML model. And hence we observed improvement in MAE at distances where these two groups differ from each other, *i.e.*, MAE reduced from 0.13 eV to 0.08 eV with descriptors up to nn4 *versus* nn5. A similar jump (decrease) in MAE was observed when nn8 was also included, as shown in Table 1; nn8 is the point at which the two classes further separated. In Fig. 5(b) ML predicted energies for Al_{13} are plotted against DFT calculated energies. The MAE in this specific case is 0.02 eV. The relation of MAE as a function of neighbor distances was seen for other clusters too. A plot for a few representative cases like Al_5, Al_7, and Al_9 are shown in Fig. SI1 (ESI†). It must be noted, that since we were dealing with fixed geometries, our

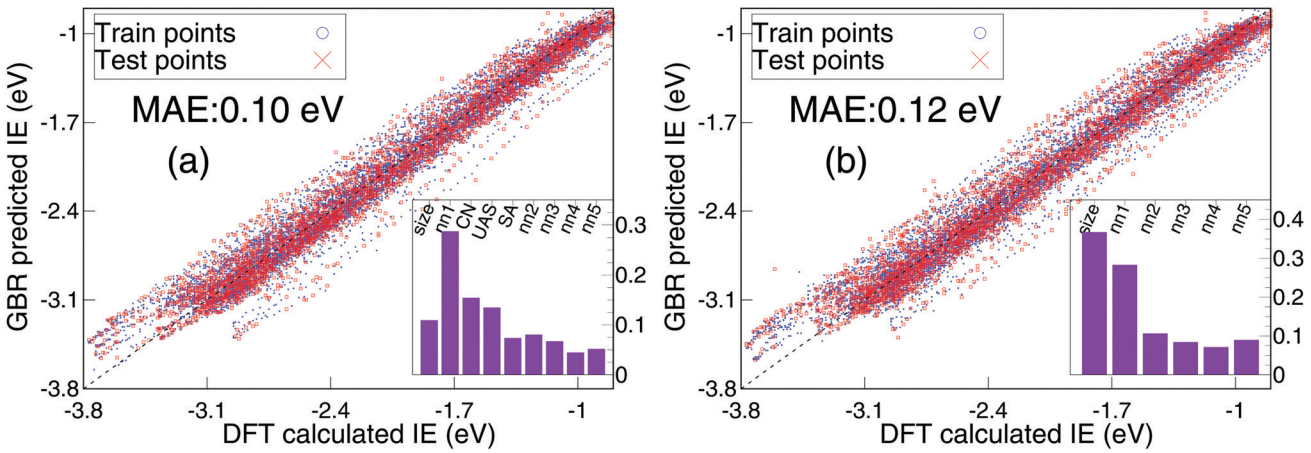

Fig. 4 ML predicted values are plotted against DFT calculated energies. In both the cases, the MAE is computed for the test set. The inset plot shows the feature ranking. (a) When all geometric descriptors were used and an MAE of 0.10 eV is reported. (b) When only the nearest neighbor distances and size of the cluster were used as descriptors. MAE of 0.12 eV is reported.

Table 1 Mean absolute errors (MAE) as a function of descriptors (interatomic distances (nn)) are shown in the table for various clusters. As seen from the table, increasing representation of the system results in improved accuracy. The overall error is 0.05 eV

Cluster size	MAE as a function of (nearest neighbors as) descriptors											
	nn2	nn3	nn4	nn5	nn6	nn7	nn8	nn9	nn10	nn11	nn12	nn13
5	0.085	0.084	0.050	0.050								
6	0.034	0.035	0.035	0.034	0.032							
7	0.179	0.180	0.178	0.059	0.051	0.051						
8	0.057	0.038	0.035	0.029	0.028	0.028	0.028					
9	0.226	0.228	0.219	0.133	0.127	0.122	0.055	0.054				
10	0.051	0.053	0.051	0.052	0.051	0.044	0.043	0.043	0.044			
11	0.281	0.268	0.194	0.169	0.114	0.096	0.096	0.079	0.060	0.060		
12	0.081	0.069	0.067	0.063	0.055	0.041	0.041	0.039	0.038	0.038	0.032	
13	0.131	0.132	0.130	0.082	0.083	0.086	0.026	0.026	0.026	0.025	0.024	0.021

Table 2 Mean absolute errors (MAE) as a function of descriptors (interatomic distances (nn)) are shown in the table for larger clusters. The last three columns present 25%, 50%, and 100% of the system representation respectively. The numbers in the brackets indicate the number of interatomic distances used to predict the interaction energy

Cluster size	MAE as a function of nearest neighbors as descriptors							
	nn2	nn3	nn4	nn5	nn10	nn 25%	nn 50%	nn 100%
25	0.137	0.111	0.113	0.107	0.071	0.095(6)	0.067(13)	0.068(24)
36	0.063	0.050	0.054	0.056	0.044	0.045(9)	0.044(18)	0.030 (35)
42	0.083	0.084	0.077	0.074	0.076	0.076(11)	0.069(21)	0.059(41)
55	0.125	0.110	0.104	0.100	0.090	0.090(14)	0.088(28)	0.086(54)
67	0.083	0.077	0.072	0.066	0.048	0.048(17)	0.045(35)	0.049(66)
74	0.103	0.073	0.068	0.065	0.061	0.060 (18)	0.059(36)	0.052(73)
75	0.119	0.125	0.100	0.081	0.081	0.080 (18)	0.071(37)	0.060(74)

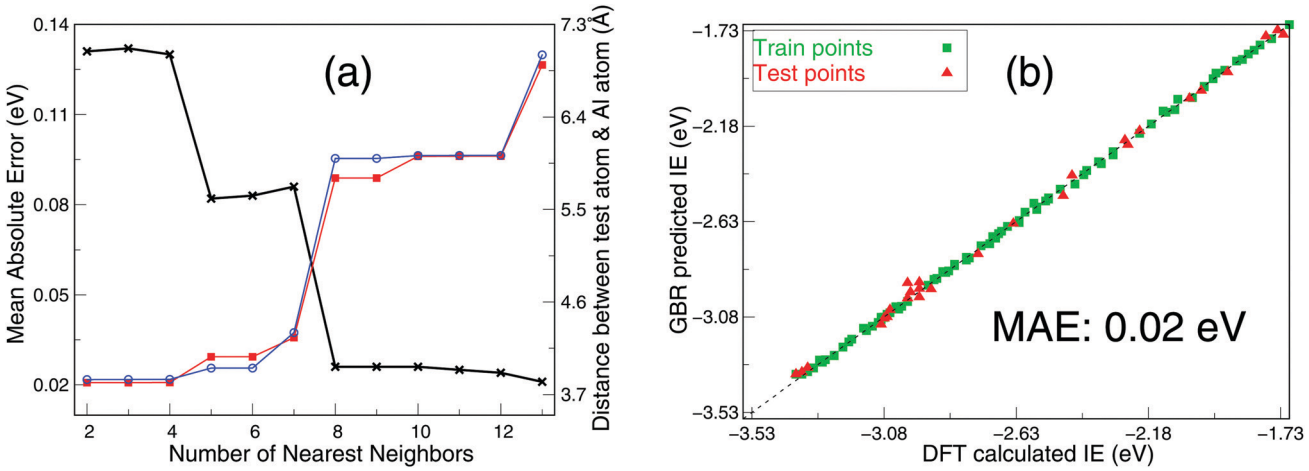

Fig. 5 (a) The decrease of errors (MAE) with increasing number of descriptors for the Al_{13} cluster plotted on the *y*1 axis. The distance between H atom and surface atoms is used as descriptors which is plotted on the *y*2 axis. When the two groups in Al_{13} are distinguished in the NN distribution, the error reduces. (b) Shows the ML predicted energies for Al_{13} plotted against the DFT calculated energies. MAE are computed over the test set.

descriptors did not violate any of the invariance (rotation, reflection, translation and atom indexing) that any set of ML descriptors are supposed to maintain. Overall, we report that the chosen set of descriptors indeed work well to predict the site as well as size specific interaction of the adsorbate with clusters.

While the correlation between nearest neighbor and variation in MAE is strikingly evident and easy to capture in smaller clusters, as can be seen from the MAE as a function of descriptors (shown in Table 2), it is not very easy to capture for larger cluster sizes. It was observed that the variation in MAE did not follow any regular trends when descriptors up to nn10 *versus* all distances (nn 100%) were used. For example, in the case of Al_{36} and Al_{75}, reduction in MAE was more than 25% for each of them when MAE with nn10 is compared to that of nn 100% (shown in Table 2). It must be noted that Al_{36} and Al_{75} are highly symmetric clusters. Whereas for asymmetric clusters like Al_{25}, Al_{55}, and Al_{67}, the reduction in errors were less than 5% as evident from Table 2. However, for other asymmetric clusters, Al_{42} and Al_{74}, the reduction is much larger, *i.e.* about 20% and 15%, respectively, which is similar to that of

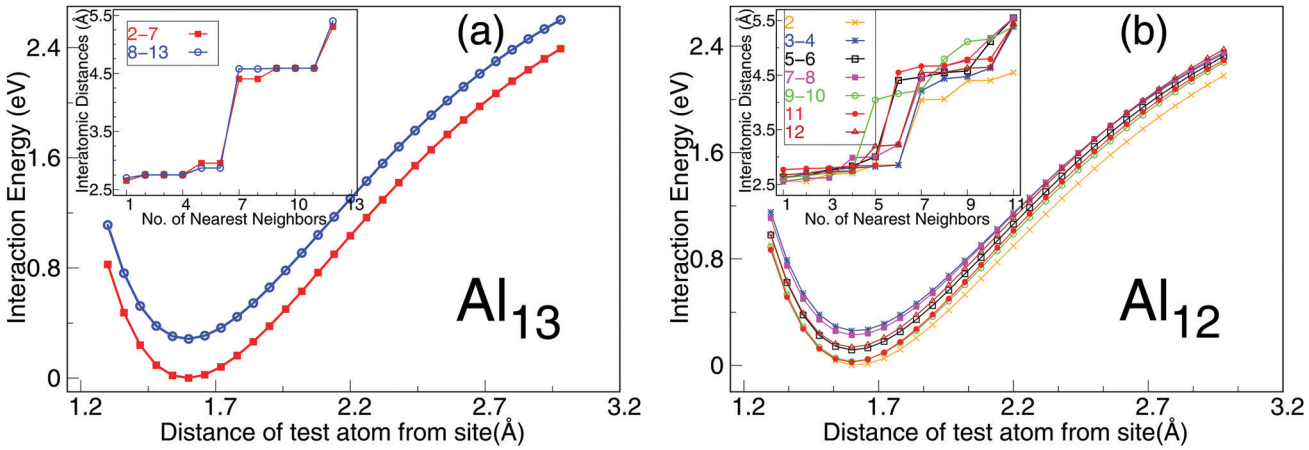

Fig. 6 Distance dependent interaction energy of all surface atoms for (a) Al_{13} and (b) Al_{12} clusters. Atoms with identical nearest neighbor distributions also exhibited an identical interaction energy pattern towards an H atom. The inset figure shows variation in the interatomic distances as a function of nearest neighbors for all surface atoms of these clusters.

symmetric clusters. Thus, generalization of results becomes difficult for larger clusters. Nonetheless, even for larger clusters, the one to one correlation between reduction in MAE and increasing system representation still holds. The overall MAE reported in our work is $\approx$0.05 eV whereas just including the first five nn distances reduces MAE to $\approx$0.1 for all the sizes.

To understand the ML results better, let's have a careful look at the DFT calculations. We observe that the interaction energy for any adsorption site in a cluster is directly related to its NN distance distribution. All atoms having identical nearest neighbor distribution within a cluster, interact identically with the incoming adsorbate. To elaborate this point further, in Fig. 6 we show the interaction energies for all the atoms within a cluster for Al_{13} and Al_{12} along with their nearest neighbor distribution (or interatomic distances) in the inset. Similar quantities are plotted for a few representative clusters in Fig. SI2 (ESI†). In the case of Al_{13} (see inset of Fig. 6a) all the surface atoms could be grouped into two classes based on their respective interatomic distances, indicating that an incoming adsorbate would experience only two different environments. Furthermore, when the interaction energy of these surface atoms with an H atom (as adsorbate) was computed, it was observed that atoms belonging to one class interact identically with the adsorbate, resulting in an identical interaction energy as shown in Fig. 6(a). As expected this is true for all the clusters that we have studied and for some of the larger clusters, as shown in Fig. SI2 (ESI†). And hence, when we provide information of NN distances wherein the two environments are separated (*i.e.*, at nn4 and then further at nn7) we correspondingly see a better performance of the ML model. While modeling the interaction of clusters with the adsorbate the (dis)similarity between the two adsorption sites had to be captured. And hence, the nearest neighbor distribution as seen by the adsorbate is a logical choice of descriptors.

Since the line of search for all the results discussed above was restricted along the radial vector, to model a real situation wherein an adsorbate can approach the cluster from any direction, all possible directions had to be scanned. The one to one correlation between identical sites and identical interaction would be difficult to quantify for this situation, as now the adsorbate was not placed only at on-top sites. Nonetheless, the same recipe of descriptors was still legitimate as the distances taken were from the adsorbate to the atoms. And so, the same set of descriptors would capture the change in chemical environment as seen by an incoming adsorbate. To validate this, we computed the interaction energy of the H atom for these randomly selected 800 points on a sphere that enclosed the Al_{13} cluster at its center (see Fig. 7(c)). Fig. 7(a and b) show PES as experienced by an incoming adsorbate computed using DFT and our trained ML model. The distance of H atom from the closest surface site of the cluster varies between 1.60 Å to 2.69 Å. This result is particularly important because through this we could predict the interaction energy of the cluster–adsorbate system at any point with MAE as low as 0.04 eV. We further plot the difference between the ML predicted vs DFT computed PES in Fig. 7(d). As can be seen, most of the area on the contour plot show errors between -0.05 to 0.05. The maximum difference in error prediction is 0.13 eV whereas the minimum difference is 0.0001 eV. The success of the ML model, in this case, is a proof of concept that the nearest neighbor distances are the correct choice of descriptors. It is important to note that the potential energy surface that we have predicted is the PES of an adsorbate when it is in the vicinity of the cluster. In other words, if the adsorbate is moved on the cluster surface, what kind of interaction it will experience as a function of its position with respect to the cluster, could be understood from the shown PES.

To further validate our model, we tested it on other clusters like Na_{10}, Al_6Ga_6, Li_4Sn_6, and $Au_{40}Cu_{40}$. To demonstrate the universality of our work, calculations performed with different adsorbing species on Al clusters are also noted below. When a single N atom was placed at the on-top positions on the Al_{13} cluster, the MAE for the interaction energy from the ML model was 0.06 eV, *i.e.*, in the same range as our previous results. The model showed transferability when trained on Al_{13} with two

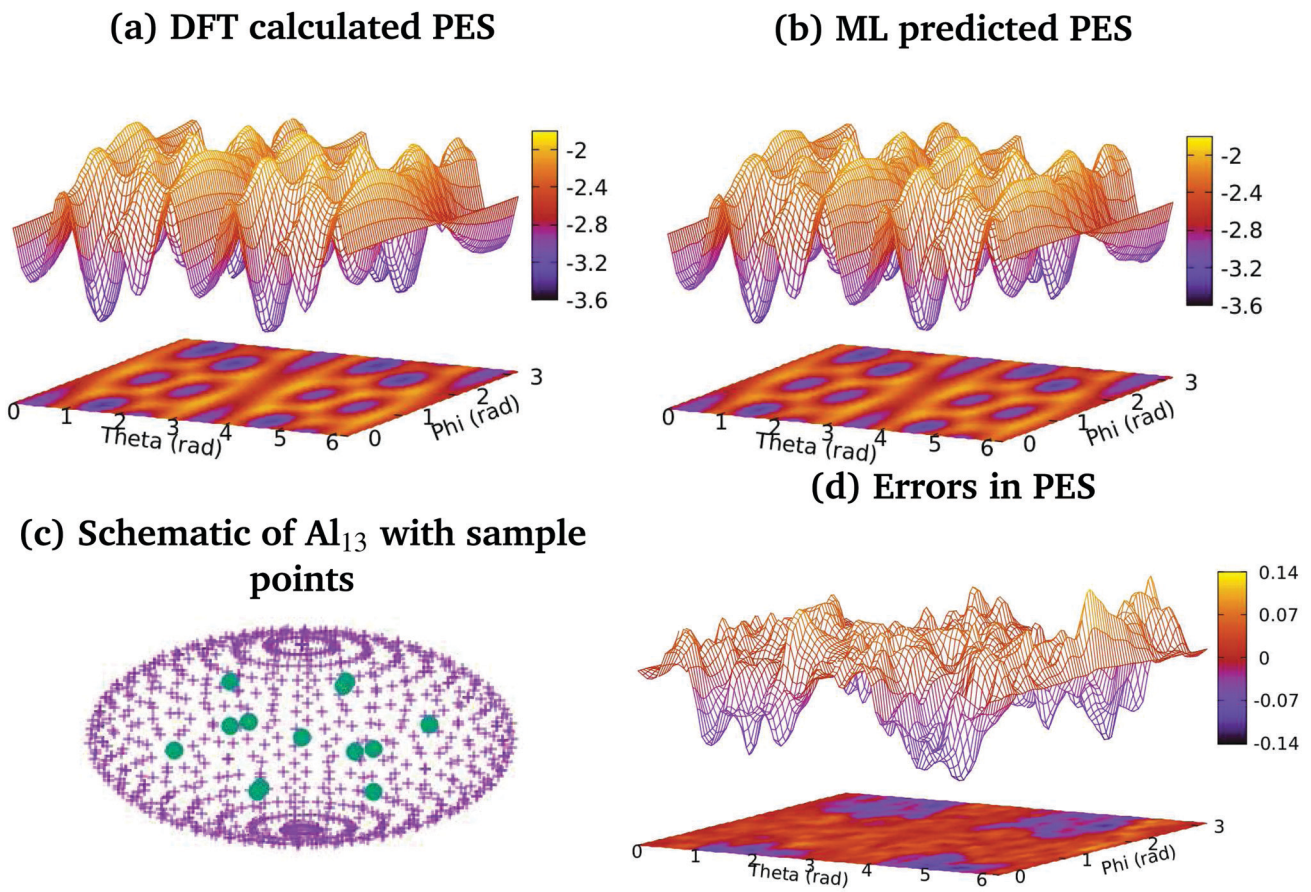

Fig. 7 (a) PES computed through DFT calculations. (b) ML generated PES. (c) Cartoon of Al_{13} cluster and points selected on a sphere to compute the interaction of the adsorbate with the cluster. (d) Difference between PES computed with DFT and ML predicted PES. It is evident that our model has picked up the variation in PES quite successfully.

adsorbates *viz.* H and N together. The resulting error in interaction energy turned out to be MAE $\approx$ 0.04 eV. We also tested the validity of our ML model on bimetallic clusters Al_6Ga_6, Li_4Sn_6, and $Au_{40}Cu_{40}$. MAE in the Interaction Energy for an H atom when placed on top of the Al_6Ga_6 cluster turned out to be 0.09 eV. This error was obtained based on only structural representation of the cluster. With the inclusion of nuclear charge/ionic radii/van der Waals radii of both the elements of the cluster, *viz.* Al and Ga, in the descriptor set, the MAE got down to 0.058 eV. Furthermore, the same set of nearest neighbor distances along with atomic charge of the nearest neighbor when used as descriptors for Li_4Sn_6 and $Au_{40}Cu_{40}$, the MAE in Interaction Energy were noted to be 0.09 eV and 0.08 eV, respectively. The reason for choosing these clusters in particular was to test the model on heterogeneous clusters of two kinds; one with elements from different groups of the periodic table (Li_4Sn_6) and the other for a larger size ($Au_{40}Cu_{40}$). As can be seen in both the cases, the developed set of descriptors worked well. For an interaction energy between the H atom (sphere calculations) and a highly asymmetric Na_{10} cluster, our ML model with the same recipe of descriptors predicted the IE with MAE $\approx$ 0.038 eV. The errors for the prediction of IE using the same ML model when molecules like N_2, O_2, and CO were adsorbed around the Al_{12} on a sphere, turned out to be, 0.045 eV, 0.049 eV, and 0.042 eV, respectively. Finally, our descriptors were transferable when trained over different adsorbing molecules as well. To prove this, we tested the model for adsorption energy prediction of N_2, O_2 and CO molecules together yielding an error of 0.05 eV.

In a nutshell, we understand the role of the nearest neighbor distances as effective descriptors to capture the interaction energy trends for not only clusters of different sizes but also for different elements. The developed set of descriptors were successful in bringing out the correlation between the IE and the nearest neighbor distances. The same was realized when the DFT results were closely analyzed. The current work demonstrates the usefulness of an ML model to develop a deeper understanding of the physical phenomena under investigation by means of simple descriptors. Transferability of our descriptors is another indicator of its effectiveness. Transferability with respect to size of the cluster, clusters of different elements, homogeneous as well as heterogeneous, and different adsorbing species was demonstrated to an extent through our detailed investigations. This transferability is particularly important as it explains the importance of developed descriptors

to understand various interactions between any cluster–adsorbate system in general. However, our model is limited by the cluster geometry under investigation. Also, for the molecular adsorbate, many complexities are not considered like various orientations of the adsorbate with respect to the site of adsorption. Similarly, for a bimetallic cluster, the current model can be improved further by introducing more descriptors depicting relative atomic arrangements with better accuracy. Nonetheless, the present work illustrates the simplicity of the developed descriptors which can be further improved to comprehend much more complex cluster–adsorbate interactions.

Conclusion

In this work we demonstrate the use of interatomic distances as descriptors to capture the cluster adsorbate interaction in Al_n clusters. Our recipe of descriptors works well when applied to 23 clusters across the size range to predict the site specific interaction with MAE $\approx$ 0.10 eV. It also performs well to predict the size specific interaction when testing and training sets were mutually exclusive, demonstrating that DFT could be completely bypassed. When the model is trained on the clusters in the size range of 5 to 20 and tested on one of the sizes in between the range (totally absent in the train set), the resulting MAE was $\approx$0.024 eV. This demonstrates the transferability of our ML model to different cluster sizes within the range on which it is trained. The chosen set of descriptors further brings down the error to MAE $\approx$ 0.05 eV when used on individual sizes. We further demonstrate the transferability of the model by applying it to clusters of different elements like Na_{10} and bimetallic clusters like Al_6Ga_6, Li_4Sn_6, and $Au_{40}Cu_{40}$ with MAE of 0.04 eV, 0.06 eV, 0.08 eV, and 0.09 eV, respectively. The model also worked for interaction energy studies between the Al_{12} cluster and various molecules as an adsorbate with an MAE of 0.04 eV. The transferability of our model, both with respect to size and different elements, leads to a direction wherein the computational cost involved due to DFT calculations can be reduced greatly. Finally, a careful examination of the interatomic distances reveals a one to one correlation between the nearest neighbor distribution and corresponding interaction energy curves. This one to one correlation indeed provides the rationale as to why our descriptors perform so well.

Conflicts of interest

There are no conflicts to declare.

Acknowledgements

The authors thank Dr Leelavati Narlikar and Professor Shobhana Narasimhan for their valuable suggestions. CSIR-4PI is gratefully acknowledged for the computational facility. KJ acknowledges DST (EMR/2016/000591) for partial financial support. SA acknowledges DST-INSPIRE for the research fellowship. SM acknowledges UGC for the research fellowship.

Notes and references

1 A. Agrawal and A. Choudhary, *APL Mater.*, 2016, **4**, 0532080.
2 K. Schwab, see https://www.foreignaffairs.com/articles/2015-12-12/fourth-industrial-revolution, 2015.
3 J. Hachmann, R. Olivares-Amaya, S. Atahan-Evrenk, C. Amador-Bedolla, R. S. Sánchez-Carrera, A. Gold-Parker, L. Vogt, A. M. Brockway and A. Aspuru-Guzik, *J. Phys. Chem. Lett.*, 2011, **2**, 2241–2251.
4 A. Jain, S. P. Ong, G. Hautier, W. Chen, W. D. Richards, S. Dacek, S. Cholia, D. Gunter, D. Skinner and G. Ceder, *et al.*, *Appl. Mater.*, 2013, **1**, 011002.
5 C. E. Calderon, J. J. Plata, C. Toher, C. Oses, O. Levy, M. Fornari, A. Natan, M. J. Mehl, G. Hart and M. B. Nardelli, *et al.*, *Comput. Mater. Sci.*, 2015, **108**, 233–238.
6 T. Mueller, A. G. Kusne and R. Ramprasad, *Rev. Comput. Chem.*, 2016, **29**, 186–273.
7 K. Hansen, F. Biegler, R. Ramakrishnan, W. Pronobis, O. A. von Lilienfeld, K.-R. Müller and A. Tkatchenko, *J. Phys. Chem. Lett.*, 2015, **6**, 2326–2331.
8 T. Xie and J. C. Grossman, *Phys. Rev. Lett.*, 2018, **120**, 145301.
9 G. Pilania, C. Wang, X. Jiang, S. Rajasekaran and R. Ramprasad, *Sci. Rep.*, 2013, **3**, 2810.
10 G. Montavon, M. Rupp, V. Gobre, A. Vazquez-Mayagoitia, K. Hansen, A. Tkatchenko, K.-R. Müller and O. A. Von Lilienfeld, *New J. Phys.*, 2013, **15**, 095003.
11 M. Rupp, R. Ramakrishnan and O. A. von Lilienfeld, *J. Phys. Chem. Lett.*, 2015, **6**, 3309–3313.
12 A. Chandrasekaran, D. Kamal, R. Batra, C. Kim, L. Chen and R. Ramprasad, *npj Comput. Mater.*, 2019, **5**, 22.
13 I. Takigawa, K.-I. Shimizu, K. Tsuda and S. Takakusagi, *RSC Adv.*, 2016, **6**, 52587–52595.
14 Z. W. Ulissi, A. J. Medford, T. Bligaard and J. K. Nørskov, *Nat. Commun.*, 2017, **8**, 14621.
15 S. Bose, D. Dhawan, S. Nandi, R. R. Sarkar and D. Ghosh, *Phys. Chem. Chem. Phys.*, 2018, **20**, 22987–22996.
16 G. Pilania, C. Wang, X. Jiang, S. Rajasekaran and R. Ramprasad, *Sci. Rep.*, 2013, **3**, 2810.
17 B. Meredig, A. Agrawal, S. Kirklin, J. E. Saal, J. Doak, A. Thompson, K. Zhang, A. Choudhary and C. Wolverton, *Phys. Rev. B: Condens. Matter Mater. Phys.*, 2014, **89**, 094104.
18 S. Bukkapatnam, M. Malshe, P. Agrawal, L. Raff and R. Komanduri, *Phys. Rev. B: Condens. Matter Mater. Phys.*, 2006, **74**, 224102.
19 T. Morawietz and J. Behler, *J. Phys. Chem. A*, 2013, **117**, 7356–7366.
20 S. K. Natarajan, T. Morawietz and J. Behler, *Phys. Chem. Chem. Phys.*, 2015, **17**, 8356–8371.
21 S. Manzhos, R. Dawes and T. Carrington, *Int. J. Quantum Chem.*, 2015, **115**, 1012–1020.
22 B. Kolb, B. Zhao, J. Li, B. Jiang and H. Guo, *J. Chem. Phys.*, 2016, **144**, 224103.
23 D. Dragoni, T. D. Daff, G. Csányi and N. Marzari, *Phys. Rev. Mater.*, 2018, **2**, 013808.
24 W. Jeong, K. Lee, D. Yoo, D. Lee and S. Han, *J. Phys. Chem. C*, 2018, **122**, 22790–22795.

25 Y.-J. Zhang, A. Khorshidi, G. Kastlunger and A. A. Peterson, *J. Chem. Phys.*, 2018, **148**, 241740.
26 T. Le, V. C. Epa, F. R. Burden and D. A. Winkler, *Chem. Rev.*, 2012, **112**, 2889–2919.
27 K. Schütt, H. Glawe, F. Brockherde, A. Sanna, K. Müller and E. Gross, *Phys. Rev. B: Condens. Matter Mater. Phys.*, 2014, **89**, 205118.
28 O. A. Von Lilienfeld, R. Ramakrishnan, M. Rupp and A. Knoll, *Int. J. Quantum Chem.*, 2015, **115**, 1084–1093.
29 A. Seko, H. Hayashi, K. Nakayama, A. Takahashi and I. Tanaka, *Phys. Rev. B*, 2017, **95**, 144110.
30 F. A. Faber, L. Hutchison, B. Huang, J. Gilmer, S. S. Schoenholz, G. E. Dahl, O. Vinyals, S. Kearnes, P. F. Riley and O. A. Von Lilienfeld, *J. Chem. Theory Comput.*, 2017, **13**, 5255–5264.
31 A. P. Bartók, S. De, C. Poelking, N. Bernstein, J. R. Kermode, G. Csányi and M. Ceriotti, *Sci. Adv.*, 2017, **3**, e1701816.
32 T. Davran-Candan, M. E. Günay and R. Yldrm, *J. Chem. Phys.*, 2010, **132**, 174113.
33 M. O. Jäger, E. V. Morooka, F. F. Canova, L. Himanen and A. S. Foster, *npj Comput. Mater.*, 2018, **4**, 37.
34 F. Musil, S. De, J. Yang, J. E. Campbell, G. M. Day and M. Ceriotti, *Chem. Sci.*, 2018, **9**, 1289–1300.
35 K. Hansen, G. Montavon, F. Biegler, S. Fazli, M. Rupp, M. Scheffler, O. A. von Lilienfeld, A. Tkatchenko and K.-R. Müller, *J. Chem. Theory Comput.*, 2013, **9**, 3404–3419.
36 K. T. Butler, D. W. Davies, H. Cartwright, O. Isayev and A. Walsh, *Nature*, 2018, **559**, 547.
37 U. Heiz, F. Vanolli, A. Sanchez and W.-D. Schneider, *J. Am. Chem. Soc.*, 1998, **120**, 9668–9671.
38 W. T. Wallace and R. L. Whetten, *J. Phys. Chem. B*, 2000, **104**, 10964–10968.
39 B. Cao, A. K. Starace, O. H. Judd and M. F. Jarrold, *J. Am. Chem. Soc.*, 2009, **131**, 2446–2447.
40 P. J. Roach, W. H. Woodward, A. Castleman, A. C. Reber and S. N. Khanna, *Science*, 2009, **323**, 492–495.
41 A. C. Reber, S. N. Khanna, P. J. Roach, W. H. Woodward and A. Castleman Jr, *J. Phys. Chem. A*, 2010, **114**, 6071–6081.
42 B. S. Kulkarni, S. Krishnamurty and S. Pal, *J. Phys. Chem. C*, 2011, **115**, 14615–14623.
43 S. Yin and E. R. Bernstein, *Int. J. Mass Spectrom.*, 2012, **321**, 49–65.
44 Z. Luo, A. Castleman Jr and S. N. Khanna, *Chem. Rev.*, 2016, **116**, 14456–14492.
45 M. Schmidt, R. Kusche, B. von Issendorff and H. Haberland, *Nature*, 1998, **393**, 238.
46 G. A. Breaux, D. A. Hillman, C. M. Neal, R. C. Benirschke and M. F. Jarrold, *J. Am. Chem. Soc.*, 2004, **126**, 8628.
47 K. Joshi, S. Krishnamurty and D. Kanhere, *Phys. Rev. Lett.*, 2006, **96**, 135703.
48 R. S. Berry and B. M. Smirnov, *Phys. Rep.*, 2013, **527**, 205–250.
49 A. A. Shvartsburg and M. F. Jarrold, *Phys. Rev. A: At., Mol., Opt. Phys.*, 1999, **60**, 1235.
50 A. Argo, J. Odzak, F. Lai and B. Gates, *Nature*, 2002, **415**, 623–626.
51 Q. Fu, H. Saltsburg and M. Flytzani-Stephanopoulos, *Science*, 2003, **301**, 935–938.
52 C. T. Campbell, *Science*, 2004, **306**, 234–235.
53 M. Chen and D. Goodman, *Science*, 2004, **306**, 252–255.
54 C. Lemire, R. Meyer, S. Shaikhutdinov and H.-J. Freund, *Angew. Chem., Int. Ed.*, 2004, **43**, 118–121.
55 J. Wei and E. Iglesia, *J. Phys. Chem. B*, 2004, **108**, 4094–4103.
56 S. Vajda, M. J. Pellin, J. P. Greeley, C. L. Marshall, L. A. Curtiss, G. A. Ballentine, J. W. Elam, S. Catillon-Mucherie, P. C. Redfern, F. Mehmood and P. Zapol, *Nat. Mater.*, 2009, **8**, 213–216.
57 H. Li, Z. Zhang and Z. Liu, *Catalysts*, 2017, **7**, 306.
58 J. R. Kitchin, *Nat. Catal.*, 2018, **1**, 230.
59 A. N. Andriotis, G. Mpourmpakis, S. Broderick, K. Rajan, S. Datta, M. Sunkara and M. Menon, *J. Chem. Phys.*, 2014, **140**, 094705.
60 R. Jinnouchi and R. Asahi, *J. Phys. Chem. Lett.*, 2017, **8**, 4279–4283.
61 R. Gasper, H. Shi and A. Ramasubramaniam, *J. Phys. Chem. C*, 2017, **121**, 5612–5619.
62 M. Rupp, A. Tkatchenko, K.-R. Müller and O. A. Von Lilienfeld, *Phys. Rev. Lett.*, 2012, **108**, 058301.
63 A. P. Bartók, R. Kondor and G. Csányi, *Phys. Rev. B: Condens. Matter Mater. Phys.*, 2013, **87**, 184115.
64 C. M. Neal, A. K. Starace, M. F. Jarrold, K. Joshi, S. Krishnamurty and D. G. Kanhere, *J. Phys. Chem. C*, 2007, **111**, 17788–17794.
65 A. K. Starace, C. M. Neal, B. Cao, M. F. Jarrold, A. Aguado and J. M. López, *J. Chem. Phys.*, 2008, **129**, 144702.
66 A. Aguado and J. M. López, *J. Chem. Phys.*, 2009, **130**, 064704.
67 A. Susan and K. Joshi, *J. Chem. Phys.*, 2014, **140**, 154307.
68 P. E. Blöchl, *Phys. Rev. B: Condens. Matter Mater. Phys.*, 1994, **50**, 17953–17979.
69 G. Kresse and D. Joubert, *Phys. Rev. B: Condens. Matter Mater. Phys.*, 1999, **59**, 1758–1775.
70 J. P. Perdew, K. Burke and M. Ernzerhof, *Phys. Rev. Lett.*, 1996, **77**, 3865–3868.
71 J. P. Perdew, K. Burke and M. Ernzerhof, *Phys. Rev. Lett.*, 1997, **78**, 1396.
72 G. Kresse and J. Hafner, *Phys. Rev. B: Condens. Matter Mater. Phys.*, 1994, **49**, 14251–14269.
73 G. Kresse and J. Furthmüller, *Phys. Rev. B: Condens. Matter Mater. Phys.*, 1996, **54**, 11169–11186.
74 G. Kresse and J. Furthmüller, *Comput. Mater. Sci.*, 1996, **6**, 15–50.
75 F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot and E. Duchesnay, *J. Mach. Learn. Res.*, 2011, **12**, 2825–2830.