Looking Beyond Adsorption Energies to Understand Interactions at Surface using Machine Learning

Sheena Agarwal*[a,b], Kavita Joshi[a,b]

[a] *Physical and Materials Chemistry Division, CSIR-National Chemical Laboratory, Dr. Homi Bhabha Road, Pashan, Pune-411008, India.*
[b] *Academy of Scientific and Innovative Research (AcSIR), Sector 19, Kamla Nehru Nagar, Ghaziabad, Uttar Pradesh- 201002, India.*

Abstract

Identifying factors that influence interactions at the surface is still an active area of research. In this study, we present the importance of analyzing bondlength activation, while interpreting Density Functional Theory (DFT) results, as yet another crucial indicator for catalytic activity. We studied the adsorption of small molecules, such as O_2, N_2, CO, and CO_2, on seven face-centered cubic (fcc) transition metal surfaces (M = Ag, Au, Cu, Ir, Rh, Pt, and Pd) and their commonly studied facets (100, 110, and 111). Through our DFT investigations, we highlight the absence of linear correlation between adsorption energies (E_{ads}) and bondlength activation (BL_{act}). Our study indicates the importance of evaluating both to develop a better understanding of adsorption at surfaces. We also developed a Machine Learning (ML) model trained on simple periodic table properties to predict both, E_{ads} and BL_{act}. Our ML model gives an accuracy of Mean Absolute Error (MAE) $\sim$ 0.2 eV for E_{ads} predictions and 0.02 Å for BL_{act} predictions. The systematic study of the ML features that affect E_{ads} and BL_{act} further reinforces the importance of looking beyond adsorption energies to get a full picture of surface interactions with DFT.

Keywords: DFT, Machine Learning, surface interactions, adsorption, bondlength activation, catalysis

Email address: sm.agarwal@ncl.res.in, shina181@gmail.com (Sheena Agarwal*)

1. Introduction

Investigating the activation of molecules on metal surfaces is a starting point for understanding catalysis.[1, 2] Several studies have been conducted to understand the various factors that govern the catalytic activity of transition metal surfaces and this subject continues to draw attention.[3, 4, 5, 6, 7] Activation of molecules is an outcome of the charge transfer taking place from the metal surface to the adsorbate. Numerous successful models and theories to this end have been proposed and are still under the ambit of active research.[8, 6, 9, 10, 11] The primary indicator to quantify the interaction of adsorbate with surfaces is often the adsorption energy, E_{ads}.Hence, efforts are geared in the direction of understanding factors that influence adsorption energies.[12, 13, 14] One such successful model is the d-band center model proposed by Hammer and Norskov, which correlates the adsorption strengths of adsorbate with the d-band center energy of the participating metal surface.[8, 2, 15, 16] Although the model is greatly useful in developing an understanding about the factors that influence surface interactions which in turn can be used to design better catalysts, there are some exceptions to it.[17, 18, 19, 20, 21] Another geometrical parameter of catalysis is the bondlength activation BL_{act} that adsorbates undergo upon adsorption.

In all these studies, although the amount of activation experienced by the adsorbate during interaction is always reported, it is seldom used as a parameter to explain the extent of interaction.[22, 23, 24, 25] Often adsorption energies are correlated with the participating d-band (or d-band center) of the metal surface and the bondlength activations are attributed to the charge transfer.[3, 26] But the question that arises is, whether the distinction between the factors that influence each of these parameters discrete ? In their study Wang et. al. investigated the key factors controlling the interaction of CO_2 with metal surfaces.[3] They found a linear correlation between the binding energies and the charge transferred from the metal surface into chemisorbed CO_2. This linear relationship was even better than the relationship observed between the binding

energy of CO_2 and the d-band centers of the metal surfaces. The two parameters viz. E_{ads} and BL_{act} are expected to be correlated for any interaction between an adsorbate and a catalyst surface. But a careful analysis of the literature proves lack of linear correlation between the two.[27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39] In our previous study of methanol activation on (220) and (311) facet of a mixed metal oxide ($ZnAl_2O_4$) surface, we observed that changing orientation of methanol resulted in different extents of bondlength activation for its O-H bond.[40] The distance from the nearest available surface oxygen was attributed for the corresponding O-H bond activation. But at the same time a clear lack of correlation between O-H bond activation of methanol and E_{ads} values was evident on both the surfaces. Although we could successfully point out the factors governing BL_{act}, the missing correlation between BL_{act} and E_{ads} remained unsolved. In another study Petersen et. al. reported that the thermodynamically most favoured adsorption site for CO on Co(221) is fcc 3-fold with E_{ads} of -1.77 eV and C-O bondlength of 1.20 Å.[41] Adsorption on another site i.e. the B5-B site is reported to be energetically less stable by 0.31 eV but C–O bondlength has an increased activation upto 1.28 Å. They also reported that the lowest energy transition state for CO dissociation is located at the B5-B site (the one with maximum BL_{act}). Both these studies not only bring out the missing correlation between E_{ads} and BL_{act} but they also highlight the importance of analyzing BL_{act} as a quantitative indicator of interaction. This makes the study of both E_{ads} and BL_{act} interesting and imperative to understand/comment on the extent of interaction. In this work, we highlight this lack of one-to-one correlation and bring out the factors that influence each of the parameters separately during interaction by implementing Machine Learning (ML) methods.

Digging out hidden patterns in the dataset is among the many areas in which ML is increasingly being applied.[42, 43, 44, 45, 46, 47, 48] Designing an appropriate set of features that accurately represents the dataset is extremely important for using ML effectively.[49, 50, 51, 52, 53] Establishing successful correlations between these input features and predicted values can advance our

understanding of a problem at hand.[54, 55, 56, 57] Yang et. al. demonstrated the use of simple substrate and adsorbate based intrinsic properties to study CO_2 reduction on transition metals and alloys.[58] The descriptors that were used translate in terms of an activity indicator that provides a basis for the design of catalysts. With tree based methods in ML, the ability to analyze and rank the features that greatly affect target value prediction is extremely useful in understanding hidden trends in the dataset.[58, 59, 60, 61] In a recent study, Liu et. al. through detailed DFT investigations show that E_{ads} of a variety of small molecules on 13 metal oxide surfaces correlates very well with adsorbate HOMO but not with LUMO. Further, they applied an ML based extra tree regressor method and demonstrate HOMO to be the top-ranking feature for adsorption energy prediction of small molecules on all surfaces. Such correlations between the predicted value and feature ranking often help in gaining insights into the factors governing catalysis. However, a problem associated with the design of features for an ML model is their complexity. If the designed set of features do not translate in terms of the physical understanding of a system, then it limits the use of ML model to a black box.[62] Therefore, developing features that can be easily calculated and understood is an ongoing area of research. Another area that remains untouched by ML methods assisting DFT is the prediction/analysis of BL_{act} along with adsorption energy studies. [61, 63, 64, 65, 66]

In this work, we performed DFT calculations for adsorption of four small molecules on twenty-one metal surfaces and investigated E_{ads} and corresponding BL_{act}. While we understand that adsorption energy acts as an indicator of catalytic activity, so does the corresponding bondlength activation. It not only tells us about the extent of interaction but also indicates the ease with which adsorbate can possibly be converted into products. Hence, analyzing factors that affect both the indicators differently is important in the process of catalyst design. Therefore we further employed ML calculations to predict both E_{ads} and BL_{act} separately but with the same set of features. The best ML model was chosen after testing it against seven other models from across three classes. Easily calculable features along with periodic table properties were used to

train our ML algorithms. Our best ML model not only performed well for the prediction of E_{ads} (MAE $\sim$ 0.2 eV) and BL_{act} (MAE $\sim$ 0.02 Å) but also highlighted the underlying factors that affect each of the target values differently.

2. Computational Details

2.1. Adsorption energy calculations using DFT

We performed adsorption studies for O_2, N_2, CO, and CO_2 on seven fcc transition metal surfaces (M = Ag, Au, Cu, Ir, Rh, Pt, and Pd) and three commonly studied facets (100, 110, and 111) for each metal. All the calculations were carried out within the Kohn-Sham formulation of DFT. Projector Augmented Wave potential [67, 68] was used, with Perdew–Burke–Ehrzenhof (PBE) [69] approximation for the exchange-correlation and generalized gradient [70] approximation, as implemented in planewave, pseudopotential based code, *VASP*[71, 72, 73]. For metal surfaces a (3×3) supercell, containing four layers of M atoms was used. The first three layers of surfaces were allowed to relax during geometrical optimization; the geometry of the bottom layer was fixed to the bulk configuration. The energies of isolated molecules were obtained using the same parameters as those in the bare surface slab calculations. Cubic simulation cell, with the image in each direction separated by 15 Å of vacuum, was used. Energy convergence criteria of 10^{-4} eV was used for SCF calculations. K-mesh of 7x7x1 was used for (100) and (111) facets of all metal surfaces and 5x7x1 for (110) surfaces. The following formula was used to calculate the adsorption energies:

$$E_{ads} = E_{system} - (E_M + E_{molecule}) \tag{1}$$

where, E_{system} is the energy of metal surface + adsorbate, E_M is the energy of the bare metal surface, and $E_{molecule}$ is the energy of the bare adsorbate.

2.2. Machine Learning methods

The data collected from DFT calculations was then used to train ML models. Data points for only adsorption (and not dissociation) were used while training

and testing the model. It resulted in a total of 296 data points. Eight algorithms from three classes were tested before selecting the final model.

1. **Linear Methods** - Ordinary Linear Regression (OLR), Ridge Regression (RR), and LASSO were tested from the linear class of methods.

Linear Regression methods assume a linear relation between the input features X_i and target values y_i while fitting the coefficients w_i.

$$y(w, x) = w_0 + w_1 x_1 + w_2 x_2 + w_3 x_3 + ... \tag{2}$$

The objective function in Linear Regression is to optimize the residual sum of squares between the true and ML predicted target values,

$$min \frac{1}{n} \sum_{i=1}^{n} (y(pred)_i - y(true)_i)^2 \tag{3}$$

Linear Regression often leads to overfitting of small datasets as it depends on the collinearity of input features. Ridge Regression and LASSO take care of this by penalising the weights of modeled coefficients. While Ridge Regressor tends to penalizes the model for the sum of squared value of the weights, LASSO penalizes the sum of absolute values of the weights.

2. **Kernel based** - Kernel Ridge Regression (KRR) and Gaussian Process Regression (GPR) methods were tested from among the kernel based methods.

Kernel based methods use the "kernel trick" in which the data is transformed to a high-dimensional feature space and then the inner products between the images of all pairs of data in the feature space are found. A "kernel" is used to transform the data in a higher dimensional feature space where a hyperplane for classification of data can be easily found. Thus, Kernel Ridge Regression combines Ridge Regression and classification with the kernel trick.

3. **Tree-based** - Gradient Boosting Regression (GBR), Random Forest (RF), and Extra Tree Regressor (ETR); three tree-based methods were tested.

Decision Trees are supervised learning methods in which the target values are predicted by learning simple decision rules from the training data. The tree-based models finally develop into a sequence of trees with the right choice of

descriptors positioned correctly at different nodes. Tree-based methods generally have a high variance associated with them. GBR is a regression technique that uses decision tree-based classifiers as weak learners. In Random Forests, two sources of randomness are introduced by drawing random samples with replacement from the training set and finding the best split from a random subset of features. The reason to introduce these two sources of randomness is to reduce the variance.

Table 1: Hyper-parameters and range tested with GridSearchCV to find the best values for different estimators used.

Class	Method	hyper-parameter [range]
Linear	OLR	None
	RR	alpha ∈ [0.001, 0.01, 0.1, 0.5, 1, 5, 10]
	LASSO	alpha ∈ [0.001, 0.01, 0.1, 0.5, 1, 5, 10]
Kernel	KRR	kernel='rbf', alpha ∈ [1e-6, 1e-5, 1e-4, 1e-3, 1e-2, 0.1, 1], gamma ∈ np.logspace(-2,2,5)
	GPR	alpha ∈ [0.001, 0.01, 0.1, 0.5, 1, 5, 10]
Tree	GBR	n_estimators ∈ [100,200,300,400,500] , loss ∈ [ls, lad, huber], min_ samples_ split ∈ [2, 3, 4, 5, 6, 7, 8, 9, 10] , learning_rate ∈ [0.001, 0.01, 0.1, 1], max_depth ∈ [2, 3, 4, 5, 6, 7, 8, 9, 10]
	RF	n_ estimators ∈ [100,200,300,400,500], max_depth ∈ [2, 3, 4, 5, 6, 7, 8, 9, 10], min_samples_split ∈ [2, 3, 4, 5, 6, 7, 8, 9, 10]
	ETR	n_estimators ∈ [100,200,300,400,500], max_depth ∈ [2, 3, 4, 5, 6, 7, 8, 9, 10], min_samples_split ∈ [2, 3, 4, 5, 6, 7, 8, 9, 10]

We used all the ML algorithms as implemented in the scikit-learn python package.[74] Linear and Kernel based algorithms require pre-processing of data before applying an ML model to them. Three scalars viz. StandardScalar,

MinMaxScalar, and MaxAbsScalar were applied for appropriate models and the results were compared. Different train-test splits (50%-50%, 75%-25%, 85%-15%) were tested and finally, the train and test datasets were split in 75%-25%. To avoid bias, random shuffling of data points was carried out. To assess the performance of the models, Monte Carlo cross-validation was performed with 100 random test/training splits, i.e., 100 random leave-n-out trials. The MAE of difference between predicted and true target values is calculated for each trial and averaged to obtain mean MAE values and standard deviation. The cross-validation assessment was performed for 100, 200, and 500 trials. It was observed that the 100 trial average was as good an estimate as the average of 200 and 500 trails. And hence, MAE averaged over 100 trials is reported. ML algorithms need to be tuned for the best values of various hyper parameters. Exhaustive GridSearchCV as implemented in sci-kit learn was carried out to find the best parameter values of various estimators. Different hyper-parameters tuned and the range tested for each is reported in table 1. Validation curves were plotted for each of them to make sure that the cross validation score remained high for the finally used values of hyper-parameters. The learning curve was plotted to test the model performance as shown in Figure SI-3.

3. Results and Discussion

3.1. DFT calculations for adsorption energies and bondlength activation

The problem under investigation was to unravel the correlation between bondlength activation of adsorbate and their adsorption energies. And hence, activation of molecules like O_2, N_2, CO, and CO_2 that are involved in most of the chemical reactions were considered for calculations. Further, the late transition metals chosen for study were the ones which have been extensively reported to show maximum catalytic activity towards these small molecules. Two parameters viz. adsorption energies (E_{ads}) and the amount of activation that adsorbate bonds undergo upon interaction (BL_{act}) were carefully analyzed. It was generally noted that one-to-one correlation between E_{ads} and BL_{act} was

Table 2: Few selected cases of O_2, N_2, CO, and CO_2 are shown to highlight the missing correlation between adsorption energies and bondlength activation. Metal surface (M), facet, and site of adsorption (site) are also mentioned. It can be seen from the last two columns in the table that BL_{act} of molecules upon adsorption and corresponding adsorption energy E_{ads} do not have a one-to-one correlation.

Ads	M	facet	site	BL_{act} (Å)	E_{ads} (eV)
	Cu	110	SB	1.39	-2.1258
	Cu	110	T	1.42	-1.4411
O_2	Rh	100	B	1.38	-3.0984
	Rh	100	T	1.38	-1.7846
	Ir	111	B	1.41	-2.4389
	Ir	111	hcp	1.46	-2.3049
	Pd	111	fcc	1.13	-0.5706
	Pd	111	hcp	1.16	-0.2266
N_2	Ag	110	T	1.12	-0.0742
	Ag	110	LB	1.12	-0.0153
	Pt	100	B	1.17	-0.6105
	Pt	100	H	1.25	-0.4142
	Ir	111	T	1.16	-2.2864
	Ir	111	hcp	1.20	-2.0085
CO	Pd	110	SB	1.18	-2.362
	Pd	110	LB	1.22	-1.9619
	Rh	100	B	1.19	-2.4766
	Rh	100	H	1.21	-2.3852
	Au	110	H	1.18	-2.0752
	Au	110	T	1.18	-1.4153
CO_2	Pt	110	H	1.18	-0.6477
	Pt	110	T	1.23	-0.3562
	Pd	111	B	1.18	-0.3449
	Pd	111	fcc	1.25	-0.0284

missing for all adsorbates i.e., cases of adsorption with the maximum value of adsorption energy did not necessarily imply corresponding maximum bondlength activation of adsorbate or vice versa. To demonstrate this we have tabulated a few cases for all molecules in table 2. Data points shown in the table are across three facets of various metal surfaces to show that the comments hold valid irrespective of the metal or facet. Examples from all seven metal surfaces have been picked up and tabulated for different molecules. In addition, while comparing trends in between E_{ads} and BL_{act} we compare values on the same facet of one metal surface to prove that the absence of linear correlation is not caused by either of these variables. For instance, as reported in table 2, adsorption energy for oxygen molecule on short bridge (SB) site of Cu (110) surface is -2.13 eV and corresponding bondlength elongation is 1.39 Å. When the molecule adsorbs on the top (T) site of Cu (110) surface, the adsorption energy reduces by about 0.8 eV but the bondlength increases to 1.42 Å. Moreover, when we evaluated the adsorption behaviour of oxygen on Rh (100) surface we observed that while the difference in adsorption energies on the bridge (B) site and top (T) site is about 1.31 eV, the corresponding bondlength activation is the same on both the sites. Similarly, for N_2 adsorption on Pt (100) surface, we note that for bridge (B) site where the adsorption energy is higher than that of the hollow (H) site, the trends in bondlength activation are reversed (H site giving more elongation compared to B site). For CO_2 on Pd (111), adsorption energy on the bridge (B) site is -0.34 eV with no bond activation between the C-O bonds. But for the fcc site where a significant average bondlength activation up to 1.25 Å is witnessed, the value of adsorption energy is lesser (-0.0284 eV).

In figure 1 we present a detailed picture to highlight the point. E_{ads} (in red) for two molecules (N_2 and CO_2) on all metal surfaces and their corresponding BL_{act} (in blue) are plotted. E_{ads} are sorted in the order of reducing strength of adsorption for each M and then plotted. The range of BL_{act} (y2 axis) is common for all metal surfaces for a molecule and is shown on rightmost plot i.e. on Pd surface plot. Similar plots for O_2 and CO are shown in SI (Figure-SI1). We can see that in figure 1(a) the bondlength activation (in blue) for N_2 on Au surface

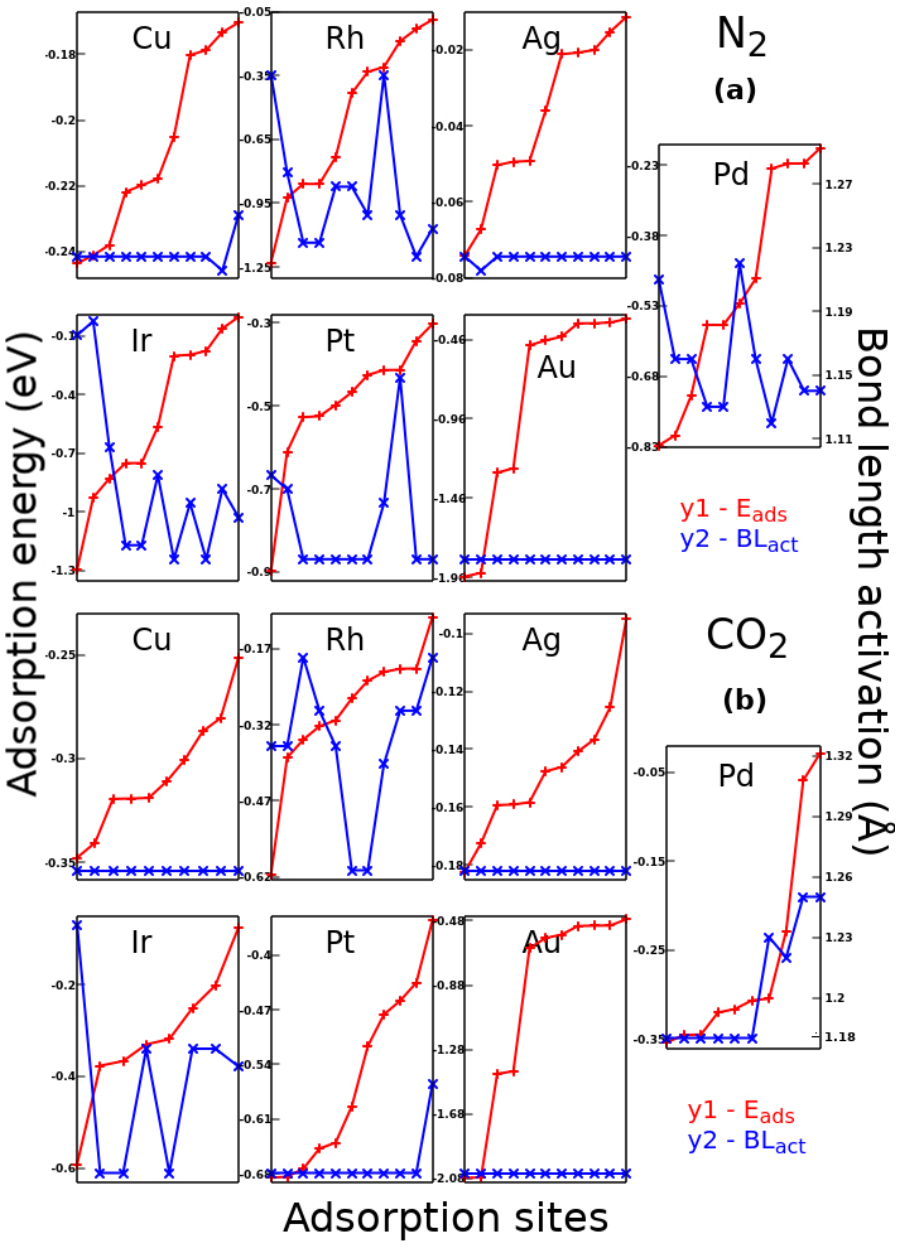

Figure 1: Adsorption energy (E_{ads}) as a function of the adsorption sites is plotted for (a) N_2 and (b) CO_2 molecules on 7 metal surfaces. E_{ads} values are sorted for each M and plotted in red on the y1 axis. Bondlength activation (BL_{act}) for corresponding E_{ads} values are plotted in blue on the y2 axis. The y2 axis for BL_{act} is kept common and shown in the plot for 'Pd' surface for both molecules. Similar plots for O_2 and CO are shown in fig. SI-1. The plots highlight the missing one-to-one correlation between E_{ads} and BL_{act}.

is negligible, but the corresponding E_{ads} (in red) has a range of about 1.5 eV. Also for Cu and Ag surfaces we notice that while there is a gradual increase in E_{ads} values, the corresponding BL_{act} does not show equal variation. In the case of CO_2 shown in figure 1(b), while Pd surface shows some correlation between E_{ads} and BL_{act}, the correlation is absolutely missing for adsorption on Ir and Rh surfaces. Even on Pt surface for CO_2 adsorption, it can be seen that while E_{ads} has values spread over the range of -0.62 eV to -0.33 eV, corresponding BL_{act} maintains nearly the same value. In the case of CO adsorption, the range for E_{ads} on Cu and Rh is small (figure SI-1) but the corresponding range for BL_{act} is still significant. Thus, we generally note that there is a lack of one-to-one correlation between the two investigated parameters viz. E_{ads} and BL_{act}. While one would expect some correlation between E_{ads} and BL_{act}, it is certainly not linear. Knowing that both parameters are important for catalysis, we need to understand the factors that influence both. Our DFT investigation brings out the fact that the factors affecting both the parameters are possibly different and a combination of factors affecting both should be taken into consideration while designing a catalyst. To test our hypothesis, we further used ML on the given dataset and analyzed the results.

3.2. Designing descriptors for ML

Designing the right set of features that correctly define the problem at hand is central to any ML problem. In this section, we discuss the strategy for designing descriptors to be used for training ML models. Three classes of descriptors were designed to handle the three variables in our problem viz. transition metal surface, facet, and adsorbate. Various periodic table properties for surface element and adsorbate along with reported values like the d-band center of an element, bond energy of adsorbate, and so on were tested. A total of thirty two descriptors were tested and finally, the descriptor set was reduced to the top twelve descriptors as mentioned below:

1. **Elemental properties** - Group of the metal atom (Group_M), Melting point (MP_M), Electronegativity (EN_M), d-band center (dbc_M), van der Waals radii

(VR_M), and Atomic mass (AM_M).

2. **Adsorbate/Molecular properties** - HOMO-LUMO gap (HL_a), electron affinity of molecule (EA_a), bond energy (BE_a), and Melting point of molecule (MP_a).

3. **Facet** - generalised cordination number (gcn_f), and atomic density (AD_f).

As mentioned before, since various classes of ML algorithms were tested before

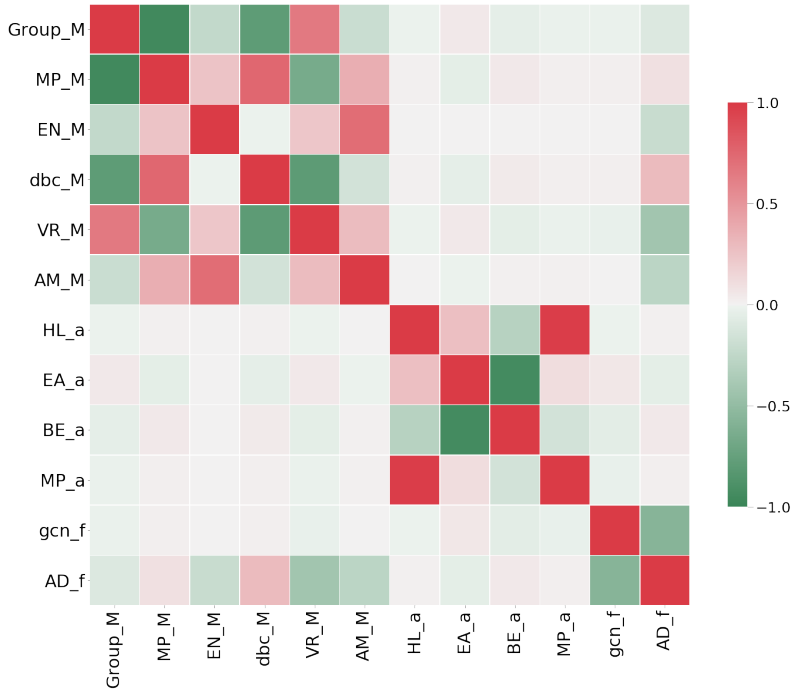

Figure 2: Pearson's correlation plot for final reduced set of 12 descriptors. Descriptors represent properties related to surface element, surface facet and adsorbate. While, some descriptors are strongly correlated, most correlated descriptors were removed unless it was important to retain both

.

finalising the best model it was important to deal with multicollinearity in the feature space. Predictions for some algorithms (like linear regression models) could be flawed if any sort of multicollinearity existed in the provided set of descriptors. Pearson correlation coefficient, a measure of linear correlation be-

tween each pair of descriptors was calculated. It is essentially the covariance of the two variables, divided by the product of their standard deviations. Although it is important to remove multicollinear features, sometimes retaining some features is equally important. For example, as shown in figure 2, the bond energy of a molecule has a high negative correlation with the electron affinity of a molecule. But both are important factors to be considered while understanding adsorption energy and also bondlength activation that molecules undergo upon adsorption. Hence, removing either of them would mean under representing the data in feature space. Nevertheless,knowing the correlated features in the dataset is important. Hence, after reducing as many collinear features as possible we came up with a final set of twelve descriptors shown in figure 2. We can see that some features like MP_M have a high correlation with Group_M and also dbc_M. MP_a has a strong positive correlation with HL_a. While training the dataset with linear models, care was taken to eliminate as many correlated features from these top twelve as possible. Since, kernel and tree-based methods are immune to multicollinearity in the feature set, having a few correlated features does not affect the working of such ML algorithms. The complete correlation plot for all the features tested is given in SI (figure SI-2).

3.3. Choosing the best ML method and parameters for E_{ads} prediction

We tested a total of 8 ML algorithms across three classes (linear, kernel, and tree based) of methods. Figure 3 presents a summary of the train and test MAE for adsorption energy predictions. The values of MAE are averaged over 100 random trials of train-test splits with 25% data held out as test data each time. We observe that the range of error in E_{ads} prediction varies from 0.44 eV to 0.24 eV. Figure 4 shows how well the models fit DFT data. We plot ML predicted values of E_{ads} versus the DFT calculated values during one of the trials with training performed on 75% data and tested on 25% data. We clearly see improved performance for prediction by GPR and RFR in comparison with the Ridge algorithm (linear method). While it is clear that linear methods give maximum error in prediction, kernel based and tree based methods perform

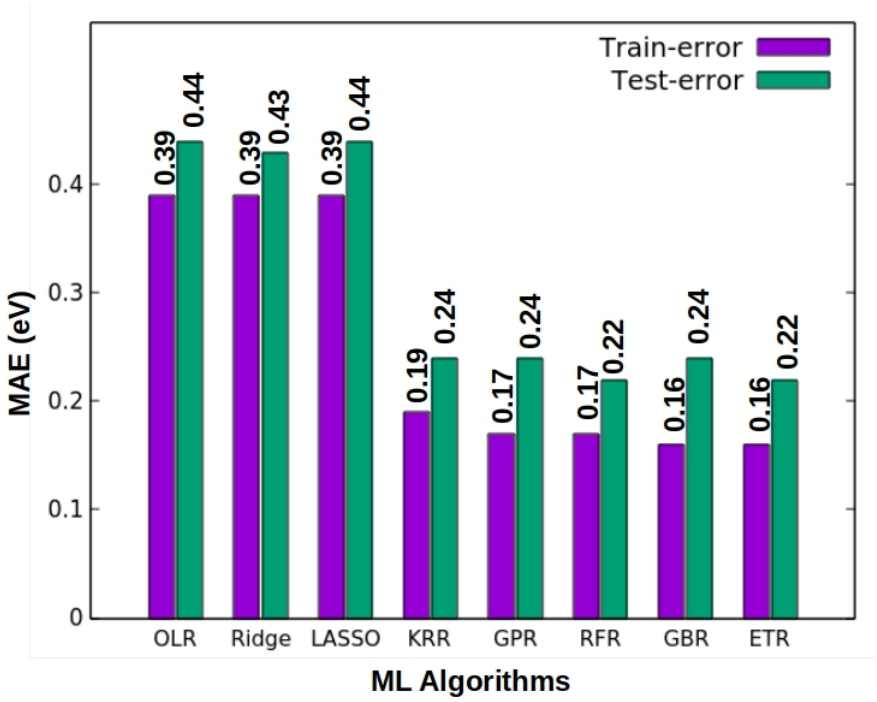

Figure 3: Average MAEs for E_{ads} prediction obtained over 100 trials for different methods. It can be seen that the linear class of methods perform the worst for the given problem. Kernel-based methods and tree-based methods perform competitively well.

competitively well. Since the bias is equally low for methods of both the classes, we further computed the explained variance for kernel-based and tree-based methods. To build a good model we need to find a good balance between bias and variance such that it minimizes the error. The accuracy for variability in model prediction for tree-based methods was 85% whereas kernel-based methods had a lower accuracy of about 80%. Tree-based methods have other benefits too, like no scaling of data is required before training the model, and algorithms like ETR and RFR are not sensitive to hyperparameter tuning which in turn reduces the training time of the model significantly. Overall we noted that the tree based methods work the best and hence they were chosen for further investigations.

Furthermore, the dataset used to train the ML model is small in size, and

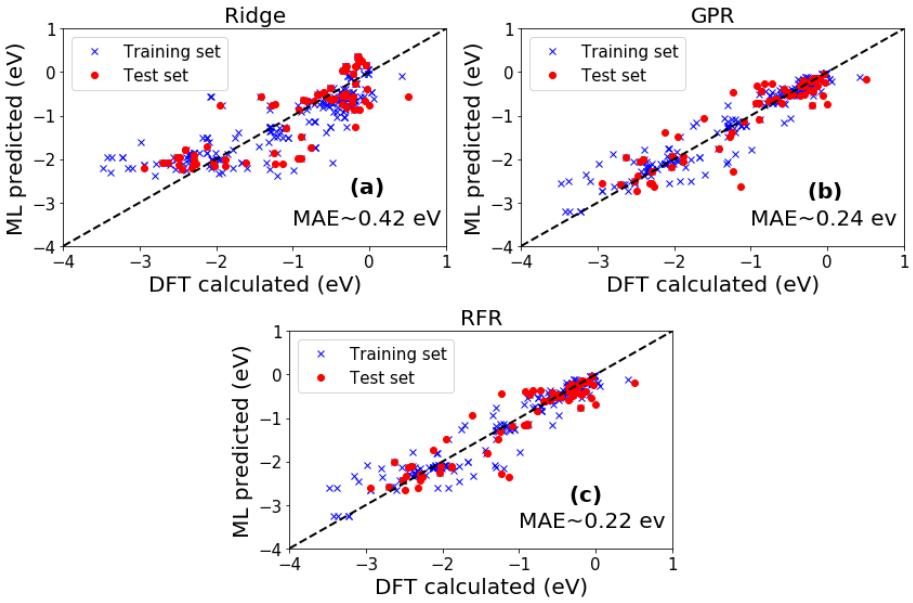

Figure 4: Comparison between DFT versus ML predicted E_{ads} values methods from each class, (a) Linear: Ridge regression, (b) Kernel: Gaussian process regressor, and (c) Tree: Random tree regressor. MAE's for each method are noted in the plot. Tree based methods perform the best overall.

hence the prediction variance of the model could highly depend on which values get classified into the test set. Upon changing the examples in the test set the model may end up performing badly due to large prediction variance. One way to deal with this is by randomly splitting the dataset into train-test sets multiple

Table 3: Average MAEs for E_{ads} prediction using Random Forest method. Averaged over 100, 200, and 500 trials and tested for different train-test split sizes viz. 50-50%, 75-25%, and 85-15%. 75-25% split gives the lowest average MAE as well as std over all their trial sizes.

Train-Test size	50-50%		75-25%		85-15%	
no. trials	mean	std	mean	std	mean	std
100	0.27	0.026	0.24	0.028	0.24	0.036
200	0.27	0.023	0.24	0.030	0.24	0.040
500	0.27	0.023	0.24	0.030	0.24	0.040

times and testing the model performance. Therefore, prediction of MAE was tested on 200 and 500 (instead of 100) random splits of dataset along with different train-test split sizes using RFR, which was the best model (see table 3).

It is observed that MAE averaged over 100, 200, and 500 trials have a higher value for 50%-50% train-test split. Although the average MAE for 85%-15% train-test split is as much as that for 75%-25% split, it has a higher standard deviation for prediction. The averaged values of MAE for all train-test splits are comparable for a greater number of trials (i.e. 200 and 500). Hence, we can safely use the train-test split of 75%-25% with MAE averaged over 100 random trials.

3.4. Comparison between E_{ads} and BL_{act} using ML

We investigated the descriptor ranking for E_{ads} as computed by the tree-based random forest algorithm, as shown in figure 5. The feature importance score as calculated by the methods can be analyzed to understand the factors that play an important role in predicting the E_{ads} values. We calculated and plotted both impurity-based (figure 6(a)) and permutation feature ranking (figure 6(b)) for random forest method. Impurity based feature ranking is calculated while training the ML model based on the decisions taken at nodes while splitting. On the other hand, the permutation importance is calculated after the model is fitted. While performing permutation importance, data for any one of the feature is randomly shuffled in the validation set keeping the rest of the input and target values as they are. In this way the model is expected to perform worse, and the drop in accuracy indicates the importance of the feature in predicting the target. Hence, permutation importance can deal with the bias that impurity-based ranking can have towards highly important features that have not seen the test data. Figure 5 shows the impurity based feature ranking computed over 100 trials. The feature ranking plot in Figure 6, is for one of the instances out of the 100 train-test splits. We observe that out of the twelve descriptors used to train the ML model, a mixture of adsorbate, element, and

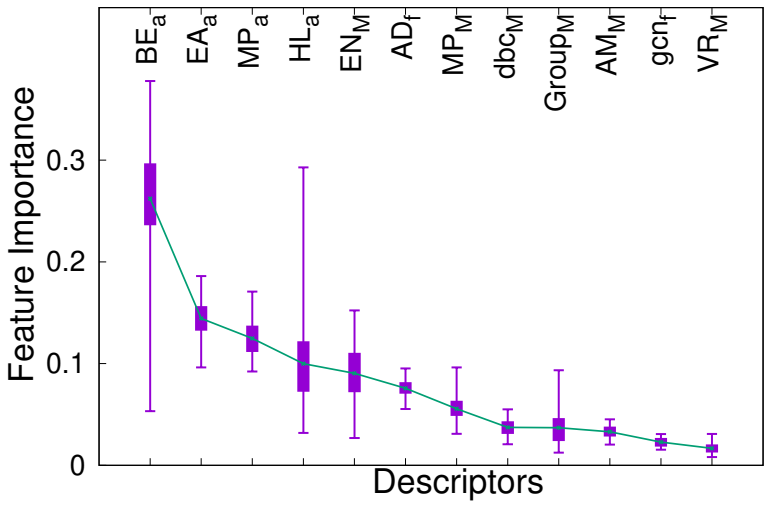

Figure 5: Impurity-based feature importance for prediction of E_{ads} values using 25% test data with Random Forest Regressor calculated over 100 trials. A mixture of adsorbate, element and facet properties rank among the top features.

facet properties ranked among the top six descriptors. Bond energy, melting point, and electron affinity of adsorbate were followed upon by electronegativity of the element. While bond energy and electron affinity of adsorbate describe its capacity to take part in bond formation and indulge in the transfer of charge, EN_M describes the same for the element. We understand that for adsorption to take place there has to be a transfer of electrons from the surface to the adsorbate indicating interaction. The extent of this interaction/adsorption will depend on both, the capacity of the element to donate electrons and at the same time the capacity of adsorbate orbitals to gain electrons. Hence, we noticed that the properties indicating charge transfer capacities of element and adsorbate rank higher than other features. We also observed that the atomic density of the given facet (AD_f), was among the top-ranking features. This number changes by changing both, metal or facet. Hence, this feature defines the distribution of atoms on the surface available for interaction. The atomic

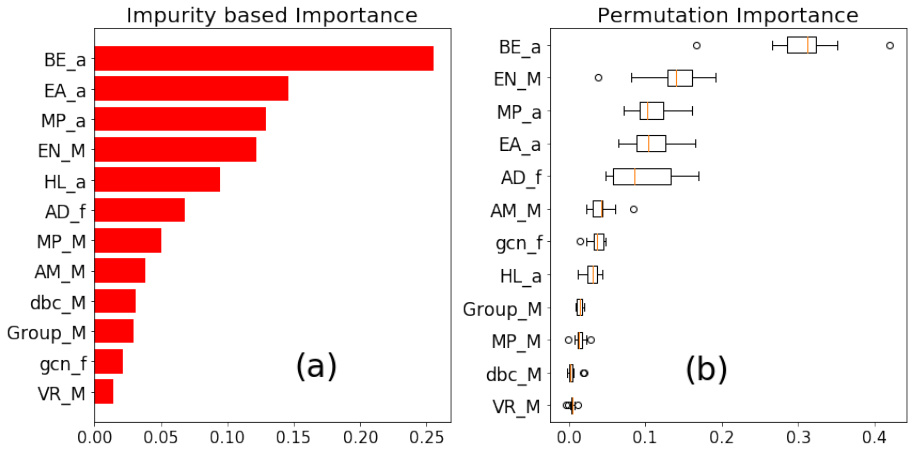

Figure 6: (a) Impurity-based and (b) Permutation feature importance for prediction of E_{ads} values using 25% test data with Random Forest Regressor. A mixture of adsorbate, element and facet properties rank among the top features. Set of top 6 features is the same for both with some shuffling in the order.

arrangement is another important factor that is known to play an active role in governing interactions at the surface. While the ranking shuffled a little, the set of top 6 descriptors were the same for impurity-based ranking (see figure 6(a)), permutation based ranking (see figure 6(b)) and also when averaged over 100 trials (see figure 5) .

Finally, to test our hypothesis about finding factors that differently affect E_{ads} and BL_{act} during adsorption, we trained our best ML model with the same set of features but a different target value this time. We used the same set of 12 chosen descriptors to train the Random Forest algorithm and predict bondlength activations for all molecules upon adsorption. It was observed that the model performed well with an MAE of 0.019 ± 0.002 Å for BL_{act} prediction. Figure 7 shows the ML prediction for corresponding DFT values. Although the prediction errors are less than 10% (since the range of BL_{act} prediction is 0.2 Å) there is still scope for further improvement. Not only predictions on the test data points but also training of the model can be improved. One way to achieve this is with the inclusion of more descriptors that fit better for BL_{act}

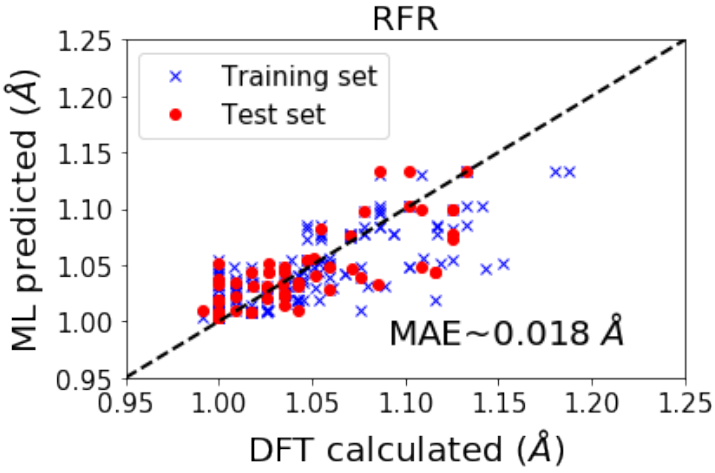

Figure 7: ML predicted versus DFT calculated values for BL_{act} using 25% test data with tree based random forest regressor. MAE for BL_{act} prediction with same set of 12 descriptors is 0.018 Å.

predictions. But since the problem under investigation was to unravel the lack of linear correlation between E_{ads} and BL_{act} we use the same set of descriptors to predict both.

Investigations for feature ranking highlighted some interesting observations. We note that bond energy of the molecule ranked invariably the highest for E_{ads} predictions, but that is not the case for BL_{act}. Although bond energy still has a significant ranking among the tested set of features, it does not retain its number one position while predicting BL_{act} values. Also, the 'HOMO-LUMO gap' which was a significant feature for E_{ads} prediction, ranked consistently low for BL_{act} prediction, as can be seen in figure 8 and figure 9. Further, element properties like the d-band center and melting point show up among the top ranking features for BL_{act} prediction. While MP had about 5% importance for E_{ads} prediction as well, the d-band center ranked very low. In contrast, for BL_{act} prediction we noticed that the d-band center of element ranks consistently high in all three feature ranking plots (figure 8, figure 9(a) and figure 9(b)). The d-band center is often related to adsorption energies of the element, but our results proved

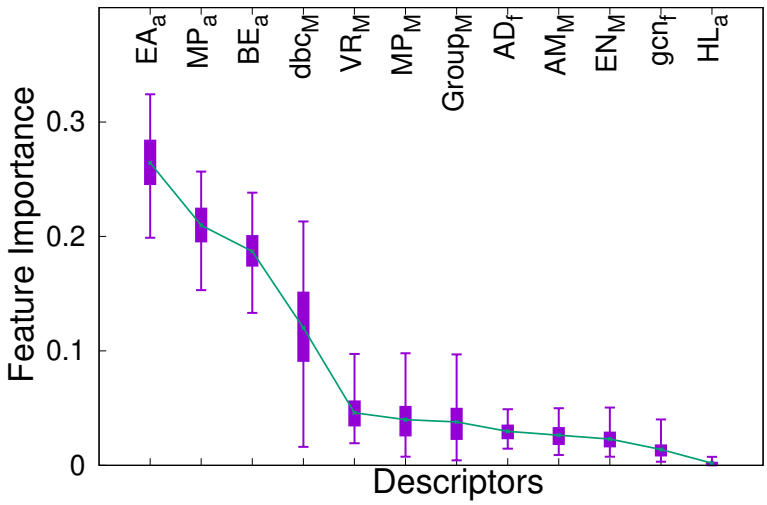

Figure 8: Impurity-based feature importance for prediction of E_{ads} values using 25% test data with the Random Forest Regressor calculated over 100 trials. Newer features like the d-band center, Van der Waal's radii, and Group of the element contributes non-trivially for BL_{act} prediction.

otherwise. Knowing that the d-band center as a descriptor has proven to be extremely successful for adsorption studies, our investigations showed that linking it with bondlength activation could probably yield better models. At this point we would also like to point out the importance that melting point (both of the adsorbate and the element) as a descriptor has for BL_{act} as well as E_{ads} predictions. Not only our study, but many other studies where attempts are made to link elemental properties with adsorption energies report high importance for the melting point of an element as a feature.[60, 57] This observation calls for further investigations to understand the hidden correlation between the two. It was also noted that an electronic descriptor VR (van der Waals radii) of the element ranks higher than what it did for E_{ads}. VR of the elements influence their nearest neighbor arrangement on the surface. Upon adsorption the availability of nearest atom from adsorption site and the local environment as seen by the adsorbate gets reflected in this descriptor. Hence, we observed that an

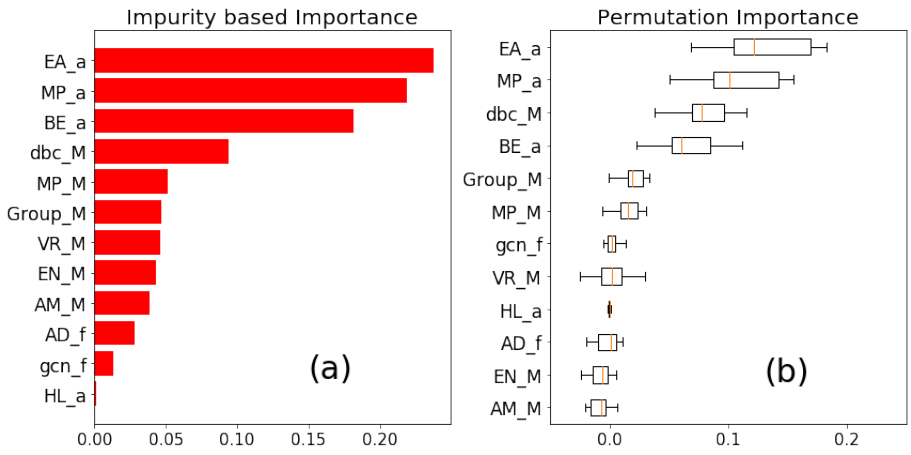

Figure 9: (a) Impurity-based and (b) Permutation feature importance for prediction of BL_{act} values using 25% test data with the Random Forest Regressor. The d-band center ranked comparatively lower for E_{ads} prediction. HLgap ranks consistently low for BL_{act} prediction, unlike E_{ads}.

electronic feature influencing the surface geometry ranked consistently low for E_{ads} prediction, but contributed non-trivially for BL_{act} prediction. Thus, we observed that the set of features that influence E_{ads} and BL_{act} share some but not all common features. Understanding that both the parameters under investigation do not follow a linear relation, our ML investigations further highlight the factors that affect each of the parameters differently. Overall we noted that understanding interaction at surfaces is a multi-variate problem. In our study, we highlighted the role of BL_{act} as another important factor to be analyzed to get a better picture of these interactions. Also, the scope of improving ML models further to predict bond length activation values calls for more attention towards the parameter.

4. Conclusion

Understanding adsorption at transition metal surfaces requires a comprehensive understanding of multiple factors that govern interactions. Adsorption energy is often considered an important parameter to understand the extent of

interaction of an adsorbate with the surface. While we do not debate the importance of analyzing E_{ads}, we emphasize the role of another important factor i.e. bondlength activation of adsorbates. Our DFT investigations for more than one molecule adsorbed on multiple surfaces demonstrated the absence of linear correlation between E_{ads} and BL_{act}. This makes the analysis of both the parameters important and interesting. ML methods were employed to understand the factors that influenced E_{ads} and BL_a and also confirm that they were different for both. Tree-based Random Forest Regression model was chosen to perform ML analysis, after checking its performance against 7 more ML algorithms. It is interesting to note that our best performing ML model worked well with simple descriptors and predicted E_{ads} with an MAE of 0.2 eV and BL_{act} with MAE of 0.02 Å. Further investigations to understand the feature ranking highlighted the difference between the factors that influence the prediction of each of the parameters using ML models. They not only point at the lacking one-to-one correlation between E_{ads} and BL_{act}, but also bring out the differences in terms of feature ranking. Charge transfer descriptors like the bond energy and electron affinity of adsorbate along with the electronegativity of element were among the top-ranking features for E_{ads} predictions. While E_{ads} correlated well with the HOMO-LUMO gap of the adsorbate, BL_{act} did not. It was also noted that the d-band center of the element correlated better with BL_{act} values (resulting in higher ranking) in comparison with E_{ads} values. These observations successfully test our hypothesis and prove that the correlation between E_{ads} and BL_{act} is not linear and analyzing both while studying interactions at the surface is imperative.

Conflicts of interest

"There are no conflicts to declare".

Acknowledgements

ParamBrahma- IISER (Pune) and CSIR-4PI are gratefully acknowledged for the computational facility. KJ acknowledges DST (EMR/2016/000591) for partial financial support. SA acknowledges DST-INSPIRE for research fellowship.

References

[1] A. Nilsson, L. G. Pettersson, J. Norskov, Chemical bonding at surfaces and interfaces, Elsevier, 2011.

[2] B. Hammer, J. K. Nørskov, Theoretical surface science and catalysis—calculations and concepts, Advances in catalysis 45 (2000) 71–129.

[3] S.-G. Wang, X.-Y. Liao, D.-B. Cao, C.-F. Huo, Y.-W. Li, J. Wang, H. Jiao, Factors controlling the interaction of co2 with transition metal surfaces, The Journal of Physical Chemistry C 111 (45) (2007) 16934–16940.

[4] M. M. Montemore, M. A. van Spronsen, R. J. Madix, C. M. Friend, O2 activation by metal surfaces: implications for bonding and reactivity on heterogeneous catalysts, Chemical Reviews 118 (5) (2017) 2816–2862.

[5] D. C. Ford, A. U. Nilekar, Y. Xu, M. Mavrikakis, Partial and complete reduction of o2 by hydrogen on transition metal surfaces, Surface Science 604 (19-20) (2010) 1565–1575.

[6] J. K. Nørskov, T. Bligaard, A. Logadottir, S. Bahn, L. B. Hansen, M. Bollinger, H. Bengaard, B. Hammer, Z. Sljivancanin, M. Mavrikakis, et al., Universality in heterogeneous catalysis, Journal of Catalysis 209 (2) (2002) 275–278.

[7] A. J. Medford, A. Vojvodic, J. S. Hummelshøj, J. Voss, F. Abild-Pedersen, F. Studt, T. Bligaard, A. Nilsson, J. K. Nørskov, From the sabatier principle to a predictive theory of transition-metal heterogeneous catalysis, Journal of Catalysis 328 (2015) 36–42.

[8] B. Hammer, J. K. Nørskov, Electronic factors determining the reactivity of metal surfaces, Surface science 343 (3) (1995) 211–220.

[9] A. Kulkarni, S. Siahrostami, A. Patel, J. K. Nørskov, Understanding catalytic activity trends in the oxygen reduction reaction, Chemical Reviews 118 (5) (2018) 2302–2312.

[10] B. W. Chen, L. Xu, M. Mavrikakis, Computational methods in heterogeneous catalysis, Chemical Reviews.

[11] J. Greeley, Theoretical heterogeneous catalysis: scaling relationships and computational catalyst design, Annual review of chemical and biomolecular engineering 7 (2016) 605–635.

[12] C. Shi, H. A. Hansen, A. C. Lausche, J. K. Nørskov, Trends in electrochemical co 2 reduction activity for open and close-packed metal surfaces, Physical Chemistry Chemical Physics 16 (10) (2014) 4720–4727.

[13] P. Ferrin, S. Kandoi, A. U. Nilekar, M. Mavrikakis, Hydrogen adsorption, absorption and diffusion on and in transition metal surfaces: A dft study, Surface Science 606 (7-8) (2012) 679–689.

[14] M. Okumura, Y. Irie, Y. Kitagawa, T. Fujitani, Y. Maeda, T. Kasai, K. Yamaguchi, Dft studies of interaction of ir cluster with o2, co and no, Catalysis Today 111 (3-4) (2006) 311–315.

[15] B. Hammer, J. K. Norskov, Why gold is the noblest of all the metals, Nature 376 (6537) (1995) 238–240.

[16] A. Groß, Theoretical surface science, A Microscopic Perspective. Originally published in the series: Advanced Texts in Physics 132.

[17] H. Xin, S. Linic, Communications: Exceptions to the d-band model of chemisorption on metal surfaces: The dominant role of repulsion between adsorbate states and metal d-states (2010).

[18] M. P. Hyman, J. W. Medlin, Effects of electronic structure modifications on the adsorption of oxygen reduction reaction intermediates on model pt (111)-alloy surfaces, The Journal of Physical Chemistry C 111 (45) (2007) 17052–17060.

[19] M. P. Hyman, B. T. Loveless, J. W. Medlin, A density functional theory study of h2s decomposition on the (1 1 1) surfaces of model pd-alloys, Surface Science 601 (23) (2007) 5382–5393.

[20] M. T. Gorzkowski, A. Lewera, Probing the limits of d-band center theory: electronic and electrocatalytic properties of pd-shell–pt-core nanoparticles, The Journal of Physical Chemistry C 119 (32) (2015) 18389–18395.

[21] S. Bhattacharjee, U. V. Waghmare, S.-C. Lee, An improved d-band model of the catalytic activity of magnetic transition metal surfaces, Scientific reports 6 (1) (2016) 1–10.

[22] Y. Tang, Z. Yang, X. Dai, Preventing the co poisoning on pt nanocatalyst using appropriate substrate: a first-principles study, Journal of Nanoparticle Research 14 (5) (2012) 1–11.

[23] Y. Qian, Y. Liu, Y. Zhao, X. Zhang, G. Yu, Single vs double atom catalyst for n2 activation in nitrogen reduction reaction: A dft perspective, EcoMat 2 (1) (2020) e12014.

[24] W. Li, H. Wang, X. Jiang, J. Zhu, Z. Liu, X. Guo, C. Song, A short review of recent advances in co 2 hydrogenation to hydrocarbons over heterogeneous catalysts, RSC Advances 8 (14) (2018) 7651–7669.

[25] W. J. Durand, A. A. Peterson, F. Studt, F. Abild-Pedersen, J. K. Nørskov, Structure effects on the energetics of the electrochemical reduction of co2 by copper surfaces, Surface Science 605 (15-16) (2011) 1354–1359.

[26] J. Ko, B.-K. Kim, J. W. Han, Density functional theory study for catalytic activation and dissociation of co2 on bimetallic alloy surfaces, The Journal of Physical Chemistry C 120 (6) (2016) 3438–3447.

[27] D. Hibbitts, E. Iglesia, Prevalence of bimolecular routes in the activation of diatomic molecules with strong chemical bonds (o2, no, co, n2) on catalytic surfaces, Accounts of Chemical Research 48 (5) (2015) 1254–1262.

[28] F. Liu, C. Wu, G. Yang, S. Yang, Co oxidation over strained pt (100) surface: a dft study, The Journal of Physical Chemistry C 119 (27) (2015) 15500–15505.

[29] C. Ren, Q. Jiang, W. Lin, Y. Zhang, S. Huang, K. Ding, Density functional theory study of single-atom v, nb, and ta catalysts on graphene and carbon nitride for selective nitrogen reduction, ACS Applied Nano Materials 3 (6) (2020) 5149–5159.

[30] C. D. Zeinalipour-Yazdi, J. S. Hargreaves, C. R. A. Catlow, Dft-d3 study of molecular n2 and h2 activation on co3mo3n surfaces, The Journal of Physical Chemistry C 120 (38) (2016) 21390–21398.

[31] L. O. Paz-Borbón, F. Baletto, A dft study on the o2 adsorption properties of supported ptni clusters, Inorganics 5 (3) (2017) 43.

[32] C.-L. Hu, J.-Q. Li, Y.-F. Zhang, X.-L. Hu, N.-X. Lu, Y. Chen, A dft study of o2 adsorption on periodic gan (0 0 0 1) and ($000\bar{1}$) surfaces, Chemical Physics Letters 424 (4-6) (2006) 273–278.

[33] F. Nasehnia, M. Seifi, Adsorption of molecular oxygen on viiib transition metal-doped graphene: A dft study, Modern Physics Letters B 28 (30) (2014) 1450237.

[34] N. H. Linh, T. Q. Nguyen, W. A. Diño, H. Kasai, Effect of oxygen vacancy on the adsorption of o2 on anatase tio2 (001): A dft-based study, Surface Science 633 (2015) 38–45.

[35] J. Ye, C. Liu, Q. Ge, Dft study of co2 adsorption and hydrogenation on the in2o3 surface, The Journal of Physical Chemistry C 116 (14) (2012) 7817–7825.

[36] L. I. Bendavid, E. A. Carter, Co2 adsorption on cu2o (111): a dft+ u and dft-d study, The Journal of Physical Chemistry C 117 (49) (2013) 26048–26059.

[37] B. Wanno, C. Tabtimsai, A dft investigation of co adsorption on viiib transition metal-doped graphene sheets, Superlattices and Microstructures 67 (2014) 110–117.

[38] M. Gajdoš, J. Hafner, Co adsorption on cu (1 1 1) and cu (0 0 1) surfaces: Improving site preference in dft calculations, Surface science 590 (2-3) (2005) 117–126.

[39] H. Orita, Y. Inada, Dft investigation of co adsorption on pt (211) and pt (311) surfaces from low to high coverage, The Journal of Physical Chemistry B 109 (47) (2005) 22469–22475.

[40] S. Mehta, S. Agarwal, N. Kenge, S. P. Mekala, V. Patil, T. Raja, K. Joshi, Mixed metal oxide: A new class of catalyst for methanol activation, Applied Surface Science 534 (2020) 147449.

[41] M. A. Petersen, J.-A. van den Berg, I. M. Ciobica, P. van Helden, Revisiting co activation on co catalysts: impact of step and kink sites from dft, ACS Catalysis 7 (3) (2017) 1984–1992.

[42] R. Jinnouchi, R. Asahi, Predicting catalytic activity of nanoparticles by a dft-aided machine-learning algorithm, The Journal of Physical Chemistry Letters 8 (17) (2017) 4279–4283.

[43] S. Back, K. Tran, Z. W. Ulissi, Toward a design of active oxygen evolution catalysts: insights from automated density functional theory calculations and machine learning, ACS Catalysis 9 (9) (2019) 7651–7659.

[44] I. Funes-Ardoiz, F. Schoenebeck, Established and emerging computational tools to study homogeneous catalysis—from quantum mechanics to machine learning, Chem 6 (8) (2020) 1904–1913.

[45] M. Sun, A. W. Dougherty, B. Huang, Y. Li, C.-H. Yan, Accelerating atomic catalyst discovery by theoretical calculations-machine learning strategy, Advanced Energy Materials 10 (12) (2020) 1903949.

[46] H. Li, Z. Zhang, Z. Liu, Application of artificial neural networks for catalysis: a review, Catalysts 7 (10) (2017) 306.

[47] G. Pilania, C. Wang, X. Jiang, S. Rajasekaran, R. Ramprasad, Accelerating materials property predictions using machine learning, Scientific Reports 3 (2013) 2810.

[48] A. Jain, T. Bligaard, Atomic-position independent descriptor for machine learning of material properties, Physical Review B 98 (21) (2018) 214112.

[49] A. Seko, H. Hayashi, K. Nakayama, A. Takahashi, I. Tanaka, Representation of compounds for machine-learning prediction of physical properties, Physical Review B 95 (14) (2017) 144110.

[50] K. Schütt, H. Glawe, F. Brockherde, A. Sanna, K. Müller, E. Gross, How to represent crystal structures for machine learning: Towards fast prediction of electronic properties, Physical Review B 89 (20) (2014) 205118.

[51] Z. Li, X. Ma, H. Xin, Feature engineering of machine-learning chemisorption models for catalyst design, Catalysis Today 280 (2017) 232–238.

[52] S. Laghuvarapu, Y. Pathak, U. D. Priyakumar, Band nn: A deep learning framework for energy prediction and geometry optimization of organic small molecules, Journal of computational chemistry 41 (8) (2020) 790–799.

[53] P. Pankajakshan, S. Sanyal, O. E. de Noord, I. Bhattacharya, A. Bhattacharyya, U. Waghmare, Machine learning and statistical analysis for materials science: stability and transferability of fingerprint descriptors and chemical insights, Chemistry of Materials 29 (10) (2017) 4190–4201.

[54] I. Miyazato, T. N. Nguyen, L. Takahashi, T. Taniike, K. Takahashi, Representing catalytic and processing space in methane oxidation reaction via

multioutput machine learning, The Journal of Physical Chemistry Letters 12 (2) (2021) 808–814.

[55] J. R. Kitchin, Machine learning in catalysis, Nature Catalysis 1 (4) (2018) 230–232.

[56] S. Kapse, S. Janwari, U. V. Waghmare, R. Thapa, Energy parameter and electronic descriptor for carbon based catalyst predicted using qm/ml, Applied Catalysis B: Environmental 286 (2021) 119866.

[57] S. Saxena, T. S. Khan, F. Jalid, M. Ramteke, M. A. Haider, In silico high throughput screening of bimetallic and single atom alloys using machine learning and ab initio microkinetic modelling, Journal of Materials Chemistry A 8 (1) (2020) 107–123.

[58] Z. Yang, W. Gao, Q. Jiang, A machine learning scheme for the catalytic activity of alloys with intrinsic descriptors, Journal of Materials Chemistry A 8 (34) (2020) 17507–17515.

[59] S. Agarwal, S. Mehta, K. Joshi, Understanding the ml black box with simple descriptors to predict cluster–adsorbate interaction energy, New Journal of Chemistry 44 (20) (2020) 8545–8553.

[60] T. Toyao, K. Suzuki, S. Kikuchi, S. Takakusagi, K.-i. Shimizu, I. Takigawa, Toward effective utilization of methane: machine learning prediction of adsorption energies on metal alloys, The Journal of Physical Chemistry C 122 (15) (2018) 8315–8326.

[61] C. Liu, Y. Li, M. Takao, T. Toyao, Z. Maeno, T. Kamachi, Y. Hinuma, I. Takigawa, K.-i. Shimizu, Frontier molecular orbital based analysis of solid–adsorbate interactions over group 13 metal oxide surfaces, The Journal of Physical Chemistry C 124 (28) (2020) 15355–15365.

[62] K. T. Butler, D. W. Davies, H. Cartwright, O. Isayev, A. Walsh, Machine learning for molecular and materials science, Nature 559 (7715) (2018) 547–555.

[63] T. Davran-Candan, M. E. Günay, R. Yıldırım, Structure and activity relationship for co and o 2 adsorption over gold nanoparticles using density functional theory and artificial neural networks, Journal Chemical Physics 132 (17) (2010) 174113.

[64] T. Mueller, A. G. Kusne, R. Ramprasad, Machine learning in materials science: Recent progress and emerging applications, Reviews in Computational Chemistry 29 (2016) 186–273.

[65] X. Ma, Z. Li, L. E. Achenie, H. Xin, Machine-learning-augmented chemisorption model for co2 electroreduction catalyst screening, The Journal of Physical Chemistry Letters 6 (18) (2015) 3528–3533.

[66] M. O. Jäger, E. V. Morooka, F. F. Canova, L. Himanen, A. S. Foster, Machine learning hydrogen adsorption on nanoclusters through structural descriptors, npj Computional Materials 4 (1) (2018) 37.

[67] P. E. Blöchl, Projector augmented-wave method, Physical Review B 50 (1994) 17953–17979.

[68] G. Kresse, D. Joubert, From ultrasoft pseudopotentials to the projector augmented-wave method, Physical Review B 59 (1999) 1758–1775.

[69] J. P. Perdew, K. Burke, M. Ernzerhof, Generalized gradient approximation made simple, Physical Review Letters 77 (1996) 3865–3868.

[70] J. P. Perdew, K. Burke, M. Ernzerhof, Generalized gradient approximation made simple [phys. rev. lett. 77, 3865 (1996)], Physical Review Letters 78 (1997) 1396–1396.

[71] G. Kresse, J. Hafner, *Ab initio* molecular-dynamics simulation of the liquid-metal–amorphous-semiconductor transition in germanium, Physical Review B 49 (1994) 14251–14269.

[72] G. Kresse, J. Furthmüller, Efficient iterative schemes for *ab initio* total-energy calculations using a plane-wave basis set, Physical Review B 54 (1996) 11169–11186.

[73] G. Kresse, J. Furthmüller, Efficiency of ab-initio total energy calculations for metals and semiconductors using a plane-wave basis set, Computational Material Science 6 (1) (1996) 15 – 50.

[74] 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, E. Duchesnay, Scikit-learn: Machine learning in Python, J. Mach. Learn. Res. 12 (2011) 2825–2830.