

# In Silico High Throughput Screening of Bimetallic and Single Atom Alloys Using Machine Learning and Ab Initio Microkinetic Modelling

Shivam Saxena<sup>1,\$</sup>, Tuhin S. Khan<sup>1,\$,\*</sup>, Fatima Jalid<sup>1,3</sup>, Manoj C. Ramteke<sup>2</sup>, M. Ali Haider<sup>1,\*</sup>

1. Renewable Energy and Chemicals Lab, Department of Chemical Engineering, Indian Institute of Technology Delhi, Hauz Khas, New Delhi, India
2. Department of Chemical Engineering, Indian Institute of Technology Delhi, Hauz Khas, New Delhi, India
3. Department of Chemical Engineering, National Institute of Technology Srinagar, Srinagar, Jammu and Kashmir, India

\$ equal contribution

\* corresponding authors email address: [tuhinsk@iitd.com](mailto:tuhinsk@iitd.com), [haider@iitd.ac.in](mailto:haider@iitd.ac.in)

## ABSTRACT

The advent of machine learning (ML) techniques in solving problems related to material science and chemical engineering is driving expectations to give faster predictions of material properties. For heterogeneous catalysis applications, relying on the age-old Sabatier principle, an ab initio in-silico high throughput screening of catalyst materials is envisaged, wherein ML based methods are showing potential to significantly reduce the experimental as well as computation cost. Availability of ML algorithms (through open source libraries like Scikit-Learn) and materials database (like CatApp, Materials Project) is further augmenting this realization. On using these resources, ML models are developed to predict the binding energies of oxygen and carbon on A<sub>3</sub>B bimetallic alloys and Cu-based single atom alloys (SAAs) using features of the metals that are readily available in periodic table and other published reports. Several ML models for predicting oxygen binding energy for AA terminated A<sub>3</sub>B alloys are analysed and gradient boosting regressor (GBR) is observed to give superior performance with a root mean square error of 0.31 eV in test. In addition, GBR based ML models is demonstrated to predict the carbon and oxygen binding energy of AA terminated A<sub>3</sub>B alloys with a test error of 0.34

eV and oxygen and carbon binding energies of AB terminated A<sub>3</sub>B alloys with a test error of 0.38 eV and 0.35 eV respectively. The binding energy of oxygen and carbon on Cu-based SAAs are also predicted with test error of 0.36 eV and 0.37 eV respectively. Moreover, the computational time for predicting the binding energy using ML is 0.0006 s on a dual-core laptop which is significantly less than the time required for DFT calculations. DFT and ML calculated carbon and oxygen binding energies for the bimetallic A<sub>3</sub>B alloys are further employed in ab-initio microkinetic model (MKM) to calculate turn over frequency (TOF) for ethanol decomposition and non-oxidative dehydrogenation reactions. The catalytic rates over the bimetallic alloys obtained using the ML calculated binding energies follow the same trends as obtained using the DFT energies, with the TOF values same or within an order of magnitude range. This shows that catalyst screening using binding energy as a descriptor can be performed using the ML model, bypassing time and resources consuming DFT calculations. This is likely to speed up the process of novel catalyst discovery.

**Keywords:** metal alloys, catalyst, machine learning, binding energy, microkinetic modelling, density functional theory (DFT)

## 1. INTRODUCTION

Binding energy of the adsorbates on a heterogeneous catalyst surface is the key parameter in determining the catalyst activity. From the time of Paul Sabatier, catalysis is thought of an interplay of surface-adsorbate interactions, in which the adsorbate is expected to bind ‘optimum’ with the catalyst surface, neither ‘weak’ nor ‘strong’, to achieve maximum turn-over of the catalytic cycle, resulting into higher reaction rates. While catalytic reactions on the surface are expected to generate several adsorbates, binding of the key adsorbates are thought to be rate controlling<sup>1–3</sup>. In essence, reaction rates are a function of descriptors linked to the binding of adsorbates. In recent times, success of ab-initio density functional theory (DFT) calculations in binding energy estimations have

propagated microkinetic models<sup>4–8</sup> in which binding energy of carbon, oxygen or nitrogen atoms on the catalyst surface are used as descriptors to determine the turnover frequencies (TOFs)<sup>1,3,4,9–15</sup>. For example, in syngas conversion to ethanol, the reaction rates on a series of transition metal catalysts have been represented as a function of the binding energy of the ‘descriptor’ atom (carbon, oxygen and/or nitrogen), in the form of a Sabatier plot<sup>15</sup>. This inherently is possible, since binding energy of all the reaction intermediates on the surface are linked with the carbon and oxygen binding energies on the surface following the linear scaling relationships<sup>4,5,9,10,16–19</sup>. From the Sabatier plots, one can rationally think of designing the catalyst surface wherein a metal showing weaker binding of the ‘descriptor’ atom may be alloyed with the metal having stronger binding to synthesize bimetallic alloys which may show desired binding of the ‘descriptor’ corresponding to the Sabatier maxima of reactivity.

The proposed thought on catalyst design, essentially is the recipe for in-silico high throughput screening of catalyst materials to come up with bimetallic alloys which are expected to produce desired reactivity in experiments, reducing the number of experiments required in a typical hit and trial approach to synthesize active catalysts<sup>4,6,11,19–25</sup>. However, in this recipe, time required for DFT simulations, in-general, is a limiting factor, since the quantum mechanical approach to scan the binding sites of an adsorbate on the catalyst surface and corresponding binding energy calculations are computationally expensive<sup>26–29</sup>. Herein, an alternative machine learning (ML) approach is presented which can provide the same information at a significantly reduced time scale.

Recent progress in integrating machine learning with data obtained from DFT calculations has opened up a possibility of exploring a whole new way of high throughput catalyst screening. Towards this, efforts to integrate ML and heterogeneous catalysis are commonly applying Artificial Neural Network (ANNs)<sup>30</sup>. However, the challenge with training conventional ANNs is the high computation time, which increases further with more number of hidden layers and neurons<sup>31</sup>. Another disadvantage of ANNs is their low

interpretability. With algorithmic developments, improvised ML algorithms can be developed that are sometimes more accurate and much faster than ANNs. One such algorithm is the Gradient Boosting Decision Trees (GBDT)<sup>32</sup> which incorporates important advantages of decision tree while using “Boosting” to overcome their biggest drawback - poor predictive performance. Decision trees are adaptable, easy to interpret, can handle different types of variables, need very less pre-processing of data and can fit nonlinear relationships accurately.<sup>33</sup> Boosting is a technique that is used to convert many weak learners to form a single strong learner.<sup>34</sup> The advantages provided by GBDTs have made it one of the most widely used ML algorithm.<sup>32</sup>

In the selection of an ML model, features are an important constituent.<sup>27</sup> Once the dataset of various catalysts is available, it has to be described by features that uniquely represent the catalyst and relate it to the target variable. There have been important contributions in this regard to predict the binding energies as a target variable in order to screen bimetallic catalysts. ANNs have been used to predict binding energies using electronic properties of alloys like d-band centre as features for the model.<sup>26,27</sup> Tree based ensemble algorithms have shown significantly accurate prediction of binding energies of CH<sub>4</sub> related species on Cu-based alloys using only readily available physical properties of metals as features.

35

In this study, an ML based model is developed to predict the binding energy of the ‘descriptor’ oxygen and carbon atoms on bimetallic alloys of the form A<sub>3</sub>B (211 surface). Various ML algorithms are evaluated to put forward the advantage of GBR over others for the supervised regression problem. Additionally, the GBR model is developed to predict the binding energies of oxygen and carbon over copper based single atom alloys (SAAs). The ML model developed using readily available properties of metals as features is observed to predict the binding energy with accuracy equivalent to that of DFT calculations. Also, the computation time required for these ML model prediction is negligible as compared to DFT calculations. The ML predicted binding energies were

further used with ab-initio microkinetic model (MKM) to calculate the catalytic rates for two important catalytic reactions; ethanol decomposition<sup>13</sup> and non-oxidative dehydrogenation (NODH) reactions<sup>15</sup> over the A<sub>3</sub>B bimetallic alloys. Both reactions were earlier studied by us in detail by constructing MKMs for understanding the trend in catalytic activity of transition metals for C-O bond scission in ethanol<sup>13</sup> and NODH of ethanol to produce acetaldehyde<sup>15</sup> on undercoordinated step (211) sites. Here in this study, the ML calculated binding energies showed similar predictions for the reactivity of bimetallic alloys as was earlier shown by ab-initio MKM. These findings can ultimately be extended to other metal alloys and catalytic reactions to provide a faster way of catalyst screening.

## 2. METHODS

The ML model is trained and tested on a dataset comprising of binding energies of oxygen and carbon over bimetallic alloys. For the A<sub>3</sub>B alloys, the binding energies for AA and AB terminated alloys are obtained from CatApp.<sup>20</sup> A model representation of (211) surface of AA terminated A<sub>3</sub>B alloy is shown in Figure 1(a) and (b). Corresponding images for AB terminated A<sub>3</sub>B alloy are shown in Figure 1(c) and (d). For Cu-based SAA, the binding energies are calculated using plane wave DFT code as implemented in Vienna ab initio simulation package (VASP-5.3.5version, University of Vienna).<sup>36</sup> The core electrons are described using Vanderbilt ultrasoft pseudopotentials.<sup>37</sup> Kohn Sham one electron valence states are expanded with the plane wave basis function and truncated at a cut-off energy of 396 eV. Revised Perdew-Burke-Ernzerhof (RPBE) exchange correlation functional developed by Hammer et al.<sup>38</sup> is used to describe the exchange and correlation contributions for single electron equation. Terrace (111) sites of Cu-based SAA are modelled using a slab of 4 layers of size 4 x 4. One Cu atom is replaced with transition metal; Sc, Ti, V, Fe, Co, Ni, Cu, Zn, Y, Zr, Nb, Mo, Ru, Rh, Pd, Ag, Hf, W, Re, Ir, Pt and Au or a p-block element; B, Al, Ga, Ge, In, Sn. The model geometry representing Cu-based SAA (111) terminated surface is shown in Figure 1(e) and (f). Bottom two layers of the model are fixed while the upper half along the adsorbate (C and O) is allowed to relax to represent the surface, sub-surface and bulk mimicking

the phenomenon of surface and subsurface restructuring. Final geometry of the optimized binding configurations for C and O over the Cu-based SAA is shown in the Supporting Information (SI), Figure SI-1 and SI-2. A Monkhorst-Pack k point grid with  $3 \times 3 \times 1$  mesh is used to sample the irreducible Brillouin zone.<sup>39</sup> Slabs are periodically repeated with 20 Å vacuum between the slabs to distinctly represent the gas phase adsorption to the surface. Convergence criteria for force and energy are set to 0.05 eV/Å and  $10^{-4}$  eV, respectively.

Each bimetallic alloy is represented by a set of features that uniquely describe it. For  $A_3B$  bimetallic alloys, a set of 27 features are chosen to include physical properties of both the metals in the alloy. These properties are readily available from the periodic table and other databases<sup>40–42</sup>. Overall, each alloy is depicted by a feature vector comprising of 27 values. For Cu-based SAA, each alloy is represented by a set of 12 features that include the physical properties of the single atom in the alloy. Features related to Cu are not included as they would be constant for all the SAAs used in the study.

Except ANN, all other ML algorithms are implemented using widely used open-source library Scikit-Learn.<sup>43</sup> ANNs are implemented using Keras<sup>44</sup> with a TensorFlow<sup>45</sup> backend. For evaluating the predictive power of the ML algorithms, the dataset is first split into two parts, train data and test data. All the ML models are built including all the features as input. The models are tested using 5-fold cross validation and by 100 times repeating the random splits of train and testing data (80%/20%) so as to avoid any data biasing. Accuracies of the predictions are calculated by averaging the root mean square errors (RMSE) of those 100 trials. Since the values of hyperparameters is expected to affect the accuracy of the model; a range of hyperparameters are tested for each model using GridSearch in Scikit-Learn. Models like Linear Regression, K-Nearest Regressor, Support Vector Regressor and Neural Networks needed feature scaling<sup>46</sup>. To implement this, features are standardized by removing the mean and scaling to unit variance for the algorithms that need feature scaling.

Final test and accuracy of the ML model in predicting the TOFs for applied catalytic reactions is directly evaluated using the MKM implemented with the descriptor based

analysis tool CatMAP<sup>6</sup>. In the CatMAP software package, a multi-dimensional Newton's root finding method from the python mpmath library is implemented to obtain steady-state solutions of the governing differential equations and the production rate is calculated as catalytic TOFs. The steady state kinetics in the ab-initio method is determined using the mean field approach by solving all the rate equation without making any assumption of the surface coverage or rate-determining step. The MKM is constructed using the reaction energetics obtained from DFT calculations and also with ML predicted binding energies. Binding energies of the adsorbed species, transition state and gas phase species used in the model are obtained from our previous MKM studies of ethanol over the transition metals<sup>13,15</sup>. The gas phase energies of hydrogen, water and methane is taken as a reference for expressing energies of all species. Similar methodology and reaction conditions are employed as in the previous DFT based MKM study of ethanol decomposition<sup>13</sup> and NODH<sup>15</sup>. Carbon and the oxygen binding energies are taken as the descriptors for both the reaction models. For ethanol decomposition reaction, the comparison with the ML based model is made at the reaction conditions;  $T = 523\text{ K}$  and  $P = 2\text{ bar}$  with 1:1 ratio of hydrogen and ethanol in the inlet stream<sup>13</sup>. For NODH, the reactions conditions considered are –  $T = 473\text{ K}$  with 10% conversion of ethanol and  $P = 1\text{ bar}$ <sup>15</sup>.

### 3. RESULTS AND DISCUSSION

#### 3.1 Machine Learning applied to Bimetallic Alloys

ML models are built over a dataset comprising of 151  $A_3B$  bimetallic alloys. Oxygen and carbon binding energy values for AA terminated  $A_3B$  alloys are shown in a matrix form in Figure 2(a) and Figure 2(b), respectively while the data for AB terminated  $A_3B$  alloys are shown in Figure 2(c) and Figure 2(d) respectively. Linear scaling relationships have been identified for transition metals for the binding energies for chemically related species.<sup>47,48</sup> However, bimetallic alloys and SAAs have shown to deviate from this linear relationship.<sup>11,49</sup> As a result, mixed trends of binding energies are also observed in our dataset. For example, the binding energy of oxygen on pure Pt is 1.3 eV. When Pt is

alloyed with an early transition metal, the binding energy values for AA terminated alloys are observed to be very high (-1.3 eV for V and -1.5 eV for Nb). Early transition metals like V and Nb are known to be oxophilic in nature, hence strong binding of oxygen is observed for bimetallic alloys having early transition metals. As we move from left to right in the periodic table, the binding energies are 0.8 eV and 0.7 eV for Pt alloys with Co and Rh (to the immediate left of Pt) and 1.5 eV and 1.9 eV for Pt alloys with Cu and Ag (to the immediate right of Pt). Thus, for transition metals, the binding energy decreases (become more positive in value) as the Pt is alloyed with element from left to right in the periodic table. However, no clear trends are observed when Pt is alloyed with a p-block element. Similar trend is also observed for Rh alloys. However, for noble metal alloys like Au, the binding energy for all AA terminated Au alloys are higher (less positive) than the binding energy of pure Au which is 2.7 eV. Alloying Au with metals of higher reactivity is expected to increase the oxygen affinity as seen here.

Thus, the prediction of binding energy is a complex non-linear problem. ML algorithms with their ability to learn complex non-linear interactions can therefore be used for predicting binding energies of these bimetallic alloys.

Supervised learning is a type of ML task where the algorithm learns an inferred function from already labelled data. This inferred function is then used to predict the target value from new data. The “No Free Lunch” theorem in ML states that there is no one model that works best for all problems<sup>50</sup>. Hence, it is always advisable to try multiple ML algorithms to identify which model works best for a particular problem. A number of widely used ML algorithms were evaluated which can be classified as linear models, distance based models, support vector machines, tree based ensemble algorithms and neural networks.

For the prediction of oxygen binding energies on AA terminated  $A_3B$  bimetallic alloys, all models were initially tested by giving all 27 features as input. The optimum hyperparameters are obtained by using grid search with 10-fold cross validation for each algorithm as given in Table 2. Mean training and testing error for each algorithm (for tuned hyperparameter values) is also shown in Table 2, along the minimum and



maximum error for 100 trials. In each of these 100 trials, the data is split randomly into train and test data. The model is built on the training data and training error is evaluated on the same training data while the test error is evaluated on the testing data. The RMSE errors in eV for all the algorithms tested are shown in Figure 3 and listed in Table 2.

Linear models tested include ordinary linear regression (OLR) and ridge regression. The OLR involves predicting the target variable as a linear function of the input features. It can be mathematically represented as,

$$y(x,w) = w_0 + w_1x_1 + \dots + w_nx_n$$

These are easy to model and form the basis of many sophisticated ML algorithms.<sup>51</sup> Since the model is a linear function of input features, it only looks for linear relationships between the features and target value. As discussed before, the prediction of binding energy for bimetallic alloy is a non-linear problem and hence a large test error of 0.55 eV and 0.53 eV is obtained for OLR and ridge regression respectively as given in Table 2.

The distance based model, k-nearest regressor is one of the simplest machine learning models. It is a non-parametric model where the principle is to predict a target by looking at the properties of its nearest neighbours in the training set.<sup>52</sup> Despite being simple and easy to interpret, these distance based models have been proven to be successful in various applications.<sup>53</sup> However, since the model computes distances every time a prediction needs to be made, it suffers from a poor run-time performance. Also, it is very sensitive to the erroneous data and irrelevant features.<sup>53</sup> We obtained a large test error of 0.54 eV (Figure 3, Table 2) for prediction of oxygen binding energies proving that distance based models do not work well for this problem.

Support Vector Regressor (SVR) is a ML algorithm that uses high dimensional feature space to predict functions using set of support vectors. Instead of minimising the training error during learning, it minimises the generalisation error.<sup>53</sup> It has been applied successfully to various problems like optical character recognition (OCR) and time series prediction.<sup>53</sup> The drawbacks of SVR include a high algorithmic complexity and extensive memory requirements.<sup>53</sup> The RMSE error observed for SVR is 0.34 eV (Table 2), which is much better as compared to linear as well as distance based models.

The ensemble based algorithms used include Extra Tree Regressor (ETR), Random Forest Regressor (RFR) and Gradient Boosting Regressor (GBR). The underlying goal of all the ensemble algorithms is to combine predictions from several weak estimators to construct a strong estimator. These algorithms differ in how they construct number of decision trees to eventually build an ensemble. RFR builds ensembles of decision trees where each tree is built on a random selection of examples from the training data. Additionally, RFR adds randomness while constructing these trees. Instead of choosing the most important feature while splitting a node, it chooses the most important feature from a random sample of features.<sup>54</sup> ETR adds an extra layer of randomness to the RFR.<sup>55</sup> The final prediction for a new input is made by averaging the predictions by each of the trees in the ensemble for both the algorithms. These decision tree based ensemble methods can capture linear as well as non-linear complex relationship<sup>54</sup> and thus RMSE error values observed for ETR and RFR are 0.32 eV and 0.35 eV respectively (Table 2). These tree based ML algorithms have shown to best in predicting binding energy of CH<sub>4</sub> related species (CH<sub>3</sub>, CH<sub>2</sub>, CH, C and H) over Cu based alloys.<sup>35</sup>

Artificial Neural Networks (ANNs) have been developed inspired by the biological neural networks in the brain.<sup>56</sup> It consists of multiple interconnected nodes that are loosely modelled on neurons. Due to their ability to fit non-linear and complex data, their robustness to noise and adaptive learning, they have proven to be predictive in solving various complex real world problems.<sup>57</sup> However, they have found criticisms because of their behaviour as a “black-box” – being hard to interpret and due to the requirement of high computational resources.<sup>58</sup> The observed RMSE error for ANN is 0.39 eV (Table 2) which is higher as compared to SVR and tree based ensemble algorithms.

It can be seen from the error values that the GBR performed better than any other model used in the study with a test error of 0.31 eV (Table 2) for predicting oxygen binding energy. A simple representation of GBR has been shown in Figure 4. The GBR is an ensemble algorithm where the decision trees are learned sequentially. Initially a weak learner (decision tree) is built and then the model is improved by adding to it another learner that is built on the error (also called as residual) of the last learner. In general, the

next learner tries to correct the errors of its predecessor.<sup>32</sup> The ensemble algorithms improve upon the biggest drawback of decision trees that is overfitting<sup>32,59</sup>. They produce models that are adaptable, easy to interpret and provide better prediction than many ML algorithms.<sup>33</sup> Since GBR was the best choice for predicting binding energy, further analysis has been done using GBR only.

The RMSE calculated over 100 trials for the GBR is comparable to the error in binding energy calculations via DFT ( $\sim 0.3$  eV)<sup>38,60,61</sup>. The training and testing errors are evaluated by increasing the number of trials from 100 to 200 and 300. In each trial, a random split of data into training and testing data is performed. The RMSE errors observed to be consistent averaged over 200 trials and 300 trials. This proves that the accuracy of model remains stable even when the number of trials is increased.

### **3.1.1 Analysis of the Computational Time**

Traditionally, all the important heterogeneous catalysts are developed using large scale screening experiments which requires a trial and error approach to identify optimum catalyst for a reaction.<sup>62</sup> For example, Alwin Mittasch and co-workers conducted over 6500 experiments and screened about 2500 catalysts<sup>63,64</sup> before arriving at the Fe-based catalyst very similar to what is industrially used for ammonia synthesis. Computationally, Nørskov and co-workers<sup>4,9,21,62,65,66</sup> have shown relevant descriptors that can be identified to describe the activity of a catalyst. There can be many variables that affect the kinetics of the reaction such as the energy of transition state or the intermediates but generally these variables are highly correlated. Thus, one or two independent descriptors can be identified in each case that can be used to predict the catalytic activity.<sup>4</sup> This descriptor based approach has helped the screening of catalyst for reactions such as methanation<sup>67</sup>, ammonia synthesis<sup>62,66</sup>, hydrogen evolution<sup>21</sup> and oxygen reduction<sup>68</sup>. Such descriptors in terms of binding energy of adsorbate over catalyst have been identified for a number of reactions.<sup>65</sup>

Experiments are generally performed to measure the binding energy using temperature programmed desorption (TPD) and adsorption isotherms. More advanced and accurate experiments are now devised to measure binding energies using single crystal surfaces, which are well-adapted in the works of Somorjai<sup>69</sup>, Ertl<sup>70,71</sup> Masel<sup>72</sup>, King<sup>73</sup> and Cambell<sup>74</sup>. However, all of them are time consuming and DFT calculations certainly helps in speeding up the process. In this regard, a significant advantage of using ML over DFT calculations and tedious experiments is obtained in the reduction of both computational time and resources for estimation of binding energies of species on catalyst surface. Average computational time taken for calculation of a 4x4 (111) surface having 64 metal atoms takes 2100 sec with 8 CPUs (2.5GHz/12-Core with 62 GB RAM), whereas time taken for calculation of systems with adsorbate is on average of 4000 sec. Time taken for calculation of gas-phase CH<sub>4</sub>, H<sub>2</sub>O and H<sub>2</sub> are in the range of ~100 sec. In summary, on an average the total computational time taken for single adsorption energy calculation is in the range of 6000 secs or 100 mins on 8 high performance CPUs. Meanwhile, the prediction of one adsorption energy value using GBR (after the model is built) is 0.0006 seconds on a dual-core laptop. Even when the complete computational time of the process is calculated, which includes the time for hyperparameter optimisation and then the time to calculate the test error for 100 random splits of test/train data, it takes about 480 seconds or 8 minutes on a dual-core laptop for the GBR model built for predicting oxygen binding on the dataset of A<sub>3</sub>B bimetallic alloys. Thus, even if we start to build a new ML model for a completely new dataset, ML models would save a great amount of time and resource.

### **3.1.2 Feature importance for GBR prediction**

Adsorbate interactions with the transition metal catalyst surface are suggested to occur via the molecular level interactions of the metal d-band and the molecular orbitals of the adsorbates<sup>4,5</sup>. Transition metal d-band itself contains features such as the d-band centre, skewness, kurtosis etc. which are important descriptors of adsorbate binding

energies. Most applicable surface science theories propagated by Nørskov and co-workers have simplified molecular level explanations of adsorbate-metal interaction with a purview to come up with simple descriptors such as the centre of the d-band of the transition metal surface.<sup>4,5</sup> However, such simplistic descriptions often ignore the richness of the surface-adsorbate interactions where individual physical properties of the surface metals such as group, period, atomic number, atomic mass, atomic radius, electronegativity, melting point, boiling point, density, heat of fusion, ionization energy and surface energy play an important role. The d-band theory has also shown disagreement with experimental results for a number of systems.<sup>75–77</sup> Nonetheless, artificial neural networks have been shown to be efficient in predicting the binding energy of CO and OH over bimetallic alloys using electronic d-band properties of bimetallic alloys as feature.<sup>26,27</sup> However, the use of these electronic properties, which in turn needs to be calculated via DFT, limits the usage of features for prediction.

Machine learning has also been used to predict the binding energy of CH<sub>4</sub> related species over Cu-based bimetallic alloys.<sup>35</sup> The features used in the study were physical properties of the other metal in the Cu-based alloy. The use of physical properties that are readily available in literature makes this model much more interpretable and universal. Moreover, it facilitates the rapid discovery of new alloys as features of every alloy are readily available. We build upon these features (which include group, period, atomic number, atomic mass, atomic radius, electronegativity, melting point, boiling point, density, heat of fusion, ionization energy and surface energy of the catalyst elements) and extend it to all bimetallic alloys of the type A<sub>3</sub>B, thus building a universal model which can be used to predict the binding energy of oxygen and carbon over any bimetallic alloy of the form A<sub>3</sub>B. Additionally, features like d-band center, Pauling electronegativity and work function have been used to describe the A element as used by Xin et al. in their work.<sup>26,27</sup> The relevance of such features is further consolidated by using them to predict

the binding energy of oxygen and carbon over single atom alloys with the example of Cu-based SAA's.

### 3.2 AA terminated A<sub>3</sub>B bimetallic alloy

For the purpose of predicting oxygen binding energy of AA terminated A<sub>3</sub>B bimetallic alloys, a total of 27 features were used to describe each bimetallic alloy. The correlation matrix for oxygen binding energy with the individual features of “A” metal and “B” metal of A<sub>3</sub>B bimetallic alloy are shown in Figure 5(a) and Figure 5(b). The “good” features are generally those that are not correlated to the other features and at the same time are highly correlated to the target variable.<sup>78</sup> The correlation plot shows that there may be a need for variable selection in order to obtain the best set of predictors. Each time a GBR model is trained, a feature importance matrix can be obtained that gives the relative importance of a feature with respect to other features for that model. The feature importance of a variable is measured based on two measures: the number of times it is used for splitting a node in the decision tree and the improvement in the model due to that split. This is averaged over all the trees in the model to calculate the final feature importance.<sup>58</sup> Surface energy has been shown to be a good descriptor of catalytic activity of 15 different alloys in a previous study.<sup>79</sup> Also, Takigawa et al. found surface energy as the most important feature in the ML prediction of binding energies of C and CH over Cu based alloys.<sup>35</sup> This is reiterated in our results where the surface energy of the B metal in the AA terminated A<sub>3</sub>B alloy is the most important feature. Figure 6(a) shows the feature importance for all features in the GBR model for the carbon binding energy averaged out over the 100 trials. We observe that the most important features belong to the dopant (B element) rather than element that forms the matrix (A element) in the bimetallic alloy. Along with the surface energy, other important features for prediction are the ionization energy, electronegativity, density and heat of fusion of the dopant.

In order to optimize the ratio of Test/Train Data split, additional analysis is performed to measure RMSE of the model for Test/Train split ratio of 15/85, 20/80, 25/75, 30/70, and

50/50. The errors obtained in above mentioned cases are tabulated in Table 3. The test error increases as the ratio of test/train data is increased. As this ratio is increased, the amount of data available for training the model decreases. This resulted in the reduction of accuracy of the model. Thus, it can be seen that the ML model improves with the availability of more training data. This also indicates that if we include more train data in our model, it should further decrease the RMSE error obtained. This increase in data could be achieved either by adding more number of alloys or including more relevant features for each alloy. The deviation of the predicted values from DFT calculated values for GBR for different ratios of test/train data is represented in Figure 7.

Another ML model was built to predict carbon binding energies for the AA terminated  $A_3B$  bimetallic alloys. Since we have established the relevance of GBR in predicting binding energy of oxygen, only a GBR model was fitted for this prediction. In the data, instead of the oxygen binding energy, the carbon binding energy of AA terminated  $A_3B$  alloy was the target variable and the rest of the input features remained the same. Again, the optimum hyperparameters were obtained using grid search with a 10-fold cross validation in Scikit-learn. The test error for optimized model averaged over 100 trials was found to be 0.34 eV. In each of these trials, the test and train data were chosen randomly. This proves that the GBR ML model can be effectively used to predict carbon binding energies for these bimetallic alloys too with accuracies equivalent to DFT calculations. The deviation of the predicted values from DFT calculated values for GBR for different ratios of test/train data can be seen in Figure SI-3 and the error obtained for different ratios of test/train splits are available in Table SI-1.

The features used were the same as used for oxygen binding energy calculation. The correlation matrix of features of “A” metal and “B” metal for  $A_3B$  bimetallic alloy with the carbon binding energy can be seen in Figure 5(a) and 5(b). The feature importance of this model was again calculated and is represented in Figure 6(b). It can be seen that the top features for predicting the carbon binding energy remain almost similar to those for

predicting oxygen binding energy. The surface energy of the dopant is still the most important feature followed by ionization energy and density. The fact that the most important features remain almost similar show that these physical features are highly correlated to the binding energy. Not only that, the ML model is able to identify this correlation and predict the binding energies for both carbon and oxygen over AA terminated  $A_3B$  bimetallic alloys.

### 3.3 AB terminated $A_3B$ bimetallic alloy

In order to demonstrate the ability of ML to predict binding energies of bimetallic alloys for different surface configurations, we built a GBR model to predict the binding energies of oxygen and carbon for AB terminating  $A_3B$  bimetallic alloys as well. The same dataset of 151  $A_3B$  bimetallic alloys is used, however the target value is the binding energies of oxygen and carbon calculated over AB terminated alloys. The same procedure to build the GBR model was used as described before and the RMSE test error for oxygen and carbon binding energy for AB terminated bimetallic alloy are 0.38 eV and 0.35 eV, respectively. The deviation of the predicted values from DFT calculated values for oxygen binding energy and carbon binding energy for different ratios of test/train data can be seen in Figure SI-4 and Figure SI-5 respectively, whereas the respective errors obtained for the splits are tabulated in Table SI-2 and Table SI-3. This again represents that with the increase in train data, the ML algorithm predicts better. The feature importance for oxygen and carbon binding energy in the GBR model are calculated and are represented in Figure 8(a) and Figure 8(b), respectively. However, in contrast with the feature importance obtained for the case of AA terminated alloys, the surface energies of the both the dopant (B element) and the matrix (A element) are the most important feature for both oxygen and carbon binding energy prediction. The other important features remain the same as for AA terminated surface and include the electronegativity, ionization energy, density and heat of fusion of the B element in the  $A_3B$  alloy.

### 3.4 Single atom alloys



We calculated the data of binding energy of oxygen and carbon over Cu-based SAAs via DFT in order to demonstrate the reliability of ML model prediction for SAA as well. The 27 metals used for forming the single point alloys have been highlighted in the periodic table, according to the oxygen and carbon binding energy as shown in Figure 9(a) and Figure 9(b), respectively. For oxygen binding energies, large negative values are observed for early transition metals (-1.71 eV for W). As we move from left to right in the periodic table, binding energy move towards the positive scale peaking at Pt (1.77 eV) and Pd (1.68 eV). A somewhat similar trend is also observed for carbon binding energy values although the binding energy values are all positive.

All the models are tested with 12 features as input which include the physical properties (group, period, atomic number, atomic mass, atomic radius, electronegativity, melting point, boiling point, density, heat of fusion, ionization energy and surface energy) of single atom in the alloy. The correlation matrix of features of “B” metal for Cu-based SAA for the carbon and oxygen binding energy can be seen in Figure 10. The features describing Cu are not included as they would remain constant for all the alloys used in the model. Again, a similar procedure is followed to get the optimized GBR models as mentioned for the case of  $A_3B$  bimetallic alloys. A set of hyper parameters are tested using 10-fold cross validation in order to obtain the best hyper parameters for the GBR model. The test error for the optimized models averaged over 100 trials for oxygen binding energies and the carbon binding energies are 0.36 eV and 0.37 eV respectively. This result again shows the effectiveness of GBR model for prediction of absorption energies. The deviation of the DFT calculated values with the predicted values for different test/train ratio for predicting oxygen and carbon binding energies can be seen in Figure SI-6 and Figure SI-7 respectively. Errors obtained in the above different test/train ratios are tabulated in Table SI-4 and SI-5 for oxygen and carbon binding energies over SAA respectively. This again illustrates that increasing the train data improves the prediction of GBR model.

The feature importance for the optimized models for the prediction of oxygen and carbon binding energies are shown in Figure 11(a) and Figure 11(b) respectively. The most

important feature for the prediction of carbon binding energy is still the surface energy of the element. The rest of the features have almost similar relative importance. This again shows the high correlation between surface energy and the binding energy and the adequacy of ML model to identify this correlation for prediction. However, for prediction of oxygen binding energies, both the group and surface energy of the single atom have similar importance. The importance of group in the prediction is also observed in the study by Takigawa et al.<sup>35</sup> where it is the most important feature for predicting the binding energies of H and CH<sub>2</sub> over Cu-based alloys.

### **3.5 In silico screening using ab initio microkinetic modelling**

Binding energy of alloys obtained from ML is further verified by microkinetic modelling. This was implemented using the CatMAP module<sup>6</sup>, wherein the reaction energetics are derived for DFT based alloy energies as well as those acquired from ML. The MKM is evaluated for two reactions for which previously alloy catalyst search has been undertaken. The two reactions are: ethanol decomposition<sup>13</sup> and NODH of ethanol<sup>15</sup>. Figure 12 depicts the turnover of the products of reaction for ethanol decomposition over DFT as well as ML based alloy binding energies. The reaction conditions are considered to be T=573 K, and P = 2 bar with an inlet stream ratio of 1:1 for ethanol and hydrogen. Ethane is formed upon C-O scission of ethanol whereas methane is the C-C scission product as shown in Figure 12(a) and (b) respectively. It can be inferred from Figure 12 that the activity trend over the alloys remains the same for DFT and ML based alloy energies. The activity trend for DFT based alloy energetics is Co<sub>3</sub>Ru ~ Ni<sub>3</sub>Fe ~ Co<sub>3</sub>Ni ~ Co<sub>3</sub>Fe > Ni<sub>3</sub>Cu > Ni<sub>3</sub>Pt > Ni<sub>3</sub>Rh<sup>13</sup>, whereas that based on ML alloy energetics is Co<sub>3</sub>Ru > Ni<sub>3</sub>Fe ~ Co<sub>3</sub>Ni ~ Co<sub>3</sub>Fe > Ni<sub>3</sub>Cu ~ Ni<sub>3</sub>Pt ~ Ni<sub>3</sub>Rh. The turnover of C-O scission product (ethane) over the alloys is given in Table 4. For Co<sub>3</sub>Ru, Ni<sub>3</sub>Rh, and Ni<sub>3</sub>Pt, the calculated TOF remains the same for both the DFT and ML energetics. However, for the rest of the alloys, the TOFs are under predicted for energies obtained from ML by an order of magnitude. An error bar

of 0.3 eV is considered for C and O binding energies on the metals and metal alloys as this is comparable to the error found in the ML model tested in the study.

Further, a similar comparison study is made for the NODH of ethanol at 473K and an initial conversion of 10%. Figure 13(a) and (b) show the volcano plots for the reaction products; acetaldehyde and ethylene over the alloys as well as transition metals for descriptor energies obtained from DFT<sup>15</sup> and ML respectively. The TOFs of acetaldehyde over the alloys are enlisted in Table 5. For Ni<sub>3</sub>Sn, Cu<sub>3</sub>Rh and Cu<sub>3</sub>Pd, the TOF remains the same for both DFT and ML, however for Cu<sub>3</sub>Ni and Cu<sub>3</sub>Pt, a difference of an order of magnitude is observed in the TOF between the two methods. Overall, it can be concluded that the obtained trend in binding energies on alloys are similar and the difference in individual binding energy values between ML and DFT predictions are not more than an order of magnitude. Therefore, increasing the input alloy data set of DFT calculations for ML models can thereby enhance the accuracy of ML.

## CONCLUSION

ML models are developed to predict the binding energy of oxygen and carbon over bimetallic and SAAs. Among all the ML models studied, GBR model is observed to be superior in giving significantly reduced errors. The model is enriched in incorporating several features of the catalyst surface, which include the periodic properties of the elements and many other electronic properties. The ML model, therefore, contains more descriptive information about adsorbate-surface interactions as compared to commonly applied simplified surface science theories based on the d-band centre of the transition metal catalyst. For example, in predicting the binding energies of carbon and oxygen atoms on all of the alloys studied; surface energy, electronegativity, ionization energy, density and heat of fusion of the alloying metals are considered to be the most significant features in exercising maximum influence on adsorption. However, as an exception, in case of single atom alloys, the periodic group of the single metal atom, alloyed on the surface of the lesser reactive metal, is observed to be the most important feature. These

features enabled molecular level insights is furthering the importance of the ML model as compared to computationally intensive DFT calculations or experiments required for the screening of catalyst materials. The applicability of the ML model in predicting the binding energies of metal alloys in an ab initio MKM is demonstrated for two different reactions of ethanol. In both the reactions, the ML model predicted the TOFs similar to what was earlier obtained from DFT calculations. With the increase in availability of alloy data from DFT calculations, ML models are expected to become more accurate. To sum up, the use of ML models will significantly reduce time-scale required for high throughput screening, with enriched and improvised understandings on metal adsorbate interactions.

#### **Acknowledgement:**

Financial support from Board of Research in Nuclear Sciences (BRNS) (37(2)/20/19/2018-BRNS/37149) to MAH is acknowledged. Computational resources of High Performance Computing Facility at IIT Delhi are acknowledged.

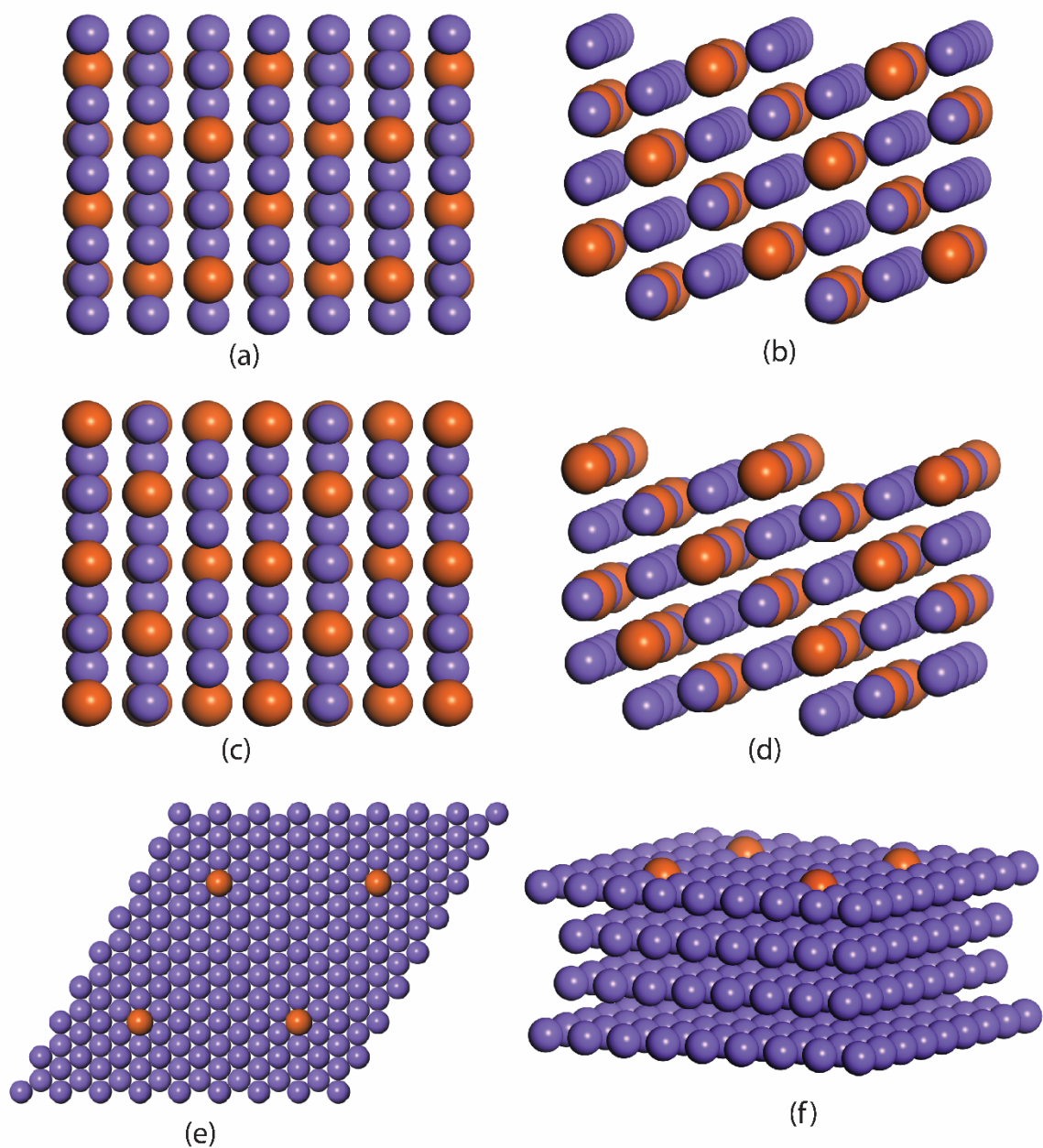
#### **REFERENCES**

- 1 C. T. Campbell, *ACS Catal.*, 2017, **7**, 2770–2779.
- 2 C. Stegelmann, A. Andreasen and C. T. Campbell, *J. Am. Chem. Soc.*, 2009, **131**, 8077–8082.
- 3 C. A. Wolcott, A. J. Medford, F. Studt and C. T. Campbell, *J. Catal.*, 2015, **330**, 197–207.
- 4 J. K. Nørskov, F. Abild-Pedersen, F. Studt and T. Bligaard, *Proc. Natl. Acad. Sci.*, 2011, **108**, 937–943.
- 5 J. K. Nørskov, F. Studt, F. Abild-Pedersen and T. Bligaard, *Fundamental concepts in heterogeneous catalysis*, John Wiley & Sons, 2014.
- 6 A. J. Medford, C. Shi, M. J. Hoffmann, A. C. Lausche, S. R. Fitzgibbon, T. Bligaard and J. K. Nørskov, *Catal. Letters*, 2015, **145**, 794–807.
- 7 L. C. Grabow, F. Studt, F. Abild-Pedersen, V. Petzold, J. Kleis, T. Bligaard and J. K. Nørskov, *Angew. Chemie - Int. Ed.*, 2011, **50**, 4601–4605.
- 8 F. Studt, F. Abild-Pedersen, Q. Wu, A. D. Jensen, B. Temel, J.-D. Grunwaldt and J. K. Nørskov, *J. Catal.*, 2012, **293**, 51–60.
- 9 A. J. Medford, A. Vojvodic, J. S. Hummelshøj, J. Voss, F. Abild-Pedersen, F. Studt, T. Bligaard, A. Nilsson and J. K. Nørskov, *J. Catal.*, 2015, **328**, 36–42.

- 10 H. Falsig, J. Shen, T. S. Khan, W. Guo, G. Jones, S. Dahl and T. Bligaard, *Top. Catal.*, 2014, **57**, 80–88.
- 11 Y. Xu, A. C. Lausche, S. Wang, T. S. Khan, F. Abild-pedersen and F. Studt, *New J. Phys.*, 2013, **15**, 125021.
- 12 A. C. Lausche, A. J. Medford, T. Suvra, Y. Xu, T. Bligaard, F. Abild-Pedersen, J. K. Nørskov and F. Studt, *J. Catal.*, 2013, **307**, 275–282.
- 13 F. Jalid, T. S. Khan, F. Qayoom and M. A. Haider, *J. Catal.*, 2017, **353**, 265–273.
- 14 T. S. Khan, S. Hussain, U. Anjum and M. A. Haider, *Electrochim. Acta*, 2018, **281**, 654–664.
- 15 T. S. Khan, F. Jalid and M. A. Haider, *Top. Catal.*, 2018.
- 16 F. Abild-Pedersen, *Catal. Today*, 2016, **272**, 6–13.
- 17 S. Wang, V. Petzold, V. Tripkovic, J. Kleis, J. G. Howalt, E. Skúlason, E. M. Fernández, B. Hvolbæk, G. Jones, A. Toftelund, H. Falsig, M. Björketun, F. Studt, F. Abild-Pedersen, J. Rossmeisl, J. K. Nørskov and T. Bligaard, *Phys. Chem. Chem. Phys.*, 2011, **13**, 20760.
- 18 S. Wang, V. Vorotnikov, J. E. Sutton and D. G. Vlachos, 2014.
- 19 J. Greeley, *Annu. Rev. Chem. Biomol. Eng.*, 2016, **7**, 605–635.
- 20 J. S. Hummelshøj, F. Abild-Pedersen, F. Studt, T. Bligaard and J. K. Nørskov, *Angew. Chemie - Int. Ed.*, 2012, **51**, 272–274.
- 21 J. K. Nørskov, T. Bligaard, J. Rossmeisl and C. H. Christensen, *Nat. Chem.*, 2009, **1**, 37.
- 22 M. Neurock, *J. Catal.*, 2003, **216**, 73–88.
- 23 M. K. Sabbe, M.-F. Reyniers and K. Reuter, *Catal. Sci. Technol.*, 2012, **2**, 2010–2024.
- 24 S. Curtarolo, G. L. W. Hart, M. B. Nardelli, N. Mingo, S. Sanvito and O. Levy, *Nat. Mater.*, 2013, **12**, 191.
- 25 J. Greeley and N. M. Markovic, *Energy Environ. Sci.*, 2012, **5**, 9246–9256.
- 26 Z. Li, S. Wang, W. S. Chin, L. E. Achenie and H. Xin, *J. Mater. Chem. A*, 2017, **5**, 24131–24138.
- 27 Z. Li, X. Ma and H. Xin, *Catal. Today*, 2017, **280**, 232–238.
- 28 J. E. Sutton and D. G. Vlachos, *Chem. Eng. Sci.*, 2015, **121**, 190–199.
- 29 A. Jain and T. Bligaard, *Phys. Rev. B*, 2018, **98**, 214112.
- 30 B. R. Goldsmith, J. Esterhuizen, J. Liu, C. J. Bartel and C. Sutton, *AIChE J.*, 2018, **64**, 2311–2323.
- 31 J. Schmidhuber, *Neural Networks*, 2015, **61**, 85–117.
- 32 J. H. Friedman, *Comput. Stat. Data Anal.*, 2002, **38**, 367–378.
- 33 J. Ye, J.-H. Chow, J. Chen and Z. Zheng, in *Proceedings of the 18th ACM Conference on Information and Knowledge Management*, ACM, New York, NY, USA, 2009, pp. 2061–2064.
- 34 Y. Freund and R. E. Schapire, *J. Japanese Soc. Artif. Intell.*, 1999, **14**, 771–780.
- 35 T. Toyao, K. Suzuki, S. Kikuchi, S. Takakusagi, K. Shimizu and I. Takigawa, *J. Phys. Chem. C*, 2018, **122**, 8315–8326.
- 36 G. Kresse and J. Furthmüller, *Comput. Mater. Sci.*, 1996, **6**, 15–50.

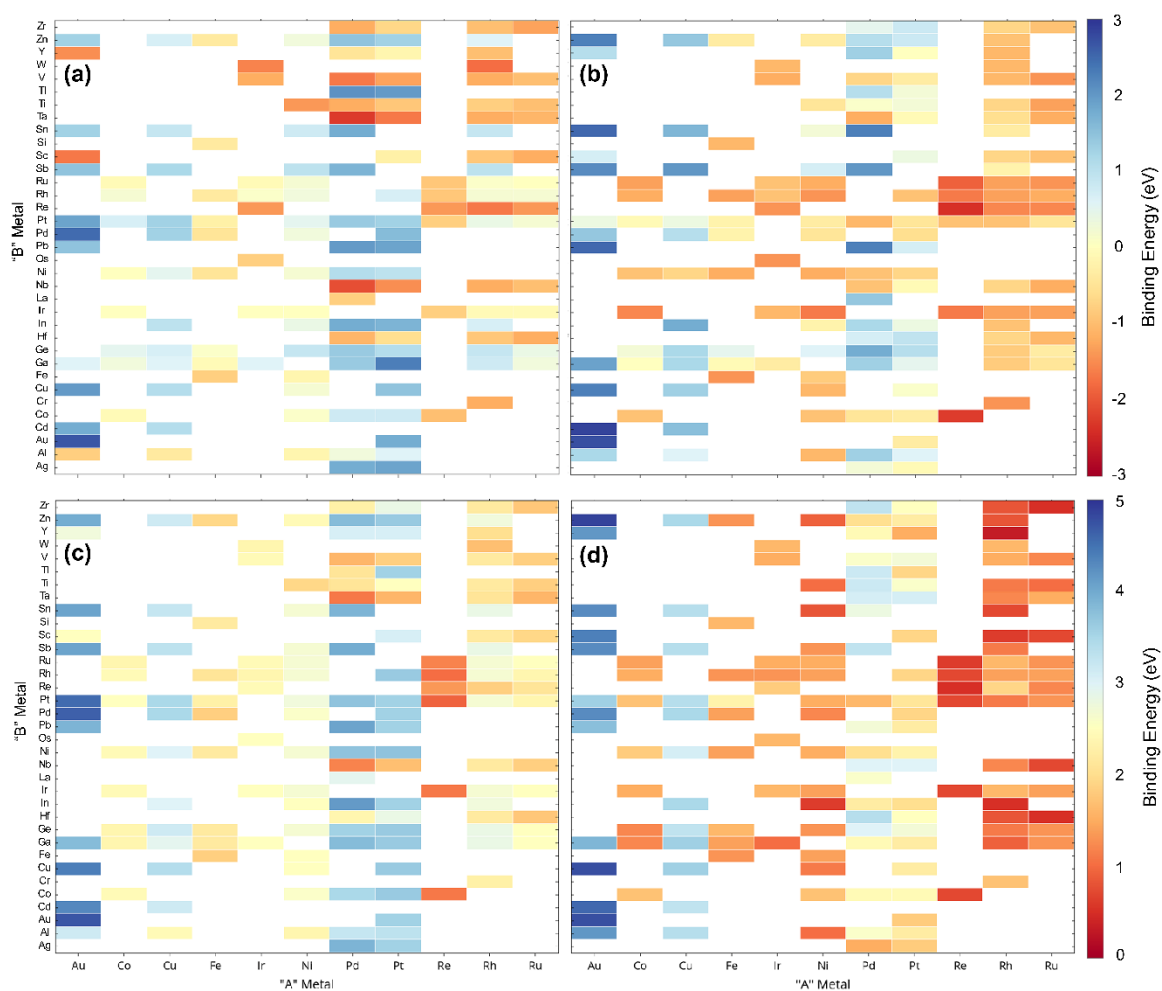
- 37 D. Vanderbilt, *Phys. Rev. B*, 1990, **41**, 7892–7895.
- 38 B. Hammer, L. B. Hansen and J. K. Nørskov, *Phys. Rev. B*, 1999, **59**, 7413–7421.
- 39 H. J. Monkhorst and J. D. Pack, *Phys. Rev. B*, 1976, **13**, 5188–5192.
- 40 R. Tran, Z. Xu, B. Radhakrishnan, D. Winston, W. Sun, K. A. Persson and S. P. Ong, *Sci. Data*, 2016, **3**, 160080.
- 41 D. R. Lide, *CRC Handbook of Chemistry and Physics*, CRC Press: Boca Raton, FL, 84th edn., 2003.
- 42 B. Hammer and J. K. Nørskov, in *Advances in Catalysis*, Academic Press, 2000, vol. 45, pp. 71–129.
- 43 F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot and É. Duchesnay, *J. Mach. Learn. Res.*, 2011, **12**, 2825–2830.
- 44 F. Chollet, <https://github.com/fchollet/keras>.
- 45 M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, M. Kudlur, J. Levenberg, R. Monga, S. Moore, D. G. Murray, B. Steiner, P. Tucker, V. Vasudevan, P. Warden, M. Wicke, Y. Yu and X. Zheng, in *Proceedings of the 12th USENIX Conference on Operating Systems Design and Implementation*, USENIX Association, Berkeley, CA, USA, 2016, pp. 265–283.
- 46 Q. V Le, M. Ranzato, R. Monga, M. Devin, K. Chen, G. S. Corrado, J. Dean and A. Y. Ng, in *Proceedings of the 29th International Conference on International Conference on Machine Learning*, Omnipress, USA, 2012, pp. 507–514.
- 47 M. M. Montemore and J. W. Medlin, *Catal. Sci. Technol.*, 2014, **4**, 3748–3761.
- 48 L. C. Grabow, *Computational Catalyst Screening*, The Royal Society of Chemistry, 1st edn., 2014.
- 49 M. T. Darby, M. Stamatakis, A. Michaelides and E. C. H. Sykes, *J. Phys. Chem. Lett.*, 2018, **9**, 5636–5646.
- 50 D. H. Wolpert, in *Soft Computing and Industry: Recent Applications*, eds. R. Roy, M. Köppen, S. Ovaska, T. Furuhashi and F. Hoffmann, Springer London, London, 2002, pp. 25–42.
- 51 N. M. Nasrabadi, *J. Electron. Imaging*, 2007, **16**.
- 52 N. S. Altman, *Am. Stat.*, 1992, **46**, 175–185.
- 53 P. Cunningham, M. Cord and S. J. Delany, eds. M. Cord and P. Cunningham, Springer Berlin Heidelberg, Berlin, Heidelberg, 2008, pp. 21–49.
- 54 L. Breiman, *Mach. Learn.*, 2001, **45**, 5–32.
- 55 P. Geurts, D. Ernst and L. Wehenkel, *Mach. Learn.*, 2006, **63**, 3–42.
- 56 B. Yegnanarayana, *Artificial neural networks*, PHI Learning Pvt. Ltd., 2009.
- 57 I. A. Basheer and M. Hajmeer, *J. Microbiol. Methods*, 2000, **43**, 3–31.
- 58 J. V Tu, *J. Clin. Epidemiol.*, 1996, **49**, 1225–1231.
- 59 B. J. H. Friedman, *Ann. Stat.*, 2001, **29**, 1189–1232.
- 60 A. J. Medford, A. C. Lausche, F. Abild-Pedersen, B. Temel, N. C. Schjødt, J. K. Nørskov and F. Studt, *Top. Catal.*, 2014, **57**, 135–142.

- 61 A. J. Medford, J. Wellendorff, A. Vojvodic, F. Studt, F. Abild-Pedersen, K. W. Jacobsen, T. Bligaard and J. K. Nørskov, *Science (80-. )*, 2014, **345**, 197–200.
- 62 C. J. H. Jacobsen, S. Dahl, B. S. Clausen, S. Bahn, A. Logadottir and J. K. Nørskov, *J. Am. Chem. Soc.*, 2001, **123**, 8404–8405.
- 63 S. R. Tennison, *Catalytic Ammonia Synthesis*, Springer US, 1991.
- 64 A. Mittasch and W. Frankenburg, *Adv. Catal.*, 1950, **2**, 81–104.
- 65 T. Bligaard, J. K. Nørskov, S. Dahl, J. Matthiesen, C. H. Christensen and J. Sehested, *J. Catal.*, 2004, **224**, 206–217.
- 66 J. K. Honkala, A. Hellman, I. Remediakis, A. Logadottir, A. Carlsson, S. Dahl, C. H. Christensen and J. K. Nørskov, *Science (80-. )*, 2005, **307**, 555–558.
- 67 A. L. Kustov, A. M. Frey, K. E. Larsen, T. Johannessen, J. K. Nørskov and C. H. Christensen, *Appl. Catal. A Gen.*, 2007, **320**, 98–104.
- 68 J. Greeley, I. E. L. Stephens, A. S. Bondarenko, T. P. Johansson, H. A. Hansen, T. F. Jaramillo, J. Rossmeisl, I. Chorkendorff and J. K. Nørskov, *Nat. Chem.*, 2009, **1**, 552.
- 69 G. A. Somorjai, *Introduction to surface chemistry and catalysis*, John Wiley & Sons, 2010.
- 70 C. T. Campbell, G. Ertl, H. Kuipers and J. Segner, *J. Chem. Phys.*, 1980, **73**, 5862–5873.
- 71 G. Ertl, *Catal. Rev.*, 1980, **21**, 201–223.
- 72 D. A. Kyser and R. I. Masel, *Rev. Sci. Instrum.*, 1987, **58**, 2141–2144.
- 73 C. E. Borroni-Bird, N. Al-Sarraf, S. Andersson and D. A. King, *Chem. Phys. Lett.*, 1991, **183**, 516–520.
- 74 O. Lytken, W. Lew and C. T. Campbell, *Chem. Soc. Rev.*, 2008, **37**, 2172–2179.
- 75 M. T. Gorzkowski and A. Lewera, *J. Phys. Chem. C*, 2015, **119**, 18389–18395.
- 76 H. Thirumalai and J. R. Kitchin, *Top. Catal.*, 2018, **61**, 462–474.
- 77 M. P. Hyman, B. T. Loveless and J. W. Medlin, *Surf. Sci.*, 2007, **601**, 5382–5393.
- 78 M. A. Hall and L. A. Smith, 1998, Volume 20 No 1, 181–191.
- 79 H. Zhuang, A. J. Tkalych and E. A. Carter, *J. Phys. Chem. C*, 2016, **120**, 23698–23706.

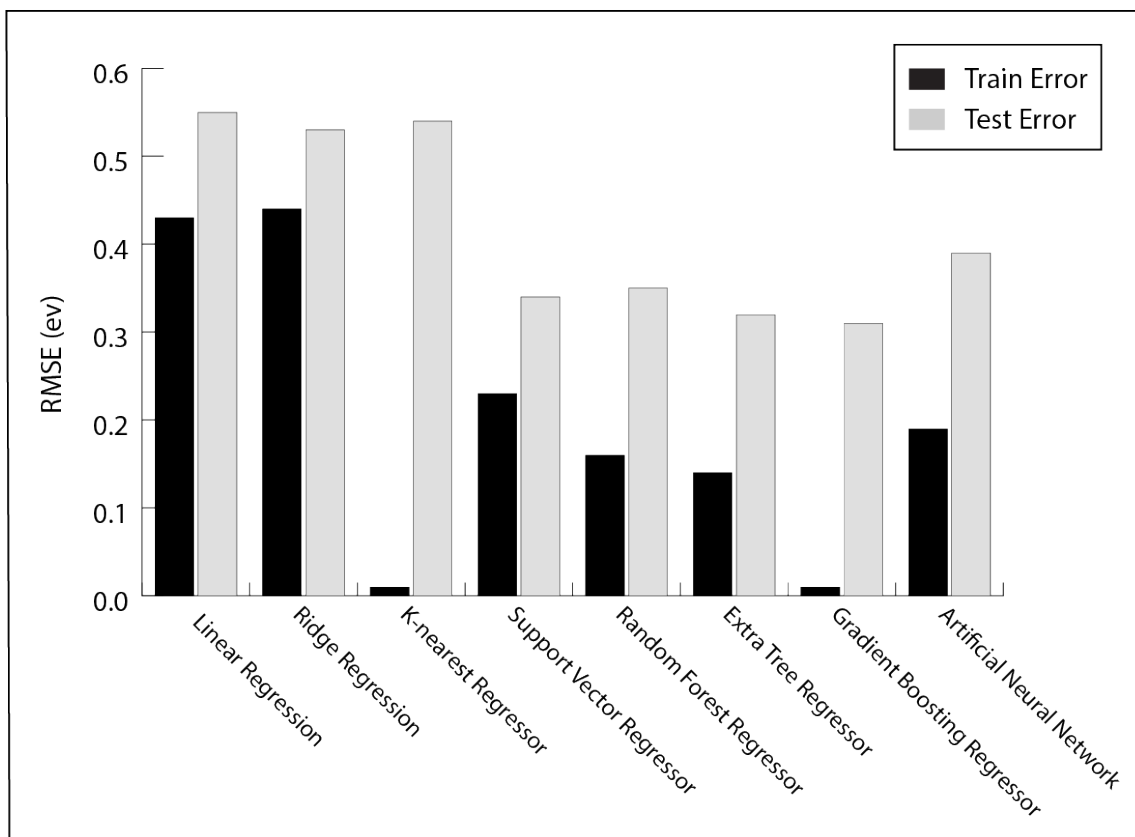


**Figure 1.** (a) Top view and (b) side view of the (211) surface of  $A_3B$  bimetallic alloy with AA termination; (c) top view and (d) side view of the (211) surface of  $A_3B$  bimetallic alloy with AB termination; (e) top view and (f) side view of (111) surface of Single Atom Alloy (SAA) AB, where B atom is embedded in the matrix of A metal. Blue color atoms represent the A transition metal and the orange color atom represents the B transition metal/non-metal from p-block.

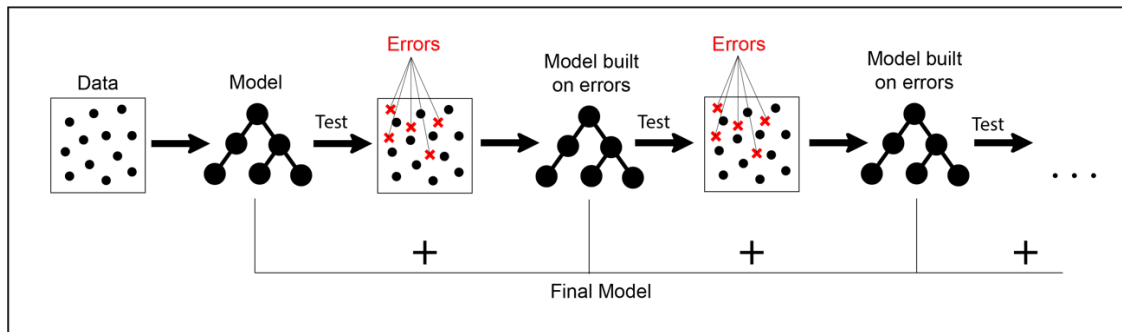




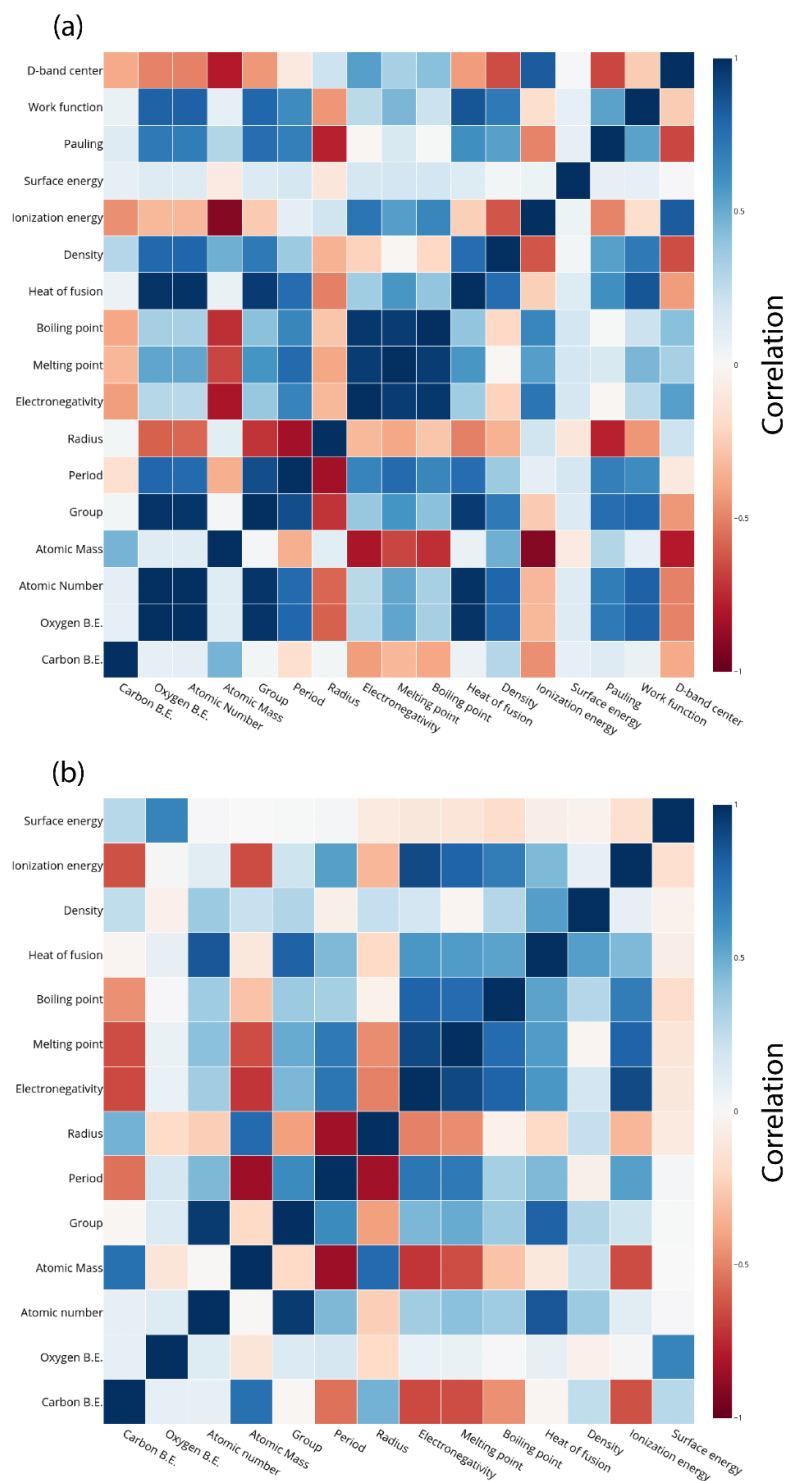
**Figure 2.** The dataset used for ML model for (a) oxygen (b) carbon for AA terminated and (c) oxygen (d) carbon for AB terminated  $A_3B$  bimetallic alloys. Highlighted cells represent the “A” metal and “B” metal in the bimetallic alloy used in the dataset and the filled color represents the binding energy of oxygen/carbon taken from CatApp database.



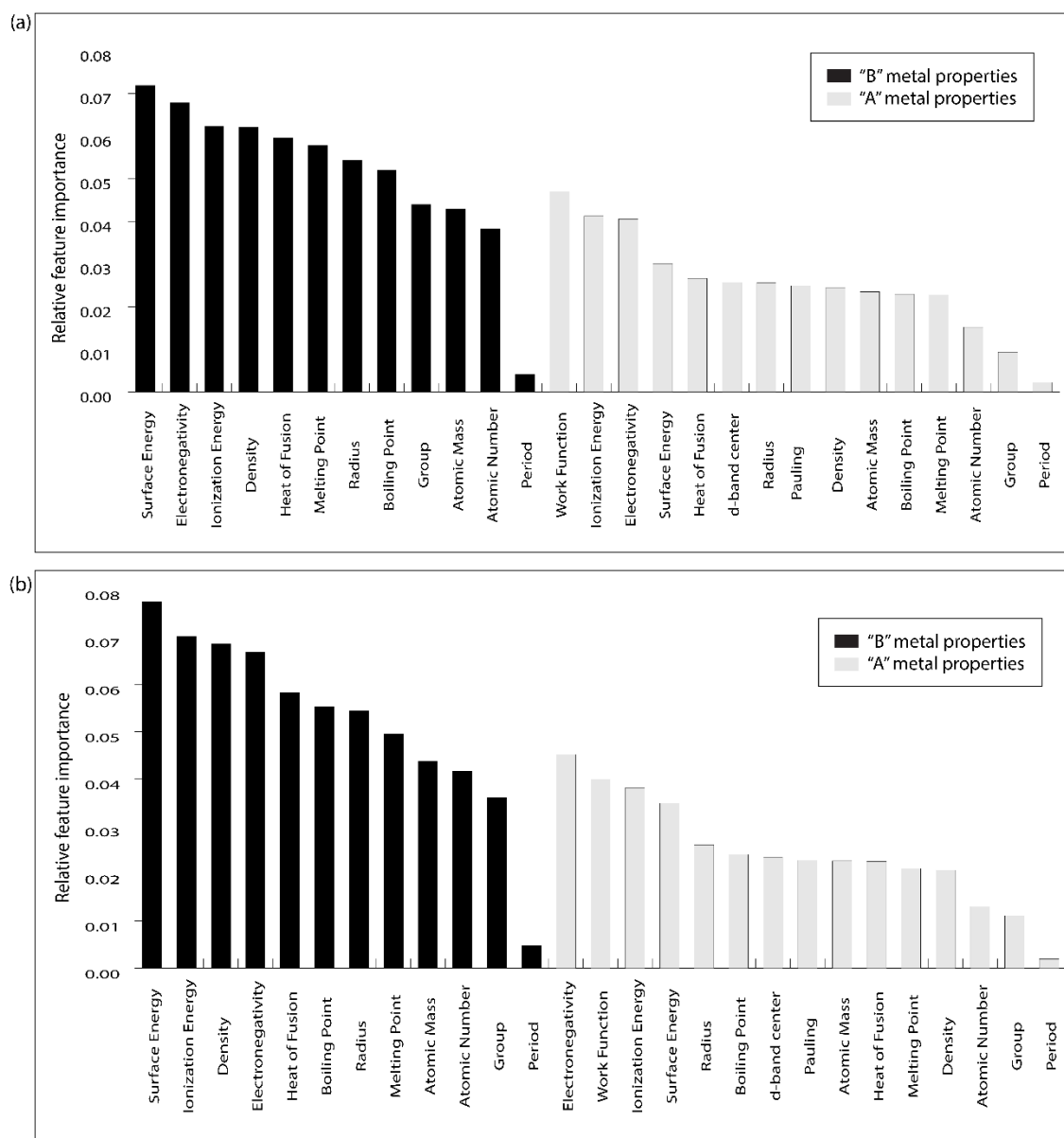
**Figure 3.** The train error and test error for the evaluated ML algorithms for predicting oxygen binding energies on  $A_3B$  bimetallic alloys



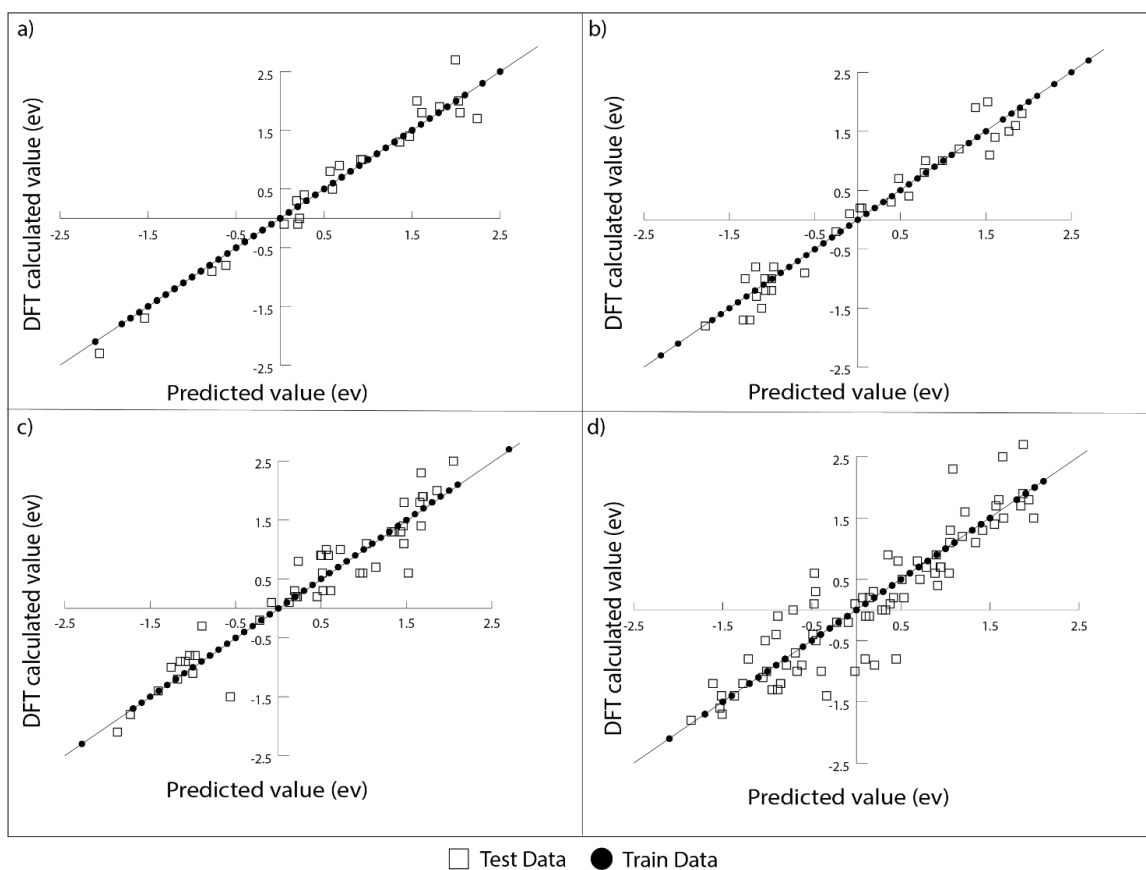
**Figure 4.** A simple representation of gradient boosting decision tree algorithm.



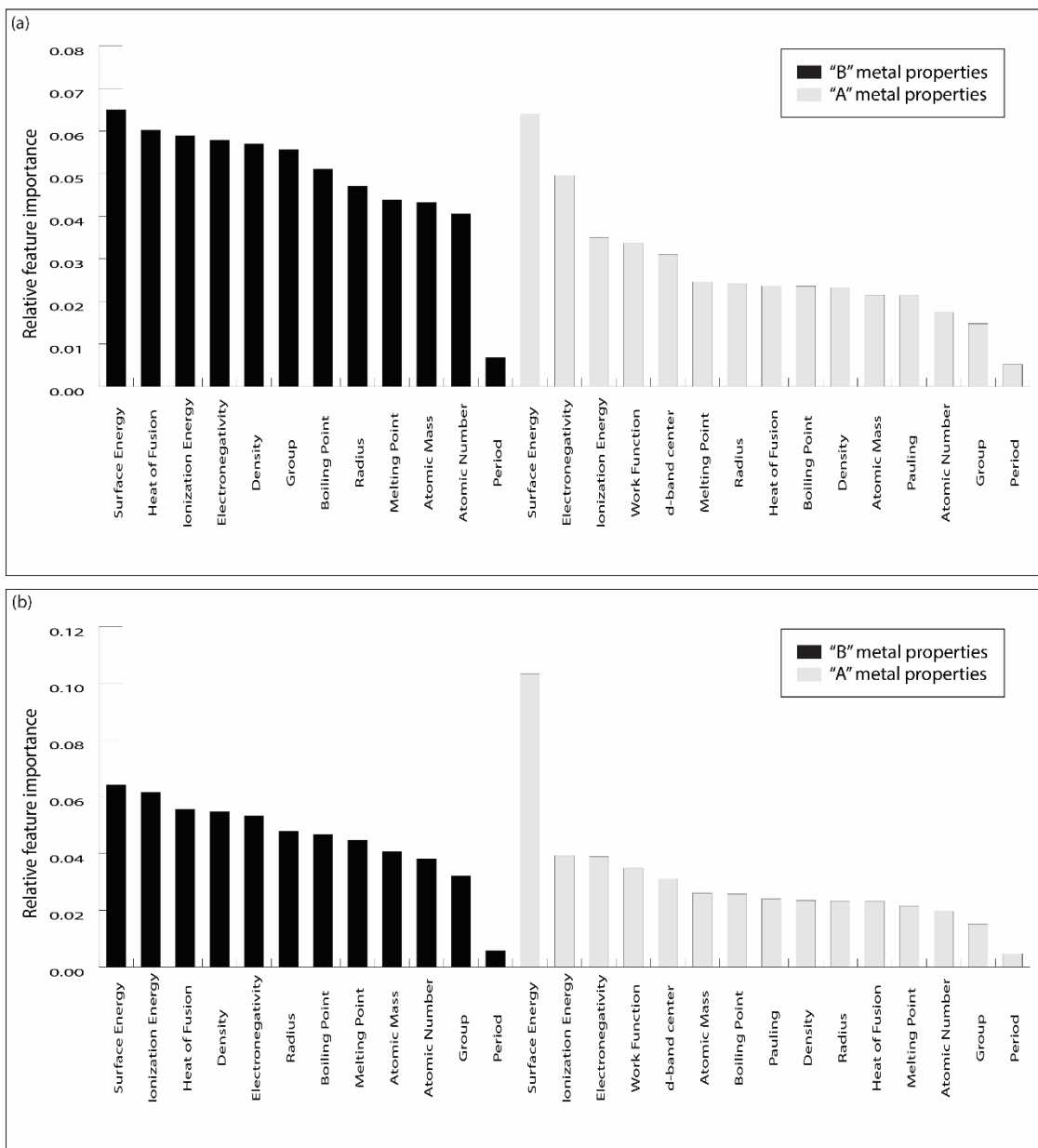
**Figure 5.** Correlation plot for oxygen and carbon binding energy with the features of the (a) “A” metal and (b) “B” metal in the  $A_3B$  bimetallic alloy AA terminated (211) surface.



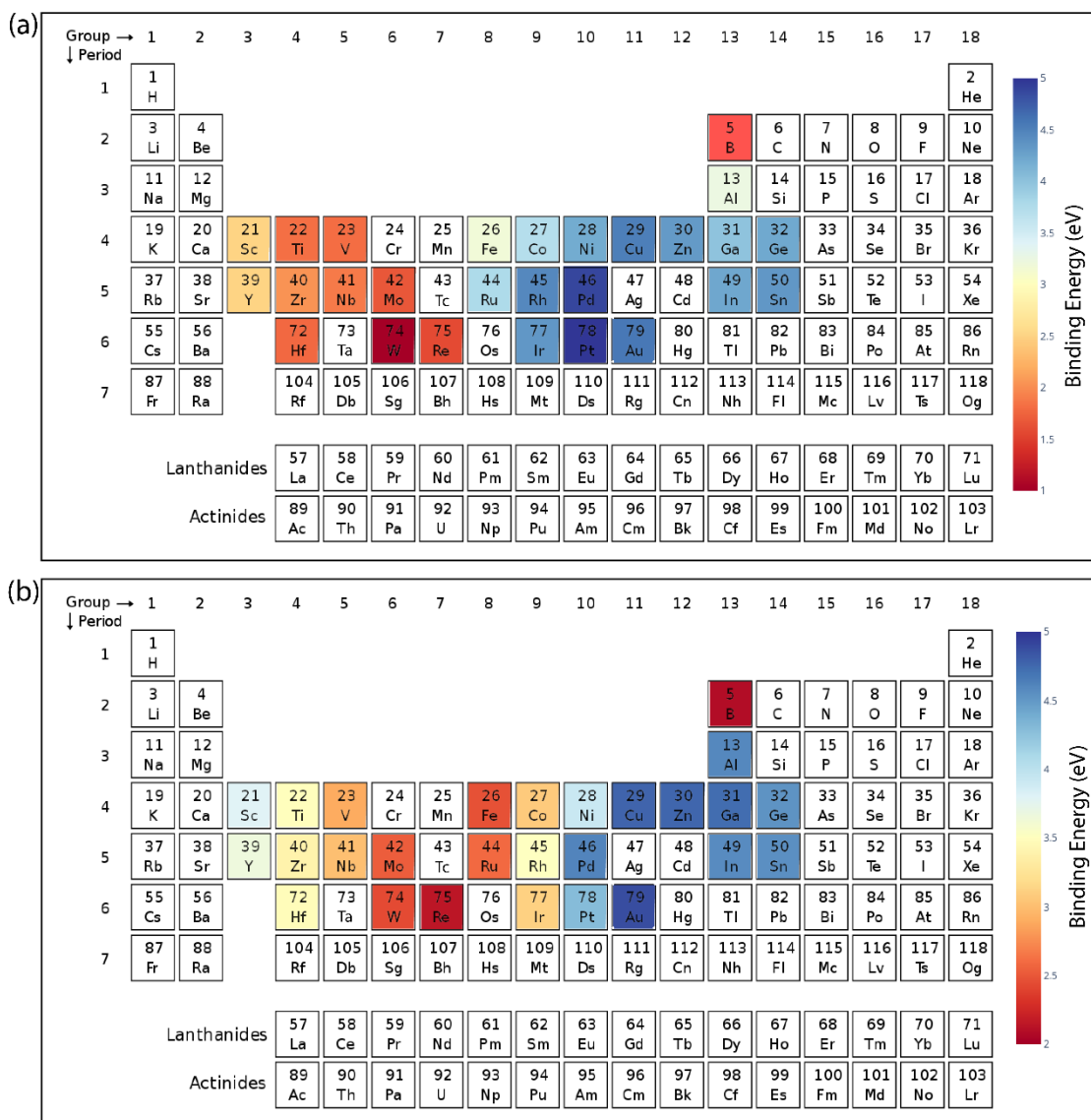
**Figure 6.** Relative feature importance for GBR model for predicting (a) oxygen and (b) carbon binding energies for AA terminated A<sub>3</sub>B bimetallic alloy (211) surface.



**Figure 7.** The deviation of DFT calculated oxygen binding energy with that predicted from the GBR model for AA terminated  $A_3B$  bimetallic alloy for a) test/train ratio of 15/85 b) test/train ratio of 20/80 c) test/train ratio of 30/70 and d) test/train ratio of 50/50

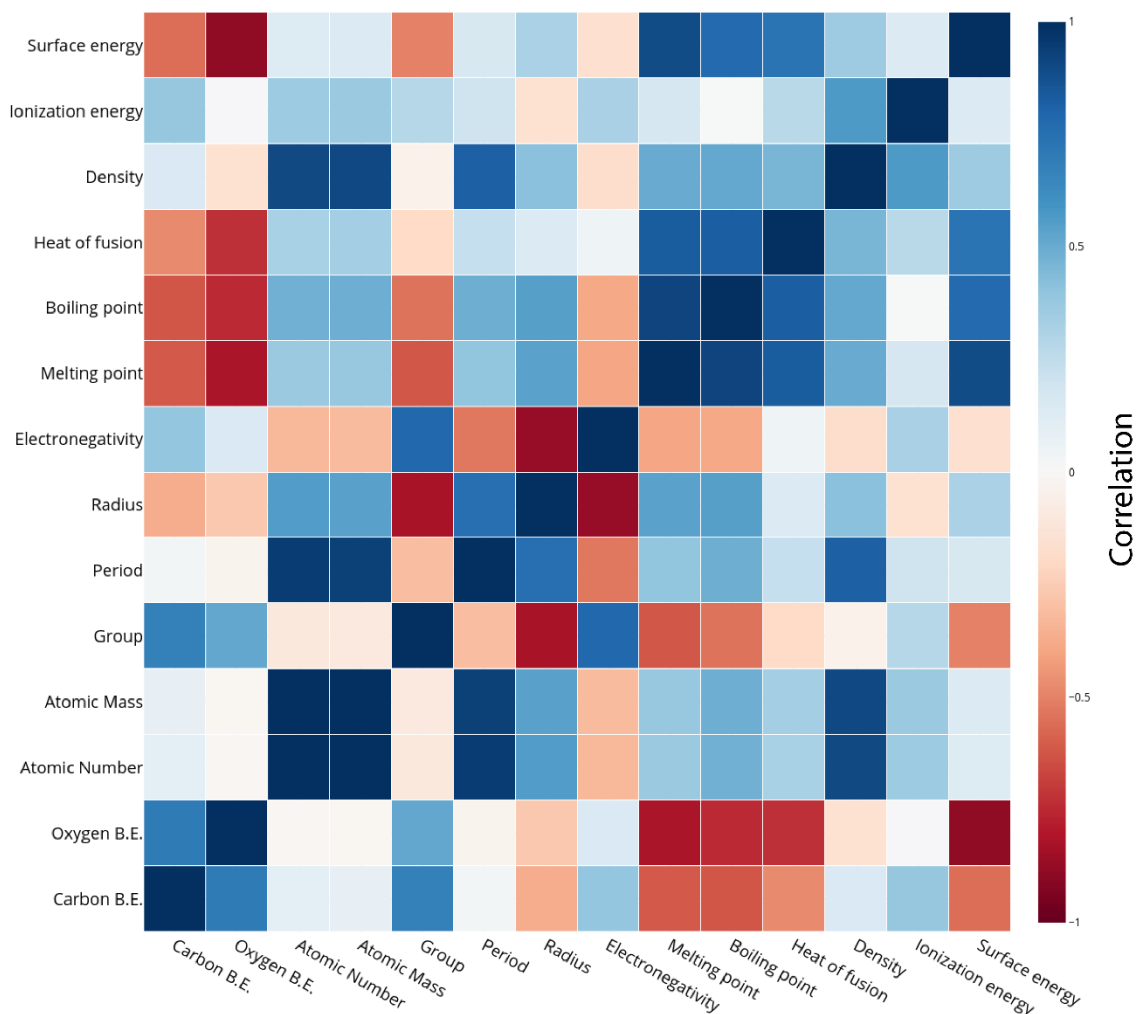


**Figure 8.** Relative feature importance for GBR model for predicting (a) oxygen and (b) carbon binding energies for AB terminated  $A_3B$  bimetallic alloy (211) surface.

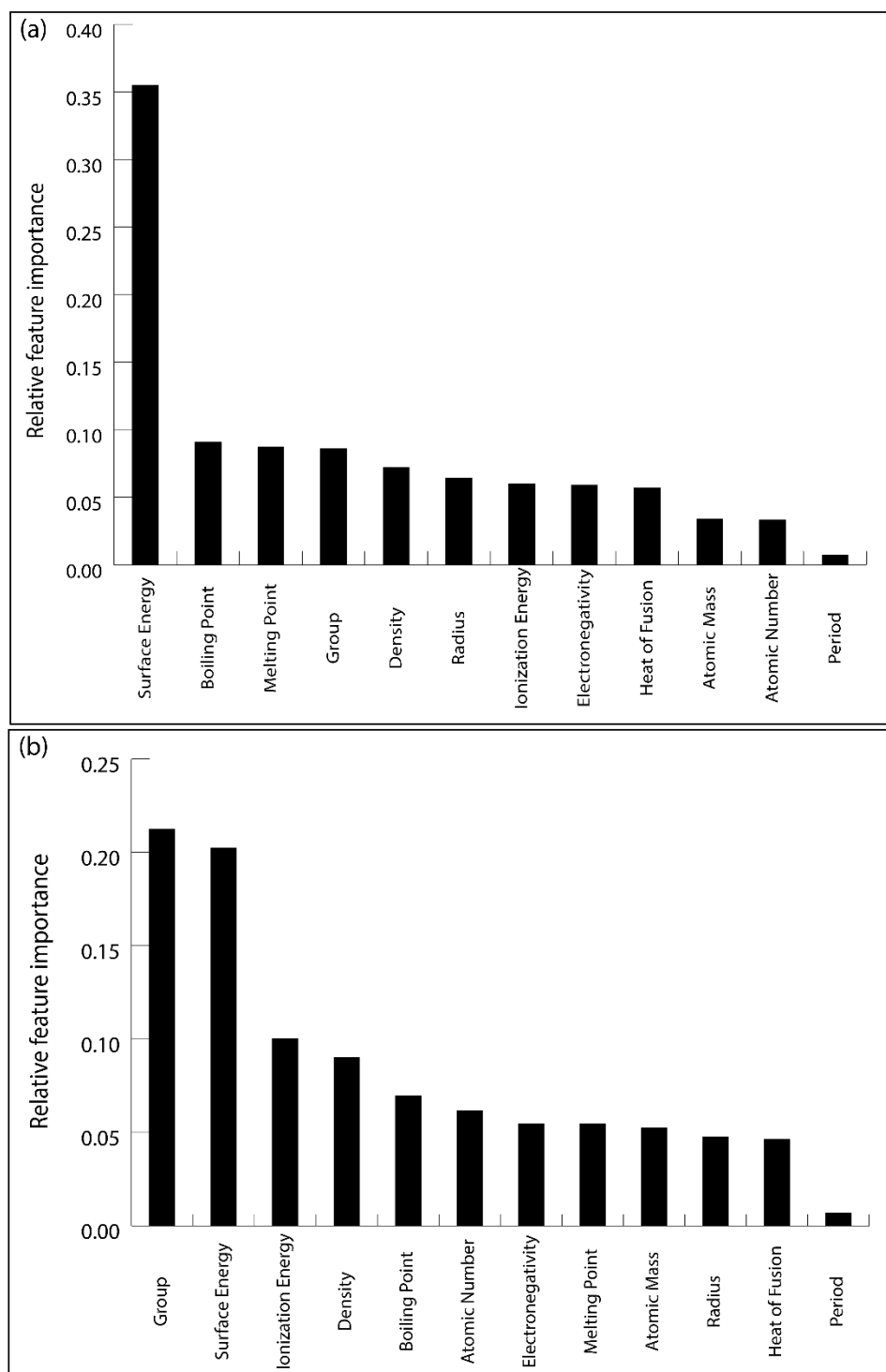


**Figure 9.** Elements used as single atoms in Cu-based SAA in the dataset. The highlighted colour represents the binding energy of (a) oxygen and (b) carbon on respective SAA as calculated from DFT calculations.

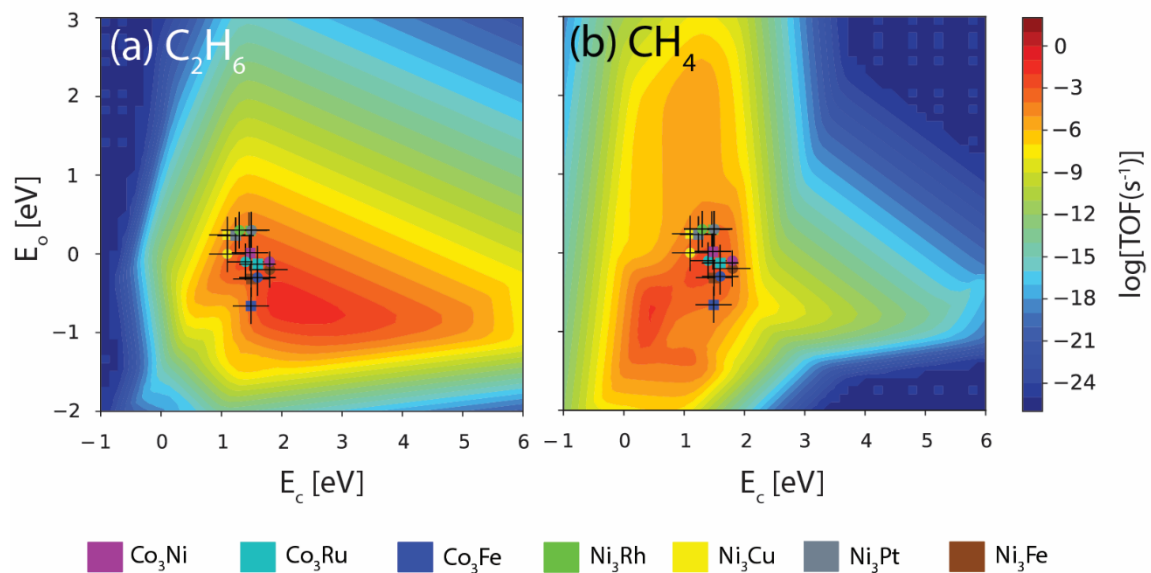
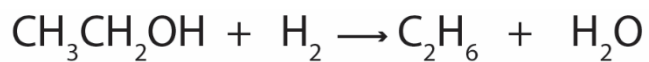




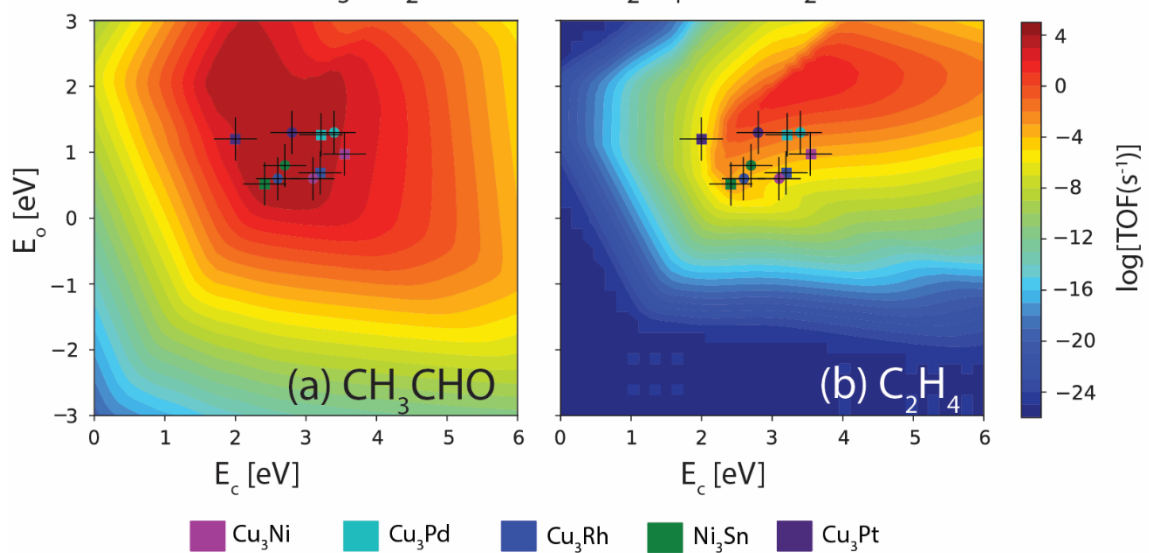
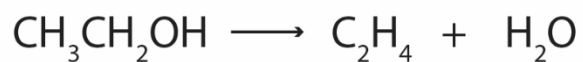
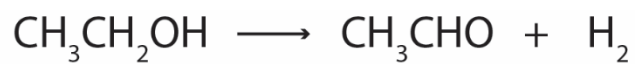
**Figure 10.** Correlation plot for oxygen and carbon binding energy with the features of the single atom metal in Cu-based SAA.



**Figure 11.** Relative feature importance for GBR model for predicting (a) carbon and (b) oxygen binding energies for Cu-based SAA.



**Figure 12.** Comparison of TOFs of ethanol decomposition to produce (a) ethane and (b) methane on Co and Ni based bimetallic alloys as obtained from DFT (circle) and ML (square). Error bar = 0.3 eV.



**Figure 13.** Comparison of TOFs for the NODH of ethanol to produce (a) acetaldehyde and (b) ethylene on bimetallic alloys as obtained from DFT (circle) and ML (square). Error bar = 0.3 eV.

**Table 1.** The range of hyperparameters tested for each ML model. (Hyperparameters not mentioned were kept at their default values as per Scikit-Learn/Keras documentations).

ML Algorithm	Hyperparameters tested
Linear Regression	non-parametric
Ridge Regression	alpha= [0.1,0.5,0.8,1,10,100]
K-nearest Regressor	n_neighbors=[5,10,20] weights=['uniform','distance']
Support Vector Regressor	kernel=['rbf'] C=[1,10,100,1000,10000,100000] Gamma=[0.1,0.01,0.001,0.0001,0.00001]
Random Forest Regressor	n_estimators=[50,100,200,300,400,500,600,700,800] max_depth=[2,3,4,5,6,7,8]
Extra Tree Regressor	n_estimators=[50,100,200,300,400,500,600,700,800] max_depth=[2,3,4,5,6,7,8]
Gradient Boosting Regressor	n_estimators=[50,100,200,300,400,500,600,700,800] max_depth=[2,3,4,5,6,7,8] learning_rate=[0.1,0.3,0.4,0.5,0.6,0.7,0.8]
Artificial Neural Network	Number of layers= 3 Neurons in each hidden layer= 50 Activation=[tanh,ReLu] Loss=[mean_squared_error,mean_absolute_error] Optimizer=[sgd,RMSprop]

**Table 2** Optimum hyperparameters evaluated for each ML algorithm along with training and test Error for Predicting oxygen binding energies for A<sub>3</sub>B bimetallic alloy

ML Algorithm	Tuned Hyperparameter Value	Train Error	Test Error
		mean (min, max)	mean (min, max)
Linear Regression	non-parametric	0.43 (0.37, 0.46)	0.55 (0.43, 0.75)
Ridge Regression	alpha= [0.5]	0.44 (0.39, 0.48)	0.53 (0.36, 0.69)
K - n e a r e s t Regressor	n_neighbors=5 weights='distance'	0 (0, 0)	0.54 (0.35, 0.77)
Support Vector Regressor	kernel=['rbf'] c=[1000] gamma=[0.001]	0.23 (0.17, 0.26)	0.34 (0.24, 0.53)
Random Forest Regressor	n_estimators=[400] max_depth=[6]	0.16 (0.13, 0.18)	0.35 (0.22, 0.52)
E x t r a     T r e e Regressor	n_estimators=[100] max_depth=[6]	0.14 (0.10, 0.17)	0.32 (0.18, 0.47)
Gradient Boosting Regressor	n_estimators=[400] max_depth=[3] learning_rate=[0.3]	0.003 (0.003, 0.003)	0.31 (0.2, 0.44)
Artificial Neural Network	Activation=[Relu] Loss=[mean_squared_error] Optimizer=[sgd]	0.19 (0.14,0.23)	0.39 (0.25,0.54)

**Table 3.** Effect on training and testing errors with change of test/train data ratio for GBR model for predicting binding energy of oxygen on A<sub>3</sub>B bimetallic alloy (211) AA terminated surface.

Test/Train Split	Train Error	Test Error
15%/85%	0.0003	0.31
20%/80%	0.0003	0.31
25%/75%	0.0003	0.33
30%/70%	0.0003	0.35
50%/50%	0.0003	0.4

**Table 4.** TOF of ethane production from ethanol decomposition over Ni and Co based bimetallic alloys obtained using DFT and ML calculated binding energies

Alloy	TOF [DFT]	TOF [ML]
Ni <sub>3</sub> Fe	10 <sup>-3</sup> s <sup>-1</sup>	10 <sup>-4</sup> s <sup>-1</sup>
Ni <sub>3</sub> Rh	10 <sup>-5</sup> s <sup>-1</sup>	10 <sup>-5</sup> s <sup>-1</sup>
Ni <sub>3</sub> Cu	10 <sup>-4</sup> s <sup>-1</sup>	10 <sup>-5</sup> s <sup>-1</sup>
Ni <sub>3</sub> Pt	10 <sup>-5</sup> s <sup>-1</sup>	10 <sup>-5</sup> s <sup>-1</sup>
Co <sub>3</sub> Fe	10 <sup>-3</sup> s <sup>-1</sup>	10 <sup>-4</sup> s <sup>-1</sup>
Co <sub>3</sub> Ni	10 <sup>-3</sup> s <sup>-1</sup>	10 <sup>-4</sup> s <sup>-1</sup>

$\text{Co}_3\text{Ru}$	$10^{-3}\text{s}^{-1}$	$10^{-3}\text{s}^{-1}$
------------------------	------------------------	------------------------

**Table 5.** TOFs of acetaldehyde production from NODH of ethanol over the bimetallic alloys obtained using DFT and ML calculated binding energies

Alloy	TOF [DFT]	TOF [ML]
$\text{Ni}_3\text{Sn}$	$10^4\text{s}^{-1}$	$10^4\text{s}^{-1}$
$\text{Cu}_3\text{Rh}$	$10^4\text{s}^{-1}$	$10^4\text{s}^{-1}$
$\text{Cu}_3\text{Ni}$	$10^4\text{s}^{-1}$	$10^3\text{s}^{-1}$
$\text{Cu}_3\text{Pt}$	$10^4\text{s}^{-1}$	$10^3\text{s}^{-1}$
$\text{Cu}_3\text{Pd}$	$10^4\text{s}^{-1}$	$10^4\text{s}^{-1}$