ARTICLE

Machine-Learning Adsorption on Binary Alloy Surfaces for Catalyst Screening

Tai-ran Wang[a], Jian-cong Li[b], Wu Shu[b], Su-lei Hu[b], Run-hai Ouyang[c], Wei-Xue Li[a,b,d*]

a. Department of Chemical Physics, School of Chemistry and Materials Science, University of Science and Technology of China, Hefei 230026, China
b. Hefei National Laboratory for Physical Sciences at the Microscale, University of Science and Technology of China, Hefei 230026, China
c. Materials Genome Institute, Shanghai University, Shanghai 200444, China
d. Dalian National Laboratory for Clean Energy, Dalian Institute of Chemical Physics, Chinese Academy of Sciences, Dalian 116023, China

(Dated: Received on April 14, 2020; Accepted on June 1, 2020)

Over the last few years, machine learning is gradually becoming an essential approach for the investigation of heterogeneous catalysis. As one of the important catalysts, binary alloys have attracted extensive attention for the screening of bifunctional catalysts. Here we present a holistic framework for machine learning approach to rapidly predict adsorption energies on the surfaces of metals and binary alloys. We evaluate different machine-learning methods to understand their applicability to the problem and combine a tree-ensemble method with a compressed-sensing method to construct decision trees for about 60,000 adsorption data. Compared to linear scaling relations, our approach enables to make more accurate predictions lowering predictive root-mean-square error by a factor of two and more general to predict adsorption energies of various adsorbates on thousands of binary alloys surfaces, thus paving the way for the discovery of novel bimetallic catalysts.

Key words: Machine learning, Heterogenous catalysis, Adsorption energy, Bimetallic catalyst

I. INTRODUCTION

Heterogeneous catalysts composed of metallic materials have an enormous impact on a vast array of technologically important applications [1–3]. Theoretical approaches to the rational design of metallic catalysts have been based on the Sabatier principle and the Brønsted-Evans-Polanyi (BEP) relation [4–6]. The former states that the activity of catalysts displays a volcano shape as a function of adsorption strength for the key intermediates of the reaction, and the latter reveals the linear relationship between energy barriers of elementary reactions and adsorption energies of molecules. Hence, the reliable calculations or predictions of adsorption energies hold the key to understanding the catalytic mechanism and screening catalysts with improved performance.

Density functional theory (DFT) has proven to be a promising strategy for the calculations of adsorption energies and the atomistic design of catalysts [7–9]. However, the total numbers of the combinations between adsorbates and sites on different surfaces grow exponentially with the diversity of catalytic materials, rendering the computational screening of new catalysts with quantum-chemical calculations costly and time-consuming. This has become one of the major challenges in catalyst discovery. Binary alloys, for instance, span a vast set of materials and have shown attractive promise for catalyzing many reactions and the potential to substitute the noble metal catalysts [10–12]. However, the high-throughput predictions of adsorption energy on bimetallic surfaces are challenging due to the intricate composition and structure. To this end, the development of the adsorption models based on statistical learning is highly required for a rapid survey of appropriate adsorption energies for reactions of interest.

Theoretical models for chemisorption on pure metal surfaces have been developed [13, 14]. When molecules were adsorbed on transition metals, the most important contribution of adsorbate-metal interactions to the adsorption strength comes from the coupling of d-states of metal with adsorbates [14, 15]. In consequence, the linear relationships were established between the adsorption energies of atoms and that of their hydrogenated species [16, 17]. Further linear dependencies were also identified between other adsorbed species and could be summarized as group additivity properties [17, 18]. In

*Author to whom correspondence should be addressed. E-mail: wxli70@ustc.edu.cn

DOI:10.1063/1674-0068/cjcp2004049

spite of the huge success of linear relationships in accelerating metal catalyst discovery [11, 12], the generalization of these simplified thermochemical models to bimetallic materials is unpractical, which will be shown in this study. In recent years, machine learning (ML) methods have emerged as a powerful approach for predicting catalytic properties in heterogeneous systems [19–24]. ML algorithm learns from existing data to find insights and map the correlation between the varieties of properties with desired prediction targets [25, 26]. ML methods have been performed in many studies to establish the predictive models for adsorption energy, most starting from predicting certain adsorbates on pure metal surfaces or some bimetallic alloys, and applying these predictions to the entire reaction network by extending to other intermediates through linear relations [24, 26–30].

In this study, we instead directly exploit the adsorption energies for various adsorbed species of interest, *e.g.* C, H, O, N, S, CH_x*, OH_x*, NH*, SH*, *etc.*, at their most stable sites on surfaces of a wide range of binary alloys using ML methods, without any assumptions of linearity. Due to different advantages and characteristics of ML algorithms, the choice of the appropriate method depends on its applicability to the problem domains and is crucial for the learning results. The performance of different ML methods is evaluated by atomic adsorption energies. We show that random forest regression achieves the best predictive performance and it was further combined with compressed-sensing method (sure independence screening and sparsifying operator) to learn the whole dataset that is composed of approximately 60000 adsorption energies for 48 species adsorbed on the surfaces of over two thousand metals and binary alloys. This approach can be used to rapidly predict adsorption energies with high accuracy. The test root-mean-square error (RMSE) for the entire dataset is 0.29 eV, which is far less than scaling relations and comparable with previous ML models, though more types of adsorbates and materials are involved in this work. Thus, our exhaustive high-throughput approach allows the fast predictions for adsorption energies of intermediates in the reaction network, which goes beyond the traditional strategies, and facilitates the discovery of novel heterogeneous catalytic materials.

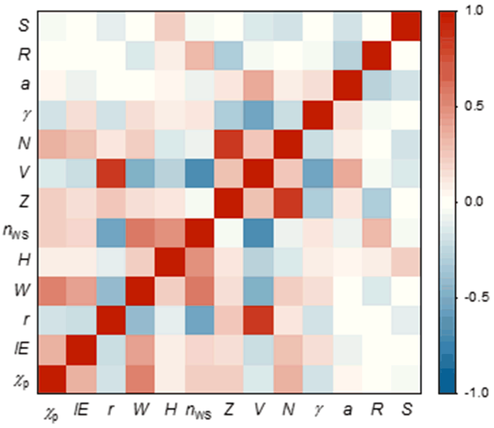

FIG. 1 The correlation map of Pearson pairwise product moment among thirteen features. If the absolute value of Pearson correlation is larger than 0.5, the correlation between the two features is obviously strong. On the contrary, the correlation is weak or even negligible.

II. METHODS

A. Dataset for machine learning

As for the dataset, we adopted a recently published database of the chemical adsorption, publicly available in Catalysis-Hub.org [31, 32]. This large-scale dataset contains more than 90000 systematic DFT calculations for adsorption energy of 48 adsorbates, which are composed of C, H, O, N and S elements, on 2035 surfaces (666 A_3B stoichiometries, 666 AB_3 stoichiometries, 666 AB stoichiometries and 37 metals) enumerated from 37 metals and transition metals [31]. The DFT calculations were performed considering all possible adsorption sites on (111) termination of surfaces. There are multiple adsorption energy values on the same surface, thus the factors of the complex surface structure were introduced. To address this problem, we used a python script to select the most stable adsorption sites (the most negative adsorption energies) in the stage of data preprocessing, remaining 59876 pieces of adsorption data for ML investigation. For each adsorbate, we randomly selected 80% of the data as the working set for model training and the validation process, and the rest 20% data were split into the test set to examine whether models generated from the working set can successfully predict these values. The initial features used in ML processing consist of the properties of metals and alloys, including Pauling electronegativity χ_p, ionization energy IE, metallic radius r, work function W, sublimation energy H, electron density at the boundary of the Wigner-Seitz cell n_ws, atomic volume V, atomic numbers Z, valence electrons numbers N, surface energy γ, lattice parameters a, ratio of two metals in binary alloys R and numerical labels of the most stable adsorption sites S. The initial features were selected based on domain knowledge or the suggestion in previous studies [19, 21, 24, 27–31, 33]. The values of these features are readily available from databases to enable rapid screening [34–36]. The correlations among the features were calculated by the Pearson product moment, and are given in FIG. 1. The weaker the correlation among the features, the more information features can be provided in ML. Noticeably, most features are not correlated with each other, while atomic volume was strongly correlated with metallic radius and electron density at the boundary of the Wigner-Seitz cell, as well as the correlated pair of valence electrons numbers and atomic numbers. We wiped off the features with the absolute value of

DOI:10.1063/1674-0068/cjcp2004049

Pearson correlation lager than 0.5 when training the ML algorithms, except in the method of sure independence screening and sparsifying operator (SISSO), where the dimensionality of high-dimensional feature space was reduced through sure independence screening (SIS) approach and those initial features would be combined to high dimensional features.

B. Machine learning methods

Four types of machine-learning regression algorithms including six methods were applied on the dataset to select the proper method for adsorption energy. (i) Compressed-sensing methods used here included least absolute shrinkage and selection operator (LASSO) [37] and SISSO [38, 39] for identifying the explicit descriptor of the adsorption strength. Compressed-sensing is a signal processing approach for effectively reproducing a high-quality signal by finding solutions to underdetermined systems. The LASSO approach is a linear model based on l_1-norm regularized minimization, efficiently decreasing the feature numbers upon which the given solution is dependent. Combining the dimensionality reduction approach SIS with sparsifying operator method, *e.g.* LASSO and l_0-norm, the method SISSO can tackle ultra-high-dimensional as well as strongly correlated feature spaces and give descriptors that yield the best prediction for the target property. SISSO could also be used in feature construction and selection. (ii) Kernel ridge regression (KRR) [40] and support vector regression (SVR) [41] were selected as kernel regressors. The form of the KRR model is identical to the SVR, but they use different loss function. (iii) In tree-ensemble methods, random forest regression (RFR) [42] was also selected. RFR constructs a multitude of decision trees in training process and outputs the mean prediction of the individual trees. By applying the bagging algorithm and random selection of features, RFR corrects for the habit of overfitting of decision trees. (iv) As one of the most popular approach in ML, neural network (NN) [43] algorithm was also applied. The network contains input, hidden, and output layers. In each layer, a set of neurons are used as processing units and connected by a series of weight parameters, which will be optimized in terms of back-propagation algorithm in the training procedure. ML methods were implemented using Scikit-learn package [44] except for SISSO method, which is not integrated into the platform of Scikit-learn.

To avoid overfitting and improve the robustness, we used the approach of exhaustive grid search combined with the repeated (ten times) ten-fold cross validation on the training set to systematically optimize the hyperparameters in each ML approach. We adjusted the structures of these methods by searching the best set of hyperparameters. The optimal model with hyperparameters that yields the lowest validation error was further used to predict the adsorption energy values in the test set. The hyperparameters in different methods optimized by cross validation include: LASSO (“alpha”, “max_iter”, “tol”, “fit_intercept”), SISSO (“desc_dim”, “maxcomplexity”, “subs_sis”, “dimclass”, “opset”), KRR (“alpha”, “kernel”, “gamma”), SVR (“C”, “kernel”, “gamma”, “epsilon”, “tol”, “max_iter”), RFR (“n_estimators”, “oob_score”, “max_features”, “max_depth”, “min_weight_fraction_leaf” , “max_leaf_nodes”), NN (“hidden_layer_size”, “activation”, “alpha”, “batch_size”, “learning_rate”, “learning_rate_init”, “max_iter”, “early_stopping”).

III. RESULTS AND DISCUSSION

A. Linear scaling relations

Inspired by the success of linear scaling relations for predicting the adsorption energies of hydrogenated species on transition metals [16, 17], we begin our discussion by considering the performance of linear scaling relations on this dataset. We use the chemisorption energies of the central bonded atoms on a particular site to predict site-specific chemisorption energies of corresponding hydrogenated species, and the RMSEs of the prediction are summarized in Table I, which will subsequently be used as the benchmark for ML investigation.

The linear scaling relations on pure metals achieve acceptable errors from 0.20 eV to 0.33 eV, while it performs poorly with RMSEs ranging from 0.40 eV to 0.84 eV when applied to the binary alloys, whose composition varies over a larger materials space. This suggests that the scaling relation derived from pure metals fails to predict adsorption strength on bimetallic materials for rational catalyst optimization. We note that the average errors in Table I increase with the valence state of the central atoms, for instance, the average RMSEs for CH^*, CH_2^*, and CH_3^* are 0.43 eV, 0.58 eV, 0.78 eV respectively. The reason for such poor correlation in high valence hydrogenated adsorbates versus corresponding atoms could be attributed to the variation of the most stable sites for adsorbates with different electron densities. The hollow site is usually the most stable site for the adsorption of C atom. On the contrary, CH_3^* is often adsorbed on the top site, because the central C atom is a highly saturated center. Also listed is the range of adsorption energies of adsorbate species, as a reference for prediction performance. Due to the great variation in surface sites, the wide range of adsorption energy values is another reason for the poor performance of linear scaling relation on this dataset. In general, the considerable deviations between predictions of linear scaling relations and DFT-calculated adsorption energies on alloys indicate the need for adsorption models with higher accuracy.

TABLE I RMSEs for adsorption energies E_{ads} (eV) of hydrogenated species on the top, bridge, and hollow sites of pure metals and binary alloys, predicted using linear scaling relations with adsorption energies of corresponding central atoms. $\Delta E_{\mathrm{ads}}=E_{\mathrm{ads,max}}-E_{\mathrm{ads,min}}$ represents the range of adsorption energies of corresponding adsorbate, listed in the last column as a reference for prediction errors.

Linear	Average	RMSE in metals			RMSE in alloys			ΔE_{ads}/eV
relations	RMSE	Top	Bridge	Hollow	Top	Bridge	Hollow	
CH^* *vs.* C	0.43	0.20	0.21	0.20	0.40	0.44	0.47	6.54
$CH_2{}^*$ *vs.* C	0.58	0.23	0.25	0.24	0.57	0.56	0.61	5.25
$CH_3{}^*$ *vs.* C	0.78	0.31	0.32	0.33	0.77	0.76	0.84	3.34
OH^* *vs.* O	0.50	0.23	0.21	0.21	0.49	0.48	0.52	5.22
$OH_2{}^*$ *vs.* O	0.71	0.30	0.29	0.31	0.71	0.70	0.73	3.49
NH^* *vs.* N	0.59	0.23	0.25	0.25	0.55	0.57	0.62	7.15
SH^* *vs.* S	0.46	0.20	0.22	0.21	0.45	0.45	0.48	4.67

B. Evaluation of ML algorithms on atomic adsorption

To assess the performance of different ML algorithms on adsorption energy, we employ the six ML methods on the atomic adsorption datasets to obtain the prediction models. There are 1776, 1836, 1795, 1796, and 1806 adsorption energies data for the adsorption energy of C, H, O, N, and S atoms, respectively. We randomly select 80% of data in each set to optimize the hyper-parameters through cross validation and fit the model parameters for each algorithm to obtain a training error, as shown in FIG. 2. RFR achieves the best fitting performance, and the training that RMSEs for C, H, O, N and S atoms are 0.17 eV, 0.07 eV, 0.16 eV, 0.15 eV, 0.12 eV respectively. In contrast, the training errors of LASSO are the highest among the six methods with RMSEs above 0.6 eV for C, O and N atoms. The training errors of the rest four methods are generally similar. The remaining 20% of data in each set is used for testing the prediction accuracy and the test RMSEs in FIG. 2 present the predictive ability of trained models. Comparing the average test error of each algorithm for five adatoms, RFR yields the best predictions with an average RMSE of 0.3 eV, followed by KRR (0.35 eV), NN (0.37 eV), SISSO (0.39 eV), SVR (0.44 eV), LASSO (0.55 eV).

In FIG. 3, we plot the predicted adsorption energies from the methods that acquire the best prediction performance according to test RMSEs in FIG. 2, *versus* DFT-calculated adsorption energies in each dataset. ML methods with the lowest test errors for C, H, O, N and S atoms are RFR, SISSO, NN, RFR, KRR, respectively. The points are all constrained on the diagonal in the figure, which illustrates the high accuracy of the ML methods. The magnitude of RMSE for each adatom is related to the variation range of the adsorption energies. For example, adsorption energy of H atom has the lowest test error of 0.16 eV, partially owing to the adsorption energies varying on a small scale from −1.2 eV to 1.2 eV, while the adsorption energy ranges of other adsorbed atoms are larger than 5 eV. Therefore, adsorption energies of C atom are the most difficult to predict and the predictive error is the highest with test RMSE of 0.43 eV.

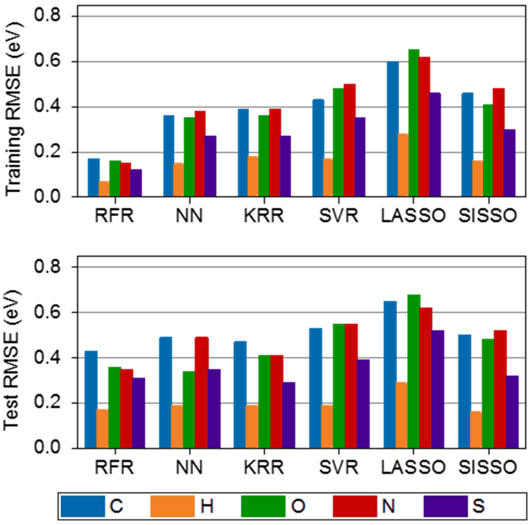

FIG. 2 Training and test RMSEs (eV) for adsorption energies of five adatoms (C, H, O, N, S) obtained using six ML methods.

Combining training and predictive performance, we can find that RFR is the optimum method for predicting the adsorption energy on alloys and we will apply it to learning the whole dataset. This might suggest that tree-ensemble methods are more suitable for this kind of task where the composition of materials spans over a large space and the features used as input are numerical. Neural network could construct a more complex model by fitting more parameters, but it usually has better performance when treating a larger dataset and using more complicated chemical representations for surface structures. SISSO and LASSO methods are based on compressed sensing and more effective in identifying descriptors for the target property, which can help to ex-

DOI:10.1063/1674-0068/cjcp2004049

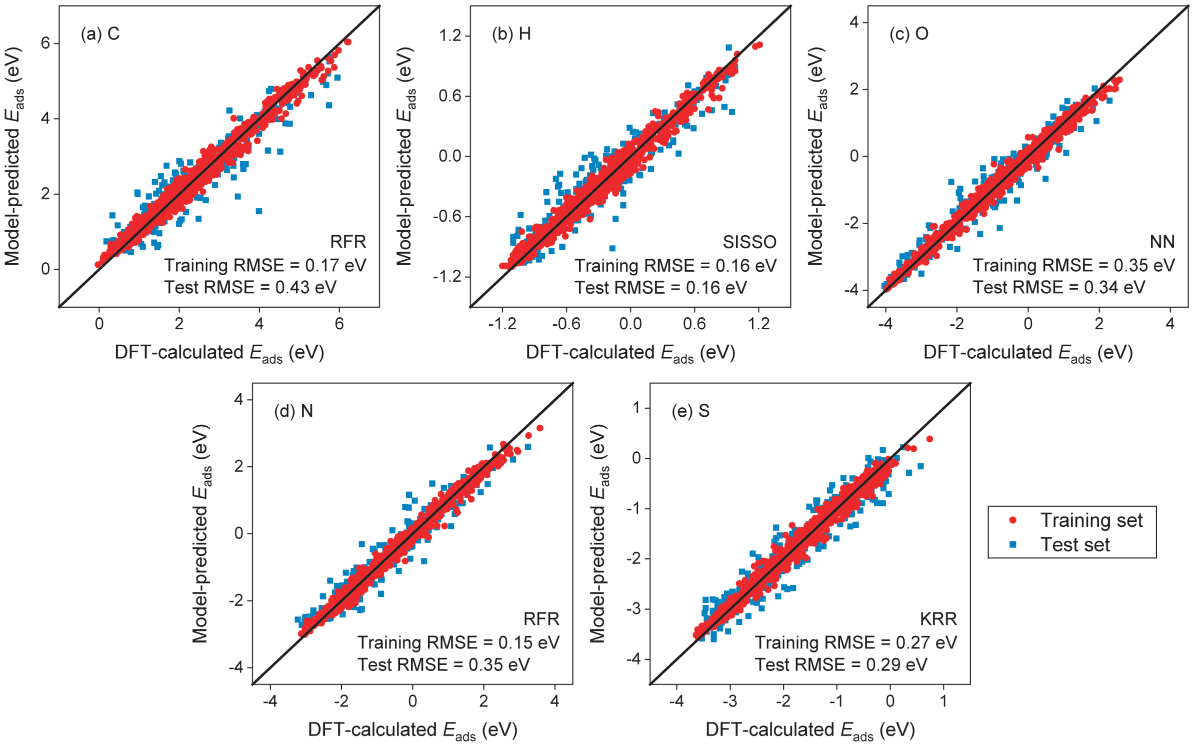

FIG. 3 Comparison between calculated and model-predicted adsorption energies in the training and test datasets of five adatoms of (a) C, (b) H, (c) O, (d) N, (e) S. The predictions are obtained by ML method (labeled in figures) that achieves the lowest test error for each adatom.

TABLE II Descriptors identified by SISSO method for adsorption energy of five adatoms. The superscripts of A and B in the property abbreviations represent the metal components of alloys.

Adatom	Predictive equation
C	$-0.12\cdot\frac{\mathrm{IE}^{\mathrm{B}}}{\mathrm{IE}^{\mathrm{A}}}\cdot H^{\mathrm{A}}\cdot r^{\mathrm{B}}+3.2\frac{\chi_{\mathrm{p}}^{\mathrm{A}}\cdot\chi_{\mathrm{p}}^{\mathrm{B}}}{H^{\mathrm{A}}+H^{\mathrm{B}}}-9.8\cdot\left(a^{\mathrm{A}}\cdot N^{\mathrm{A}}+a^{\mathrm{B}}\cdot N^{\mathrm{B}}\right)+2.8$
H	$0.12\cdot(W^{\mathrm{A}}\cdot r^{\mathrm{A}}+W^{\mathrm{B}}\cdot r^{\mathrm{B}})+29\cdot\frac{\chi_{\mathrm{p}}^{\mathrm{A}}}{\chi_{\mathrm{p}}^{\mathrm{B}}\cdot(H^{\mathrm{A}}+H^{\mathrm{B}})}-4.4\cdot\frac{H^{\mathrm{A}}}{W^{\mathrm{A}}\cdot(H^{\mathrm{A}}+H^{\mathrm{B}})}-2.12$
O	$0.19\cdot\chi_{\mathrm{p}}^{\mathrm{A}}\cdot\left(W^{\mathrm{A}}-W^{\mathrm{B}}\right)-9.1\cdot\frac{H^{\mathrm{A}}\cdot R}{W^{\mathrm{A}2}}-0.031\cdot\frac{H^{\mathrm{B}}\cdot\mathrm{IE}^{\mathrm{A}}}{R\cdot\chi_{\mathrm{p}}^{\mathrm{B}}}-1.4$
N	$0.47\cdot\left(W^{\mathrm{A}}\cdot r^{\mathrm{A}}+W^{\mathrm{B}}\cdot r^{\mathrm{B}}\right)+1.1\cdot\frac{\chi_{\mathrm{p}}^{\mathrm{A}}\cdot\chi_{\mathrm{p}}^{\mathrm{B}}}{H^{\mathrm{A}}+H^{\mathrm{B}}}-16\cdot\frac{H^{\mathrm{A}}\cdot\gamma^{\mathrm{B}}}{W^{\mathrm{A}}\cdot\mathrm{IE}^{\mathrm{A}}}+1.8$
S	$-9.3\cdot\frac{H^{\mathrm{A}}R}{W^{\mathrm{A}2}}+10\cdot\frac{\chi_{\mathrm{p}}^{\mathrm{A}}\cdot\chi_{\mathrm{p}}^{\mathrm{B}}}{V^{\mathrm{A}}\cdot n_{\mathrm{ws}}^{\mathrm{A}}}+0.010\cdot\frac{\left(H^{\mathrm{A}}-H^{\mathrm{B}}\right)\cdot\mathrm{IE}^{\mathrm{A}}}{S}-1.3$

tract explicit physical insights into the problem. SISSO might, by contrast, have better fitting ability as shown in our study. It will also be combined with RFR to obtain a better performance below.

Method SISSO starts with enlarging the feature space by iteratively combining initial features into more complex new combinations using mathematical operations, so that the target property can be well expanded in the feature space: $y=\sum_i c_i d_i$, where c_i are coefficients and d_i are feature combinations, *i.e.* identified descriptors [38]. Within SISSO, two hyperparameters are optimized. The first is descriptor complexity (the number of mathematical operators in a descriptor) and the second is model dimension (the number of descriptors). In Table II, we present the predictive equations (descriptor complexity is 3 and model dimension is 3 as well) for atomic adsorption energy found by SISSO. There are some common descriptors in different predictive equations, such as $W^{\mathrm{A}}\cdot r^{\mathrm{A}}+W^{\mathrm{B}}\cdot r^{\mathrm{B}}$, $\chi_{\mathrm{p}}^{\mathrm{A}}\cdot\chi_{\mathrm{p}}^{\mathrm{B}}/(H^{\mathrm{A}}+H^{\mathrm{B}})$, and $H^{\mathrm{A}}\cdot R/W^{\mathrm{A}2}$, suggesting that they are important quantities for atomic adsorption and the adsorption

DOI:10.1063/1674-0068/cjcp2004049

strengthes of different atoms are correlative. Based on the SISSO equations, we find that electronic properties are determinant for adsorption energy due to the multiple occurrences of some relative features, *e.g.* electronegativity, ionization energy and work function. This also accords with the general understanding of chemisorption that the adsorption strength is mainly contributed by electronic coupling between adsorbates and catalysts. Although it is difficult to directly give a physical interpretation of SISSO identified descriptors, we can combine it with RFR method, which has demonstrated better predictive ability, by using SISSO to exploit relationships between adsorption energy and initial features. Specifically, we first use SISSO to enlarge the initial feature space and select some important feature combinations. These combined features are then used as input in RFR training procedure together with initial features. In this way, we can employ SISSO to accomplish the design of fingerprint features as discussed below.

C. Prediction for the entire dataset

With the aim of predicting the adsorption strengths for various species in this massive dataset, we apply RFR method that performs the best in evaluations to learning the adsorption energy values in the whole dataset. The entire dataset is classified into 48 subsets in terms of adsorbate types and each subset is trained individually using the same hyperparameters. Before constructing the holistic models, we first test the effect of data partitioning on test RMSE for the entire dataset. To understand the relative performance, FIG. 4 compares RFR test errors using a certain randomly chosen fraction of the dataset. The height of error bars or boxes in the figure represent the deviation of prediction errors between multiple tests. The mean values of RMSEs in ten independent tests decrease when increasing the amount of training data, until the training set size reaches around 60% of the dataset. This suggests that the models trained by such amount of data can accurately predict the rest. Though there is no statistically significant difference of the average RMSEs on the training data over 60% of the dataset, the size of the error bars, as well as the height of boxes, reaches the minimum when the data ratio of training to test is set 8:2. If we decrease the amount of test data or decrease the size of training set, the uncertainty of prediction performance will increase.

To construct predictive models for the entire dataset of adsorption energy, we begin with developing fingerprint features by employing method SISSO to enlarge the initial feature space. Here we denominate the set of initial features that we introduced in METHODS as feature space 1. These initial features are combined and selected by SISSO to generate twenty descriptors (the outcome is the model with dimension of 1 and descriptor complexity of 1) that are more relative to adsorption energy than initial features. Selected descriptors and initial features compose feature space 2. We then use two feature spaces as input to train RFR models separately. In FIG. 5, we present an overall picture for training and test performance of RFR from both feature spaces. After appending those features identified from SISSO, the training RMSE decreases from 0.27 eV to 0.21 eV, and the test RMSE decreases from 0.36 eV to 0.29 eV. The lower error of predictions from feature space 2 demonstrates that it improves the prediction accuracy by combining two methods. Note that the test error from both feature spaces are far less than that of the linear scaling relations as discussed. This test error for predicting 11975 data is also comparable with the results in previous ML studies [29, 30, 45], even though the amount of adsorption data used in this work is much larger and more types of adsorbents and materials are involved. We have to clarify that this test RMSE, far less than the considerable scale of adsorption energy for about 40 eV, is obtained on the whole dataset with the composition of bimetallic materials over a large chemical space, resulting in the higher predictive error than DFT precision. However, in actual catalyst screening, the candidate catalytic materials account for a small proportion of metals, mostly composed of late transition metal elements. The accuracy can be systematically improved by reducing the materials space both in the training and testing dataset. Taking the catalyst screening for methanol electro-oxidation reaction as an example, our approach can achieve test RMSEs of 0.13 eV and 0.15 eV for adsorption energies of CO^* and OH^*, which are key factors to the activity of this reaction, on catalysts comprised of late transition metals,

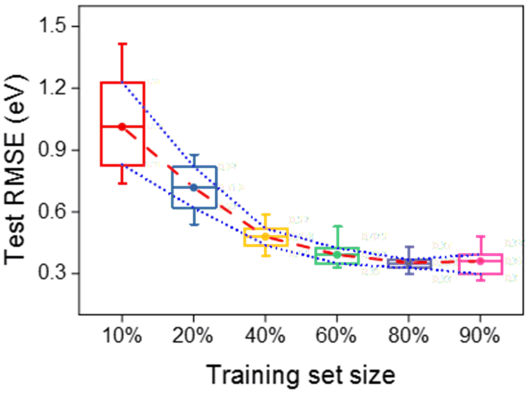

FIG. 4 Box plots of test RMSE values for the effect of training/test data ratio on the entire dataset. We perform ten independent tests for each training set size. The error bars mark the maximum and minimum values, the upper and lower limits of the boxes mark the 75% and 25% percentiles, and the horizontal lines and the points in boxes mark the mean values. The red dashed line represents the trend for the average error of ten independent calculations on a random fraction of the data set.

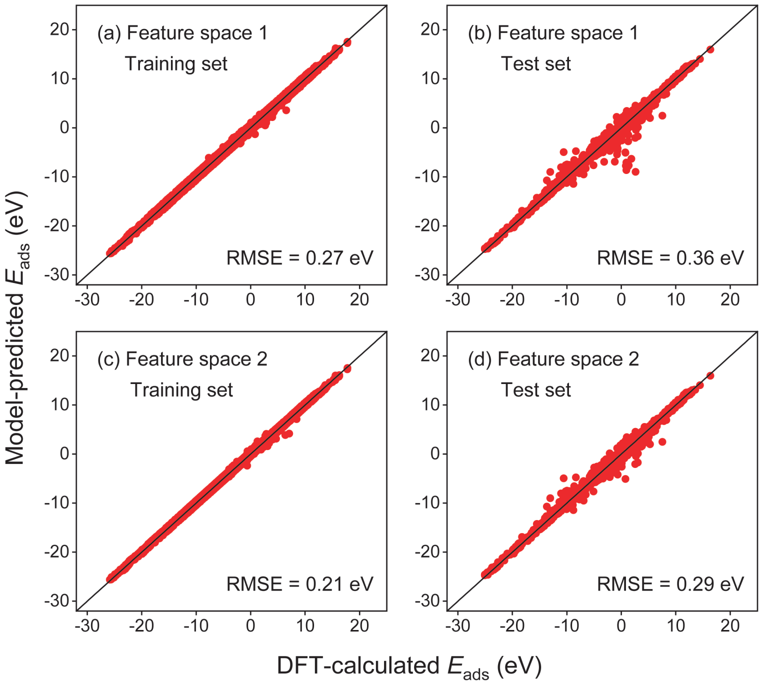

FIG. 5 Comparison between calculated adsorption energies and predicted adsorption energies using RFR algorithm. (a) and (b) show training and predicting quality, respectively, on feature space 1 that consists of initial features. (c) and (d) show training and predicting quality, respectively, on feature space 2 that consists of initial features and descriptors identified by SISSO.

thus it meets accuracy requirement for catalyst screening successfully.

For many chemical reactions, the catalytic property has hitherto been exclusively described by adsorption energies of singular species [24]. One particularly significant attempt is to disperse the adsorption strength of other key intermediates previously regarded as fixed in terms of scaling relations. For instance, in the reaction of electrolysis of water to produce molecular oxygen and hydrogen, it would be desirable to search for a catalyst where the adsorption energy difference of OOH* and O* is as large as possible [46], and in the reaction of CO_2 electrocatalytic reduction, it would be desirable to break the linear scaling relation between CHO* and CO* [47]. Our approach provides a perspective for screening bimetallic catalyst that enables to directly predict adsorption energies of interest species. As an example, we present maps of ML-predicted adsorption energies of OH*, an important intermediate in oxygen reduction reaction (ORR) [48], and NH*, which is often studied in nitrogen reduction reaction (NRR) [49, 50]. $E_{\rm ads}$(OH*) and $E_{\rm ads}$(NH*) are usually regarded as critical variables to screen the optimal electrocatalyst for ORR and NRR, which are all essential for sustainable conversion technologies to relieve increasing energy demands and impending climate change. As shown in FIG. 6, we predict adsorption energies of both OH* and NH* at their most stable sites (703 compositions including 37 pure metals and 666 alloys with AB stoichiometry of 1:1). The numbers of DFT-calculated adsorption energies for OH* and NH* are 97 and 127 in the dataset. This direct prediction goes beyond the linear scaling relations ($E_{\rm ads}$(OH*)=$a\cdot E_{\rm ads}$(O*)+b and $E_{\rm ads}$(NH*)=$c\cdot E_{\rm ads}$(N)+d, where a, b, c, and d are coefficients), we can thereby screen proper catalysts through predictions for various adsorbates.

Compared with computationally expensive DFT calculation, our framework is able to quickly identify the capacity of underlying catalytic materials without much loss of accuracy. Here we present a comparison of computational cost between DFT calculations and ML approach. Measurements of adsorption energy as we presented in FIG. 6 are possible with DFT, but are much more time-consuming, approximately 10^5-10^6 cpu·h^{-1}. By contrast, we make these predictions through ML methods using only 10−100 cpu·h^{-1}. Starting from DFT calculations and constructing the ML predictor generates the most likely pathway with fewer calculations than what would be necessary to screen the catalyst from huge materials space. Through this approach, we can realize the high-throughput predictions that are faster and lower-cost than traditional strategies. Further efforts will be made to discover novel bimetallic catalysts through this ML approach.

IV. CONCLUSION

We report a holistic framework based on machine-learning methods to explore the catalytic property over

DOI:10.1063/1674-0068/cjcp2004049

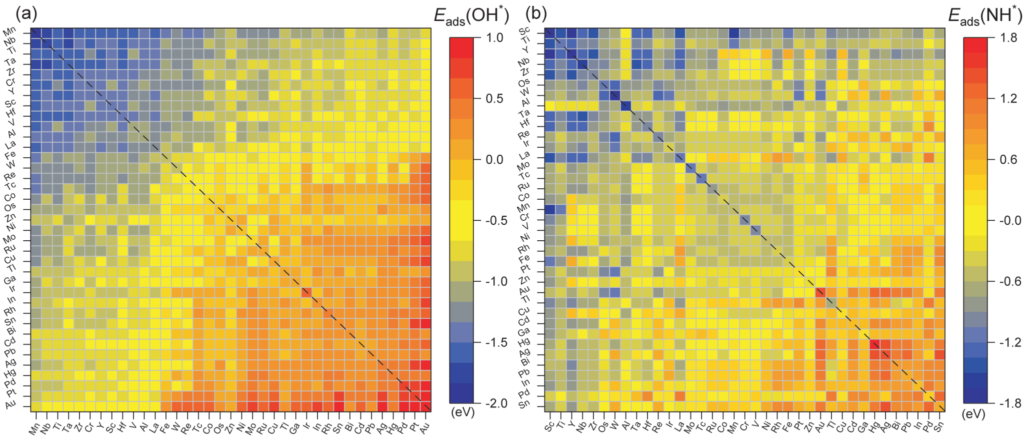

FIG. 6 Predicted adsorption energies across pure metals (37 metals, located on the dotted line) and binary alloys (666 AB alloys, symmetric across the dotted line). (a) The map of $E_{\mathrm{ads}}(\mathrm{OH}^*)$, which is calculated according to chemical equation: $H_2O(g)\rightarrow H_2(g)+OH^*$. (b) The map of $E_{\mathrm{ads}}(\mathrm{NH}^*)$, which is calculated according to chemical equation: $1/2H_2O(g)+1/2N_2(g)\rightarrow NH^*$. The metal elements are ordered in terms of the mean values of predicted adsorption energies.

a broad range of chemical space. By comparing different ML methods and utilizing RFR and SISSO together, we construct predictive models that are trained with about 60000 adsorption energies on binary alloys. Notably, based on the framework exhibited in this work, our approach can acquire the promising predictions with RMSE of 0.29 eV, far less than scaling relations and comparable with previous ML studies, over an extensive adsorbates and materials space. We can therefore quickly predict tens of thousands of adsorption energies on the most stable sites with humongous compositional and configurational degrees of freedom. To illustrate prediction examples, the adsorption energies of OH* and NH*, commonly used descriptors in catalysis, are predicted for enumerated combinations of binary alloy at the end of the study.

V. ACKNOWLEDGMENTS

This work was supported by the National Key R&D Program of China (No.2017YFB0602205, No.2018YFA0208603, No.2018YFB0704400), the Natural Science Foundation of China (No.91645202, No.91945302, No.21903077), the Chinese Academy of Sciences (No.QYZDJ-SSW-SLH054), the Fundamental Research Funds for the Central Universities, the China Postdoctoral Science Foundation, the Program of Shanghai Youth Oriental Scholars, and the DNL Cooperation Fund CAS (No.DNL201920).

[1] A. T. Bell, Science **299**, 1688 (2003).
[2] G. Ertl, Angew. Chem. Int. Edit. **47**, 3524 (2008).
[3] Q. Fu, W. X. Li, Y. Yao, H. Liu, H. Y. Su, D. Ma, X. K. Gu, L. Chen, Z. Wang, and H. Zhang, Science **328**, 1141 (2010).
[4] A. Balandin, Adv. Catal. **19**, 1 (1969).
[5] J. K. Nørskov, T. Bligaard, J. Rossmeisl, and C. H. Christensen, Nat. Chem. **1**, 37 (2009).
[6] A. Vojvodic and J. K. Nørskov, Natl. Sci. Rev. **2**, 140 (2015).
[7] R. O. Jones, Rev. Mod. Phys. **87**, 897 (2015).
[8] J. K. Nørskov, F. Abild-Pedersen, F. Studt, and T. Bligaard, Proc. Natl. Acad. Sci. USA **108**, 937 (2011).
[9] P. P. Chen, B. Y. Zhang, X. K. Gu, and W. X. Li, Chin. J. Chem. Phys. **32**, 437 (2019).
[10] S. S. Wang, M. Z. Jian, H. Y. Su, and W. X. Li, Chin. J. Chem. Phys. **31**, 284 (2018).
[11] M. Escudero-Escribano, P. Malacrida, M. H. Hansen, U. G. Vej-Hansen, A. Velázquez-Palenzuela, V. Tripkovic, J. Schiøtz, J. Rossmeisl, I. E. Stephens, and I. Chorkendorff, Science **352**, 73 (2016).
[12] W. Yu, M. D. Porosoff, and J. G. Chen, Chem. Rev. **112**, 5780 (2012).
[13] S. W. Benson, F. Cruickshank, D. Golden, G. R. Haugen, H. O'neal, A. Rodgers, R. Shaw, and R. Walsh, Chem. Rev. **69**, 279 (1969).
[14] B. Hammer, Y. Morikawa, and J. K. Nørskov, Phys. Rev. Lett. **76**, 2141 (1996).
[15] B. Hammer and J. K. Norskov, Nature **376**, 238 (1995).
[16] F. Abild-Pedersen, J. Greeley, F. Studt, J. Rossmeisl, T. Munter, P. G. Moses, E. Skulason, T. Bligaard, and J. K. Nørskov, Phys. Rev. Lett. **99**, 016105 (2007).
[17] F. Calle-Vallejo, J. Martnez, J. M. Garca-Lastra, J. Rossmeisl, and M. Koper, Phys. Rev. Lett. **108**, 116103 (2012).
[18] M. Salciccioli, S. Edie, and D. Vlachos, J. Phys. Chem. C **116**, 1873 (2012).
[19] Z. W. Ulissi, A. J. Medford, T. Bligaard, and J. K. Nørskov, Nat. Commun. **8**, 1 (2017).
[20] P. De Luna, J. Wei, Y. Bengio, A. Aspuru-Guzik, and E. Sargent, Nature **552**, 23 (2017).
[21] J. R. Kitchin, Nat. Catal. **1**, 230 (2018).
[22] B. R. Goldsmith, J. Esterhuizen, J. X. Liu, C. J. Bartel, and C. Sutton, AlChE J. **64**, 2311 (2018).
[23] Z. Li, S. Wang, and H. Xin, Nat. Catal. **1**, 641 (2018).
[24] M. Andersen, S. V. Levchenko, M. Scheffler, and K.

DOI:10.1063/1674-0068/cjcp2004049

Reuter, ACS Catal. **9**, 2752 (2019).

[25] P. Wang, T. Weise, and R. Chiong, Evol. Intell. **4**, 3 (2011).

[26] X. Li, R. Chiong, Z. Hu, D. Cornforth, and A. J. Page, J. Chem. Theory Comput. **15**, 6882 (2019).

[27] R. García-Muelas and N. López, Nat. Commun. **10**, 1 (2019).

[28] X. Ma, Z. Li, L. E. Achenie, and H. Xin, J. Phys. Chem. Lett. **6**, 3528 (2015).

[29] Z. W. Ulissi, M. T. Tang, J. Xiao, X. Liu, D. A. Torelli, M. Karamad, K. Cummins, C. Hahn, N. S. Lewis, and T. F. Jaramillo, ACS Catal. **7**, 6600 (2017).

[30] Z. Li, S. Wang, W. S. Chin, L. E. Achenie, and H. Xin, J. Mater. Chem. A **5**, 24131 (2017).

[31] O. Mamun, K. T. Winther, J. R. Boes, and T. Bligaard, Sci. Data **6**, 1 (2019).

[32] K. T. Winther, M. J. Hoffmann, J. R. Boes, O. Mamun, M. Bajdich, and T. Bligaard, Sci. Data **6**, 1 (2019).

[33] K. Tran and Z. W. Ulissi, Nat. Catal. **1**, 696 (2018).

[34] J. A. Dean, *Lange's Handbook of Chemistry*, New York, London: McGraw-Hill, Inc. (1999).

[35] D. R. Lide, *CRC Handbook of Chemistry and Physics*, 85th Edn., Eoca Raton: CRC Press LLC, 59 (2004).

[36] WebElements (https://www.webelements.com)

[37] L. M. Ghiringhelli, J. Vybiral, E. Ahmetcik, R. Ouyang, S. V. Levchenko, C. Draxl, and M. Scheffler, New J. Phys. **19**, 023017 (2017).

[38] R. Ouyang, S. Curtarolo, E. Ahmetcik, M. Scheffler, and L. M. Ghiringhelli, Phys. Rev. Mater. **2**, 083802 (2018).

[39] R. Ouyang, E. Ahmetcik, C. Carbogno, M. Scheffler, and L. M. Ghiringhelli, J. Phys. Mater. **2**, 024002 (2019).

[40] C. Robert, Machine Learning, *a Probabilistic Perspective*, Abingdon: Taylor & Francis (2014).

[41] A. J. Smola and B. Schölkopf, Stat. Comput. **14**, 199 (2004).

[42] L. Breiman, Mach. Learn. **45**, 5 (2001).

[43] H. D. Beale, H. B. Demuth, and M. Hagan, Pws, Boston (1996).

[44] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, and V. Dubourg, J. Mach. Learn. Res. **12**, 2825 (2011).

[45] T. Toyao, K. Suzuki, S. Kikuchi, S. Takakusagi, K. I. Shimizu, and I. Takigawa, J. Phys. Chem. C **122**, 8315 (2018).

[46] J. Rossmeisl, A. Logadottir, and J. K. Nørskov, Chem. Phys. **319**, 178 (2005).

[47] Y. Li and Q. Sun, Adv. Energy Mater. **6**, 1600463 (2016).

[48] J. K. Nørskov, J. Rossmeisl, A. Logadottir, L. Lindqvist, J. R. Kitchin, T. Bligaard, and H. Jonsson, J. Phys. Chem. B **108**, 17886 (2004).

[49] C. Guo, J. Ran, A. Vasileff, and S. Z. Qiao, Energy Environ. Sci. 11, 45 (2018).

[50] Z. W. Seh, J. Kibsgaard, C. F. Dickens, I. Chorkendorff, J. K. Nørskov, and T. F. Jaramillo, Science **355**, eaad4998 (2017).

DOI:10.1063/1674-0068/cjcp2004049