npj Computational Materials

www.nature.com/npjcompumats

ARTICLE OPEN

A Bayesian framework for adsorption energy prediction on bimetallic alloy catalysts

Osman Mamun[1,2]✉, Kirsten T. Winther[1,2], Jacob R. Boes[1,2] and Thomas Bligaard[2,3]✉

For high-throughput screening of materials for heterogeneous catalysis, scaling relations provides an efficient scheme to estimate the chemisorption energies of hydrogenated species. However, conditioning on a single descriptor ignores the model uncertainty and leads to suboptimal prediction of the chemisorption energy. In this article, we extend the single descriptor linear scaling relation to a multi-descriptor linear regression models to leverage the correlation between adsorption energy of any two pair of adsorbates. With a large dataset, we use Bayesian Information Criteria (BIC) as the model evidence to select the best linear regression model. Furthermore, Gaussian Process Regression (GPR) based on the meaningful convolution of physical properties of the metal-adsorbate complex can be used to predict the baseline residual of the selected model. This integrated Bayesian model selection and Gaussian process regression, dubbed as residual learning, can achieve performance comparable to standard DFT error (0.1 eV) for most adsorbate system. For sparse and small datasets, we propose an ad hoc Bayesian Model Averaging (BMA) approach to make a robust prediction. With this Bayesian framework, we significantly reduce the model uncertainty and improve the prediction accuracy. The possibilities of the framework for high-throughput catalytic materials exploration in a realistic setting is illustrated using large and small sets of both dense and sparse simulated dataset generated from a public database of bimetallic alloys available in Catalysis-Hub.org.

npj Computational Materials (2020)6:177 ; https://doi.org/10.1038/s41524-020-00447-8

INTRODUCTION

Mean-field microkinetic models—developed by combining electronic structure properties with macroscopic reaction parameters, such as reaction temperature and pressure[1,2]—are used to obtain fundamental insights into the reaction kinetics occurring on the solid/gas interfaces. However, the success of such physics based models is critically dependent upon reliable estimate of the adsorption energetics of various elementary reactions[3–6]. In the last decade, improvement in exchange-correlation functionals made it possible to estimate adsorption energies with high fidelity, which enabled computational catalysis modeling a surrogate scheme to replace time-consuming experimental methods[7,8]. However, the composition and structural space of potentially active catalysts is vast, and machine-learning assisted high-throughput computational screening is the most viable systematic strategy to discover novel catalysts with superior activity to replace existing catalysts.

A popular approach to high-throughput computational discovery of heterogeneous catalytic materials is a descriptor based approach[9,10] where suitable descriptors, e.g., *d*–band center, width, etc., is chosen to efficiently compute the chemisorption energy of all the reaction intermediates without performing a full DFT computation. To this end, Nørskov and Hammer, proposed a simplified theory for adsorbate bonding on transition metal surfaces based on the electronic interaction of adsorbate *sp*-band with the metal *d*-band[11–13]. In their work, the *d*-band center were identified as an excellent descriptor to predict the chemisorption energies on transition metal surfaces. One of the major breakthrough in computational catalysis and surface science research came about when Abild-Pedersen and co-workers identified a linear scaling relation to determine the adsorption energy relying only on the adsorbate valency (*sp*–band of the adsorbate) together with metallic *d*–band properties[3]. Due to these underlying mechanisms, linear scaling relationships are found between the adsorption energy of similar species. For any molecular fragment AH_x, the adsorption energy is generally linearly correlated with the adsorption energy of *A*, which can be expressed mathematically as,

$$\Delta E_{AH_x} = \gamma \Delta E_A + \xi \quad (1)$$

This simple yet elegant model captures the important factors that determines the adsorption strength of any hydrogenated species given the adsorption energy of the central binding atoms. For a dataset containing chemisorption energies of AH_x on elemental pure metals, the mean absolute error is reported to be 0.13 eV for the prediction of the most stable structures and 0.06 eV when the site specificity is taken into account[3]. Based on the extensive DFT calculation of nitrogen adsorption energies as a descriptors for the ammonia synthesis reaction on several monometallic transition catalysts, Nørskov and co-workers successfully identified and later experimentally verified that *CoMo* catalyst outperforms *Ru* catalysts by exhibiting optimal nitrogen binding energy[14].

Another potentially lucrative approach is data-driven computational discovery of novel catalytic materials where rich and diverse databases are used as a basis for efficient prediction of catalytic properties of unknown materials. However, this approach is still in its infancy, mainly due to the lack of well-curated databases that satisfies the 3*V's* (volume, variety, and veracity) of big data pertinent to computational catalysis research[15,16]. In computational catalysis, these 3*V's* can be ensured, (1) by compiling a dataset with substantial data, (2) by including a wide array of

[1]SUNCAT Center for Interface Science and Catalysis, Department of Chemical Engineering, Stanford University, Stanford, CA 94305, USA. [2]SUNCAT Center for Interface Science and Catalysis, SLAC National Accelerator Laboratory, 2575 Sand Hill Road, Menlo Park, CA 94025, USA. [3]DTU Energy, Department of Energy Conversion and Storage, Anker Engelunds Vej, Building 301, 2800 Kgs. Lyngby, Denmark. ✉email: mamun.che06@gmail.com; tbli@dtu.dk

Published in partnership with the Shanghai Institute of Ceramics of the Chinese Academy of Sciences

nature partner journals

materials and composition, and (3) by carefully monitoring the computation and data collection process to preserve the data integrity. Development of a database is a complex and assiduous task that includes multitude of challenges, i.e., data generation, verification, quality maintenance, data accessibility, reusability etc. With the development of high-quality computer code and database management system (DBMS), some rich databases have emerged which are playing a crucial role for the high-throughput materials exploration through materials informatics[17–20]. To expedite catalysis informatics with such databases, we have generated a database of chemisorption energies on a wide range of catalytic surfaces, available at Catalysis-Hub.org[21,22]. Specifically, we generated a dedicated database of chemisorption energies of single-atom (such as C, H, N, O, and S) and multi-atoms (such as CH, CH_2, CH_3, OH, NH, and SH) adsorbates on 2035 binary alloy materials in their A_1, $L1_0$, and $L1_2$ strukturbericht designation. Having access to such a rich and diverse dataset for single-atom adsorbate systems, one can ask: "Can we go beyond a single descriptor linear scaling relation that will provide robust estimation of chemisorption energies of multi-atoms systems, without the need for running expensive quantum chemical computation?" In this regard, machine-learning models such as Gaussian Process and neural networks have emerged as powerful tools to make efficient, fast, and reliable prediction in a fraction of time in comparison to DFT/QC computation[23–25]. In a recent work by Xin et al., an artificial neural network was trained based on an extensive number of DFT calculations to compute the chemisorption energies on second generation core-shell alloy surfaces. In their work, {100}-terminated Cu-based alloys were identified to be highly active catalysts for CO_2 electroreduction[26]. A similar approach by Ulissi and co-workers identified previously overlooked *NiGa* to be a highly active catalysts for CO_2 reduction, which suggests an exhaustive high-throughput study is the most efficient approach to discover novel heterogeneous catalytic materials for challenging chemical reactions[23].

One can naturally envision a spectrum of mathematical models where on one end we have a very simple one descriptor scaling relation (pure physics based models) and on the other end we have multidimensional nonlinear machine-learning models (pure data-driven models). Following the "no free lunch theorem"[27], naturally the single descriptor model will be less accurate but very fast to implement due to the small number of data points needed to fit the scaling line, and the multidimensional nonlinear machine-learning model will be more accurate but time consuming, since an order of 10^2–10^4 data points is needed to fit the model parameters. Despite considerable progress in generation of machine-learning models and atomic position independent descriptors (fingerprints) used for model training, the applicability of machine-learning models is still limited, mainly due to the lack of high-quality and structured data needed to train the model. On the other hand, scaling relations where the data requirement is not as intensive, are limited in their accuracy and generalizability, rendering them unreliable for accurate high-throughput exploration. Considering the sparsity and variety of the modern databases, such as catalysis-hub.org, we have developed a scaling relation-like multi-descriptor linear regression model by extracting meaningful subset of data.

Given a set of adsorbed species (A, B, AH_x, BH_x, …) with computed adsorption energies, we can apply a linear regression model to express the chemisorption energy of one specimen in terms of the others, using the remaining chemisorption energies as predictors:

$$\Delta E_{AH_x} = \beta_0 + \sum_{i=A,B,\ldots} \beta_i \Delta E_i \tag{2}$$

given in terms of the linear coefficients (β_0, β_A, β_B, etc.).

When data for several adsorbed species are available, different models can be constructed by choosing different subsets of adsorbate species as descriptors. When we have multiple competing models and need to select the best subset of descriptors to build the best linear regression model, there are two important factors to consider: (1) The evidence problem: What is the metric to use as evidence to favor one model over the others? (2) The prediction problem: How accurate is the selected model for the future prediction on unseen data[28]?

To tackle the evidence problem, we propose Bayesian information criteria (*BIC*) as the model evidence to select the best model that optimizes the bias-variance trade-off[29,30], the approach described in detail later in this manuscript. The second problem arises from the model uncertainty due to the conditioning on data that are poor representative of the underlying relation between predictor and descriptors. Specifically, in the limit of only a handful of data points to fit the model, choosing one model will lead to high model uncertainty and in turn it will result in poor predictive performance. To address the prediction problem in small datasets, we propose Bayesian model averaging (BMA) to be a robust solution where instead of choosing a single linear regression model, we use a small set of the best models to come up with a better prediction[28,30,31]. In this paper, we present and validate a Bayesian model selection and averaging framework to find the best multi-descriptor linear regression model to predict the chemisorption energies of hydrogenated species on a vast set of bimetallic alloy catalysts. Furthermore, we developed and validated Gaussian Process based machine-learning models to predict the residual of the best selected model—the difference between the actual DFT energy and the energy predicted by the best model as identified by our Bayesian model selection approach—to further improve the chemisorption energy prediction for large datasets. With the single- and multi-atom adsorption data available in the Surface Reactions database of Catalysis-Hub.org, our Bayesian framework integrated with residual-learning approach will facilitate fast calculation of the catalytic properties of vast set of bimetallic alloy materials which is a prerequisite for high-throughput materials screening.

RESULTS AND DISCUSSION

Scaling relations

Inspired by the success of scaling relation for the prediction of adsorption energies of hydrogenated species on pure elemental transition metal surfaces[3], we have used scaling relations for adsorption energy prediction on bimetallic alloy surfaces to predict the adsorption energy on a particular site based on the descriptor energy on the same site. This way we ensure that our scaling relations are capable of site-specific chemisorption energy prediction. In Fig. 1, we illustrate the scaling relations for the chemisorption energy prediction of hydrogenated species as a function of the corresponding chemisorption energy of the central bonded atom.

When applied to our bimetallic alloy dataset, linear scaling shows excellent correlation between chemisorption energy of AH_x and central atom A, except for CH_3 * vs. C *, which is quite evident by the $r^2 = 0.22$ value reported in the scaling plots in Fig. 1. The reason for such poor correlation in CH_3 * vs. C * plot can be traced back to the conservation of electron density around the central atom. Being a highly saturated center, the top site is usually the most stable adsorption site for CH_3 *; in contrast, the hollow site is often the most stable adsorption site for C * adsorption. In CH_3 * vs. C * plot, we see that all the red circles are below the scaling lines indicating the top site is the preferred site for CH_3 * while green circles are skewed towards the left of the plot indicating the hollow site is the preferred site for C * adsorption. As a result, the one to one correspondence between adsorption sites is very weak in this case, which, combined with the overall variance present in the data, results in a very poor correlation coefficient for CH_3 * vs.

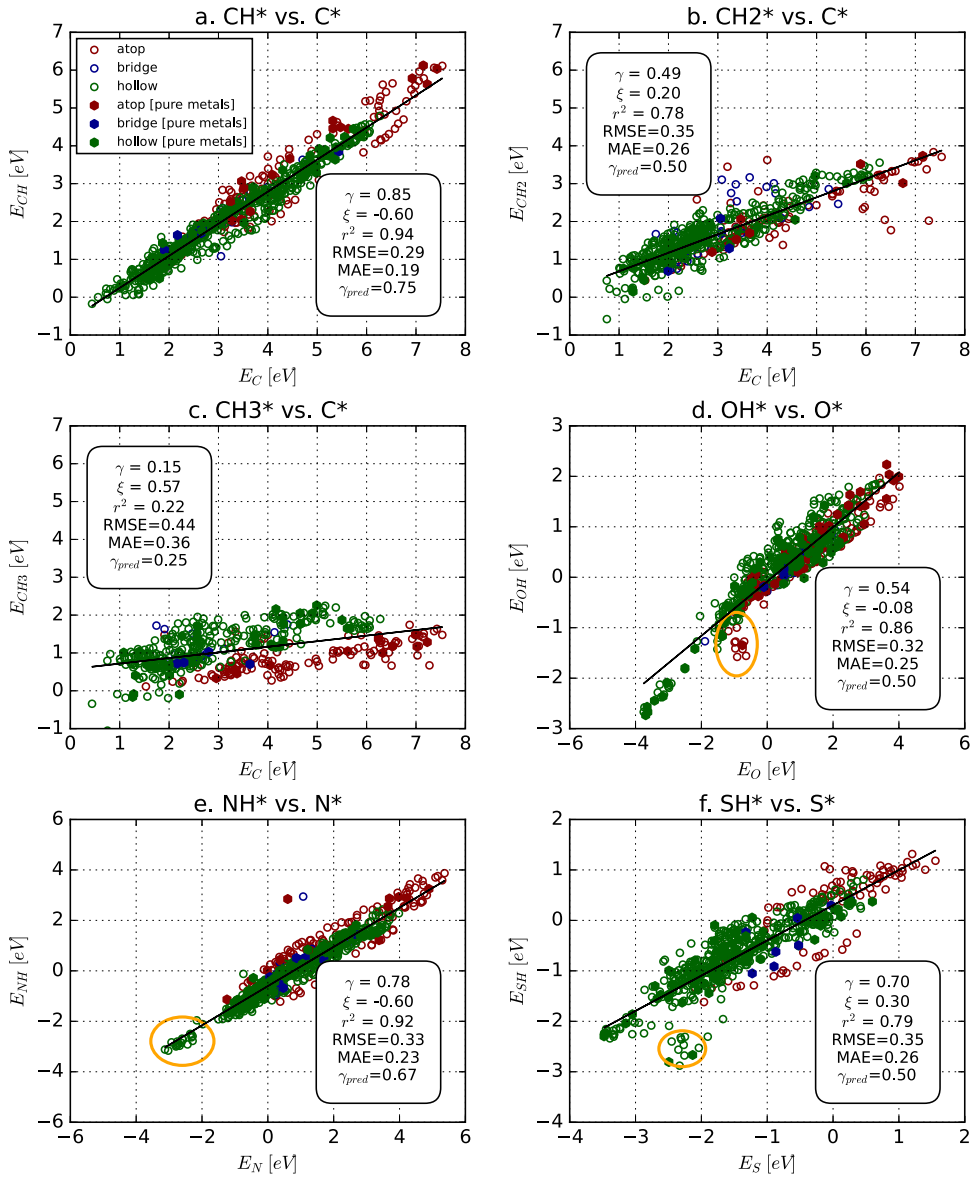

Fig. 1 Chemisorption energies of CH_x, OH, NH, and SH plotted against the chemisorption energies of C, O, N, and S, respectively. a CH * vs C *, **b** CH_2 * vs C *, **c** CH_3 * vs C *, **d** OH * vs O *, **e** NH * vs N *, and **f** SH * vs S *. The chemisorption energies are computed as $E_{ads} = E_{slab+ads} - E_{slab} - E_{reference}$. $E_{reference}$ for different adsorbates are listed in the Supplementary Table 1. Pure metals data are shown as filled hexagonal while bimetallic alloy data are shown as open circle. Also red, blue, and green color is used to denote different site types, i.e., top, bridge, and hollow sites, respectively. The orange circle indicates the cluster for non-d metallic alloy which displays a poor scaling correlation.

C * chemisorption energies. Our results show that the slope for scaling plots in our bimetallic dataset is exactly as predicted by the mathematical formula (Eq. (3)) with some aberration due to the presence of non-d metallic alloy. Here, we note that the data clusters present in the plots (shown within orange circle) are due to the poor scaling correlation present within the non-d metallic surface data.

$$\gamma = \frac{x_{max} - x}{x_{max}} \tag{3}$$

Despite the excellent correlation, the root mean squared error (*RMSE*) of prediction is quite high for the bimetallic alloy dataset, ranging from 0.29 eV for CH prediction to 0.44 eV for CH_3 adsorption energy prediction, which we deem not suitable for a reliable high-throughput screening for novel catalytic materials. Next, we discuss multi-descriptor linear regression model, as opposed to single descriptor linear scaling relations, to improve the chemisorption energy prediction.

Bayesian model selection

For any adsorbates in our dataset, we can use a combination of available descriptors of varying length to make a multivariate linear regression model; e.g., to predict E_{CH}, we can use any combination of the following 10 predictors—E_C, E_H, E_N, E_S, E_O, E_{CH_2}, E_{CH_3}, E_{NH}, and E_{OH}. In the first step, we use all the possible combinations of these descriptors to make linear models by collecting the appropriate set of data from the database. Next, we compute the model parameters and performance metrics, e.g., *RMSE*, *BIC* etc., and store them in a database. For our dataset, we have 1-023 $\sum_{i=1}^{10} {}^{10}C_i$, where C is the combination operator) linear regression models for each of the multi-atom adsorbate systems. Here, we note that we use 5-fold cross validation for *RMSE*

computation, meaning we choose 80% random training data to fit the model and the remaining 20% to evaluate the model, and we repeat this procedure 20 times to ensure that the sample variance is approximately close to the population variance. When many models are initially considered, it is oftentimes found that few models are equally good (in terms of *RMSE* values) but lead to different model predictions for the quantities of interest. To ensure robust model selection, we use *BIC* as the model evidence and select the model with the minimum *BIC* value, as *BIC* selected model ideally corresponds to the model which is *a* posteriori most probable model with the data at hand[30]. Also, the presence of the single descriptor linear scaling relation in our linear regression collection ensures that *BIC* will choose the most parsimonious quasi-true model when all the other descriptor has no correlation with the target property or chemisorption energy. Another important advantage of *BIC* for our particular model selection problem is that it doesn't require any prior information, i.e., it can work equally well for models with non-informative priors. Since the error of prediction of our models are normally distributed, the Bayesian Information Criteria is computed using the following formula,

$$\mathrm{BIC} = n\ \ln\left(\frac{\sigma^2}{\sigma_0^2}\right) + k\ \ln(n) \tag{4}$$

where, n, σ, σ_0, and k are number of data points, model variance, reference variance (set to the best variance obtained from all the models under consideration), and number of model parameters. The Bayesian information criteria is developed based on the assumption that all models have the same amount of data; However, for our practical model selection approach, we stretch this idea a little further and compute the modified Bayesian Information Criteria (*mBIC*) as,

$$mBIC = \ln\left(\frac{\sigma^2}{\sigma_0^2}\right) + \frac{k\ \ln(n)}{n} \tag{5}$$

normalized with respect to n, which permits us to compare models with varying amount of data points. With *mBIC* as the model selection metric, we do not need to compare different competing models against a baseline model (as done in hypothesis testing), rather we can compute the *mBIC* for each model and take the one with the lowest *mBIC*.

With all the data points available in our database, we found the best selected model to have performance *RMSE* of 0.11 eV, 0.13 eV, 0.15 eV, 0.20 eV, 0.09 eV, and 0.33 eV for CH, CH_2, CH_3, OH, NH, and SH, respectively. In the Supplementary Fig. 1, we show the *mBIC* of each model and the plane connecting the minimum *mBIC's* are the limit of our simple multi-descriptor model, which we dub the "*mBIC envelope*". Though the performance obtained is quite impressive, it does not address the data sparsity issue commonly encountered in modern databases. To select the best model under the constraint that we have only limited data available, we obtain the *mBIC* for different adsorbates fitted with 100 random data points for each model, shown in Fig. 2. From the *mBIC envelope* plots, we see the *mBIC envelope* is parabolic in shape, which is typical model behavior as characterized by the bias-variance trade-off. The *mBIC* provides us with a first evaluation of the relative model performance—the evidence problem. However, it ignores the model uncertainty—the prediction problem. To understand the true merit of the model, we must assess the model performance on unseen data. This is addressed in Fig. 3, where we show the *RMSE* comparison plot for in-sample and out-of-sample predictions. Interestingly, the best model selected, indicated by the blue star in the figure, performs almost equally well on the out-of-sample data as for the in-sample. This demonstrates that the *mBIC* does indeed provide a meaningful assessment of the model quality. This, with only 100 data points for each model we can achieve significant performance as evidenced by Fig. 3.

However, as we reduce the number of data points in our synthetic dataset, the best model no longer performs the best for out of sample prediction, suggesting the model uncertainty is becoming a predominant factor governing the model performance. In order to illustrate the reliability of the best selected model as the number of data points is reduced, we perform the simulation with varying amount of data for 10 runs and in the Supplementary Table 2, we tabulate the statistics obtained from this 10 runs. The table demonstrates that as the number of data points used to train the model is reduced the model uncertainty becomes significant, i.e., the difference between in sample *RMSE* and out of sample *RMSE* is larger. In the next section, we discuss Bayesian model averaging approach as a possible scheme to estimate adsorption energies when the model uncertainty is significant, and the best selected model no longer performs satisfactorily.

Bayesian model averaging

Having a robust model selection criterion established for models with significant data points, the question arises what should be the best scheme to make prediction when model uncertainty is significant. Here, we propose Bayesian model averaging (BMA) as an excellent choice. BMA has two important advantage, (1) It accounts for uncertainty in the model selection process, and (2) By averaging over the models, BMA is somewhat robust to the model uncertainty. BMA has been successfully applied to make reliable inference to many statistical model classes, including linear regression and generalized linear models[31]. Since *mBIC* can be thought of as an estimator of the relative model performance, we use the following formula for model averaging[30]:

$$E_{pred} = \frac{\sum_i \exp\left(-\frac{\Delta mBIC_i}{2}\right) E_i}{\sum_i \exp\left(-\frac{\Delta mBIC_i}{2}\right)} \tag{6}$$

thus, weighting each model prediction with the exponential of the $\Delta mBIC_i$, where $\Delta mBIC_i$ is the *mBIC* of model i w.r.t the *mBIC* of the best model. With this averaging scheme, we ensure that model contribution is proportional to the fitness of the models, as characterized by the exponential of the negative $\frac{\Delta mBIC_i}{2}$ value.

In Fig. 4, we plot the average out of sample RMSE for different number of data points for 10 random synthetic dataset for different amount of data used to train the model for both Bayesian model selection and Bayesian model averaging approach. Here, we note that BMA performs significantly well with just a few data points, i.e., (≤50), surpassing the single descriptor accuracy, and even for a large number of data points, i.e., >50 data points, prediction accuracy is slightly better than the BMS scheme. This analysis suggest that the BMA is a reliable approach to estimating adsorption energy for small datasets.

Based on our simulation results described in the above two sections, we now have a framework that: (1) Selects the best model when the model uncertainty is not significant, e.g., >200 data points are used to train the model, and (2) Averages over the all models when the model uncertainty interferes with the prediction accuracy. In the next section, we simulate a case study to show how to obtain reliable inference with this framework for real dataset with varying amounts of data available to build the models.

Prediction under varying amounts of data to fit the models

In our Bayesian model selection and model averaging discussion, we used idealized datasets where we have equal amount of data to build each model; more often than not, in a practical case different models will have varying amounts of data available in the database. To further illustrate the predictive power of our Bayesian

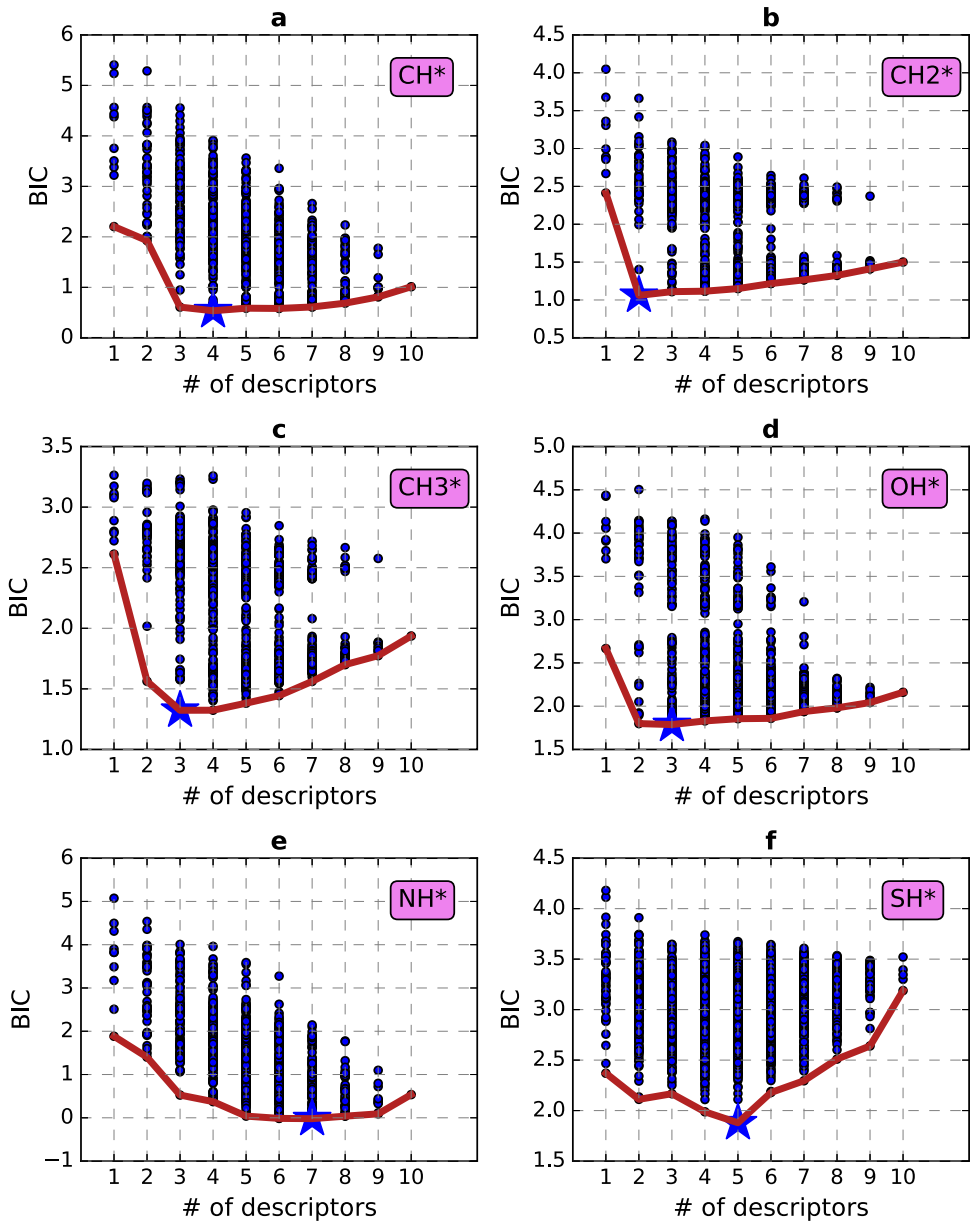

Fig. 2 Bayesian information criteria's (BIC) are plotted against the number of parameters used for different multi-atoms adsorbates for a synthetic dataset containing 100 data points for each model. a CH *, **b** CH_2 *, **c** CH_3 *, **d** OH *, **e** NH *, and **f** SH *. The red line connects the minimum of each descriptors (BIC envelope). The blue star indicates the best model (with the lowest BIC value).

model averaging approach, we simulate a dataset with varying amount of data. Specifically, we used 80, 70, 60, 50, 40, 30, 25, 25, 20, and 20 random data points to build 1, 2, 3, 4, 5, 6, 7, 8, 9, and 10 descriptor models, respectively. Typically, simple one descriptor models will have a lot of available data to train the models and the data availability will decrease with increasing model complexity, i.e., model parameters. To get significant statistics about the merit of Bayesian model averaging approach, we built the models with five different random sets of data points and the models are tested on five different unseen datasets with 20 points each. The mean *RMSE* in *eV* for all the different model runs are 0.11, 0.12, 0.11, 0.20, 0.10, and 0.35 for CH, CH_2, CH_3, OH, NH, and SH, respectively, indicating Bayesian model averaging to be equally powerful for building predictive models from datasets with varying amounts of data. This combined model selection and averaging framework is preferred over other conventional sparsification approaches, e.g., LASSO, SISSO, greedy elimination, etc., in cases where we only have a few data points for some descriptors and a larger dataset for other descriptors. One of the motivations behind developing this framework lies in the fact that in computational catalysis we often face with situation where we have lots of data for some adsorbates and only a small number of data points for others. In order to apply conventional sparsification approach to such cases, we would have to include only systems/surfaces with all adsorption energies, thus not taking advantage of all the information available.

Residual learning

So far, we have discussed the predictive capability of multivariate linear regression models to improve the performance of chemisorption energy prediction in respect to the predictive capacity of scaling relations. Now, we present a Gaussian process based machine-learning model to further improve the chemisorption energy prediction. Fingerprint generation is the most crucial

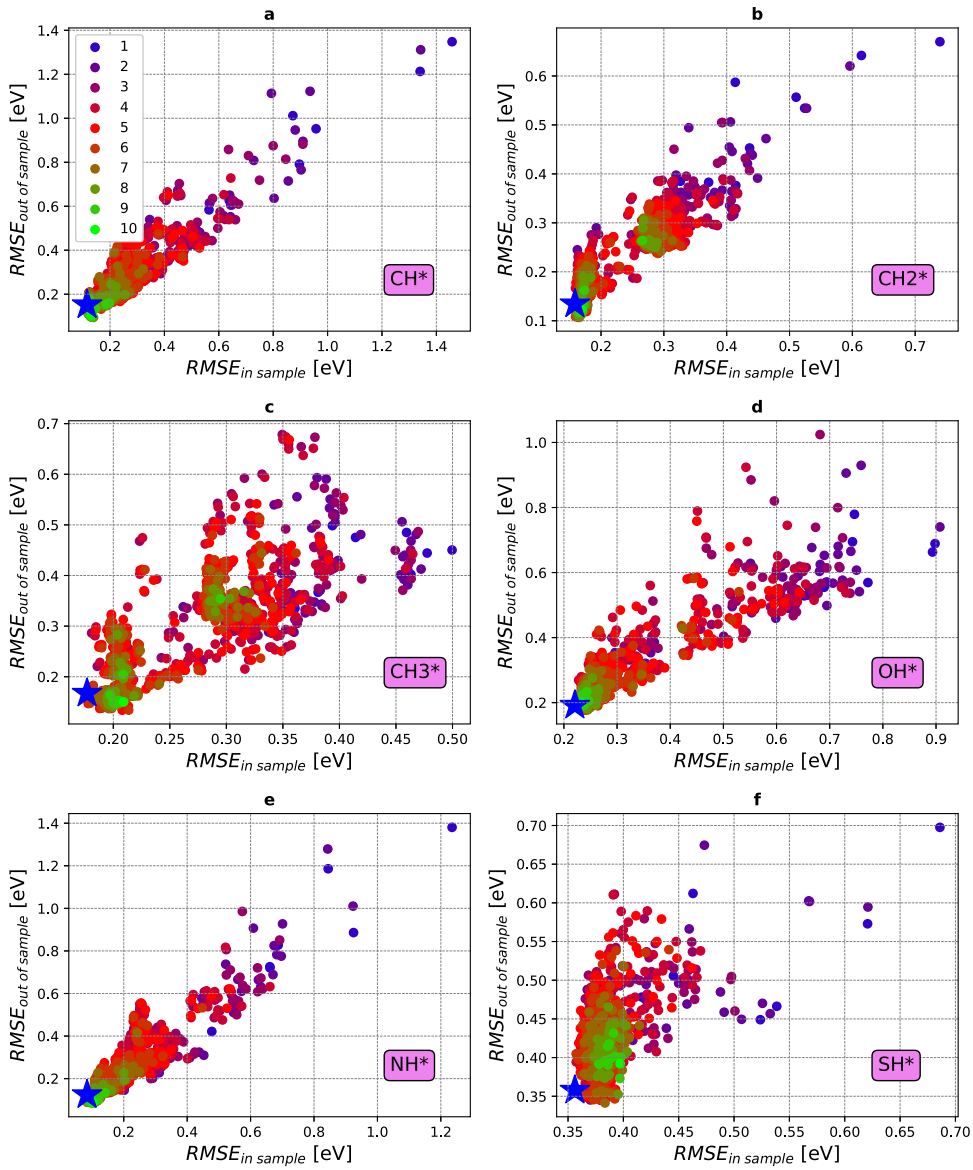

Fig. 3 Comparison of RMSE values for in sample and out of sample prediction for all models. The blue star indicates the best model (with the lowest BIC value). **a** CH *, **b** CH_2 *, **c** CH_3 *, **d** OH *, **e** NH *, and **f** SH *. The points are color coded to show the number of parameters used in that model.

component of machine-learning model generation workflow. One of the prerequisites of our fingerprint generation scheme is that it should not depend on the absolute coordinates of the atomic structure, which are unknown a priori. In order to tackle this constraint, we use the atomic connectivity to generate the fingerprints, which does not depend on the atomic positions. In this scheme, the fingerprints are generated based on the local connectivity of the atoms and then summed together.

$$f = \sum_i \sum_j P_i P_j \delta_{ij}(d_{ij}, d) \tag{7}$$

In this equation, properties of atom i are multiplied to properties of atom j, given they are d distance away where d denotes the neighboring distance (e.g., 1 for 1st nearest neighbor distance)[32]. In our fingerprinting scheme, we use $d = 1$ for the connection of adsorbed atoms to slab atoms and $d = 0$, 1, 2 otherwise, thus eliminating the dependence of the absolute atomic positions. This convolution of physical properties based on the nearest neighbor distance can be seen as analogous to cluster expansion. Starting from the nearest neighbor distance zero (no interaction between adjacent atoms) we can sequentially include more nearest neighbor interaction to improve our prediction accuracy. Calculations also suggest similar trends, i.e., prediction accuracy increases as more nearest neighbors are included in the fingerprint vector. As for the physical properties used to generate the fingerprints, we list them in the Supplementary Notes. In total, 10 physical properties were used for 3 nearest neighbor distance, resulting in a fingerprint vector of length 30.

In the first stage we used all the data to develop a Bayesian machine-learning regression models based on Gaussian Processes[33] to predict the chemisorption energies on bimetallic alloy surfaces relying only on the fingerprints generated using the 2D connectivity matrix, $C_{ij} = \delta(d_{ij}, 1)$. To train the Gaussian Process, the Radial Basis Function (RBF) kernel is used together with a white kernel to map the underlying correlations of the fingerprints and target properties. We add the white kernel as a regularizer, which accounts for noise by adding a constant to the diagonal elements of the co-variance matrix to prevent over-fitting.

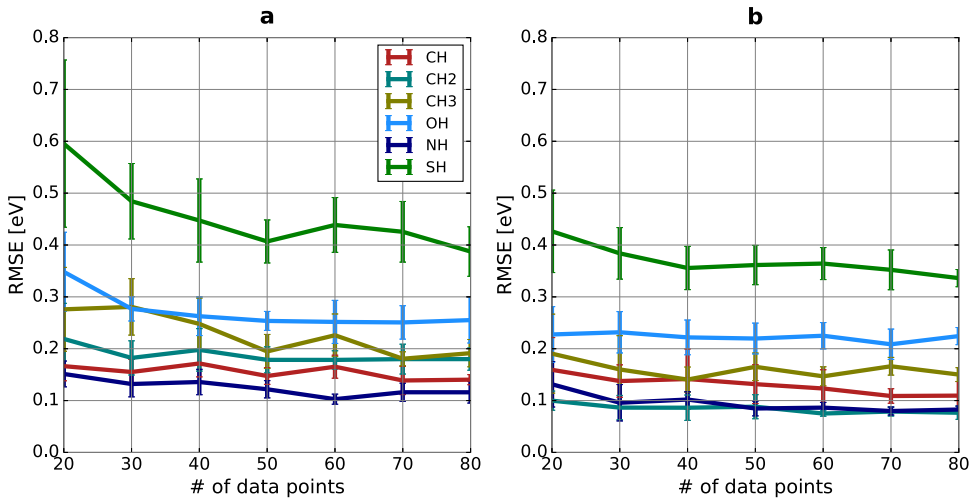

Fig. 4 Bayesian model selection (BMS) and Bayesian model averaging (BMA) performance plot. a BMS and **b** BMA. Mean RMSE in eV for 10 different runs are plotted against the number of data points used to train the model.

The hyperparameters of both kernel components are optimized by maximizing the log marginal likelihood of the Gaussian Process within the *scikit–learn* program package[34]. In Fig. 5, we show the parity plots for the Gaussian Process prediction computed with 5-fold cross validation. In the Supplementary Figs. 2 and 3, we provide the learning curves, and the distribution of residuals for all the adsorbates. The *RMSE* for the testing set are 0.22, 0.41, 0.23, 0.24, 0.24, and 0.40 for CH, CH_2, CH_3, OH, NH, and SH, respectively. Overall, our connectivity based fingerprints performs quite well for all the datasets given, where standard DFT error is estimated to be ~0.10 eV. Mean predicted error and uncertainties are both uniformly small, suggesting the model is robust to prediction. We see excellent improvement of the machine-learning models over the scaling relations, partly due to the better correlation of convolution properties for the *non–d* metallic alloys and partly due to the more flexible nature of the fitted function. Moreover, CH_2 and *SH* errors are larger than the scaling relations which is due to the poor selection of prior for the Gaussian Process Regression.

In our Gaussian process, we used all the fingerprints generated; however, it is oftentimes reported that only a subset of fingerprints contributes to the prediction and the rest usually add noise to the prediction, thus deteriorating machine-learning model performance. Also, GP regression doesn't scale well (~N^3) and require fingerprint reduction for efficient and fast computation. Here, LASSO CV is used to identify the important fingerprints (see Supplementary Figs. 4–9) and used the subsets of fingerprints in our subsequent analysis.

The success of a Gaussian Process relies largely on the prior assigned to the process; however, in the absence of any physical interpretation of the convolution properties with the target properties, we employ a $\mathcal{N}(0, \sigma^2)$ prior distribution to the model. If we consider that chemisorption energy is the sum of one linear term and one nonlinear term, we can use the scaling relation or multivariate linear regression model to predict the linear term and machine-learning model to predict the nonlinear term (Eq. (8)). One significant advantage of this method is that after subtracting the linear term from the chemisorption energy, the remaining nonlinear term can better be represented as a Gaussian distribution with zero mean prior. This way we can also significantly reduce the time and data points required to train the model. To test our hypothesis, we formulate a workflow for a residual-learning process wherein a Gaussian process based machine-learning model is used to predict the residual chemisorption energy of a system.

$$\Delta E = \Delta E_{\text{linear}} + \Delta E_{\text{nonlinear}} \tag{8}$$

In Fig. 6, we illustrate the significant performance gain of the residual learning using scaling relation as a basis over both scaling relation and Gaussian Process Regression. The *RMSE* for the testing set are 0.20, 0.33, 0.22, 0.16, 0.20, and 0.31 for CH, CH_2, CH_3, OH, NH, and SH, respectively. For CH adsorption energy prediction, residual learning improves the prediction accuracy by 0.04 eV in comparison to machine learning only models. More interestingly, the uncertainty of prediction is uniformly reduced for all the data points leading to a higher confidence in the regression prediction. Similarly, for CH_2, OH, SH, and NH, we see similar 0.05 eV to 0.1 eV improvement in the machine-learning prediction of chemisorption energies. In contrast, residual learning has a less perceptible effect on the prediction performance of E_{CH_3}, with sizable scattering in the parity plot, which can be rationalized by the poor scaling correlation coefficient observed in the CH_3 * vs. C * data. In light of our residual-learning prediction, it is our observation that without a high degree of linearity, characterized by the *r*-squared value, in the data between target energies and descriptor energies, residual-learning models will still perform as well as a pure Gaussian Process model, if not better. Another significant advantage of residual-learning method over the Gaussian Process only model is that it requires significantly fewer DFT calculations to train, typically half (see Supplementary Figs. 2, 3, 10, and 11 for the learning curve and histogram of the residuals for both the Gaussian process regression only model and residual-learning model), which makes residual learning strictly superior for an iterative framework to explore a well-defined enumerated space exhaustively. Finally, the most important contribution of our Gaussian Process model and residual-learning method is the use of 2D connectivity matrix based fingerprints which allow us to efficiently make prediction for different surface sites without the need to specify the 3D atomic coordinate information. As a result, with just ~400–500 data points we can accurately predict the adsorption energies on all the unique surface sites of all the 2035 surfaces considered in this study, which means computational efforts required for exhaustive exploration can be reduced by a factor ~40. Though this approach seems promising for low index bi- or multi-metallic catalyst surfaces; however, application of this approach to other complex catalysis phenomena, such as low index single-atom catalysis, reaction condition effects, multidentate adsorbates, etc., might be quite challenging and will require design of new features in the Gaussian Process Regression to capture those effects.

Next, we analyze the feasibility of such residual-learning technique to improve the performance of Bayesian model

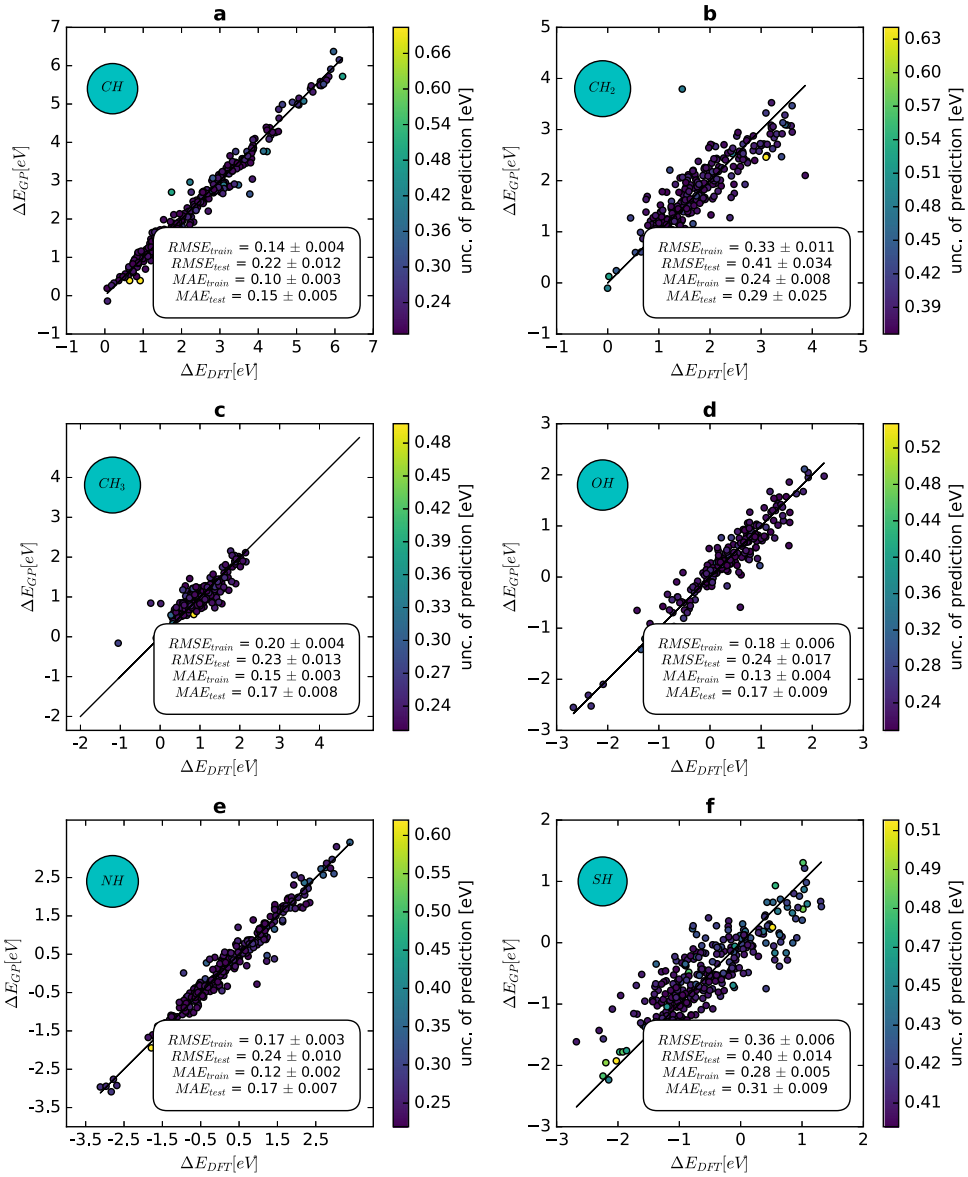

Fig. 5 Parity plot showing the Gaussian Process predicted Chemisorption energies of AH_x plotted against the DFT-computed chemisorption energies for the testing set. a CH *, **b** CH_2 *, **c** CH_3 *, **d** OH *, **e** NH *, and **f** SH *. We show the uncertainty in the prediction using a color bar shown in the right bar of each plot.

selection approach. In our analysis, when all the data are considered, we found the *RMSE* for the testing set prediction for CH, CH_2, CH_3, OH, NH, and SH to be 0.09, 0.14, 0.13, 0.17, and 0.27 *eV*, respectively, in our residual-learning approach with Bayesian model selection as the baseline model. For CH, CH_3, OH, NH, and SH, the performance gain of the residual learning with Bayesian model selection is impressive. For CH_2, the performance gain of the residual learning is not as excellent as we would expect, mainly due to the insufficient amount of training data available for model generation. For the selected model, our database contains only a limited amount of data for CH_2. Only when we have sufficient data available to train the model, we see significant improvement (0.05–0.1 eV) of residual learning over the model selection or other schemes. In Fig. 7, we summarize our results for different schemes when all the data available are used to make models. For all adsorbates, both Bayesian model selection and residual learning with Bayesian model selection perform significantly well in comparison to the other competing approaches.

In this article, we demonstrate a Bayesian framework for the model selection and model averaging for efficient and robust prediction of chemisorption energies of a few important multi-atom mono-dentate adsorbates on a bimetallic alloy dataset with the goal of novel catalytic materials discovery for hydrocarbon or nitrogen containing chemical reaction processes, such as, methanation, Fischer–Tropsch synthesis, and ammonia synthesis. When we have access to a heterogeneous sparse dataset, our analysis suggests Bayesian model averaging scheme—averaging over models rather than taking the best model—to be a reliable method to estimate the chemisorption energy. The model test *RMSE* varies from 0.15–0.4 eV with only 20 data points used and 0.10–0.30 eV for 80 data points used to build the models (except for *SH*). For a dense database, Bayesian model selection approach is the best scheme to select a single best model to make future prediction, provided we have sufficient data to build the models—i.e., model uncertainty is not significant. Based on the analysis of Bayesian model selection and averaging scheme, we conclude that it is possible to build accurate machine-learning models just

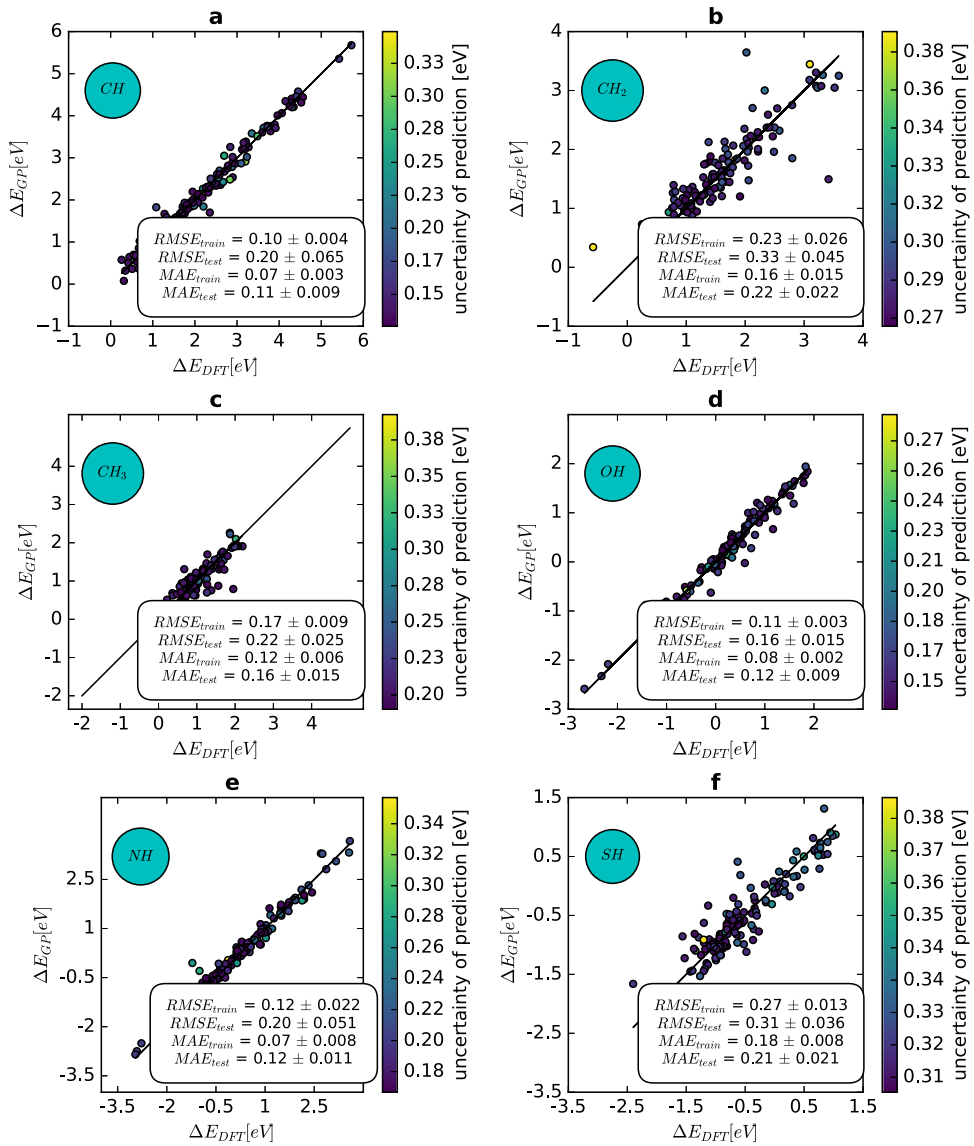

Fig. 6 Parity plot showing the residual learning using scaling relation model predicted Chemisorption energies of AH_x plotted against the DFT-computed chemisorption energies for the testing set. a CH *, **b** CH_2 *, **c** CH_3 *, **d** OH *, **e** NH *, and **f** SH *. We show the uncertainty in the prediction using a color bar shown in the right bar of each plot.

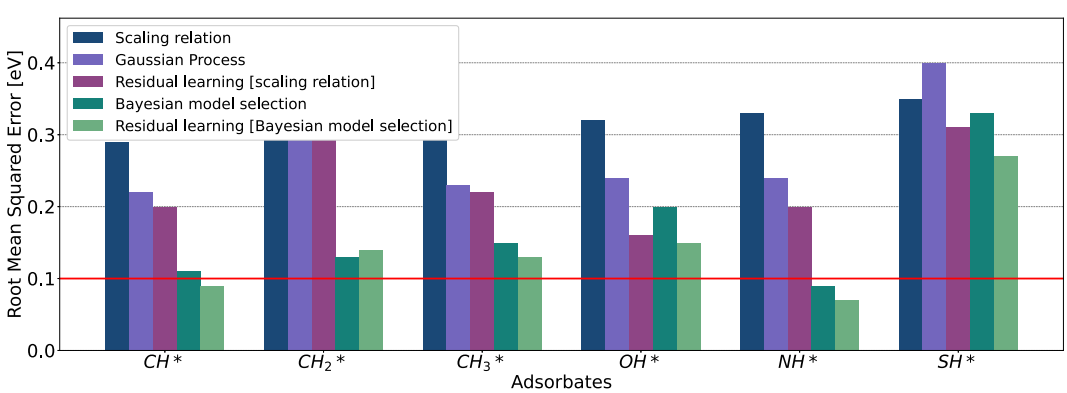

Fig. 7 Root mean squared error (RMSE) in eV for the holdout test set for different methods. RMSE for scaling relations, Gaussian process regression, residual learning with scaling relation as the baseline model, Bayesian model selection, and residual learning with Bayesian model selection as the baseline model, respectively. We add a horizontal guiding line at 0.1 eV to show the error with respect to the commonly accepted DFT error.

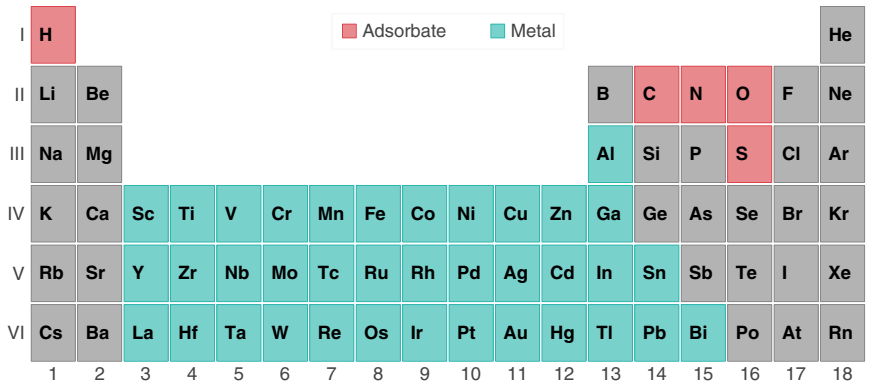

Fig. 8 The periodic table showing the 37 transition metals and 5 adsorbates used in this study. Elements highlighted in light blue color indicates 37 transition metals that were used to form A_1, $L1_0$, and $L1_2$ alloys and pink color indicates the atoms used as the adsorbate.

by developing linear regression models from the meaningful subsets of data. Furthermore, Gaussian process based regression model can be used to further improve the model fidelity. Specifically, we show that with only 2*D* connectivity information we can build very accurate machine-learning model to predict the residual of the linear models. We show that residual learning can improve the overall predictive performance by 0.05–1.0 eV depending on the adsorbates. This framework is particularly aimed at making reliable estimate of chemical and physical properties for computational materials design by leveraging databases containing varying amount of data of different systems such that the whole space of interest can be explored efficiently and fast. By sharing our code and data through our cloud server, we aim to accelerate the high-throughput exploration of catalytic materials.

METHODS

We use a recently published dataset for the chemical adsorption of monoatomic and hydrogenated species on bimetallic alloy surfaces, publicly available in the open-source computational database Catalysis-Hub.org[21,22]. In Fig. 8, we show the 37 metals used in this study to enumerate the 2035 surfaces (1332 $L1_2$ structures, 666 $L1_0$ structures-, and 37 A_1 structures) along with the five adsorbate elements considered in this study. All the first principles calculations for adsorption energy computation were performed using the ASE[35] interface to Quantum Espresso software package[36] with the BEEF-vdw exchange-correlation functional[7]. The surfaces were cleaved as three layers *fcc*(111) structures for A_1 and $L1_2$ metals and *fcc*(101) structures for $L1_0$ metals. The topmost layer was relaxed while bottom two layers were kept fixed until the forces were converged to 0.05 eV/Å in all directions of the relaxed atoms, including the adsorbate. Structure generation, job submissions and data collections were managed using a high-throughput framework available in the catalytic research toolkit *CatKit*[37]. We use the automatic structure generation submodule *CatGen* which automatically builds the slab from bulk structure, then identifies all the unique surface sites available for surface adsorption, and finally returns all the enumerated structure. In our *CatKit* enumeration, we found 4,10, and 9 unique adsorption sites for A_1, $L1_2$, and $L1_0$ surfaces, respectively. Next, we use the *CatFlow* submodule, which provides an interface to *Fireworks*[38] to automatically submit and manage our throughput computational workflow between users in various supercomputer facilities. After the jobs are finished, calculation details were stored in a centralized MongoDB database for future data collection and analysis. Later, we used a customized Python script to filter out the reconstructed surfaces—i.e., horizontal sliding, or atom dissociation from top layers, and to identify the final adsorption site after relaxation. The database contains 8856, 5487, 7457, 5690, 7556, 1508, 1458, 1285, 1235, 1743-, and 1559 adsorption energies for H, C, N, O, S, CH, CH_2, CH_3, OH, SH, and NH, respectively.

DATA AVAILABILITY

The datasets used to generate the results in this work are available at https://www.catalysis-hub.org/publications/MamunHighT2019.

CODE AVAILABILITY

A code to implement the Bayesian framework is available at https://github.com/mamunm/BayesianFramework.

Received: 1 July 2020; Accepted: 26 October 2020;
Published online: 23 November 2020

REFERENCES

1. Medford, A. J., Kunz, M. R., Ewing, S. M., Borders, T. & Fushimi, R. Extracting knowledge from data through catalysis informatics. *ACS Catal.* **8**, 7403–7429 (2018).
2. Mamun, O., Walker, E., Faheem, M., Bond, J. Q. & Heyden, A. Theoretical investigation of the hydrodeoxygenation of levulinic acid to γ-valerolactone over Ru (0001). *ACS Catal.* **7**, 215–228 (2017).
3. Abild-Pedersen, F. et al. Scaling properties of adsorption energies for hydrogen-containing molecules on transition-metal surfaces. *Phys. Rev. Lett.* **99**, 16105 (2007).
4. Ulissi, Z. W., Medford, A. J., Bligaard, T. & Nørskov, J. K. To address surface reaction network complexity using scaling relations machine learning and DFT calculations. *Nat. Commun.* **8**, 14621 (2017).
5. Walker, E., Ammal, S. C., Terejanu, G. A. & Heyden, A. Uncertainty quantification framework applied to the water–gas shift reaction over Pt-based catalysts. *J. Phys. Chem. C* **120**, 10328–10339 (2016).
6. Döpking, S. et al. Addressing global uncertainty and sensitivity in first-principles based microkinetic models by an adaptive sparse grid approach. *J. Chem. Phys.* **148**, 34102 (2018).
7. Wellendorff, J. et al. Density functionals for surface science: exchange-correlation model development with Bayesian error estimation. *Phys. Rev. B* **85**, 235149 (2012).
8. Mallikarjun Sharada, S., Bligaard, T., Luntz, A. C., Kroes, G.-J. & Nørskov, J. K. SBH10: a benchmark database of barrier heights on transition metal surfaces. *J. Phys. Chem. C* **121**, 19807–19815 (2017).
9. Nørskov, J. K., Abild-Pedersen, F., Studt, F. & Bligaard, T. Density functional theory in surface chemistry and catalysis. *Proc. Natl. Acad. Sci. USA* **108**, 937 LP–943 (2011).
10. Greeley, J. et al. Alloys of platinum and early transition metals as oxygen reduction electrocatalysts. *Nat. Chem.* **1**, 552–556 (2009).
11. Hammer, B. & Norskov, J. K. Why gold is the noblest of all the metals. *Nature* **376**, 238–240 (1995).
12. Hammer, B. & Nørskov, J. K. In *Impact of Surface Science on Catalysis* vol. 45, pp. 71–129 (Academic Press, 2000).
13. Hammer, B. & Nørskov, J. K. Electronic factors determining the reactivity of metal surfaces. *Surf. Sci.* **343**, 211–220 (1995).
14. Jacobsen, C. J. H. et al. Catalyst design by interpolation in the periodic table: bimetallic ammonia synthesis catalysts. *J. Am. Chem. Soc.* **123**, 8404–8405 (2001).
15. Himanen, L., Geurts, A., Foster, A. S. & Rinke, P. Data-driven materials science: status, challenges, and perspectives. *Adv. Sci.* **6**, 1900808 (2019).
16. Curtarolo, S. et al. The high-throughput highway to computational materials design. *Nat. Mater.* **12**, 191–201 (2013).
17. Landis, D. D. et al. The computational materials repository. *Comput. Sci. Eng.* **14**, 51–57 (2012).

18. Kirklin, S. et al. The open quantum materials database (OQMD): assessing the accuracy of DFT formation energies. *npj Comput. Mater.* **1**, 15010 (2015).
19. Jain, A. et al. Commentary: The Materials Project: a materials genome approach to accelerating materials innovation. *APL Mater* **1**, 11002 (2013).
20. Curtarolo, S. et al. AFLOW: an automatic framework for high-throughput materials discovery. *Comput. Mater. Sci.* **58**, 218–226 (2012).
21. Winther, K. T. et al. Catalysis-Hub.org, an open electronic structure database for surface reactions. *Sci. Data* **6**, 75 (2019).
22. Mamun, O., Winther, K. T., Boes, J. R. & Bligaard, T. High-throughput calculations of catalytic properties of bimetallic alloy surfaces. *Sci. Data* **6**, 76 (2019).
23. Ulissi, Z. W. et al. Machine-learning methods enable exhaustive searches for active bimetallic facets and reveal active site motifs for CO2 reduction. *ACS Catal.* **7**, 6600–6608 (2017).
24. Back, S. et al. Convolutional neural network of atomic surface structures to predict binding energies for high-throughput screening of catalysts. *J. Phys. Chem. Lett.* **10**, 4401–4408 (2019).
25. Li, Z., Wang, S. & Xin, H. Toward artificial intelligence in catalysis. *Nat. Catal.* **1**, 641–642 (2018).
26. Ma, X., Li, Z., Achenie, L. E. K. & Xin, H. Machine-learning-augmented chemisorption model for CO2 electroreduction catalyst screening. *J. Phys. Chem. Lett.* **6**, 3528–3533 (2015).
27. Wolpert, D. H. & Macready, W. G. No free lunch theorems for optimization. *IEEE Trans. Evol. Comput.* **1**, 67–82 (1997).
28. Wasserman, L. Bayesian model selection and model averaging. *J. Math. Psychol.* **44**, 92–107 (2000).
29. Schwarz, G. Estimating the dimension of a model. *Ann. Stat.* **6**, 461–464 (1978).
30. Neath, A. A. & Cavanaugh, J. E. The Bayesian information criterion: background, derivation, and applications. *WIREs Comput. Stat* **4**, 199–203 (2012).
31. Hoeting, J. A., Madigan, D., Raftery, A. E. & Volinsky, C. T. Bayesian model averaging: a tutorial (with comments by M. Clyde, David Draper and E. I. George, and a rejoinder by the authors. *Stat. Sci.* **14**, 382–417 (1999).
32. Hansen, M. H. et al. An atomistic machine learning package for surface science and catalysis. *arXiv*. Preprint at arXiv1904.00904 (2019).
33. Williams, C. K. I. & Rasmussen, C. E. Gaussian processes for regression. In *Advances in Neural Information Processing Systems*. 514–520 (1996).
34. Pedregosa, F. et al. Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
35. Hjorth Larsen, A. et al. The atomic simulation environment—a Python library for working with atoms. *J. Phys. Condens. Matter* **29**, 273002 (2017).
36. Giannozzi, P. et al. QUANTUM ESPRESSO: a modular and open-source software project for quantum simulations of materials. *J. Phys. Condens. Matter* **21**, 395502 (2009).
37. Boes, J. R., Mamun, O., Winther, K. & Bligaard, T. Graph theory approach to high-throughput surface adsorption structure generation. *J. Phys. Chem. A* **123**, 2281–2285 (2019).
38. Jain, A. et al. FireWorks: a dynamic workflow system designed for high-throughput applications. *Concurr. Comput. Pract. Exp.* **27**, 5037–5059 (2015).

ACKNOWLEDGEMENTS

This research was supported by the U.S. Department of Energy, Office of Science, Office of Basic Energy Sciences, Chemical Sciences, Geosciences, and Biosciences Division, Catalysis Science Program to the SUNCAT Center for Interface Science and Catalysis.

AUTHOR CONTRIBUTIONS

O.M. and T.B. conceived the study. J.B., K.W., and O.M. computed and compiled the dataset. O.M. performed machine-learning training. O.M., K.W., and J.B. analyzed the data. O.M. drafted the manuscript. T.B., K.W., and J.B. reviewed the manuscript.

COMPETING INTERESTS

The authors declare no competing interests.

ADDITIONAL INFORMATION

Supplementary information is available for this paper at https://doi.org/10.1038/s41524-020-00447-8.

Correspondence and requests for materials should be addressed to O.M. or T.B.