THE JOURNAL OF PHYSICAL CHEMISTRY C

pubs.acs.org/JPCC Article

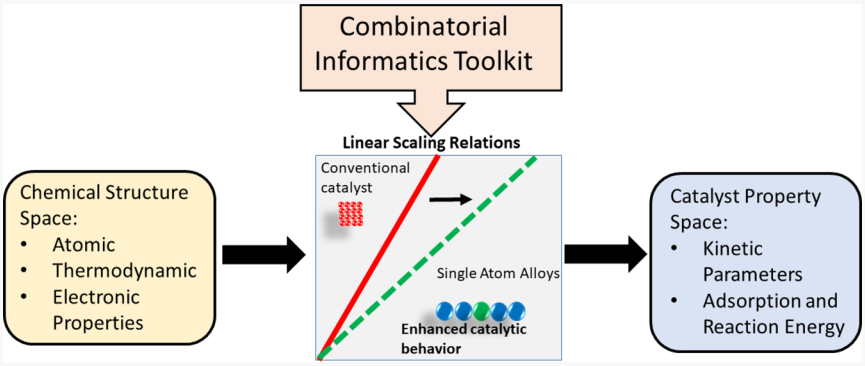
Machine Learning-Aided Identification of Single Atom Alloy Catalysts

Published as part of The Journal of Physical Chemistry virtual special issue "Machine Learning in Physical Chemistry".

Aparajita Dasgupta, Yingjie Gao, Scott R. Broderick, E. Bruce Pitman, and Krishna Rajan*

Cite This: *J. Phys. Chem. C* 2020, 124, 14158–14166 Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information



ABSTRACT: In metal catalytic design, there is a well-established linear scaling relationship between reaction and adsorption energies. However, owing to the challenges of performing experimental and/or computational experiments, there is a paucity of empirical data regarding these systems. In particular, there is little experimental evidence suggesting how the linear scaling law might be overcome in order to discover catalysts with more desirable properties. In this paper, we employ machine-learning techniques in order to predict reaction and adsorption energies for 300 hypothetical binary compounds. We then apply outlier detection methods to identify which of these predicted compounds do not follow the known scaling law. These outlier compounds, which would not have been identified through traditional design rules, are the most likely to have unexpected and potentially transformative catalytic behavior. Thus, this paper proposes a data-driven screening methodology to identify those metallic compounds (as a function of gaseous environment) which are most likely to have targeted catalytic behavior.

INTRODUCTION

Building upon the Sabatier optimization principle introduced over a century ago,[1] Nørskov and co-workers have elucidated a scaling law between reaction and adsorption energies to predict the catalytic behavior for new chemistries.[2,3] These studies use the catalytic design concept of the "volcano plot", which describes a relationship between the catalytic behavior and binding energy of a catalyst. Generally, the reactions being studied can be described in terms of a single descriptor such as the adsorption energy of a simpler reaction intermediate. Consequently, the volcano plot greatly reduces the number of catalytic descriptors which need to be defined for a given reaction. Certain notable exceptions exist. For example, in an oxygen reduction reaction the scaling relationship is described using the binding energy of the intermediates OH and OOH. In these cases, a three-dimensional volcano plot is used.[4] This relationship has been used extensively in a wide variety of complex reaction networks.[5]

Other relations provide information on energetics. For example, the Brønsted–Evans–Polanyi relationship implies a linear relationship between the activation barrier for a given reaction and binding energies of a few key intermediates.[6] Similarly, the binding energy can be calculated using the d-band model, which establishes a relationship between the d-band center of a catalytic metal and its adsorption strength.[7] However, even employing these relationships, we have at our disposal only a small set of chemistries for which we have appropriate descriptors.

Received: February 20, 2020
Revised: May 9, 2020
Published: May 14, 2020

https://dx.doi.org/10.1021/acs.jpcc.0c01492
J. Phys. Chem. C 2020, 124, 14158–14166

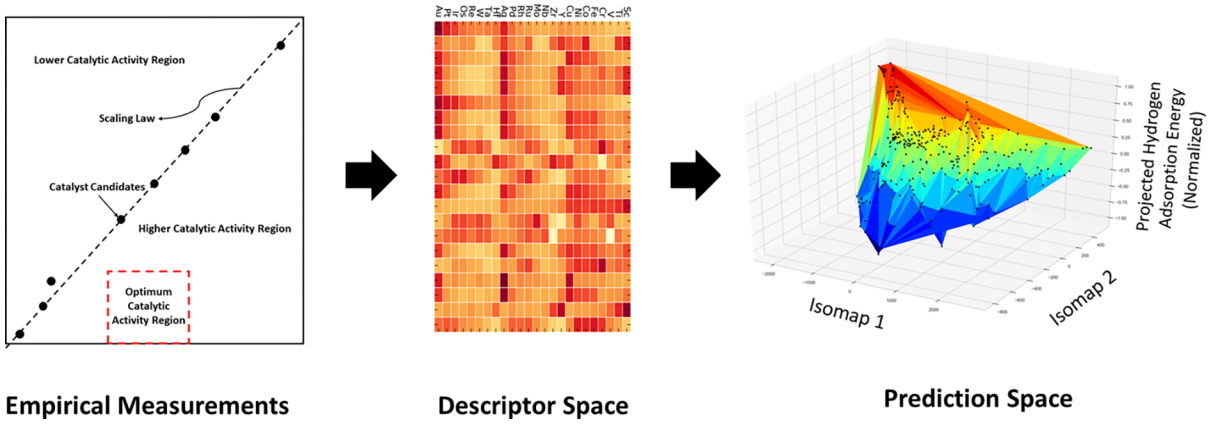


Figure 1. Logic of the approach to identify binary chemistries which have the most promising catalytic properties. We assume the known scaling laws of single-element catalysts and convert this relationship to a descriptor space where we describe the elements based on their characteristics (size, charge, bonding, etc.). Using this descriptor space, extended for binary compounds and coupled with a hybrid informatics approach, we predict the catalytic properties of the alloys. Tracking those binary alloys which are outliers relative to this scaling relationship then identifies the compounds which have the greatest potential for transformative performance (i.e., those compounds which have high kinetic reactivity in the left image).

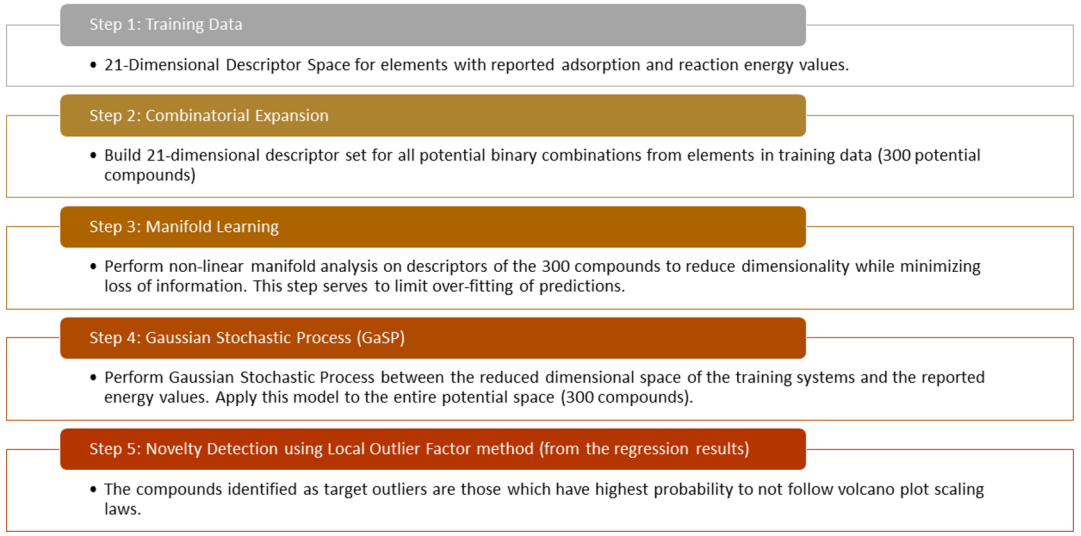


Figure 2. Description of the approach developed here for identifying alloys which have the potential to break from known scaling laws. We first collect known empirical measurements and break them down into basic descriptors. Next, assuming physics is represented by the known scaling laws, we expand these descriptors into a "virtual" material space and then follow a hybrid ML approach which predicts the catalytic properties for this "virtual" space. The final stage uses outlier detection to identify the most promising alloys, namely, those compounds which do not follow the scaling laws. This logic provides a data-driven approach to catalysis screening, which we demonstrate here for catalytic alloys. This methodology is transferable to all chemical design issues where limited data exists.

A different effect was introduced by Darby et al., who proposed the idea of the "single-atom alloy" (SAA). This notion is based on the ensemble effects that come into play due to the low concentration doping of active transition metals onto inert coinage metals. Darby et al. recognized that with "...alloys one can access areas in the "volcano plot" that would be inaccessible for pure metals, thereby developing catalysts with improved performance...".[8] In a later paper, they remarked that "...to date, only a small subset of SAAs has been synthesized experimentally and it is unclear which metallic combinations may best catalyze which chemical reactions."[9]

Recent studies have used machine-learning (ML) methods to screen the large combinatorial landscapes that must be searched to discover new compounds within the catalyst domain.[10−12] SAAs are materials of particular interest due to the wide range of applications and high reactivities they have been found to display. The primary objective in these studies is to circumvent the need for time-consuming and computationally expensive DFT calculations or experiments. Thus, ML-based approaches have been applied to microkinetic-model-based reaction networks as well as existent DFT calculations.[13] Descriptor-based approaches typically operate on the principle that the property of a compound is directly associated with the properties of the individual constituent elements.[14,15] Scaling laws and properties calculated using DFT[16] are expressed as combinations of descriptors or intermediates for input into ML based regression models.[17,18] Although key advances have been made, the legacy data that exists remains largely unexplored, with minimal knowledge regarding key considerations such as thermodynamic stability and the processing plausibility of these compounds.[19]

https://dx.doi.org/10.1021/acs.jpcc.0c01492
J. Phys. Chem. C 2020, 124, 14158−14166

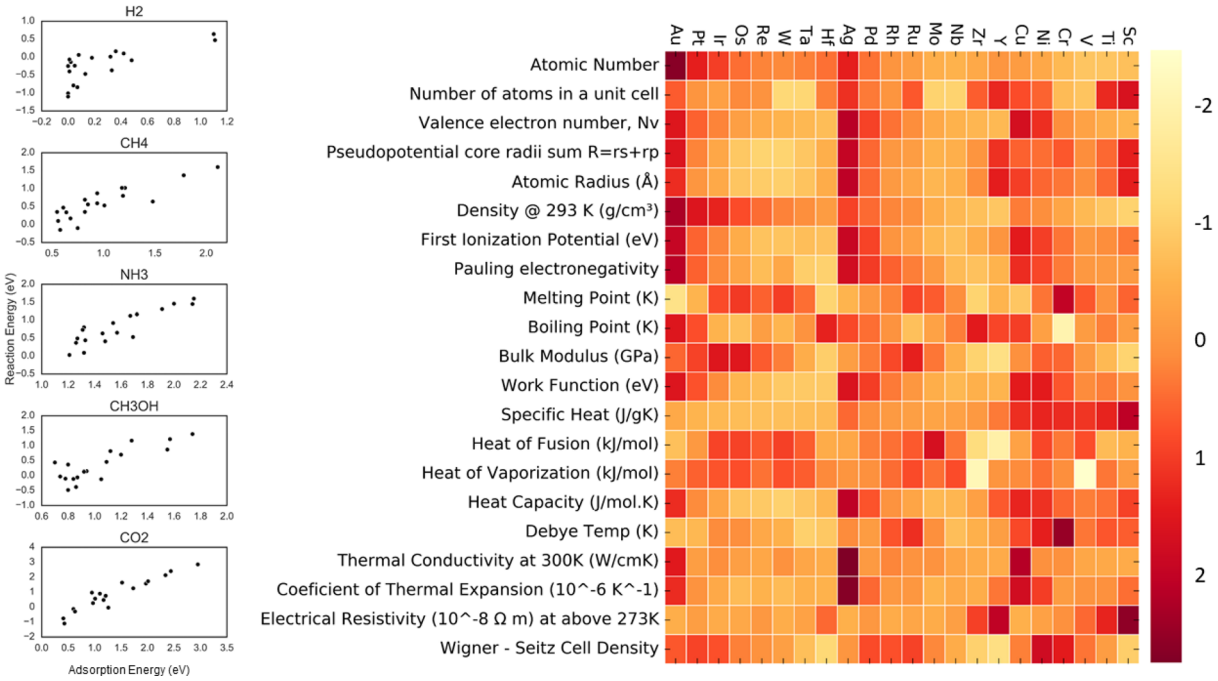


Figure 3. Training data space. The left figures show the empirical reaction and adsorption energies (in eV) for single-metal systems and previously calculated SAA systems for the (111) surface facet, which serve as our training (target) data. Of particular note is the relationship between the properties; the relationship has a roughly linear relationship which provides the starting point of our analysis. The right-hand figure is our conversion of these elements into a normalized descriptor space (this panel labels all of the input descriptors in the analysis described in the following sections). That is, the elements on the left are described by properties; however, in this paper we seek to understand how the underlying characteristics of the materials (i.e., parameters representing the elemental contributions) contribute to the properties. This is governed by the logic that if we can understand how the characteristics of the elements contribute to the catalytic properties then we can develop models which are based on the governing physics and are not solely statistical correlations.

These studies provide the motivation for our data-driven approach to binary catalysts. We break the limitations imposed by a small set of experimental measurements, all of which follow the known scaling law described by Nørskov.[2] Indeed, we assume that the scaling law does exist, and we model binary systems assuming this relationship and the associated physics. However, we introduce a new twist, by tracking the outlier behaviors of chemistries relative to this scaling law. That is, chemistries which we identify as not following the scaling law are those which are most likely to have improved performance, as defined by Darby et al.[9] Using this ML approach to identify important features of the binary compounds, we reduce the chemical search space by 2 orders of magnitude. The logic of our approach is laid out in Figure 1, and the rest of the paper follows this logic.

METHODS

Our methodology integrates several ML techniques including nonlinear manifold analysis to identify the most important dimensions in the high-dimensional descriptor space, graph theory to define close connections among the various binary compounds, Gaussian process regression which provides a nonparametric method for interpolating a quantity of interest, and local outlier factor analysis. This approach is outlined in Figure 2. We begin with a 21-dimensional space of descriptors for each of the 19 metal/SAA catalysts. In prior work, we discussed this descriptor space as it represents the underlying physics of material properties[20,21] and described approaches for scaling these descriptors.[22−26] In the current work, scaling and normalization were performed on the entire descriptor space. For simulating the expanded virtual library of compounds, the descriptors of the imagined compounds were calculated with the assumption of an equimolar concentration of both the dopant atom and metal host for each property. This was done to assign equal weight to both site types and not introduce an inherent bias based on the site type.

Input Data. Owing to the difficulty of experimental generation of SAAs,[9] legacy data within this class of materials is extremely limited, resulting in a high-dimensional but sparse data set. That is, although the number of descriptors used to study catalytic reactions spans a high-dimensional data space akin to the large number of factors that contribute to reaction kinetics, the number of sample points within this space is small. This "curse of dimensionality" is a challenge to optimization and prediction, for which a large data set is required in order to stabilize any model predictions.[27]

As an initial data set, we used the data described by Darby et al. containing details of the spillover energy associated with two site types, activation energies, and reaction energies for 7 elements and 12 SAAs for 5 adsorbates: H_2, CO_2, CH_3OH, NH_3, and CH_4[9] (i.e., 19 data points per adsorbate). Each of these 19 data points correspond to the (111) surface facet, and all predictions made within our model also correspond to the (111) surface facet. These energies were calculated using periodic plane wave DFT calculations using the Vienna ab initio Simulation Package (VASP). Details of the calculations are described in Darby et al. The activation energy is consequently derived as the difference in the DFT total energy in the transition state and initial state of the reaction, whereas the reaction energy is derived as the difference in total energy between the final and initial states of the reaction. Reactions

https://dx.doi.org/10.1021/acs.jpcc.0c01492
J. Phys. Chem. C 2020, 124, 14158−14166

studied are the dissociation of the adsorbate on the metal/SAA alloy substrate with the elementary steps being an initial state, transition state and final state of bond dissociation.[8]

Because we have a relatively small set of data residing in a large descriptor space (Figure 3), we need to reduce the descriptor dimensionality in order to ensure that we do not overfit the data. The challenge is that if we remove descriptors then we also remove data from our descriptor space. For this reason, we employ a nonlinear manifold analysis on our descriptor space, which reduces the dimensionality with minimal loss of information. The result from this dimensionality reduction provides a simpler parametrization of the data. This lower-dimensional data description serves as an input into the Gaussian process regression. The following subsections describe the relevant mathematics.

IsoMap for Parametrization via Nonlinear Manifold Analysis. Through the application of the IsoMap algorithm[28] on the descriptor space and the potential 300 chemistry space (21 descriptors/dimensions and 300 chemistries) of the virtual combinatorial library of hypothetical SAA alloys, we are able to develop a parametrization of the data space. That is, every compound is described as a function of a few IsoMap dimensions with minimal loss in information encoded in the data. This step is critical because otherwise we must either (i) make a significant and not necessarily justified assumption regarding the importance (i.e., bias) in the underlying data structure or (ii) confront the likely overfitting of the data.

IsoMap is a nonlinear dimensionality reduction method; thus, a nonlinear analogue to principal component analysis. In addition to reducing the dimensionality of the data, IsoMap determines the nearest neighbors of every data point, thereby preserving the local geometry of the data.

A parametrization of data in the first two IsoMap dimensions is shown in Figure 4. This shows the distribution of data, demonstrating a distribution which is not overly impacted by outliers. This parametrization will serve as the input into a Gaussian process regression. That is, the parameters in this figure serve as the inputs for the training reaction through the adsorption energies.

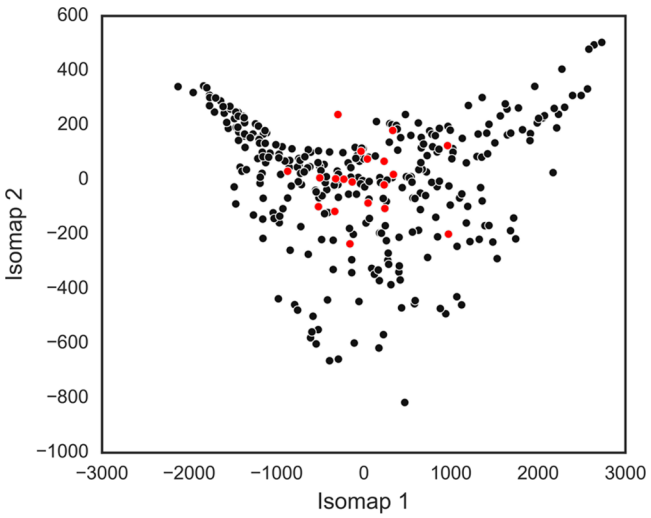


Figure 4. Development of reduced descriptor space. The red circles are the training systems from step 1, and the black circles are the chemistries from step 2. We use these two dimensions as the input into the ML model. By performing this step, we reduce the overfitting of the predictions while maintaining the majority of the physics of the input descriptor space.

A required input for the IsoMap is to prescribe the desired number of nearest neighbors; one can attempt to find a near optimal value of this nearest-neighbor number. The number of nearest neighbors was calculated based on a derivation of the proposed protocol by Samko et al.[29] The optimal number of neighbors was assumed to be the point at which reconstruction error is minimized in each case. For the large data set (i.e., the combinatorial expansion of the training data), 7 was determined to be the "optimal" number of neighbors. For the smaller training data set, the optimal number was calculated to be 11. The computed transformed data was then used to calculate and predict adsorption and reaction energies using Gaussian process regression.

Gaussian Process Regression for Predicting Catalytic Properties Assuming Scaling Laws. A Gaussian stochastic process (GaSP) is a nonparametric methodology for regressing data points. The data are assumed to be the realizations of a Gaussian process, and the posterior distribution of the output variable over the input parameters μ, conditioned on the data. Thus, the output is

$$y \sim N(\psi, \sigma^2 R)$$

where ψ is an assumed mean trend in the data and R is a correlation matrix giving the relation between the outputs for inputs i and j, often assumed as $R = (r_{ij}) = \exp(-(1/\gamma)(\mu_i - \mu_j)^2)$. ψ and the correlation length γ are determined from the data.[30]

Notably, the objective of this regression step is not to identify chemistries with target adsorption and reaction energies but rather to develop a data set which captures the relationship between the chemical descriptors and the catalytic reactions. That is, the regression is an intermediate step within our larger developed framework for outlier detection.

Local Outlier Factor Analysis for Identifying the Chemistries Which Do Not Follow Scaling Laws. Local outlier factor (LOF) analysis computes a score that indicates how likely a specified data point (say z) is to be an anomaly or outlier. The local outlier factor examines the density of near neighbors of z and compares this to the neighbor density of other not-too-distant points. Specifically, the k-distance is defined as the distance from z to its kth nearest neighbor. Next, the average distance from z to all these k near neighbors is determined. Finally, this average distance is compared to the average distance of all the k neighbors.[31] If z has a much larger average distance than its neighbors, then it is more likely to be an outlier.

Assessing the Correlation of Descriptors with Adsorption and Reaction Energies. Our current computational framework uses the IsoMap algorithm to reduce the dimensionality of the feature space before applying the Gaussian process regression algorithm to determine the predictive surface of the adsorption and reaction energies for a given adsorbate. Due to the nonlinear manifold projection of the IsoMap algorithm, it is not possible to assess an interpretation of the physical meaning of each IsoMap dimension in terms of the initial descriptors used. In order to assess the correlation of the descriptors with the adsorption and reaction energies of each of the adsorbates, a principal component analysis (PCA) was performed. Since PCA is a linear method, it is far easier to track the distribution of the initial data space descriptors into the orthogonal components

https://dx.doi.org/10.1021/acs.jpcc.0c01492
J. Phys. Chem. C 2020, 124, 14158−14166

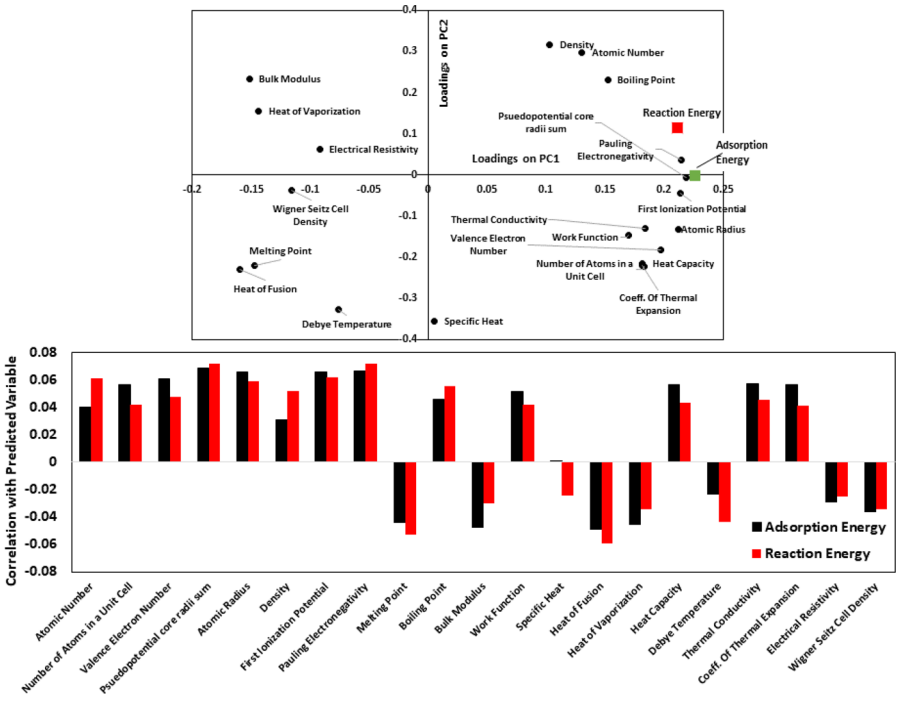


Figure 5. (top) Loadings plot for PCA and (bottom) assessed descriptor correlation with adsorption energy and reaction energy. We observe that pseudopotential core radii sum and the Pauling electronegativity are the most correlated features with the adsorption energy and reaction energy indicating that a consideration of the electronic structure is necessary when designing SAA catalysts. Furthermore, we observe that the other variables also seem to play an important role in the design of new catalysts, thus confirming our initial assumption of this being a high-dimensional data space. Thus, our approach to initialize the data space by projecting the variables within a lower dimensional nonlinear manifold helps us to preserve the maximum amount of information while significantly reducing the dimensionality.

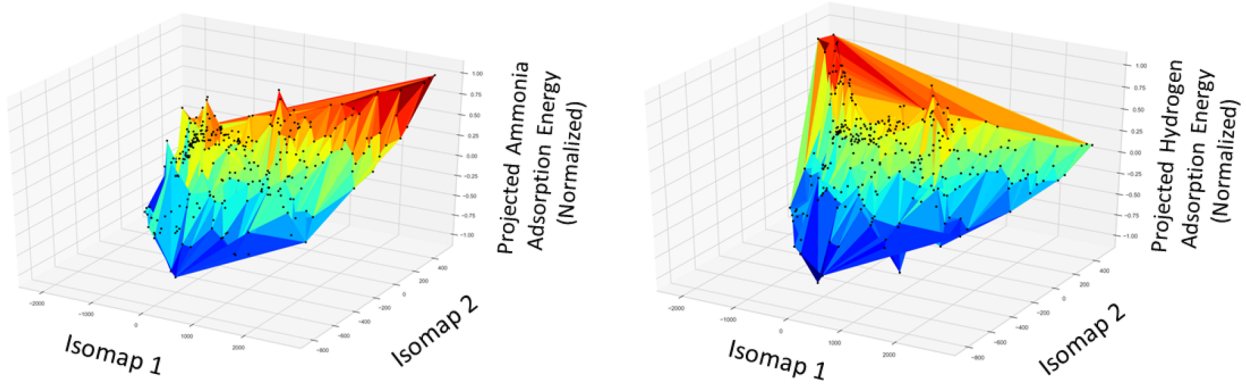


Figure 6. Prediction of adsorption energies for two different environments, (left) ammonia and (right) hydrogen, as a function of the IsoMap parameters. The points represent the 300 potential chemistries (the data values and compound names are provided in the Supporting Information). The coloring corresponds with the predicted energy values and is a visual aid to assess the data landscape. This figure serves to highlight that we have developed a dimensionally reduced set of parameters (i.e., the IsoMap parameters) which can be used for modeling different types of reactions. Beyond representing the relationship between model input and output, this figure also demonstrates the impact on the outliers on defining the data landscapes. The approach for capturing the outliers is discussed in the following sections.

of the PCA. We discuss the underlying mathematical concepts and the physical meaning of the obtained results within this section.

Mathematically, PCA is an eigenvector decomposition of the covariance of a matrix X with n samples and N features such that $\mathrm{cov}(X) = X^T X/n - 1$. Thus, it is possible to decompose the data matrix X into the sum of the outer product of principal components/score vectors (t_i) and loadings vector (p_i) plus a residual matrix (E) such that

$$X = t_1 p_1^T + t_2 p_2^T + \ldots + t_A p_A^T + E$$

Generally, the decomposition of X into the first few orthonormal prinicipal components is sufficient to explain the data set with minimal loss of information.[32,33] In the current analysis, the correlation of the descriptors with each of the adsorption and reaction energies was assessed as the weighted sum of the relative weighting of the loads of the descriptors and the percentage of variance explained by each principal component. Figure 5 depicts the loads plot and the correlation of descriptors with the adsorption energy and reaction energy of ammonia.

PCA was performed on the entire data set (i.e., 21 descriptors and 10 predicted variables, 5 adsorption energies,

https://dx.doi.org/10.1021/acs.jpcc.0c01492
J. Phys. Chem. C 2020, 124, 14158−14166

and 5 reaction energies). The first five principal components (explaining 96% of the data set variance) were used to assess the correlation of descriptors. The first principal component explained 60%, and the second principal component explained 23%. From the variable importance projection (VIP), we find that the two most critical among the numerous descriptors having some significance are the pseudopotential radii sum and the Pauling electronegativity. Interestingly, both of these descriptors are based on the electronic structure, with the descriptors describing the behavior of nonvalence electrons and then valence electrons, respectively. This highlights the complexity of the problem, where all aspects of the electronic structure require consideration. Although this analysis provides insight into what is governing the relationships we study here, an in-depth study of what is physically unique among the most promising compounds is beyond the scope of the paper. Our objective is to narrow the search space for chemistries which warrant an in-depth analysis, whether through DFT or experiments, and thereby provides guidance of future analyses.

RESULTS AND DISCUSSION

Results Based on Scaling Law Physics. In this section and the next, we primarily discuss results for ammonia reactions. However, the results for hydrogen, methane, methanol, and carbon dioxide reactions can be obtained by a similar analysis and are provided in the Supporting Information. As discussed earlier, our predictions are predicated on modeling the physics by the known scaling laws. The GaSP regression is based on predicting the measured reaction and adsorption energies as a function of the IsoMap parameters of the red points in Figure 4. That is, we have two models: one for reaction energy as a function of IsoMap parameters and a second for adsorption energy as a function of the same IsoMap parameters. On the basis of this regression, we can then input the IsoMap parameters for other chemistries (i.e., the black points in Figure 4) and predict the reaction and adsorption energies. Figure 6 shows the GaSP energy predictions. The Gaussian process regression surfaces were trained using the Dot Product[34] and White Noise kernel[35] for training and prediction. For the resultant surfaces, the coefficient of determination[36] was used to determine how many components of the IsoMap would be most suited for surface fitting. We see a relatively continuous surface, demonstrating that the input descriptors and the dimensionality reduction both represent the general catalytic information and physics. Of particular note, Figure 6 shows the predicted data landscape for two different reactions: ammonia and hydrogen. The results for the other reactions have been included in the Supporting Information. The x- and y-axes for all of the predicted surfaces (that is, the model input) are the same, highlighting that we are making predictions for different reactions with the same parameter set, even though the compounds with highest energy values are not the same across reactions.

Given that the empirical measurements were relatively small, our approach was developed to address a different set of challenges than in ML analyses which have large input data. To limit the overfitting of the data, the IsoMap approach was used to reduce the dimensionality to a minimum amount without bias, while the selection of descriptors and the multiple iterations of modeling were used to ensure consistency.

Building on our predicted energy landscapes, we seek to assess the divergence from the expected relationship between adsorption and reaction energies (Figure 7). The residual of this divergence is simply defined as the difference between the adsorption energy predicted from the Gaussian process regression and the adsorption energy based on the predicted reaction energy. A larger residual value indicates a chemistry which deviates more from a linear relationship between adsorption and reaction energies. The right panel of Figure 7 shows the residual distribution for the ammonia reaction. We see that the largest distribution is around residual value of 0, showing that we are indeed largely representing a linear relationship between adsorption and reaction energies. Those points with the largest magnitude residual values are those which are of principle interest for this paper, that is, those chemistries which are most likely not to follow the scaling laws and which have the highest likelihood for unique catalytic behavior. The assessment of these compounds is discussed in the next section.

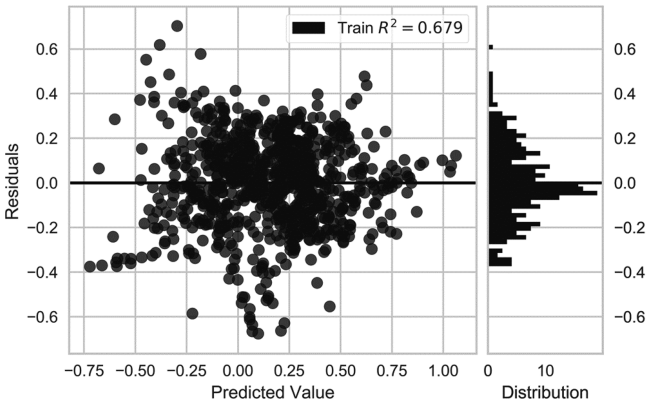


Figure 7. (Left) Predicted adsorption energy versus residuals for ammonia reactions. This shows a roughly linear relationship, as expected from the scaling law assumed in our models. (Right) Distribution of residual values for adsorption energy, where the residual is defined as the difference between the predicted adsorption energy via Gaussian process regression and the adsorption energy based on a purely linear relationship. Therefore, those chemistries with the highest magnitude residual are those which are most likely not to follow the previously described scaling laws used in catalytic design.

As an added point of interest, we compared the accuracy of the Gaussian process regression models (based on the coefficients of determination) with the residual distributions. Our reasoning was to test whether reactions with larger residual spreads might be due to greater inaccuracy in the model; this comparison is able to assess whether the outlier behaviors are likely real or solely resulting from the predictions. Interestingly, no systematic relationship between the residual spread and model accuracy was found. The prediction gave the lowest accuracy for a methanol reaction, which also had the least spread in residual values; in contrast, the most accurate model was for CH_4 reaction, which had the second smallest spread in residual values. This implies that the identification of outlier chemistries is governed by physics and not by data methodology.

Outlier Identification for Systems Least Likely to Follow Scaling Law. In this section, we perform novelty detection using the LOF method. The objective was to identify the chemistries which have the highest outlier behavior, determine a ranking of the outlier behavior, and assess the

https://dx.doi.org/10.1021/acs.jpcc.0c01492
J. Phys. Chem. C 2020, 124, 14158−14166

resulting distributions. We employ an LOF analysis, rather than simply examining the residual values determined in the previous section, in order to define outliers as those chemistries that do not follow the distribution with respect to the rest of the chemistries.

The result of the LOF analysis is shown in Figure 8, with the identification of outliers. This analysis identifies 29 of the 300

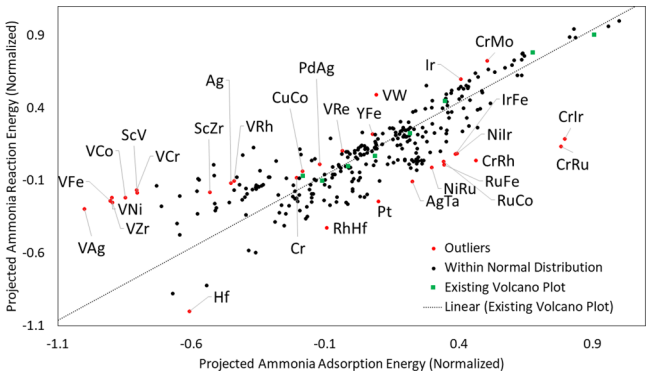


Figure 8. Predicted reaction and adsorption energies for ammonia reaction with the outliers identified via novelty detection method. The red points (which are also labeled) represent the novel cases, and the black points represent points belonging to the distribution. The green points represent the current known data points used in calculating the scaling law based on the volcano plot, and the dotted line represents the existing scaling law. The adsorption and reaction energies are shown here as normalized values for clarity. The outlier detection method operates on the relative values, and the normalization of values is to avoid any confusion due to regression extrapolation. From our starting search space of 300 chemistries, we identify the 29 chemistries which are most likely to not follow the known catalytic scaling laws and instead to operate as SAAs. The notation for the chemistries denotes the possibility of adding one element into the other and the possibility that some combination of these metal species as SAAs could provide enough of a perturbation within the catalyst domain so as to be able to break the scaling law. Thus, the order of element names is not specific. In terms of the LOF score values, CrIr and CrRu are the most prominent outliers. This result therefore identifies the compounds which are most likely not to follow known scaling laws and which traditional modeling approaches would not differentiate as unique.

chemistries as potential SAA chemistries which do not follow the expected scaling relationships. The objective of this stage is to identify those compounds which have unique physics that is not captured in the current input data space. A few pure metal systems have been identified as outliers. However, we note that this is a reflection of having few single-metal systems with that particular catalytic behavior; thus, the focus of this figure is instead on the labeled SAAs which have the potential to break the linear scaling laws.

It should be noted that we would not expect all of these compounds to have improved performance, but rather would expect some outliers to have particularly poor performance. Along this line and following the existing figure of merit from Figure 1, we anticipate the compounds falling below the relative trendline (i.e., with reaction energy below the trendline) would be the most promising, while those above the trendline would likely have poor performance. Thus, we have identified the compounds which would introduce additional physics into the models while also identifying those which would be most promising for further experimentation.

To track the distributions, a Q−Q plot for adsorption energy was considered (Figure 9). The ordering of points for the

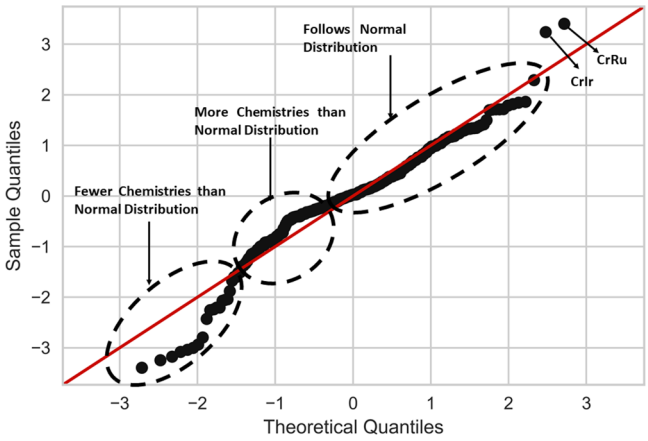


Figure 9. Q−Q plot for the predicted adsorption energies for ammonia reaction. The red line shows the normal distribution. From this, we find that the chemistries with lower adsorption energy are more likely not to follow the normal distribution (agreeing with our result that the majority of outliers were at negative adsorption energy). Additionally, CrIr and CrRu have the largest difference from the normal distribution.

sample quantiles corresponds with the ordering of predicted adsorption energy, and the theoretical quantiles are based assuming a normal distribution. From this, we identify the regions of the predictions which do not follow a normal distribution. The regions which do not follow closely with normal distribution correspond roughly with the chemistries with negative predicted adsorption energy, while those compounds with positive predicted values follow a normal distribution. Interestingly, this is consistent with the novelty detection analysis, where the outlier chemistries primarily have predicted negative adsorption energy. From the Q−Q plot, CrIr and CrRu were the two chemistries most not following the normal distribution, in agreement with the LOF analysis which identified those as having the largest LOF score value. Therefore, the analysis of the values relative to an expected distribution helps to explain the basis for defining the outlier chemistries.

The results for reactions other than ammonia are provided in the Supporting Information. We summarize the findings of these here. The other most promising SAA binary chemistries were found to be VCr and VZr for hydrogen, VW and AgAu for methane, ScTi and VAu for methanol reactions, and CuAg and NiAg for CO_2 reactions.

Feasibility and Synthesis of Proposed Candidates. While SAAs have been studied due to their potential catalytic applications and unique properties, synthesis of this class of compounds presents a significant challenge. Currently, two major routes of synthesis exist, each leading to a slightly different variant of the SAA. The first method is an effective doping wherein one atom is dispersed into another metal atom's nanocluster allowing for more delicate control within the microscopic material and an ability to accurately model the resulting compound and efficient characterizations using first-principles methods. Conversely, synthesis using this route has proved challenging and has been demonstrated in very limited

https://dx.doi.org/10.1021/acs.jpcc.0c01492
J. Phys. Chem. C 2020, 124, 14158−14166

applications in nanocrystals such as Au_{25}[37] and Ag_{25}.[38] Gold nanostructures are typically the preferred host and are the most studied for these types of synthesis routes. Generally, wet chemistry methods have been employed (e.g., the Brust method)[39] to accomplish the synthesis of this type of compound.

The second method constitutes atomically dispersing one metal into another metal's nanocrystals. This is usually achieved using surface chemistry methods. Multiple SAAs have been synthesized using this route including PtCu[40] and PdCu.[41,42] Within the current list of proposed SAAs for further investigation, we would suggest the use of the single-atom doping method for synthesis wherever it is feasible (for example in Cu-doped Ag). The second method can be utilized for the synthesis of Ag dispersed in Cu. Due to the very limited literature describing the synthesis of other SAAs, further investigations using model-based methods must be made before an attempt to synthesize some of the more unique outliers we have suggested in this work.

CONCLUSION

We have presented a computational framework for the investigation of possible combinations of SAAs for five adsorbates. From an analysis integrating nonlinear manifold learning, Gaussian process regression and novelty detection, we identified a limited number of "virtual" compounds which have the highest likelihood of not following established scaling laws. This approach introduces a new framework for significantly shrinking the chemical search space, even in cases where limited data is available, and provides a guide for the next series of experiments to perform.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jpcc.0c01492.

Results of regression analysis and novelty detection algorithms (PDF)

AUTHOR INFORMATION

Corresponding Author

Krishna Rajan − *Department of Materials Design and Innovation, University at Buffalo, Buffalo, New York 14260, United States;* orcid.org/0000-0001-9303-2797; Phone: +1 (716) 645 1380; Email: krajan3@buffalo.edu

Authors

Aparajita Dasgupta − *Department of Materials Design and Innovation, University at Buffalo, Buffalo, New York 14260, United States*

Yingjie Gao − *Department of Materials Design and Innovation, University at Buffalo, Buffalo, New York 14260, United States*

Scott R. Broderick − *Department of Materials Design and Innovation, University at Buffalo, Buffalo, New York 14260, United States*

E. Bruce Pitman − *Department of Materials Design and Innovation, University at Buffalo, Buffalo, New York 14260, United States*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jpcc.0c01492

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

The authors acknowledge support from the Toyota Research Institute Accelerated Materials Design and Discovery program. Support from the Collaboratory for a Regenerative Economy (CoRE) at the University at Buffalo is gratefully acknowledged. A.D., S.R.B., and K.R. acknowledge support from the National Science Foundation (NSF) under Grant No. 1640867. E.B.P. acknowledges support from the NSF under Grant No. 1821311. K.R. acknowledges the Erich Bloch Endowed Chair at the University at Buffalo.

REFERENCES

(1) Sabatier, P. Hydrogénations et déshydrogénations par catalyse. *Ber. Dtsch. Chem. Ges.* **1911**, *44*, 1984−2001.

(2) Nørskov, J. K.; Bligaard, T.; Hvolbæk, B.; Abild-Pedersen, F.; Chorkendorff, I.; Christensen, C. H. The nature of the active site in heterogeneous metal catalysis. *Chem. Soc. Rev.* **2008**, *37*, 2163−2171.

(3) Greeley, J. Theoretical Heterogeneous Catalysis: Scaling Relationships and Computational Catalyst Design. *Annu. Rev. Chem. Biomol. Eng.* **2016**, *7*, 605−635.

(4) Wan, H.; Østergaard, T. M.; Arnarson, L.; Rossmeisl, J. Climbing the 3D Volcano for the Oxygen Reduction Reaction Using Porphyrin Motifs. *ACS Sustainable Chem. Eng.* **2019**, *7*, 611−617.

(5) Hammer, B.; Nørskov, J. K. Electronic factors determining the reactivity of metal surfaces. *Surf. Sci.* **1995**, *343*, 211−220.

(6) Logadottir, A.; Rod, T.; Nørskov, J.; Hammer, B.; Dahl, S.; Jacobsen, C. The Brønsted−Evans−Polanyi Relation and the Volcano Plot for Ammonia Synthesis over Transition Metal Catalysts. *J. Catal.* **2001**, *197*, 229−231.

(7) Xin, H.; Vojvodic, A.; Voss, J.; Nørskov, J. K.; Abild-Pedersen, F. Effects of d-band shape on the surface reactivity of transition-metal alloys. *Phys. Rev. B: Condens. Matter Mater. Phys.* **2014**, *89*, 115114.

(8) Darby, M. T.; Réocreux, R.; Sykes, E. C. H.; Michaelides, A.; Stamatakis, M. Elucidating the Stability and Reactivity of Surface Intermediates on Single-Atom Alloy Catalysts. *ACS Catal.* **2018**, *8*, 5038−5050.

(9) Darby, M. T.; Stamatakis, M.; Michaelides, A.; Sykes, E. C. H. Lonely Atoms with Special Gifts: Breaking Linear Scaling Relationships in Heterogeneous Catalysis with Single-Atom Alloys. *J. Phys. Chem. Lett.* **2018**, *9*, 5636−5646.

(10) Ma, X.; Li, Z.; Achenie, L. E. K.; Xin, H. Machine-Learning-Augmented Chemisorption Model for CO2 Electroreduction Catalyst Screening. *J. Phys. Chem. Lett.* **2015**, *6*, 3528−3533.

(11) Goldsmith, B. R.; Esterhuizen, J.; Liu, J.-X.; Bartel, C. J.; Sutton, C. Machine learning for heterogeneous catalyst design and discovery. *AIChE J.* **2018**, *64*, 2311−2323.

(12) García-Muelas, R.; López, N. Statistical learning goes beyond the d-band model providing the thermochemistry of adsorbates on transition metals. *Nat. Commun.* **2019**, *10*, 4687.

(13) Chowdhury, A. J.; Yang, W.; Walker, E.; Mamun, O.; Heyden, A.; Terejanu, G. A. Prediction of Adsorption Energies for Chemical Species on Metal Catalyst Surfaces Using Machine Learning. *J. Phys. Chem. C* **2018**, *122*, 28142−28150.

(14) Toyao, T.; Suzuki, K.; Kikuchi, S.; Takakusagi, S.; Shimizu, K.-i.; Takigawa, I. Toward Effective Utilization of Methane: Machine Learning Prediction of Adsorption Energies on Metal Alloys. *J. Phys. Chem. C* **2018**, *122*, 8315−8326.

(15) Saxena, S.; Khan, T. S.; Jalid, F.; Ramteke, M.; Haider, M. A. In silico high throughput screening of bimetallic and single atom alloys using machine learning and ab initio microkinetic modelling. *J. Mater. Chem. A* **2020**, *8*, 107−123.

(16) Thirumalai, H.; Kitchin, J. R. Investigating the Reactivity of Single Atom Alloys Using Density Functional Theory. *Top. Catal.* **2018**, *61*, 462−474.

(17) Jinnouchi, R.; Asahi, R. Predicting Catalytic Activity of Nanoparticles by a DFT-Aided Machine-Learning Algorithm. *J. Phys. Chem. Lett.* **2017**, *8*, 4279−4283.
(18) Andersen, M.; Levchenko, S. V.; Scheffler, M.; Reuter, K. Beyond Scaling Relations for the Description of Catalytic Materials. *ACS Catal.* **2019**, *9*, 2752−2759.
(19) Lu, Z.; Yadav, S.; Singh, C. V. Predicting aggregation energy for single atom bimetallic catalysts on clean and O* adsorbed surfaces through machine learning models. *Catal. Sci. Technol.* **2020**, *10*, 86−98.
(20) Andriotis, A. N.; Mpourmpakis, G.; Broderick, S.; Rajan, K.; Datta, S.; Sunkara, M.; Menon, M. Informatics guided discovery of surface structure-chemistry relationships in catalytic nanoparticles. *J. Chem. Phys.* **2014**, *140*, 094705.
(21) Srinivasan, S.; Broderick, S. R.; Zhang, R.; Mishra, A.; Sinnott, S. B.; Saxena, S. K.; LeBeau, J. M.; Rajan, K. Mapping Chemical Selection Pathways for Designing Multicomponent Alloys: an informatics framework for materials design. *Sci. Rep.* **2015**, *5*, 17960.
(22) Villars, P. A three-dimensional structural stability diagram for 1011 binary AB2 intermetallic compounds: II. *J. Less-Common Met.* **1984**, *99*, 33−43.
(23) Villars, P. A three-dimensional structural stability diagram for 998 binary AB intermetallic compounds. *J. Less-Common Met.* **1983**, *92*, 215−238.
(24) Villars, P. Three-dimensional structural stability diagrams for 648 binary AB3 and 389 binary A3B5 intermetallic compounds: III. *J. Less-Common Met.* **1984**, *102*, 199−211.
(25) Mooser, E.; Pearson, W. B. On the crystal chemistry of normal valence compounds. *Acta Crystallogr.* **1959**, *12*, 1015−1022.
(26) Miedema, A. The electronegativity parameter for transition metals: Heat of formation and charge transfer in alloys. *J. Less-Common Met.* **1973**, *32*, 117−136.
(27) Donoho, D. L. High-dimensional data analysis: The curses and blessings of dimensionality. *AMS Conference on Math Challenges of the 21st Century*; AMS: Los Angeles, CA, August 6−12, 2000; Abstract No. 956-62-02.
(28) Tenenbaum, J. B.; Silva, V. d.; Langford, J. C. A Global Geometric Framework for Nonlinear Dimensionality Reduction. *Science* **2000**, *290*, 2319−2323.
(29) Samko, O.; Marshall, A.; Rosin, P. Selection of the optimal parameter value for the Isomap algorithm. *Pattern Recognition Letters* **2006**, *27*, 968−979.
(30) Williams, C. K. I.; Rasmussen, C. E. Gaussian Processes for Regression. In *Advances in Neural Information Processing Systems 8: Proceedings of the 1995 Converence*; Touretzky, D. S., Mozer, M. C., Hasselmo, M. E., Eds.; MIT Press, 1996; pp 514−520.
(31) Breunig, M. M.; Kriegel, H.-P.; Ng, R. T.; Sander, J. LOF: Identifying Density-Based Local Outliers. *SIGMOD Rec.* **2000**, *29*, 93−104.
(32) Cho, H.-W.; Kim, S.; Jeong, M.; Park, Y.; Miller, N. G.; Ziegler, T.; Jones, D. Discovery of metabolite features for the modelling and analysis of high-resolution NMR spectra. *IJDMB* **2008**, *2*, 176−92.
(33) Johnson, R. A.; Wichern, D. W. *Applied Multivariate Statistical Analysis*; Prentice Hall: Upper Saddle River, NJ, 2002; Vol. *5*.
(34) Kar, P.; Karnick, H. Random feature maps for dot product kernels. *PMLR* **2012**, 583−591.
(35) Wilson, A.; Adams, R. Gaussian process kernels for pattern discovery and extrapolation. *ICML* **2013**, *28*, 1067−1075.
(36) Draper, N. R.; Smith, H. *Applied Regression Analysis*; John Wiley & Sons, 1998; Vol. *326*.
(37) Wang, S.; Abroshan, H.; Liu, C.; Luo, T.-Y.; Zhu, M.; Kim, H. J.; Rosi, N. L.; Jin, R. Shuttling single metal atom into and out of a metal nanoparticle. *Nat. Commun.* **2017**, *8*, 848.
(38) Bootharaju, M. S.; Sinatra, L.; Bakr, O. M. Distinct metal-exchange pathways of doped Ag25 nanoclusters. *Nanoscale* **2016**, *8*, 17333−17339.
(39) Brust, M.; Walker, M.; Bethell, D.; Schiffrin, D. J.; Whyman, R. Synthesis of thiol-derivatised gold nanoparticles in a two-phase Liquid-Liquid system. *J. Chem. Soc., Chem. Commun.* **1994**, *0*, 801−802.
(40) Lucci, F. R.; Lawton, T. J.; Pronschinske, A.; Sykes, E. C. H. Atomic Scale Surface Structure of Pt/Cu(111) Surface Alloys. *J. Phys. Chem. C* **2014**, *118*, 3015−3022.
(41) Kyriakou, G.; Boucher, M. B.; Jewell, A. D.; Lewis, E. A.; Lawton, T. J.; Baber, A. E.; Tierney, H. L.; Flytzani-Stephanopoulos, M.; Sykes, E. C. H. Isolated Metal Atom Geometries as a Strategy for Selective Heterogeneous Hydrogenations. *Science* **2012**, *335*, 1209−1212.
(42) Han, J.; Lu, J.; Wang, M.; Wang, Y.; Wang, F. Single Atom Alloy Preparation and Applications in Heterogeneous Catalysis. *Chin. J. Chem.* **2019**, *37*, 977−988.

https://dx.doi.org/10.1021/acs.jpcc.0c01492
J. Phys. Chem. C 2020, 124, 14158−14166