JACS
JOURNAL OF THE AMERICAN CHEMICAL SOCIETY

Article

Machine-learning Prediction of CO Adsorption in Thiolated Ag-alloyed Au Nanoclusters

Gihan Panapitiya, Guillermo Avendaño Frano, Pengju Ren, Xiao-Dong Wen, Yongwang Li, and James P. Lewis

J. Am. Chem. Soc., **Just Accepted Manuscript** • DOI: 10.1021/jacs.8b08800 • Publication Date (Web): 08 Nov 2018

ACS Publications

is published by the American Chemical Society. 1155 Sixteenth Street N.W., Washington, DC 20036

Machine-learning Prediction of CO Adsorption in Thiolated Ag-alloyed Au Nanoclusters

Gihan Panapitiya[1], Guillermo Avendaño Frano[1], Pengju Ren[2,3], Xiaodong Wen[2,3], Yongwang Li[2,3], and James P. Lewis[2,1*]

1. Department of Physics and Astronomy, West Virginia University, Morgantown, WV, 26506-6315
2. State Key Laboratory of Coal Conversion, Institute of Coal Chemistry, Chinese Academy of Sciences, Taiyuan, Shanxi 030001, China
3. Synfuels China Co. Ltd., Huairou, Beijing 101407, China

Corresponding Authors

*Dr. James P. Lewis: State Key Laboratory of Coal Conversion, Institute of Coal Chemistry, Chinese Academy of Sciences, Taiyuan, Shanxi 030001, China, email: james.p.lewis.phd@gmail.edu

Gihan Panapitiya: Department of Physics and Astronomy, West Virginia University, Morgantown, WV, 26506-6315, email: gihanuthpala@gmail.com

ABSTRACT

We propose a machine-learning model, based on the random-forest method, to predict CO adsorption in thiolate protected nanoclusters. Two phases of feature selection and training, based initially on the Au_{25} nanocluster, are utilized in our model. One advantage to a machine-learning approach is that correlations in defined features disentangle relationships among the various structural parameters. For example, in Au_{25}, we find that features based on the distribution of Ag atoms relative to the CO adsorption site are the most important in predicting adsorption energies. Our machine-learning model is easily extended to other Au-based nanoclusters and we demonstrate predictions about CO adsorption on Ag-alloyed Au_{36} and Au_{133} nanoclusters.

KEYWORDS: machine learning, CO adsorption, Ag-alloyed Au nanoclusters

1. INTRODUCTION

Thiolated gold nanoclusters are one of the most widely studied systems in contemporary research. The increased interest is attributed to several promising applications proven for thiolated gold nanoclusters in a variety of fields including catalysis, sensing, electronics, and bio-medicine.[1–14] Given recent advancements in synthesis techniques, bimetallic counterparts of these systems are also produced; thereby enabling further tuning of electrochemical properties of Au-based nanoclusters and subsequently widening the scope of their applicability.[4,15–17] Unfortunately, combinatorial barriers exist which prevent fast identification of correct cluster compositions for chosen applications; subsequently, the enormity of the number of alloyed systems provides challenges for characterization. Furthermore, no theoretical methods exist to explore all combinatorically-possible alloyed configurations within a practically sensible time frame. For example, the number of bimetallic configurations for the smallest known thiolated nanocluster, $Au_{15}(SR)_{13}$, is combinatorically over 32,000 which yields a significant computational challenge to characterize all potential structures. The existence of over thirty different thiolated nanoclusters composed of 15 to 144 Au atoms increases the importance for developing smart approaches capable of making reasonable property predictions.[18]

Several researchers have presented reports where they have successfully employed modern machine-learning models to predict many electrochemical properties, including electrophilicity parameters, atomization energies, dielectric constants, condensed Fukui indices, atomic charges, band gaps, gas adsorption, and HOMO/LUMO energies in various nano-systems.[19–26] The basic workflow of these models are to: (1) energetically relax the molecular/crystalline systems, (2) create a dataset containing the target property and a set of numerical fingerprints of these systems, and (3) train and test machine learning algorithms on these datasets to arrive at predictions. A major drawback of these approaches is the difficulty to deal with large systems. Energetically relaxing thiolated gold nanocluster systems is computationally challenging which compromises the advantages of machine learning.

In recent years, different approaches of defining coordination numbers were proposed to predict adsorption properties in metal nanoclusters without ligand passivation.[27–31] However, the presence of ligands containing different non-metallic atoms and oriented in different directions gives rise to extra complexity in the adsorbent-adsorbate interactions. This complexity is difficult to capture with coordination numbers alone. In this work, we propose an efficient machine-learning model aided by a fast ab-initio density functional theory (DFT) approach to accomplish adsorption energy predictions in alloyed thiolated nanoclusters. In our model, we

use only the structural properties of the un-adsorbed, non-relaxed systems. We do not use the knowledge on adsorbent-adsorbate interactions or structural/electronic properties of the optimized configurations (which requires performing many additional DFT calculations). While this can affect the prediction accuracy, our model can serve as a very fast technique to filter out a set of candidates for further testing. We demonstrate performance of the model by predicting CO adsorption energies in Ag alloyed $Au_{25}(SR)_{18}$. CO oxidation has long been the preferred reaction to study the catalytic properties of Au-based systems.[2,4,32–34] Also, catalytic CO oxidation has been tested as a mechanism to remove poisonous CO from H_2 in fuel cells.[35] In our recent work, we found that CO-adsorption on Ag alloyed Au_{25} is sensitive to the number of dopants and the adsorption energies do not have a predictable trend.[36] To the best of our knowledge, this is the first machine-learning study done for thiolated Au-based nanoclusters.

2. COMPUTATIONAL METHODS

The success of any machine learning method depends on the quality of the numerical representation of the system under study. Such representations, realized as a collection of numerical fingerprints, are commonly called *features*. The features in our case fall into four main categories based on (*i*) the distance between atoms, (*ii*) the atomic-bonding counting, (*iii*) the graphical representation of the cluster, and (*iv*) the volume enclosed by atoms.

We begin our investigation by examining the Au_{25} structure which consists of a core of 13 Au atoms surrounded by six SR-Au-SR-Au-SR staple-like units, where SR stands for thiolate ligands (see Fig. S1 for a detailed illustration of the Au25 structure). The main adsorption sites in thiolated nanoclusters are the Au atoms on the surface and the facets formed by surface Au atoms.[3,32,37,38] Our features are applicable to any adsorption site. For the demonstration of our method, we only considered adsorption on Au atoms on the surface of Au_{25} (Blue atoms in Fig. 1). Since there are 12 Au sites in the staple units, the number of CO-adsorbed structures considered were 12×25×15. All these CO-adsorbed structures were energetically relaxed using the same force tolerance used for adsorbate-free Au_{25} structures to facilitate adsorption energy calculation.

The distance between Ag atoms and the adsorbent site ($|\boldsymbol{r}_{AS-Ag}| = |\boldsymbol{r}_{Ag} - \boldsymbol{r}_{AS}|$) forms the basis of defining the *distance* features (see Fig. 1). The mean and standard deviation of $|\boldsymbol{r}_{AS-Ag}|$ form two simply-defined features (d1, d2). And, the features labeled d3 - d7 are based on the centroid position of the Ag atoms relative to the adsorbent site as discussed in the Supporting

Information. By identifying the atoms as belonging to different layers based on their distance to the adsorbent site, we define another class of features. As marked by the red-dashed curves in Fig. 1, boundaries of each layer are defined as concentric circles having radii equivalent to the distance between the adsorbent site and neighbors nearest to the adsorbent site. Basic building blocks of a nanocluster are atoms. We can also define higher order building blocks such as A-B bonds (bonds connecting A and B atoms), or A-B-C and A-B-C-D fragments in each layer '*l*'; accordingly, for Au_{25}, we have created 81 atomic-bonding features. For each feature, we can subsequently make a normalized count. For example, if a particular layer contains six C, three H and five S atoms, normalized counts are 6/(6+3+5), 3/(6+3+5) and 5/(6+3+5) for C, H and S, respectively.

A molecule can also be defined via graphical representation. Atoms and bonds in the molecule become nodes and edges of a graph. Using this type of representation, the following features were defined: (1) number of Ag atoms within a path length of three units from AS (g1), (2) total number of atoms within a path length of two units from AS (g2), (3) number of edges connecting two metal atoms within a path length of three units from AS (g3), (4) shortest path lengths between AS and the eight nearest neighbor Au or Ag atoms (g4-g11). We add a constant to the path length when the metal atom is Ag. This gives a sequence of numbers representing the metallic environment and the relative AS-metallic atom distance. Volumetric features include the volume enclosed by AS and fifteen of its nearest neighbors (v1) and the volume enclosed by the AS and all the Ag sites (v2).

Now that we have defined a set of features to describe the Au-based nanocluster and the absorbent site, we generate data from DFT calculations to discover correlations between different features and the absorbent site. Of course, machine-driven algorithms are only feasible when very large datasets are available; small datasets increase statistical anomalies. To prepare the dataset, we generated Ag_xAu_{25-x} nanoclusters with x ranging from 1 to 25. We created at least 500 isomers with random Ag sites for each alloy case x=4-21. For the remaining alloying levels (x=1-3 and x=22-25), we created all possible Ag-alloyed configurations. All the calculations presented in this work were performed using a local-orbital density functional theory code, called FIREBALL.[39] We chose Becke[40] exchange with Lee-Yang-Parr correlation functional[41] (BLYP) to perform structural relaxations. The basis set is made of optimized numerical local atomic orbitals which were confined to regions limited by the corresponding cutoff radii r_c. The r_c values used for our study are given in the Supporting information. The chosen basis set and the DFT functional as implemented in FIREBALL have been successfully validated by several previous studies on Au nanoparticle systems.[36,42–45] The adsorbate-free Ag-alloyed structures

were energetically relaxed until the root-mean-square error of the force on the atoms was less than 0.05 eV/Å.

An inspection of structural energies of Ag_xAu_{25-x} revealed jumps in the energy profiles around the 15th lowest energy for most of the alloying levels (see Fig. S2 and S3 for energy profiles of alloying levels 1-24). These jumps will be hereafter referred to as 'gaps' in the energy profiles. It is likely that the higher energy structures above these energy gaps have a low probability to exist. This assumption is based on the experimental studies of 1-Ag alloyed Au_{25}, which have shown that preferable Ag sites are on the surface of Au_{13} core.[46,47] Interestingly, all the structures with a Ag atom on the surface of the Au_{13} core are the ones having energies below the gap in Fig. S2 (a). Therefore, a statistically meaningful dataset may consist of structures with energies less than the energy corresponding to the gap. However, such a selection criteria results in poor representation from alloying levels having low-lying energy gaps. This lack of representation reduces the generalizability of the model. Hence, to maximize the number of data points, while avoiding the effect from the high energy structures, we selected fifteen lowest energy structures from every alloying level to determine the CO adsorption-energy data.

We should note that lowest-energy adsorbate-free Au_{25} may not always be associated with lowest-energy CO-adsorbed structures. However, we did not restrict the number of CO-adsorbed structures based on their energies. This is because, when a reaction takes place, only the most energetically favorable Au_{25} have the likelihood to exist and CO gets adsorbed on these stable Au_{25}. Removing the higher energy CO-adsorbed data from the model results in our model being applicable only to a small niche of adsorbate-free Au_{25} associated with low energy CO-adsorbed Au_{25}. Such limitations reduce the practical importance and the generalizability of our model as it will fail to predict the adsorption energies of most of the experimentally synthesized Au_{25}.

The adsorption energy is calculated according to ($E = E_{CO/Ag_xAu_{25-x}} - E_{Ag_xAu_{25-x}} - E_{CO}$) for CO binding at the Ag/Au sites (at the outer shell of the nanocluster). The final dataset is a (number-of-configurations) × (number-of-features+1) matrix, with each row corresponding to a different calculated Ag alloyed configuration. One of the columns contains the adsorption energies and the rest of the columns correspond to different features.

3. RESULTS AND DISCUSSION

A distribution of our calculated adsorption energies is shown in Fig. 2. These energies have a mean of -0.72 eV and a skewness of 0.99. The positive skewness indicates the presence of values deviated from the mean towards the positive energies. Machine-learning algorithms generally prefer normal distributions. Thus we transformed all the adsorption energies according to ln(E + *constant*). Here, the *constant* has to be chosen so that the skewness of the distribution is minimized. We found that a minimum skewness of -0.002 results when 2.5 is chosen as this constant.

Developing and testing our model consists of two phases. In the first phase, we evaluate the performance of different machine learning algorithms using all the defined features. Tree ensemble methods like random forests[48,49] and xgboost[50] performed better than neural networks and other regression methods like linear, ridge, kernel ridge and support vector regression. For this study, we chose random forests as implemented in Scikit-learn Python library (discussed in the Supporting Information).[51] Random forests is a supervised learning algorithm containing a forest of decision trees. The final prediction is a weighted average of predictions of all the trees.

In the *first phase*, prior to feeding the features to the random forest algorithm, each feature was scaled such that each had a zero mean and unit variance. This is to ensure that no single feature will dominate the objective function of the algorithm, ignoring the effects of other features with lower variance. Next, we randomly split the whole dataset as training and testing sets according to the 80:20 ratio. Using a random forest with 100 trees, the data in the training set was used with five-fold cross validation. We used the built-in feature ranking method in random forests to determine the best features based on their contribution to the prediction accuracy. Fig. 3 shows the top 10 features along with their Pearson correlation coefficient and mutual information values with respect to the adsorption energy. The Pearson correlation coefficient quantifies the linear dependence between two variables while mutual information is a measure of both linear and non-linear dependence. The top 10 features have decent linear dependencies with adsorption energy, even though our features are based on the geometric properties of non-relaxed structures. As the adsorption energies were calculated using DFT-relaxed structures, these correlation values indicate that our features are likely related to the local chemical environment at the vicinity of adsorbent sites of the synthesized structures. Not surprisingly, the simplest possible feature for an alloyed system, the number of dopant Ag atoms (nAg) has the greatest effect on the prediction accuracy. However, as will be shown, nAg alone is not enough to achieve the maximum accuracy. It is also worth noting that all the distance-based features are among the top 10 features.

The feature-feature correlations in Fig. 4 indicate that the number of dopant Ag atoms has the greatest correlation to the other top 14 features (the feature v2 has the next highest correlation to the other features). One interesting aspect is that Ag dependency is directly inherent in many of the defined features, particularly features based on centroids and bonding information. However, the number of dopant Ag atoms are not directly defined within the feature HCH1 which corresponds to the normalized number of H-C-H fragments in the 1st layer of neighbors nearest to the adsorbent site as it is independent of the number of dopant Ag atoms. Surprisingly, despite the lack of defined dependency between the two features nAg and HCH1; there is a strong correlation between these two features (value of 0.68) as indicated in the correlation map of Fig. 4.

The features specifically designed to measure the clustering of the Ag (d5, d6 and d7) are among the top 5 features. It is surprising to find that d6 and d7 contain almost the same information on linear dependence (correlation coefficient of 0.99), even though they are based on different centroids. This strong correlation indicates that Ag atoms are clustered around a similar central position. Despite the close relationship between d6 and d7, random forests has considered both as highly important. Even though d5 has the lowest correlation values compared with the top 5, it has been ranked as the second most important feature. This could be due to d5 having interactions with the other features. Even v2 is an indirect measure of the clustering, as the volume enclosed by the adsorbent site and the Ag atoms decreases with Ag atoms being clustered within a proximity to each other.

The counts of building blocks were previously shown to be effective in other systems and we were inspired to add these as features for our investigation;[21–23] however, we note relatively poor scores for building block features with one exception. The occurrence of HCH1 (number of HCH units in layer 1) as one of the important features (seen in Fig. 3) show that orientation of ligands also plays a significant role in predicting the adsorption energies. This is because H-C-H units are only found in the ligands which are more readily located near CO adsorption sites.

There is no strong agreement between the trends in Pearson correlation and mutual information with that of the feature importance, even though for features between nAg and d1 in Fig. 3, high mutual information values are generally associated with high feature importance. This was further confirmed by constructing an extended correlation map by considering more features (for example, see Fig. S4 in Supporting Information). We notice that the features d5-d7 and v2 are among the highest scoring mutual information values, but not among the features with highest Pearson correlation coefficients. This may indicate that non-linear relationships

between the adsorption energy and the features that an important role in the prediction. We also note that the features having a negative linear correlation with adsorption energy (for example, AgAuAu1 and AuAgAu2) have low effect on the prediction.

Inside the cross validation loop, we train a second random forest model (also with 100 trees) with the first *n* top features to determine optimum *n* as discussed in the Supporting Information. Determination of the new top *n* features marks the end of the first phase for training our model. Using the new set of features selected in the cross validation, we create additional features in the *second phase* - known as feature engineering. These features were created by taking the square root, raising them to power two and three and taking the log of the selected features as discussed in the Supporting Information. Creating more complicated features, like adding and multiplying features in 2 or 3 feature combinations and calculating the Euclidian distance between feature values and their cluster centers, did not result in improvement of the prediction accuracy. Following the same approach as the first phase, the best features were filtered out from the ones selected in the first phase and those engineered based on them. Additionally, parameter tuning of the random forest method was also performed to determine the optimal parameters of the final model. However, no significant gain in the accuracy was obtained with different parameters. Therefore, the final model contains only 100 trees with default parameters.

The features in the final model are the ones of the best cross validation step. These are nAg, nAg^2, nAg^3, d6, $d6^3$. This result shows that adsorption energies in Au_{25} can be predicted with only two features along with their non-linear counterparts. This model was then trained once with the whole training set and tested with the testing set to arrive at the final prediction accuracy.

The accuracy of our model prediction is shown in the prediction performance plot (log transformed), Fig. 5(a) with the distribution of residual error shown in Fig. 5(b). Prediction accuracies of the log transformed data are, R2 = 0.77869, MAE = 0.13196 and RMSE = 0.17348. The higher the residual error, the more erroneous the result. For negative and positive adsorption energies, the mean of the absolute residual errors are 0.20 eV and 0.44 eV, respectively. Predicted values corresponding to positive adsorption energies have more deviation from the actual values compared to the negative adsorption energies (see Fig. S6). Positive adsorption energies correspond to weak CO/Au_{25} attraction. The potential energy surface consists of many local minima with similar energies which results in structurally different CO adsorbed isomers having similar adsorption energies. Shallow potential energy surfaces do make it difficult for the learning algorithm to accurately predict positive adsorption

energies. Another major factor affecting the accuracy of our model may be that the ligands are oriented in different directions. Ligands create a complex chemical environment for adsorbate molecules. As shown in Fig. 3 and Fig. 4, adsorption energies are mainly affected by the location of Ag atoms and the H-C-H fragments closest to the adsorption site. To capture the combined effect of ligands and dopants, we fed our model with compound features created by multiplying features. However, these new features did not result in improvements for the model accuracy. Further, the use of non-relaxed isomers to generate features is also detrimental to the prediction accuracy. This is because the presence of dopants significantly affects bonding and atom-atom distances of a relaxed structure may be highly deviated from those of the non-relaxed one. Our model was unable to learn without ambiguity; however, there is a definite relationship between the positive adsorption energies and structural parameters. We also tried to build two models for positive and negative adsorption energies. However, this did not result in an increased accuracy, which could be due to the reduced number of data points in each sample.

We extend our model to $Au_{36}(SR)_{24}$ and $Au_{133}(SR)_{52}$ nanoclusters; thereby demonstrating the versatility of our model. For $Au_{36}(SR)_{24}$, we used only 1360 isomers having either 1, 2, 3, 6, 9, 12, 13, 18, or 24 of Ag atoms. At the second phase of feature selection, the best features were found to be d7, $d7^3$, $HC1^2$, $d7^2$, HC1, $HC1^3$ which gave accuracies of 0.65388 (R2), 0.21582 (RMSE) and 0.1856 (MAE) for predictions (prediction performance plot shown in Fig. 6(a)). These accuracies are encouraging, given the less symmetric geometry of Au_{36} in comparison to Au_{25} and the smaller sample size used for the training. As Au_{36} is relatively a small cluster, the number of samples can be increased easily to achieve a possible increase in the prediction accuracy. These accuracies also show that the features defined based on the nearly spherical Au_{25} cluster can be readily used for the clusters like Au_{36} which have different symmetries. The $Au_{133}(SR)_{52}$ nanocluster is the largest thiolated nanocluster of which the crystal structure has been verified. This nanocluster contains 393 atoms with -SCH_3 as the ligand. We generated 1898 CO adsorbed isomers having 5, 10, 15, 20, 30, 48, 80 and 120 Ag atoms. Even with a small number of samples we were able to achieve accuracies of 0.75063 (R2), 0.17128 (RMSE) and 0.11882 (MAE) for predictions (prediction performance plot shown in Fig. 6(b)). This shows that our model may be well suited for spherical-like clusters. The best features selected at the second phase of feature selection were, nAg^3, nAg, nAg^2, d5, $d5^2$, $d5^3$, $d6^3$, $d6^2$, d6, $d7^3$, d7, $d7^2$, v2, $g3^3$, $g3^2$, g3, $d1^2$, $g2^2$, $v2^2$, $d2^2$.

Interestingly, in Au_{36}, features based on the orientation of the ligands have a significant effect on the prediction accuracy, whereas in Au_{25} and Au_{133}, only the dopant-based features have the

strongest influence. We hypothesize that this is due to the ligands in Au_{36} causing a higher steric hindrance than the ligands in either Au_{25} or Au_{133}. To test this hypothesis, we can use the feature HC1, the normalized count of H-C bonds within ten nearest neighbors of the adsorbent sites. The higher the number of H-C bonds close to the adsorbent site, the larger is the steric hindrance. For each adsorbate-free nanocluster, we calculated the average of HC1 over all the adsorbent sites. The averages obtained for Au_{25}, Au_{36}, and Au_{133} are 0.10, 0.14, and 0.09 respectively. Thus, there are more H-C bonds closer to the adsorbent sites in Au_{36} compared to other two nanoclusters, resulting in greater steric hindrance to approaching CO adsorbates. The steric hindrance from H-C bonds is likely related to the sphericity[52] of the core of the nanoclusters. The closer the value of sphericity to 1, the more spherical is the cluster. More spherical surfaces tend to cause less steric hindrance as illustrated in Fig. S8. As shown in Fig. S9 (a), the core of Au_{36} consists of planar-like surfaces with a sphericity of 0.735. Sphericities of Au_{25} and Au_{133} cores are 0.940 and 0.943 respectively.

4. CONCLUSION

Overall, we have developed a machine-learning model based on the random forest method to predict CO adsorption energies for Au-based nanoclusters, starting by training our model with the Au_{25} nanocluster alloyed with Ag. We have defined approximately 100 features to model nanoclusters as numerical representations. Over 2,000 data points are contained in our model. Using a two-step feature-selection process, and features engineering approaches, we predicted the adsorption energies with accuracies of 0.78 (R2) and 0.17 (RMSE). Our chosen features are based merely on the structural properties of the unoptimized adsorbate system (to enable rapid prediction); our model is an excellent filtering tool to select first round candidates for further, more accurate, analysis. The validity of our model was also tested by predicting CO adsorption energies in the less symmetric Au_{36} nanocluster and the larger Au_{133} nanocluster, using the same defined features as the Au_{25} model, with prediction accuracies (R2) of 0.65 and 0.75, respectively.

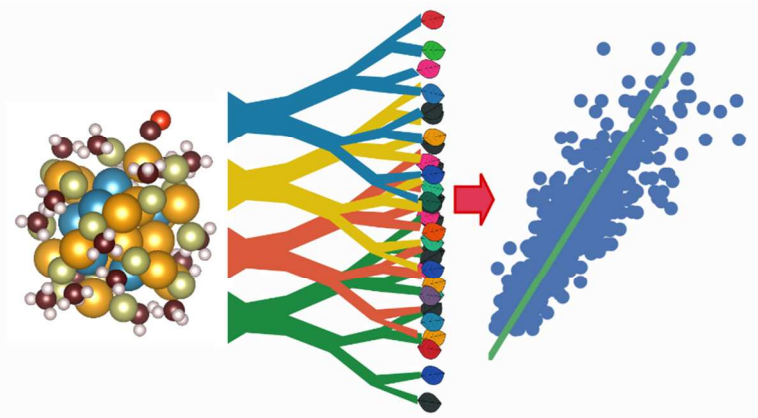
Supporting Information

More details on the computational methodology and additional calculations

ACKNOWLEDGMENTS

J.P.L. acknowledges funding from DOE SC-0004737 and funding from the Chinese Academy of Sciences President's International Fellowship Initiative (PIFI) for 2017-2018. He is also funded through the National Talent Program of the Chinese Academy of Sciences as a National Distinguished Scholar. X.W. is grateful for the financial support from the National Natural Science Foundation of China (No. 21473229, 91545121), and Synfuels China, Co. Ltd., also acknowledge National Thousand Young Talents Program of China, Hundred-Talent Program of Chinese Academy of Sciences and Shanxi Hundred-Talent Program. R.P. is grateful to the funding support from China Postdoctoral Science Foundation (No. 2016M590216). G.P and J.P.L also acknowledge the supercomputing resources from the High Performance Computing facility of the West Virginia University.

TOC GRAPHIC



Figures and Tables:

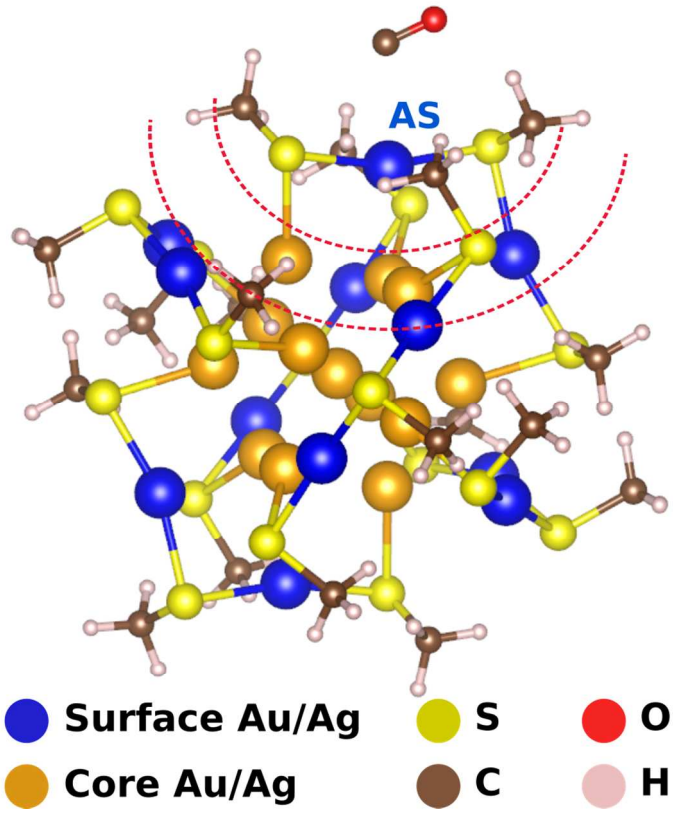


Figure 1. The CO/Au_{25} system (A detailed description on the structure of Au_{25} is given in Fig. S1). Hypothetical boundaries of two nearest-neighbor layers are shown with dashed red curves. **AS** stands for the CO adsorbent site. Adsorbent Au/Ag atomic sites on the surface are colored in blue. Orange is used for Au/Ag sites in the core inaccessible for CO due to steric hindrance.

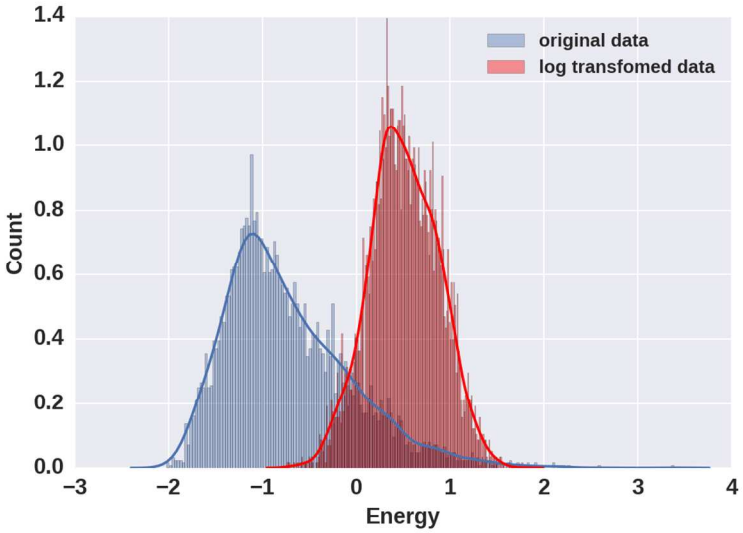


Figure 2. Variation of the calculated CO adsorption energies.

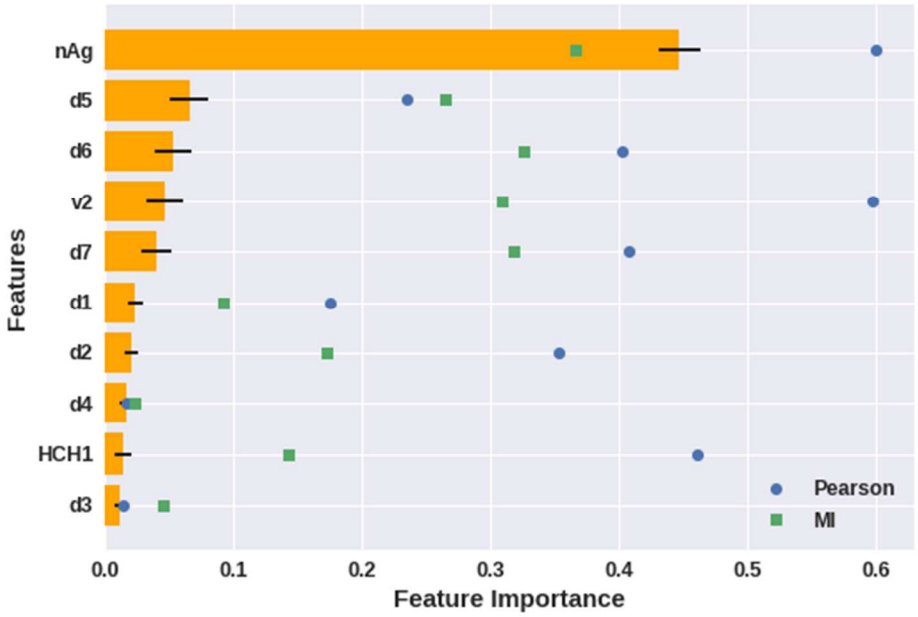


Figure 3. Highest ranking important features selected by random forests and the corresponding Pearson correlation coefficient and mutual information values; nAg is the number of Ag atoms, d3-d5 are measures of how closely the Ag atoms are clustered around the centroid of the Ag locations relative to the adsorbent site, HCH1 is the normalized number of H-C-H fragments in the 1st layer of neighbors nearest to the adsorbent site, d1 and d2 are mean and standard deviation of $|\mathbf{r}_{AS-Ag}|$, and v2 is the volume enclosed by the adsorbent site and the Ag sites.

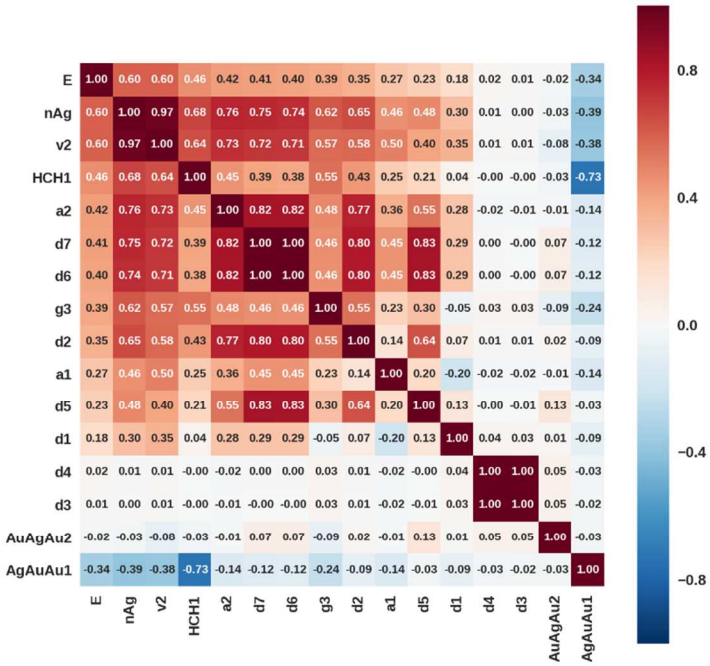


Figure 4. Feature-feature correlation map of the top features.

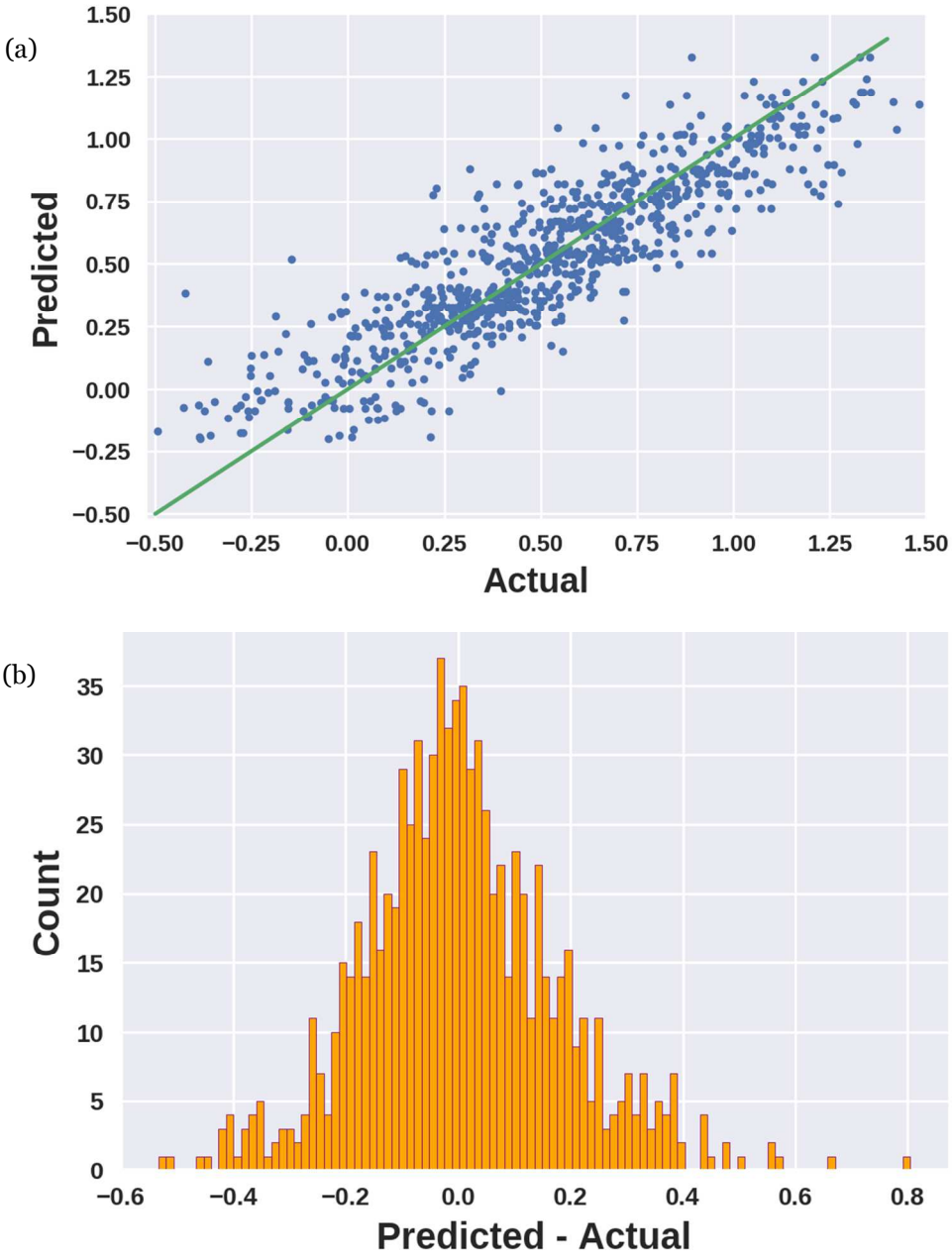


Figure 5. Prediction performance plots of the final model actual vs. predicted adsorption energies: (a) actual vs predicted on regression plot and (b) histogram counts of (predicted – actual) adsorption energies.

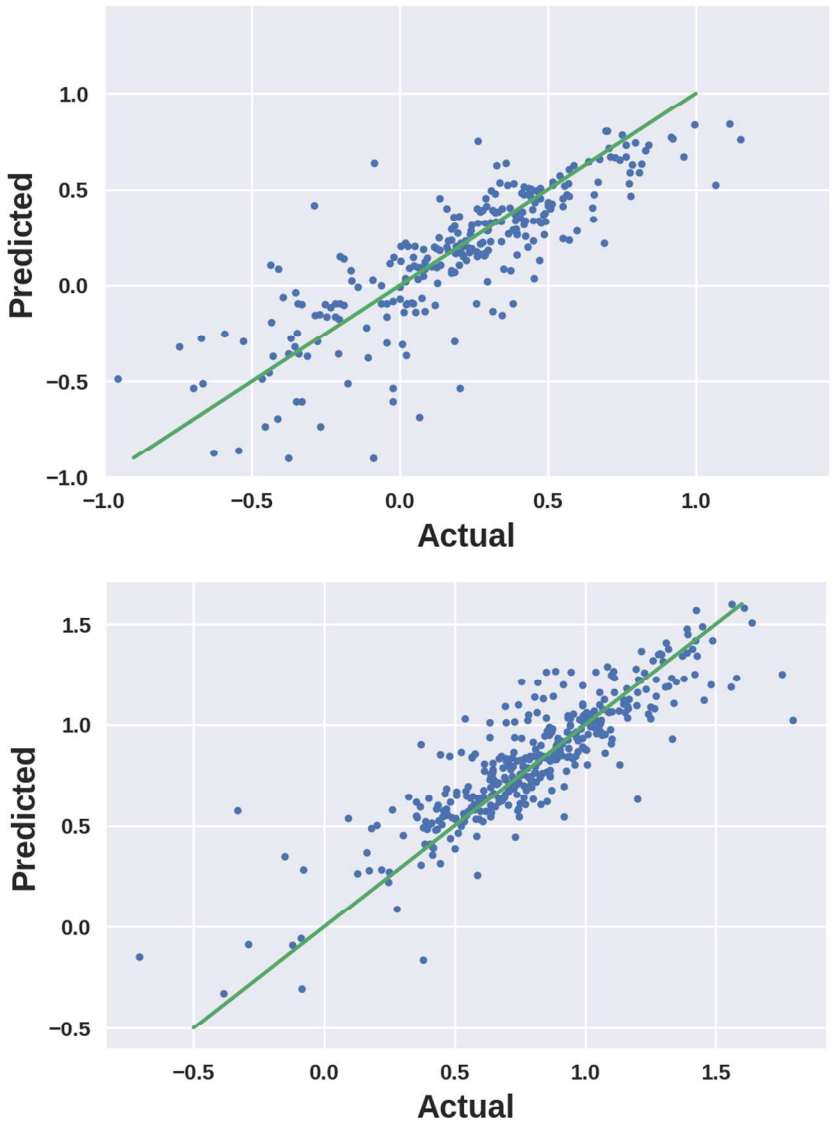


Figure 6. Prediction performance plots for (a) Au_{36} and (b) Au_{133} nanoclusters.

References

(1) Liu, Y.; Tsunoyama, H.; Akita, T.; Tsukuda, T. Efficient and Selective Epoxidation of Styrene with TBHP Catalyzed by Au25 Clusters on Hydroxyapatite. *Chemical Communications* **2010**, *46* (4), 550–552.

(2) Nie, X.; Qian, H.; Ge, Q.; Xu, H.; Jin, R. CO Oxidation Catalyzed by Oxide-Supported $Au_{25}(SR)_{18}$ Nanoclusters and Identification of Perimeter Sites as Active Centers. *ACS Nano* **2012**, *6* (7), 6014–6022.

(3) Li, G.; Jin, R. Gold Nanocluster-Catalyzed Semihydrogenation: A Unique Activation Pathway for Terminal Alkynes. *Journal of the American Chemical Society* **2014**, *136* (32), 11347–11354.

(4) Li, W.; Liu, C.; Abroshan, H.; Ge, Q.; Yang, X.; Xu, H.; Li, G. Catalytic CO Oxidation Using Bimetallic M_xAu_{25-x} Clusters: A Combined Experimental and Computational Study on Doping Effects. *The Journal of Physical Chemistry C* **2016**, *120* (19), 10261–10267.

(5) Shichibu, Y.; Negishi, Y.; Tsunoyama, H.; Kanehara, M.; Teranishi, T.; Tsukuda, T. Extremely High Stability of Glutathionate-Protected Au25 Clusters Against Core Etching. *Small* **2007**, *3* (5), 835–839.

(6) Ghosh, A.; Udayabhaskararao, T.; Pradeep, T. One-Step Route to Luminescent $Au_{18}SG_{14}$ in the Condensed Phase and Its Closed Shell Molecular Ions in the Gas Phase. *J. Phys. Chem. Lett.* **2012**, *3* (15), 1997–2002.

(7) Das, A.; Li, T.; Li, G.; Nobusada, K.; Zeng, C.; Rosi, N. L.; Jin, R. Crystal Structure and Electronic Properties of a Thiolate-Protected Au_{24} Nanocluster. *Nanoscale* **2014**, *6* (12), 6458–6462.

(8) Yu, Y.; Luo, Z.; Chevrier, D. M.; Leong, D. T.; Zhang, P.; Jiang, D.; Xie, J. Identification of a Highly Luminescent $Au_{22}(SG)_{18}$ Nanocluster. *J. Am. Chem. Soc.* **2014**, *136* (4), 1246–1249.

(9) Zhu, M.; Aikens, C. M.; Hendrich, M. P.; Gupta, R.; Qian, H.; Schatz, G. C.; Jin, R. Reversible Switching of Magnetism in Thiolate-Protected Au_{25} Superatoms. *J. Am. Chem. Soc.* **2009**, *131* (7), 2490–2492.

(10) McCoy, R. S.; Choi, S.; Collins, G.; Ackerson, B. J.; Ackerson, C. J. Superatom Paramagnetism Enables Gold Nanocluster Heating in Applied Radiofrequency Fields. *ACS Nano* **2013**, *7* (3), 2610–2616.

(11) Ramakrishna, G.; Varnavski, O.; Kim, J.; Lee, D.; Goodson, T. Quantum-Sized Gold Clusters as Efficient Two-Photon Absorbers. *J. Am. Chem. Soc.* **2008**, *130* (15), 5032–5033.

(12) Russier-Antoine, I.; Bertorelle, F.; Vojkovic, M.; Rayane, D.; Salmon, E.; Jonin, C.; Dugourd, P.; Antoine, R.; Brevet, P.-F. Non-Linear Optical Properties of Gold Quantum Clusters. The Smaller the Better. *Nanoscale* **2014**, *6* (22), 13572–13578.

(13) Knoppe, S.; Vanbel, M.; van Cleuvenbergen, S.; Vanpraet, L.; Bürgi, T.; Verbiest, T. Nonlinear Optical Properties of Thiolate-Protected Gold Clusters. *J. Phys. Chem. C* **2015**, *119* (11), 6221–6226.

(14) Day, P. N.; Pachter, R.; Nguyen, K. A.; Bigioni, T. P. Linear and Nonlinear Optical Response in Silver Nanoclusters: Insight from a Computational Investigation. *J. Phys. Chem. A* **2016**, *120* (4), 507–518.

(15) Negishi, Y.; Iwai, T.; Ide, M. Continuous Modulation of Electronic Structure of Stable Thiolate-Protected Au_{25} Cluster by Ag Doping. *Chemical Communications* **2010**, *46* (26), 4713.

(16) Christensen, S. L.; MacDonald, M. A.; Chatt, A.; Zhang, P.; Qian, H.; Jin, R. Dopant Location, Local Structure, and Electronic Properties of $Au_{24}Pt(SR)_{18}$ Nanoclusters. *The Journal of Physical Chemistry C* **2012**, *116* (51), 26932–26937.

(17) Kauffman, D. R.; Alfonso, D.; Matranga, C.; Qian, H.; Jin, R. A Quantum Alloy: The Ligand-Protected $Au_{25-x}Ag_x(SR)_{18}$ Cluster. *The Journal of Physical Chemistry C* **2013**, *117* (15), 7914–7923.
(18) Jin, R.; Zeng, C.; Zhou, M.; Chen, Y. Atomically Precise Colloidal Metal Nanoclusters and Nanoparticles: Fundamentals and Opportunities. *Chem. Rev.* **2016**, *116* (18), 10346–10413.
(19) Pereira, F.; Latino, D. A. R. S.; Aires-de-Sousa, J. Estimation of Mayr Electrophilicity with a Quantitative Structure–Property Relationship Approach Using Empirical and DFT Descriptors. *J. Org. Chem.* **2011**, *76* (22), 9312–9319.
(20) Rupp, M.; Tkatchenko, A.; Müller, K.-R.; von Lilienfeld, O. A. Fast and Accurate Modeling of Molecular Atomization Energies with Machine Learning. *Phys. Rev. Lett.* **2012**, *108* (5), 058301.
(21) Huan, T. D.; Mannodi-Kanakkithodi, A.; Ramprasad, R. Accelerated Materials Property Predictions and Design Using Motif-Based Fingerprints. *Phys. Rev. B* **2015**, *92* (1), 014106.
(22) Zhang, Q.; Zheng, F.; Zhao, T.; Qu, X.; Aires-de-Sousa, J. Machine Learning Estimation of Atom Condensed Fukui Functions. *Mol. Inf.* **2016**, *35* (2), 62–69.
(23) Mannodi-Kanakkithodi, A.; Pilania, G.; Huan, T. D.; Lookman, T.; Ramprasad, R. Machine Learning Strategy for Accelerated Design of Polymer Dielectrics. **2016**, *6*, 20952.
(24) Pilania, G.; Mannodi-Kanakkithodi, A.; Uberuaga, B. P.; Ramprasad, R.; Gubernatis, J. E.; Lookman, T. Machine Learning Bandgaps of Double Perovskites. *Scientific Reports* **2016**, *6*, 19375.
(25) Borboudakis, G.; Stergiannakos, T.; Frysali, M.; Klontzas, E.; Tsamardinos, I.; Froudakis, G. E. Chemically Intuited, Large-Scale Screening of MOFs by Machine Learning Techniques. *npj Computational Materials* **2017**, *3* (1), 40.
(26) Pereira, F.; Xiao, K.; Latino, D. A. R. S.; Wu, C.; Zhang, Q.; Aires-de-Sousa, J. Machine Learning Methods to Predict Density Functional Theory B3LYP Energies of HOMO and LUMO Orbitals. *Journal of Chemical Information and Modeling* **2017**, *57* (1), 11–21.
(27) Mpourmpakis, G.; Andriotis, A. N.; Vlachos, D. G. Identification of Descriptors for the CO Interaction with Metal Nanoparticles. *Nano Lett.* **2010**, *10* (3), 1041–1045.
(28) Calle-Vallejo, F.; Loffreda, D.; Koper, M. T. M.; Sautet, P. Introducing Structural Sensitivity into Adsorption-Energy Scaling Relations by Means of Coordination Numbers. *Nature Chemistry* **2015**, *7*, 403.
(29) Federico, C.-V.; I, M. J.; M, G.-L. J.; Philippe, S.; David, L. Fast Prediction of Adsorption Properties for Platinum Nanocatalysts with Generalized Coordination Numbers. *Angewandte Chemie International Edition* **2014**, *53* (32), 8316–8319.
(30) Calle-Vallejo, F.; Tymoczko, J.; Colic, V.; Vu, Q. H.; Pohl, M. D.; Morgenstern, K.; Loffreda, D.; Sautet, P.; Schuhmann, W.; Bandarenka, A. S. Finding Optimal Surface Sites on Heterogeneous Catalysts by Counting Nearest Neighbors. *Science* **2015**, *350* (6257), 185–189.
(31) Ma, X.; Xin, H. Orbitalwise Coordination Number for Predicting Adsorption Properties of Metal Nanocatalysts. *Physical Review Letters* **2017**, *118* (3), 036101.
(32) Wu, Z.; Jiang, D.; Mann, A. K. P.; Mullins, D. R.; Qiao, Z.-A.; Allard, L. F.; Zeng, C.; Jin, R.; Overbury, S. H. Thiolate Ligands as a Double-Edged Sword for CO Oxidation on CeO_2 Supported $Au_{25}(SCH_2CH_2Ph)_{18}$ Nanoclusters. *J. Am. Chem. Soc.* **2014**, *136* (16), 6111–6122.
(33) Haruta, M.; Kobayashi, T.; Sano, H.; Yamada, N. Novel Gold Catalysts for the Oxidation of Carbon Monoxide at a Temperature Far Below 0 °C. *Chemistry Letters* **1987**, *16* (2), 405–408.

(34) Green, I. X.; Tang, W.; Neurock, M.; Yates, J. T. Spectroscopic Observation of Dual Catalytic Sites During Oxidation of CO on a Au/TiO_2 Catalyst. *Science* **2011**, *333* (6043), 736.
(35) Gao, F.; Wood, T. E.; Goodman, D. W. The Effects of Water on CO Oxidation over TiO2 Supported Au Catalysts. *Catalysis Letters* **2010**, *134* (1), 9–12.
(36) Panapitiya, G.; Wang, H.; Chen, Y.; Hussain, E.; Jin, R.; Lewis, J. P. Structural and Catalytic Properties of the $Au_{25-x}Ag_x(SCH_3)_{18}$ (x = 6, 7, 8) Nanocluster. *Phys. Chem. Chem. Phys.* **2018**, *20* (20), 13747–13756.
(37) Li, G.; Jiang, D.; Liu, C.; Yu, C.; Jin, R. Oxide-Supported Atomically Precise Gold Nanocluster for Catalyzing Sonogashira Cross-Coupling. *Journal of Catalysis* **2013**, *306*, 177–183.
(38) Xu, W. W.; Gao, Y.; Zeng, X. C. Unraveling Structures of Protection Ligands on Gold Nanoparticle $Au_{68}(SH)_{32}$. *Sci Adv* **2015**, *1* (3).
(39) Lewis, J. P.; Jelínek, P.; Ortega, J.; Demkov, A. A.; Trabada, D. G.; Haycock, B.; Wang, H. H.; Adams, G.; Tomfohr, J. K.; Abad, E.; et al. Advances and Applications in the FIREBALL Ab Initio Tight-Binding Molecular-Dynamics Formalism. *physica status solidi (b)* **2011**, *248* (9), 1989–2007.
(40) Becke, A. D. Density-Functional Exchange-Energy Approximation with Correct Asymptotic Behavior. *Phys. Rev. A* **1988**, *38* (6), 3098–3100.
(41) Lee, C.; Yang, W.; Parr, R. G. Development of the Colle-Salvetti Correlation-Energy Formula into a Functional of the Electron Density. *Phys. Rev. B* **1988**, *37* (2), 785–789.
(42) Jin, R.; Zhao, S.; Liu, C.; Zhou, M.; Panapitiya, G.; Xing, Y.; Rosi, N. L.; Lewis, J. P.; Jin, R. Controlling Ag-Doping in $[Ag_xAu_{25-x}(SC_6H_{11})_{18}]^-$ Nanoclusters: Cryogenic Optical, Electronic and Electrocatalytic Properties. *Nanoscale* **2017**, *9* (48), 19183–19190.
(43) Karki, I.; Wang, H.; Geise, N. R.; Wilson, B. W.; Lewis, J. P.; Gullion, T. Tripeptides on Gold Nanoparticles: Structural Differences between Two Reverse Sequences as Determined by Solid-State NMR and DFT Calculations. *J. Phys. Chem. B* **2015**, *119* (36), 11998–12006.
(44) Ranasingha, O.; Wang, H.; Zobač, V.; Jelínek, P.; Panapitiya, G.; Neukirch, A. J.; Prezhdo, O. V.; Lewis, J. P. Slow Relaxation of Surface Plasmon Excitations in Au_{55}: The Key to Efficient Plasmonic Heating in Au/TiO_2. *J. Phys. Chem. Lett.* **2016**, *7* (8), 1563–1569.
(45) Carr, J. A.; Wang, H.; Abraham, A.; Gullion, T.; Lewis, J. P. L-Cysteine Interaction with Au_{55} Nanoparticle. *J. Phys. Chem. C* **2012**, *116* (49), 25816–25823.
(46) Kumara, C.; Aikens, C. M.; Dass, A. X-Ray Crystal Structure and Theoretical Analysis of $Au_{25-x}Ag_x(SCH_2CH_2Ph)_{18}^{(-)}$ Alloy. *The Journal of Physical Chemistry Letters* **2014**, *5* (3), 461–466.
(47) Gottlieb, E.; Qian, H.; Jin, R. Atomic-Level Alloying and de-Alloying in Doped Gold Nanoparticles. *Chemistry (Weinheim an der Bergstrasse, Germany)* **2013**, *19* (13), 4238–4243.
(48) Breiman, L. Random Forests. *Machine Learning* **2001**, *45* (1), 5–32.
(49) Ho, T. K. The Random Subspace Method for Constructing Decision Forests. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **1998**, *20* (8), 832–844.
(50) Chen, T.; Guestrin, C. XGBoost: A Scalable Tree Boosting System. *arXiv:1603.02754 [cs]* **2016**, 785–794.
(51) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; Duchesnay, E. Scikit-Learn: Machine Learning in Python. *Journal of Machine Learning Research* **2011**, *12*, 2825–2830.
(52) Wadell, H. Volume, Shape, and Roundness of Quartz Particles. *The Journal of Geology* **1935**, *43* (3), 250–280.