Supporting Information

Spectra-based Machine Learning for Predicting the Statistical Interaction Properties of CO Adsorbates on Surface

Shuang Jiang, #,1 Xijun Wang, #,1 Yuanyuan Chong,1 Yan Huang,1 Wei Hu,2 Pieter E. S. Smith,3 Jun Jiang1 and Shuo Feng*,1

1Key Laboratory of Precision and Intelligent Chemistry, School of Chemistry and Materials Science, University of Science and Technology of China, Hefei, Anhui 230026, China.
2School of Chemistry and Chemical Engineering, Qilu University of Technology (Shandong Academy of Science), Jinan 250353, China.
3Hefei JiShu Quantum Technology Co. Ltd., Hefei 230026, China.

# S.J. and X.W. contributed equally to this work.

*To whom correspondence should be addressed.
Email: sfeng18@ustc.edu.cn (S.Feng)

KEYWORDS: Spectra-Properties Relationship, Machine Learning, Surface Adsorption, Statistical Method
1. Computational Details
1.1 DFT calculations

Both structural optimization and electronic properties calculation were performed using the Vienna Ab initio Simulation Package (VASP). The plane wave pseudopotential PAW was used to describe the interaction between the nucleus and the electron, and the Perdew-Burke-Ernzerhof (PBE) functions were used to describe the exchange correlation. The cutoff energy was 400 eV, and the convergence standards of force and energy were set as 0.01 eV Å$^{-1}$ and $10^{-5}$ eV, respectively. A Gaussian smearing of 0.1 eV was used for optimization, and the Monkhorst-Pack $3 \times 3 \times 1$ k-point grid was selected to sample the first Brillouin zone. The alloy structure was constructed with a ratio of Au: Ag atoms of 1:1 in two ways. Considering the Au-based binary alloy, we used the lattice constant of fcc Au and replaced 50% atoms in Au unit cell with Ag (called AuAg unit cell). A similar operation was performed on Ag-based binary alloy (called AgAu unit cell). $8 \times 8 \times 8$ k-point mesh was used for cell optimization, and AuAg unit cell with slightly lower energy than AgAu unit cell was selected. A $2 \times 2 \times 1$ supercell was used to build four layers of (111) planes. Atomic configurations of AuAg unit cell and AuAg(111) surface are shown in Figure S1. During the optimization process, the atoms in the bottom two layers were fixed to simulate the bulk phase, and a 20 Å vacuum layer on the top of atoms was used to avoid artificial interaction. The adsorption energy is calculated by $E_{ads} = E_{sur-mol} - E_{sur} - E_{mol}$, where $E_{sur-mol}$, $E_{sur}$ and $E_{mol}$ represent the energy of the adsorption configuration, the initial alloy surface and an individual CO molecule, respectively. The density derived electrostatic and chemical (DDEC06) atomic population analysis method was used to analyze the charge.

Clusters cut from periodic models were used for spectral calculations, as shown in Figure S2. The infrared (IR) and Raman spectra of clusters were calculated by Gaussian 16 using PBEPBE functional. 6-31+G* basis set was used for main-group elements, and pseudo-LANL2DZ basis set was used for metal atoms. Lorentz broadening of spectral lines was performed by Multiwfn code with FWHM of 8 cm$^{-1}$, and the broadened frequency range was set to 0-3500 cm$^{-1}$. In the infrared or Raman spectrum, 2000 intensities were taken at equal intervals, respectively.

![Figure S1](image.png)

**Figure S1.** Configurations of (a) AuAg unit cell and (b) AuAg(111) surface.
**Figure S2.** Six vibrational modes most relevant to CO, including two wagging modes, a CO-metal stretching mode, two metal-C-O bending modes and a C–O stretching mode (from left to right).

**1.2 Structure construction**

**Figure S3.** Schematic diagram of CO adsorption configuration sampling.

In order to generate a large number of adsorption structures, we selected five adsorption sites on the surface of AuAg(111), as shown in **Figure S3**. For each site, we combined four adsorption parameters of CO (θ, β, d_{MC} and d_{CO}). Adsorption angles of CO in the horizontal plane at five sites are recorded as θ_1-θ_5. Seven θ angles were selected at different sites according to the symmetry of clusters, among which θ_1, θ_4, and θ_5 were 0°, 30°, 60°, 90°, 120°, 150°, 180°, θ_2 and θ_3 were 0°, 51.4°, 102.8°, 154.2°, 205.6°, 257.0°, 308.4°. The six values of β were 0°, 15°, 30°, 45°, 60°, and 75°. The five values of d_{MC} were d_1-0.2Å, d_1-0.1Å, d_1, d_1+0.1Å, and d_1+0.2Å, where d_1 is the distance between C atom and the alloy surface in the optimized adsorption configuration. Similarly, the five values of d_{CO} were d_2-0.2Å, d_2-0.1Å, d_2, d_2+0.1Å, d_2+0.2Å, where d_2 is the bond length of CO in the optimized adsorption configuration. There are 4500 adsorption configurations generated by this sampling method. A few of the 4500 configurations may have abnormal Raman intensity due to inappropriate initial structures. Apart from these structures, 4381 configurations are used for subsequent processes. A few structures are likely to be far from the equilibrium state, these configurations are used to simulate extreme conditions and extend the data space.
Figure S4. Adsorption energy ($E_{ads}$), bond energy ($E_b$) and spectra of CO@AuAg(111) as a function of four adsorption parameters of CO ((a) $\theta$, (b) $\beta$, (c) $d_{MC}$ and (d) $d_{CO}$). $\omega_6$, $\text{IR}_6$ and $\text{raman}_6$ represent the frequency, IR intensity and Raman intensity of C-O stretching vibration, respectively.

We plotted the relationship between adsorption parameters, properties and spectra at top-Au site. It was observed that tuning the CO bond length ($d_{CO}$) can significantly affect $E_{ads}$, $E_b$ and $\omega_6$. This is exemplified by the $E_{ads}$ panel in Figure S4d, where a more positive $E_{ads}$ is accompanied by a greater $d_{co}$ change. Conversely, $E_{ads}$ becomes more negative as $d_{co}$ approaches the equilibrium state. Additionally, IR and Raman intensities are simultaneously affected by four adsorption parameters.

1.3 The machine learning protocol

All machine learning algorithms were implemented by Scikit-learn$^9$ and Keras$^{10}$. The one-dimensional convolutional neural network (CNN) we used contains 9 convolutional units. Each convolutional unit consists of a convolutional layer (Conv), a batch normalization layer (BN), an activation function (Act), and a pooling layer. The number of channels is 32 for the first five convolutional layers, and 64 for the last four layers. The batch normalization layer was used to avoid the change of data distribution and improve training efficiency. Downsampling was performed by max-pooling with sliding window size of 2 and stride of 2. The output of nine convolutional units was flattened and fed into a fully connected layer of 64 neurons, and the final outputs were the predicted properties. In the training process, Adam optimizer was used, and a dropout layer with a ratio of 0.5 was added before the fully connected layer to prevent overfitting if necessary. In order to avoid the interference of different data ranges in the
training process, we standardized the infrared and Raman spectra. The data were transformed by \( \chi' = \frac{\chi - \mu}{\sigma} \), where \( \chi' \), \( \chi \), \( \mu \) and \( \sigma \) represent the standardized data, input data, mean value and standard deviation of the input data, respectively. The training set and the test set were randomly divided by 8:2. The prediction accuracy and error of the model were measured by Pearson correlation coefficient (r) and root mean square error (RMSE).

We first selected the number of convolutional units roughly based on the 4000 sampling points of the IR and Raman spectra. In order to extract spectral information as much as possible, the convolutional kernel size was set to 3, and most of the convolutional stride were set to 1. However, the increase in the number of network layers will also lead to the increase of training cost, so the convolutional stride of the last two layers was 2. The feature length is halved after each pooling operation. If too many features are fed to the fully connected layer by flattening, the parameters of the entire CNN network will increase dramatically, thus increasing the training difficulty and overfitting risk. With such parameter configuration, the length of feature reduced to 1 after 9 convolutional units, and the training parameters of the whole network were controlled at an acceptable level.

We also tested the performance of CNN networks with different numbers of convolutional units. Figure S5 shows the predictive performances of models as the number of convolutional units increases from 6 to 10. The IR and Raman spectra of 4381 CO@AuAg(111) were used as features and \( E_{\text{ads}} \) as target property. The performance of CNN network is evaluated by Pearson correlation coefficient (r) and root mean square error (RMSE) between the predicted and DFT calculated values. The r values of all CNN models are close to 0.99, and all RMSEs are between 0.20 and 0.21 eV, which means the performances of networks with 6 to 10 convolutional units on the test set are almost the same. For the selection of activation function for each model, we tested ReLU and ELU. For the dataset of 4381 CO@AuAg(111), ReLU was used for \( E_{\text{ads}} \) and \( \Delta e \), ELU was used for \( E_{\text{b}} \) and \( d_{\text{MC}} \). A learning rate of 0.001 was employed. We also tested the batch_size of 8, 16, 32. Our models were not sensitive to the change of the batch_size, so we chose a medium batch_size=16. For each epoch in the training process, 80% data in the training set was used for training and the other 20% was used for validation. MSE was used as the loss function, and the loss curves for training and validation are shown in **Figure S6**. After 300 epochs, the loss curves of training and validation for each property converged and were very close, indicating that the models did not overfit the training set.
Figure S5. Performances of convolutional neural networks with different numbers of convolutional units in predicting $E_{\text{ads}}$ of test set based on IR and Raman spectra for CO@AuAg(111). The red dotted line represents the average value of five RMSEs, and RMSEs float around the reference line in a small range.

Figure S6. The loss curves for training and validation of four properties ((a) $E_{\text{ads}}$, (b) $\Delta e$, (c) $E_b$, and (d) $d_{\text{MC}}$).

2. Generation of statistical average data sets
2.1
Figure S7. (a) The flow diagram of data sets superposition. (b) The process of selecting two monomolecular adsorption structures from training set and then calculating the adsorption energy ($E_{\text{ads}}$) and IR spectrum of their superposed configuration.

Figure S7a shows the overall process of data sets superposition. The training set or test set of single CO adsorption was superposed to generate new training sets or test sets, respectively. Boltzmann weight and Gaussian weight ($p_i$) were calculated based on the energy of conformations, and the formulas for calculating weight are presented in Section 2.2 (1) (2) and Section 2.3 (4). According to $p_i$, we selected n structures ($n=2,3...9,10,100,200,300$). The average properties and average spectra under different distributions are obtained by calculating the weighted average of these n conformations. The superposition formulas are shown in Section 2.2 (3), Section 2.3 (5) and Section 2.4 (6). Taking the training set as an example, the process of selecting two monomolecular adsorption structures to form a bimolecular adsorption configuration is shown in Figure S7b. The probability of selecting each conformation in the training set is the weight $p_i$. After the two structures were selected, the average adsorption energy (or average IR) was calculated according to the weighted average of adsorption energy (or IR) and $p_i$.

2.2 Boltzmann average

The configurations we used to calculate the weights were derived from a combination of CO adsorption sites and adsorption parameters (Figure S3), which were not obtained by Ab-initio molecular dynamics (AIMD). The distribution ratio of conformation was calculated according to the relative energy between different adsorption configurations at 300K, and the calculation formula was:

$$e^{-E_i/RT} = e^{-(\Delta E_i+E_{\text{rel}})/RT} = Ce^{-\Delta E_i/RT}$$

$$p_i = \frac{e^{-\Delta E_i/RT}}{\sum_j e^{-\Delta E_j/RT}} = \frac{Q_i(\text{Relat})}{Q(\text{Relat})}$$

(1) (2)
Where $E_i$, $E_{Ref}$, $\Delta E_i$, $R$, $T$, $C$, $p_i$ respectively represent the free energy of the $i$th structure, the lowest free energy in all conformations of the training set (or test set), the relative value between the free energy of the $i$th structure and the lowest free energy, the ideal gas constant, temperature, constant, and Boltzmann weight of the $i$th structure. $C$ is a constant related to the reference energy and independent of the specific configuration, so it has no influence on the results. The Boltzmann weight was calculated using the Shermo code\textsuperscript{11}.

The probability of each single CO adsorption structure being selected was its Boltzmann weight. Therefore, structures with lower energy and higher Boltzmann weight were used more often in the superposed data sets, and $n$ configurations were superposed by:

$$f_{\text{Boltzmann}} = \frac{\sum_{i=1}^{n} p_i f_i}{\sum_{i=1}^{n} p_i}$$

Where $f_{\text{Boltzmann}}$, $p_i$, $f_i$ are the properties or spectra after conformational average, the Boltzmann weight of the $i$th conformation, the adsorption properties or spectra of the $i$th conformation, $n=2,3,...,9,10,100,200,300$. As part of our initial data is far from the equilibrium state, the Boltzmann weight of these conformations at 300K is small, and the number of samples in the training set and test set after superposition is 1072 and 268, respectively.

### 2.3 Gaussian average

The weight of each conformation was calculated by the probability density function of Gaussian distribution:

$$p_i = \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Where $p_i$, $x$, and $\mu$ represent the weight of the Gaussian distribution of the $i$th conformation, the energy of the $i$th conformation, and the average energy of all conformations in the training set (or test set), respectively. We tested different values of $\sigma$ to make the data set as scattered as possible over the whole Gaussian distribution curve, and finally $\sigma=0.02$. Similarly, the probability of each initial adsorption structure being selected was its Gaussian weight, and $n$ structures were superposed by:
\[ f_{\text{Gaussian}} = \frac{\sum_{i=1}^{n} p_i f_i}{\sum_{i=1}^{n} p_i} \]  

(5)

Where \( f_{\text{Gaussian}} \), \( p_i \), \( f_i \) are the properties or spectra after the conformational average, the Gaussian distribution weight of the \( i_{\text{th}} \) conformation, the adsorption properties or spectra of the \( i_{\text{th}} \) conformation, \( n=2,3...9,10,100,200,300 \). After superposition, the number of samples in the training set and test set is 4000 and 1000, respectively.

2.4 Arithmetic average

The weight of each conformation was equal, and \( n \) initial adsorption structures were selected with the same probability for the conformational average:

\[ f_{\text{Arithmetic}} = \frac{\sum_{i=1}^{n} f_i}{n} \]  

(6)

Where \( f_{\text{Arithmetic}} \), \( f_i \) are the properties or spectra after the conformational average, the adsorption properties or spectra of the \( i_{\text{th}} \) conformation, \( n=2,3...9,10,100,200,300 \). After superposition, the number of samples in the training set and test set is 4000 and 1000, respectively.

References