Supporting information

Probing Active Sites in Cu_xPd_y Cluster Catalysts by Machine Learning – Assisted X-ray Absorption Spectroscopy

Yang Liu[1†], Avik Halder[2†], Soenke Seifert[3†], Nicholas Marcella[1], Stefan Vajda[2,4,5], Anatoly I. Frenkel[1,6]**

[1]Department of Materials Science and Chemical Engineering, Stony Brook University, Stony Brook, New York 11794, United States

[2]Materials Science Division, Argonne National Laboratory, 9700 South Cass Avenue, Argonne, Illinois 60439, United States

[3]X-ray Science Division, Argonne National Laboratory, 9700 South Cass Avenue, Argonne, Illinois 60439, United States

[4]Institute for Molecular Engineering, The University of Chicago, 5640 South Ellis Avenue, Chicago, Illinois 60637, United States

[5]Department of Nanocatalysis, J. Heyrovský Institute of Physical Chemistry, Czech Academy of Sciences, 18223 Prague 8, Czech Republic

[6]Chemistry Division, Brookhaven National Laboratory, Upton, New York 11973, United States

† Equally Contributing Authors

Corresponding Authors

* Email address: anatoly.frenkel@stonybrook.edu, stefan.vajda@jh-inst.cas.cz

Section 1. Mass spectrum of CuPd clusters

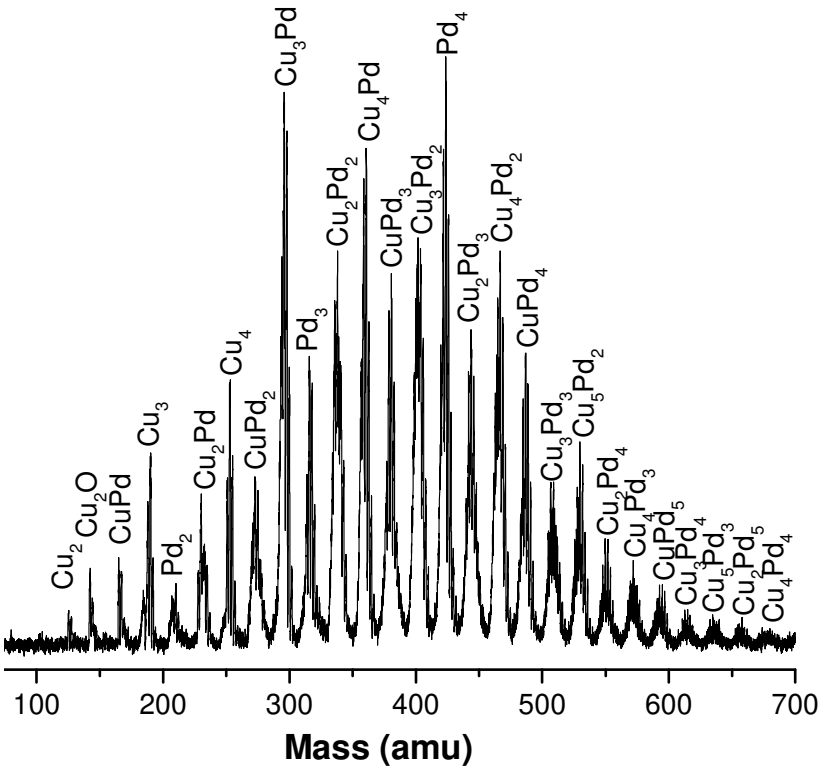

Figure S1. Typical mass spectrum of CuPd clusters produced in vacuum in molecular beans. The apparently noisy peaks reflect the resolution of isotopic distributions.

Section 2. MCR-ALS analysis of the XANES.

To determine the number of components for MCR-ALS analysis, Principal components analysis (PCA) is initially applied to the experimental XANES of Cu_3Pd, Cu_4 and Cu_4Pd in Figure 2a, 3a and 4a. The absorption coefficients are selected for Cu_3Pd, Cu_4 and Cu_4Pd under the same energy range from E_{min}=8899.84eV to E_{max} = 9095.83eV. The package of our PCA analysis was from scikit-learn library.[1] The results of our PCA analysis are listed in Figure 2c, 3c, 4c and Figure 5.

Multivariate Curve Resolution – Alternating Least Square (MCR-ALS) was performed with the code[2,3] for XANES study. The number of components were set as three based on previous PCA analysis. The initial estimation of the C concentration profile, S resolved pure species profile are prepared via purest variable detection method.[4] Two constraints are included in the fitting: 1) the non-negative values for both C and S, 2) the sum of weighting factors of all pure spectra at each reaction step equals to 1. During the iteration process, the number of iterations were set to 200 and the convergence criterion is set as 0.1.

Section 3. FEFF simulation and neural network analysis:

Similar to our previous work,[5–7] we use ab initio code FEFF[8] for XANES simulation. The non-structural parameters for XANES simulations were chosen to ensure the best agreement between the simulated spectrum for the bulk of Cu_2O, CuO and the corresponding experimental XANES data. The results are shown in Figure S3. FEFF version 9.6.4 was used for self-consistent calculation within full multiple scattering (FMS) and muffin-tin (MT) approximations. The cluster's radius 6 Å is for self-consistent calculation and cluster of radius of 7 Å for FMS calculations. S02 is optimized as 0.8 for Cu_2O and 0.9 for CuO. Random phase approximation (RPA) was used to model core-hole, as well as complex exchange-correlation Hedin-Lundqvist potential. A pure imaginary "optical" potential is added to the energy dependent exchange correlation potential 1.5 for Cu_2O and 2.0 for CuO. The detailed description of the $Cu_xPd_{1-x}O$ and $Cu_xPd_{2-x}O$ models' construction has been given the manuscript.

The non-equivalent sites in all cluster models were selected for XANES calculation. All theoretical XANES spectra were shifted in energy by ΔE to align the energy scale of theoretical calculations with experimental data. The value of ΔE for Cu_2O is calculated by the alignment of

the theoretical XANES for the Cu_2O bulk with experimental spectra with energy scale ranging from Emin =8979.97 eV to Emax = 9060.21eV. The value of ΔE for CuO is calculated by the alignment of the theoretical XANES for the CuO bulk with experimental spectra with energy scale ranging from Emin =8981.5 eV to Emax = 9059.3 eV.

To maximize the variability and increase the size of dataset, linear combinations of randomly selected site-specific XANES, based on the equations of $\mu(E) = \sum_i \frac{\mu_i(E)}{N}$, $CN_{CuCu} = \sum_i CN^i_{CuCu}/N$ and $CN_{CuPd} = \sum_i CN^i_{CuPd}/N$, is utilized to enlarge the XANES-CN datasets. Here, we randomly selected three XANES-CNs pair from our original site-specific dataset to create 20000 data points for Cu_2O and 30000 data points for CuO. The correlation between the size of dataset and the mean squared error of the validation dataset has been shown in Figure S6. Similarly to our previous works,[5] we use Wolfram Mathematica 11.3. to construct and train the neural network (NN). Our NN takes discretized XANES spectrum as input and output a vector that describes relevant structural parameters (metal-metal coordination numbers (CNs) of Cu-Cu and Cu-Pd on the first shell). The detailed NN architecture is described in Table S2. The output layer of our NN contains two nodes and the input layer contains 98 points which are determined by the number of points in the discretized XANES spectrum.

For NN training, we use "ADAM" optimization algorithm with default parameters ($\beta 1 = 0.9$ and $\beta 2 = 0.999$). Batch size was 800 for both Cu_2O and CuO, and the training rounds for NN training depends on the training loss and validation loss. Loss function was defined as the L2-norm between output and target vectors averaged across the batch. Three neural networks were trained to overcome the bias and exhibit the stability of NN.

To validate our NN, we used a set of theoretical spectra that were not used for NN training. The theoretical spectra are particle-averaged XANES spectra obtained by averaging the site-specific XANES at each site of the copper oxide models, which corresponded to the $Cu_xPd_{2-x}O$ and $Cu_xPd_{1-x}O$ models with different shapes and sizes. Results of validation for the first Cu-Cu coordination shell are shown in Figure 8. The parameters of our NN models including the number of nodes, activation function, batch size and number of iterations for training are also optimized according to the performance on the validation set.

Section 4. Error estimation:

To estimate the error of our neural network, we have trained three neural networks to get the mean values and uncertainties from the prediction. However, there is also a systematic error in Figure 8 between the true and predicted coordination numbers on the validation dataset. To incorporate the systematic error shown in the prediction of validation dataset into experimental evaluation, we first calculate the mean absolute error based on the difference between true and predicted results shown in Figure 8. Then, we estimate the min and max of the mean absolute error in a 95% confidence interval where the maximum of the mean absolute error represents the maximum error for our neural network predictions. The maximum error is calculated according to the equation below:

$$\max error = \bar{x} + \frac{1.96\sigma}{\sqrt{n}},$$

where $\bar{x}$ is the mean of absolute error between true and predicted CN for the validation dataset, n is the total number of validation dataset, σ is the standard deviation of the absolute error between true and predicted CN for the validation dataset and 1.96 is the Z value for a 95% confidence interval. After the calculation, the max errors are 0.1, 0.1, 0.2, 0.2 for CN(Cu-Cu) in $Cu_xPd_{1-x}O$, CN(Cu-Pd) in $Cu_xPd_{1-x}O$, CN(Cu-Cu) in $Cu_xPd_{2-x}O$ and CN(Cu-Pd) in $Cu_xPd_{2-x}O$. These uncertainties have been added to the prediction results of experimental XANES in Table 1.

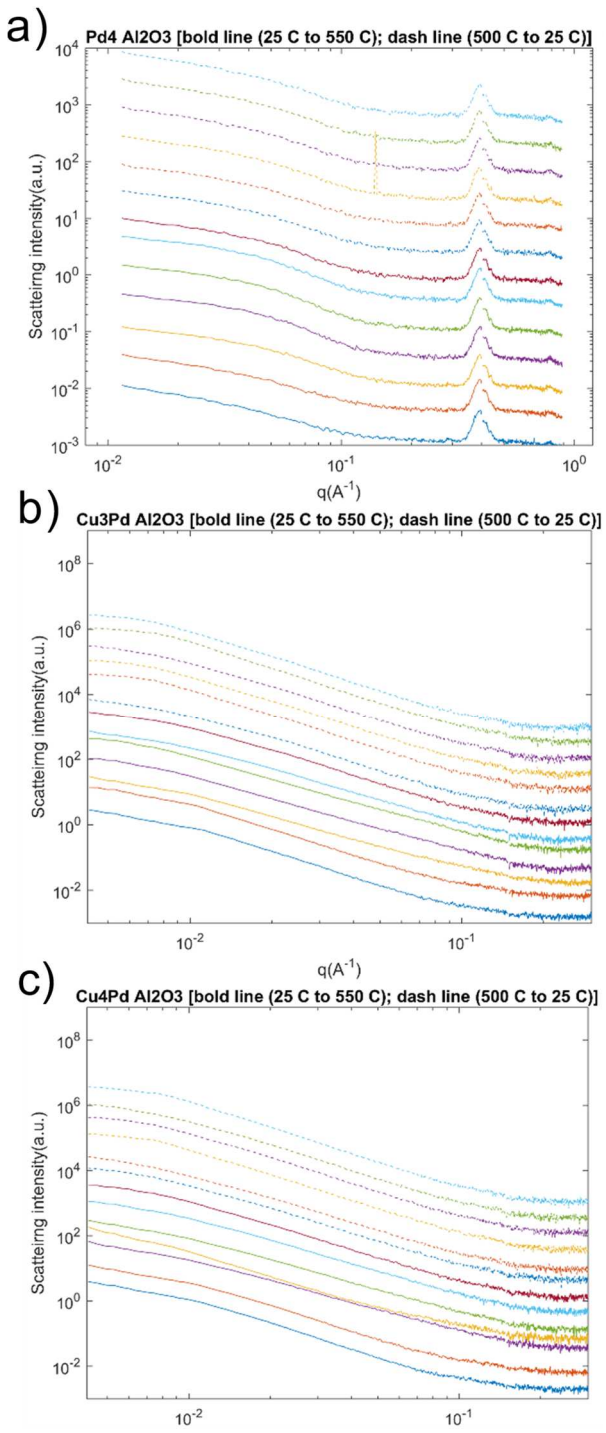

Figure S2. Horizontal cuts of two-dimensional GISAXS data obtained for the heating and cooling ramp for Pd_4, Cu_3Pd and Cu_4Pd clusters under reaction conditions, showing no patterns attributable to larger aggregates.

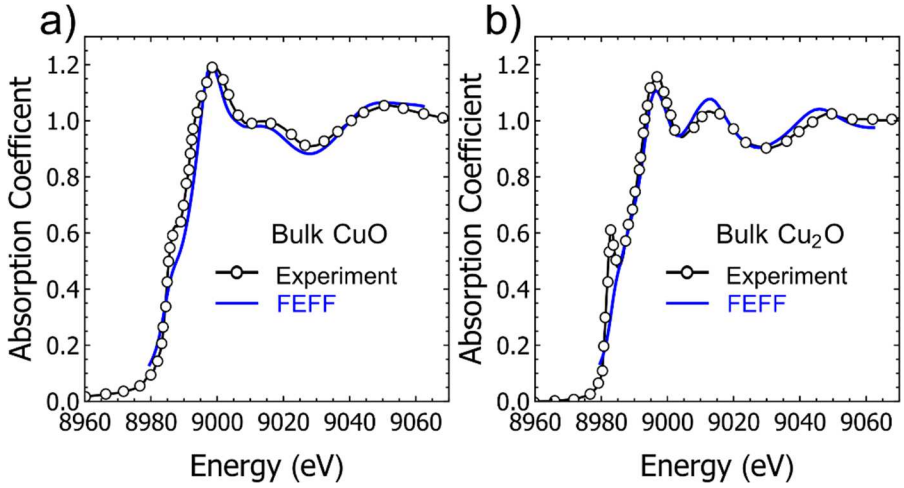

Figure S3. The comparison between the experimental standard and theory of CuO and Cu2O.

Table S1. The lattice parameters for CuO, Cu_2O, PdO and Pd_2O from materials project.[9]

Lattice parameters	Cu_2O	CuO	Pd_2O	PdO
a	4.288 Å	2.933 Å	4.542 Å	3.096 Å
b	4.288 Å	2.933 Å	4.542 Å	3.096 Å
c	4.288 Å	5.133 Å	4.542 Å	5.442 Å
α	90°	90°	90°	90°
β	90°	90°	90°	90°
γ	90°	90°	90°	90°

Table S2. The specific NN architecture as implemented in Mathematica 11.3. The layers are connected sequentially.

NN architecture	$Cu_xPd_{2-x}O$	$Cu_xPd_{1-x}O$
Input	Input (1*98)	Input (1*98)
Hidden layer	1D-Convolutional layer: 64 channels, 16 kernel size	1D-Convolutional layer: 64 channels, 16 kernel size
Hidden layer	Batch Normalization layer	Batch Normalization layer
Hidden layer	ReLu	ReLu
Hidden layer	Pooling layer 2	Pooling layer 2
Hidden layer	Dropout layer 0.2	Dropout layer 0.2
Hidden layer	1D-convolutional layer: 32 channels, 16 kernel size	1D-Convolutional layer: 32 channels, 16 kernel size
Hidden layer	Batch Normalization layer	Batch Normalization layer
Hidden layer	ReLu	ReLu

Hidden layer	Pooling layer 2	Pooling layer 2
Hidden layer	Dropout 0.2	Dropout layer 0.2
Hidden layer	Linear layer 128	Linear layer 128
Hidden layer	tanh	tanh
Output	Output (1*2)	Output (1*2)

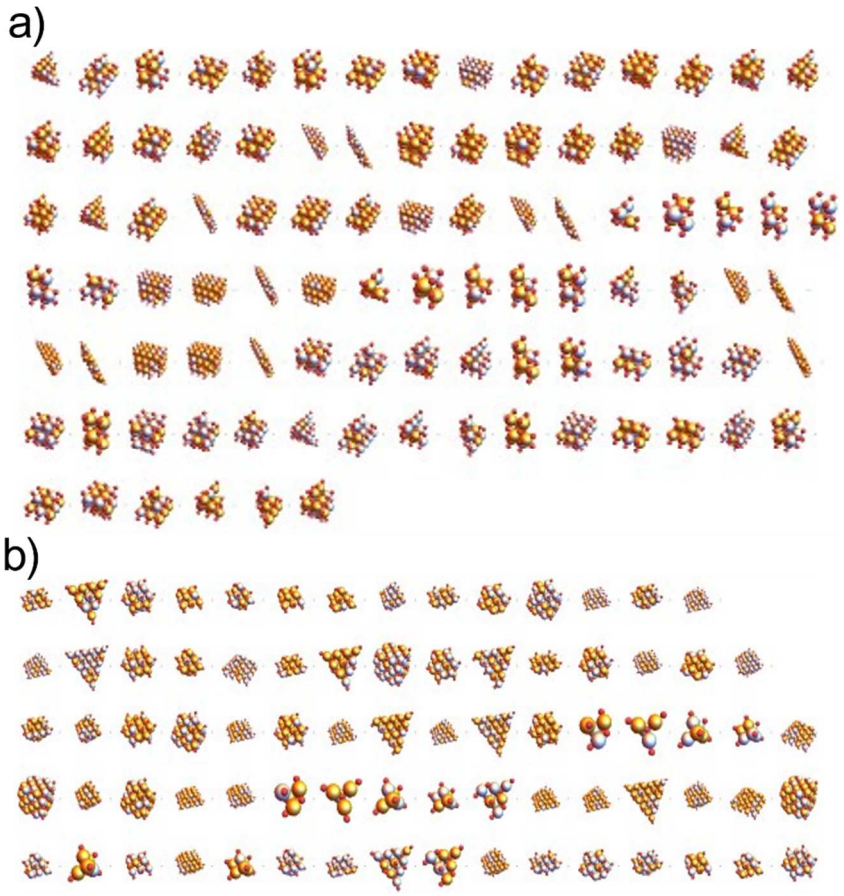

Figure S4. The $Cu_xPd_{2-x}O$ and $Cu_xPd_{1-x}O$ for the FEFF simulation. The ratio of Cu:Pd varied from 4:1, 3:1, 1:1 and 1:3.

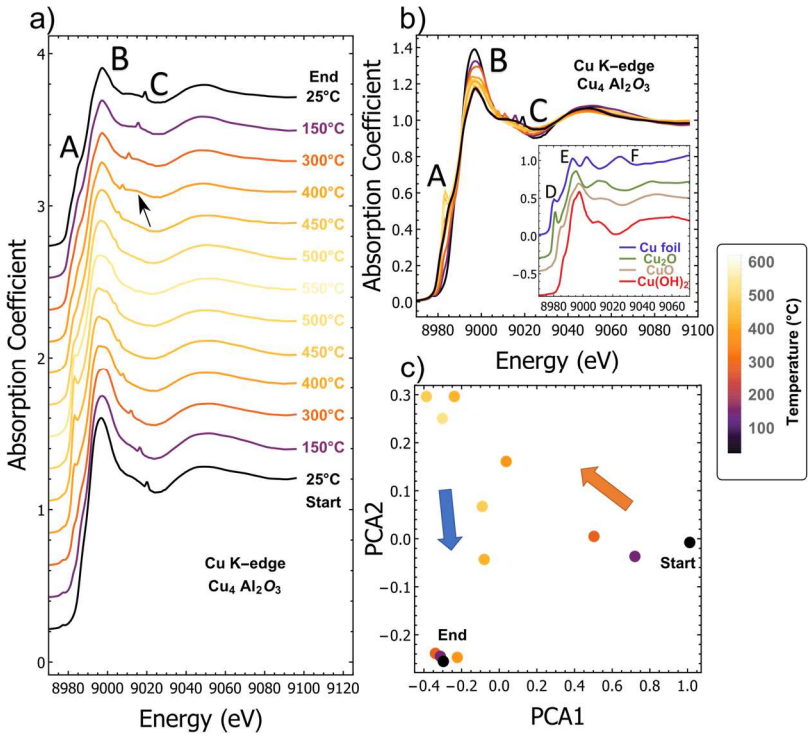

Figure S5. a, b) XANES of Cu_4 from 25℃ to 550℃ (heating cycle) and 500℃ to 25℃ (cooling cycle). The experimental XANES of Cu foil, the bulk of Cu_2O, CuO and $Cu(OH)_2$. c) The visualization of the data distribution from PCA analysis.

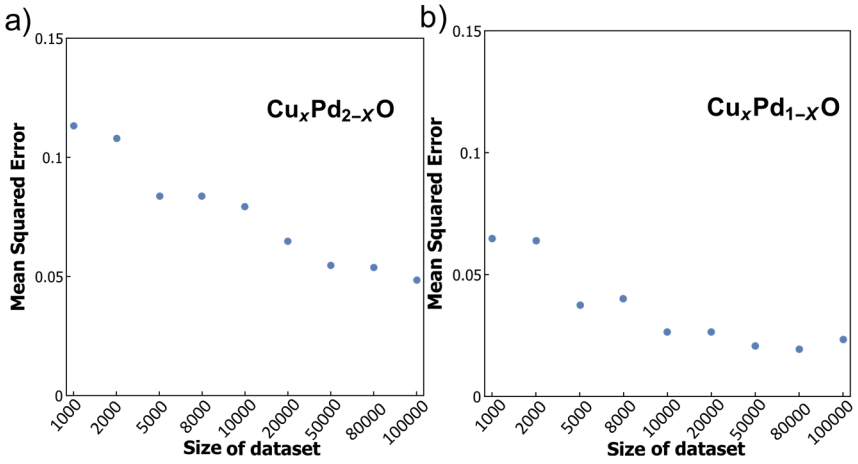

Figure S6. The correlation between the number of dataset and the mean squared loss of validation dataset, a) $Cu_xPd_{2-x}O$, b) $Cu_xPd_{1-x}O$.

References:

(1) 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. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.

(2) Jaumot, J.; Gargallo, R.; De Juan, A.; Tauler, R. A Graphical User-Friendly Interface for MCR-ALS: A New Tool for Multivariate Curve Resolution in MATLAB. *Chemom. Intell. Lab. Syst.* **2005**, *76* (1), 101–110.

(3) Jaumot, J.; de Juan, A.; Tauler, R. MCR-ALS GUI 2.0: New Features and Applications. *Chemom. Intell. Lab. Syst.* **2015**, *140*, 1–12.

(4) Windig, W.; Heckler, C. E.; Agblevor, F. A.; Evans, R. J. Self-Modeling Mixture Analysis of Categorized Pyrolysis Mass Spectral Data with the SIMPLISMA Approach. *Chemom. Intell. Lab. Syst.* **1992**, *14* (1–3), 195–207.

(5) Timoshenko, J.; Lu, D.; Lin, Y.; Frenkel, A. I. Supervised Machine-Learning-Based Determination of Three-Dimensional Structure of Metallic Nanoparticles. *J. Phys. Chem. Lett.* **2017**, *8* (20), 5091–5098.

(6) Timoshenko, J.; Halder, A.; Yang, B.; Seifert, S.; Pellin, M. J.; Vajda, S.; Frenkel, A. I. Subnanometer Substructures in Nanoassemblies Formed from Clusters under a Reactive Atmosphere Revealed Using Machine Learning. *J. Phys. Chem. C* **2018**, *122* (37), 21686–21693.

(7) Liu, Y.; Marcella, N.; Timoshenko, J.; Halder, A.; Yang, B.; Kolipaka, L.; Pellin, M. J.; Seifert, S.; Vajda, S.; Liu, P.; Frenkel, A. I. Mapping XANES Spectra on Structural Descriptors of Copper Oxide Clusters Using Supervised Machine Learning. *J. Chem. Phys.* **2019**, *151* (16), 164201.

(8) Rehr, J. J.; Kas, J. J.; Vila, F. D.; Prange, M. P.; Jorissen, K. Parameter-Free Calculations of X-Ray Spectra with FEFF9. *Phys. Chem. Chem. Phys.* **2010**, *12* (21), 5503–5513.

(9) Jain, A.; Ong, S. P.; Hautier, G.; Chen, W.; Richards, W. D.; Dacek, S.; Cholia, S.; Gunter, D.; Skinner, D.; Ceder, G.; Persson, K. a. The Materials Project: A Materials Genome Approach to Accelerating Materials Innovation. *APL Mater.* **2013**, *1* (1), 11002.