Electronic Supplementary Material (ESI) for Physical Chemistry Chemical Physics.

Supplementary information for

Solving the structure of "single-atom" catalysts using machine learning - assisted XANES analysis

Shuting Xiang,[a] Peipei Huang,[b] Junying Li,[a] Yang Liu,[a] Nicholas Marcella,[a] Prahlad K. Routh,[a] Gonghu Li,*[b] and Anatoly I. Frenkel*[a,c]

a. Department of Materials Science and Chemical Engineering, Stony Brook University, Stony Brook, New York 11794, United States.

b. Department of Chemistry, University of New Hampshire, Durham, New Hampshire 03824, United States.

c. Chemistry Division, Brookhaven National Laboratory, Upton, New York 11973, United States.

Corresponding authors

*Email: anatoly.frenkel@stonybrook.edu, gonghu.li@unh.edu.

Table of Content

Fig. S1. Schematic of experimental setup at the beamline S3

Table S1. Additional details on XAFS measurements S3

Fig. S2. XANES spectra for the control experiment without TEOA S4

Note S1. Control experiment with N_2 S4

Fig. S3. XANES for the control experiment with N_2 gas instead of CO_2 S5

Fig. S4. LCF results of the control experiment with N_2 instead of CO_2 S5

Fig. S5. XANES spectra of the CO_2 in-situ photoreduction and its repeating experiment S6

Fig. S6. LCF results and the reconstructions S7

Fig. S7. EXAFS spectra S8

Note S2. EXAFS fitting S8

Fig. S8. EXAFS fitting result for Co-cyclam-CO S9

Table S2. EXAFS fitting results S9

Fig. S9. XANES spectra showing the sensitivity study of d_C S10

Note S3. Details of FEFF parameters used in XANES simulation S10

Note S4. Neural network architecture S10

Fig. S10. DFT result of Co-cyclam complex reported by Huang P. et al.[1] S12

Fig. S11. MSE of training and validation as the function of epochs S12

Fig. S12. A schematic of Co-cyclam complex showing eight descriptors S13

Fig. S13. Test error distribution of d_C S14

Fig. S14. Test error distribution of d_O S15

References S15

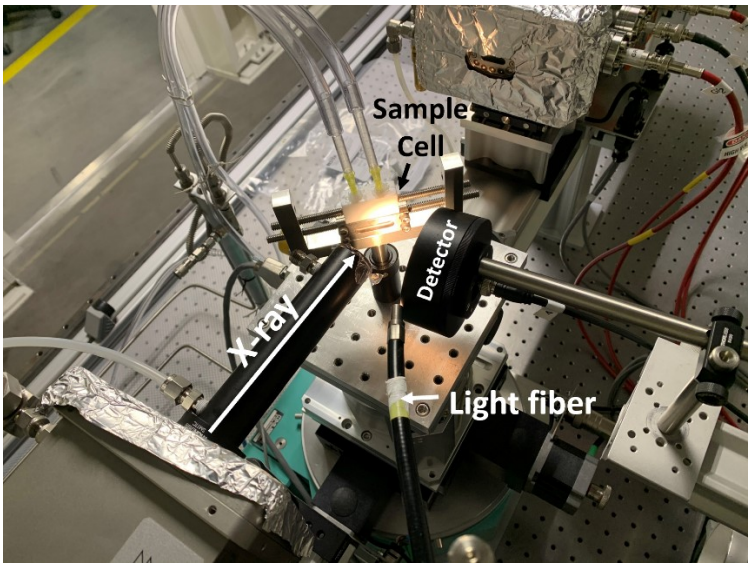

Figure S1. Schematic of experiment setup at the beamline.

Table S1. Additional details on the experimental conditions of the sample for which XAFS measurements were taken (XANES data are shown in Fig. 1). For the TEOA column and CO_2 purge column, the check mark indicates whether the sample is under these conditions. For the column of light, the time for the photocatalysis is referred to the start of the reaction.

Sample Number	Description	TEOA	CO_2 Purge	Duration of Light (Cumulative)
Reference 1	Fresh sample	--	--	--
Reference 2	With TEOA	✓	--	--
1	Before light	✓	✓	--
2	Photocatalysis	✓	✓	36 min
3	Photocatalysis	✓	✓	52 min
4	Photocatalysis	✓	✓	67 min
5	Photocatalysis	✓	✓	82 min
6	Photocatalysis	✓	✓	97 min
7	Photocatalysis	✓	✓	112 min
8	Photocatalysis	✓	✓	128 min
9	Photocatalysis	✓	✓	143 min
10	Post Photocatalysis (Light off)	✓	✓	--
11	Post Photocatalysis (Light off)	✓	✓	--

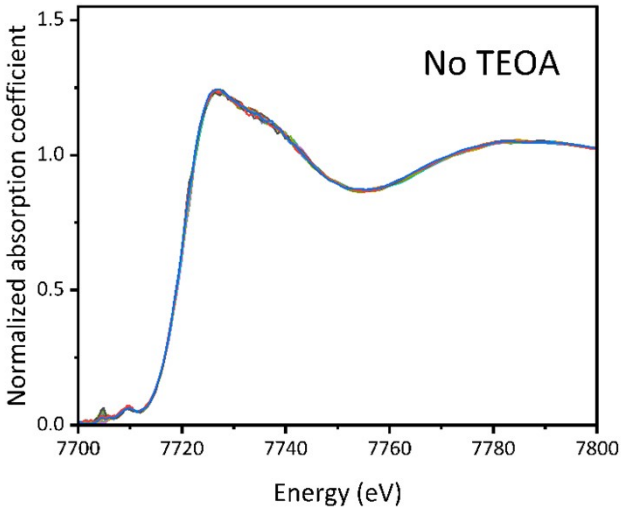

Figure S2. XANES spectra for the control experiment without TEOA, including two spectra of the sample before reaction, eight spectra of the sample with the light turned on (a total of 124 mins), and two spectra of the sample with the light turned off.

Note S1. **Control experiment with N_2**

One control experiment was conducted, in which N_2 purge was used instead of CO_2 purge. However, the reaction still happened based on Fig. S6, and the reaction reached a steady state around 58 min because there was some CO_2 from the atmosphere trapped in the sample cell. The LCF results are shown in Fig. S4(b) with the same standards as used in Fig. S5. As shown in Fig. S3(a), around 50% of the structure was converted.

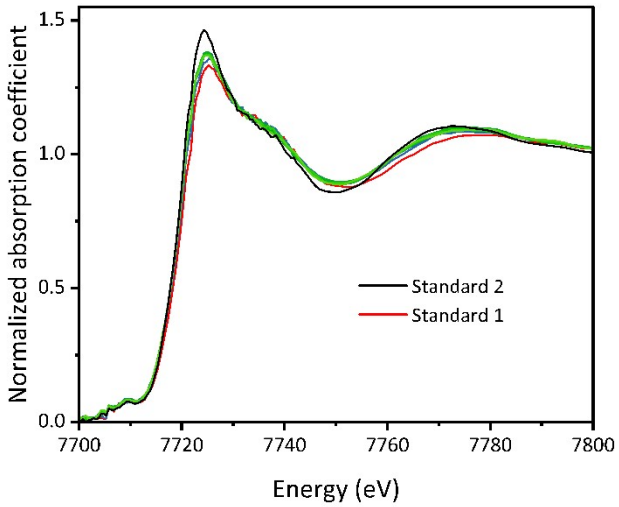

Figure S3. XANES for the control experiment with N_2 gas instead of CO_2. The reaction still happened with N_2 purge instead of CO_2 purge, and the reaction reached steady state around 58 min, because there was some CO_2 from atmosphere trapped in the sample in the presence of TEOA.

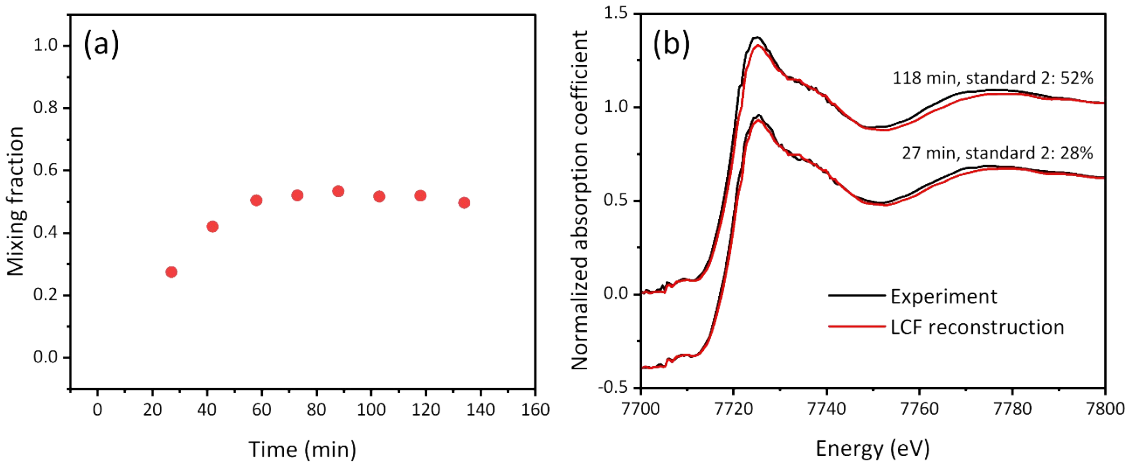

Figure S4. (a) LCF analysis of the control experiment with N_2 instead of CO_2. Around 50% of the structure were converted. (b) A comparison between the experimental XANES spectrum and the corresponding reconstruction XANES spectrum from LCF results.

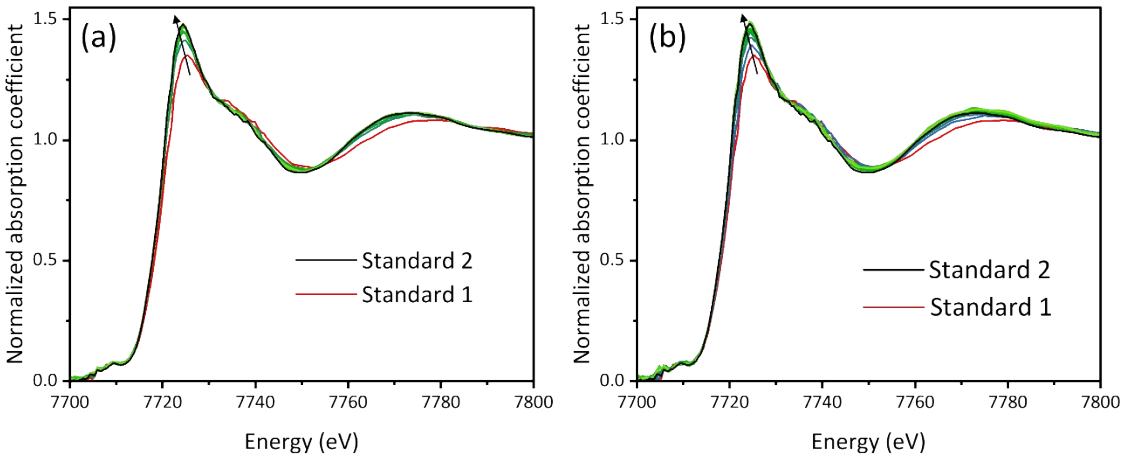

Figure S5. (a) XANES spectra of the CO_2 in-situ photoreduction. Standard one is the sample with TEOA + CO_2 purge before light. Standard two is the sample with light turned on for total of 143 min for reaction. (b) XANES spectra of the repeat experiment with the same standard as (a).

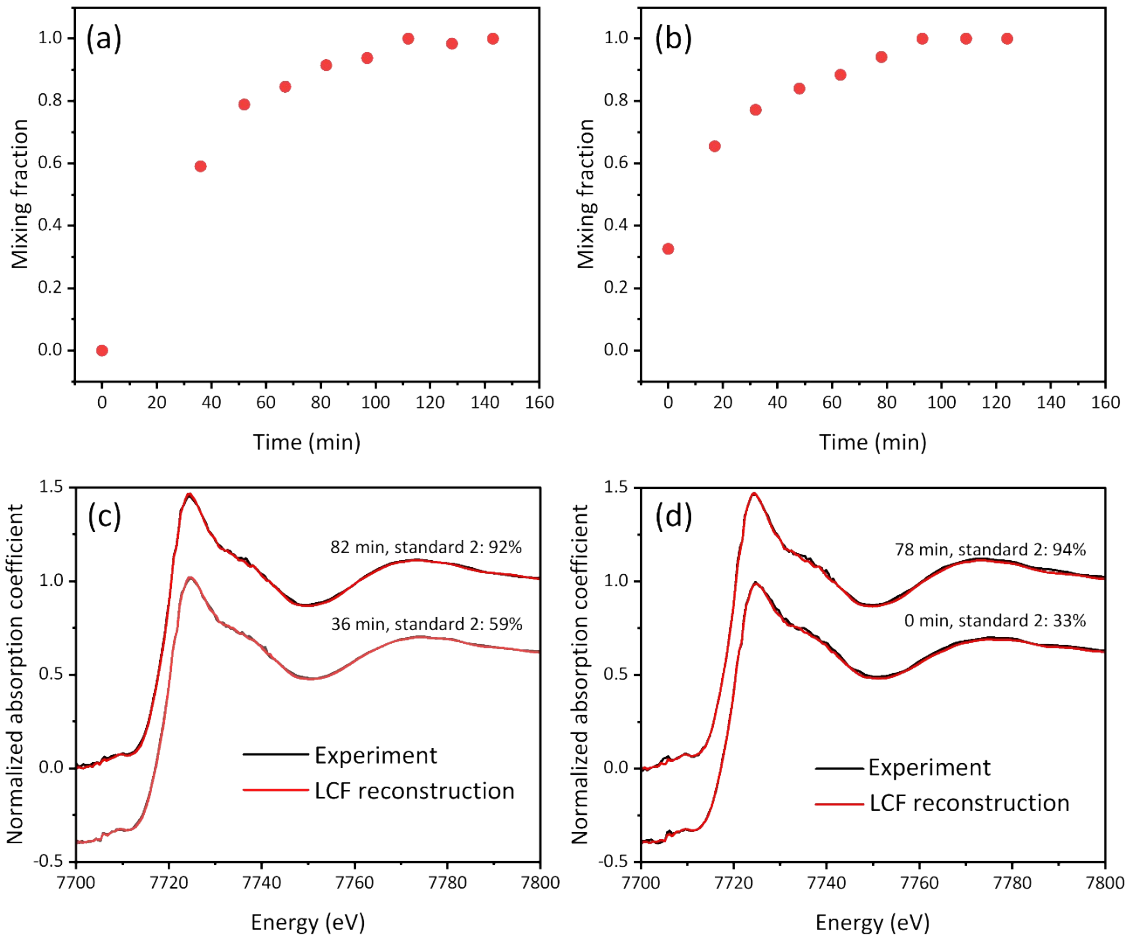

Figure S6. (a) LCF results from the XANES spectra in Fig. S5(a) showing the mixing fraction of standard two as the function of time. (b) LCF results of the repeat experiment from the XANES spectra shown in Fig. S5(b) showing the mixing fraction of standard two as the function of time. (c) and (d) show a comparison between the experimental XANES spectrum and the corresponding reconstruction XANES spectrum from LCF results at two different reaction times for the CO_2 in-situ photoreduction and repeating experiment, respectively.

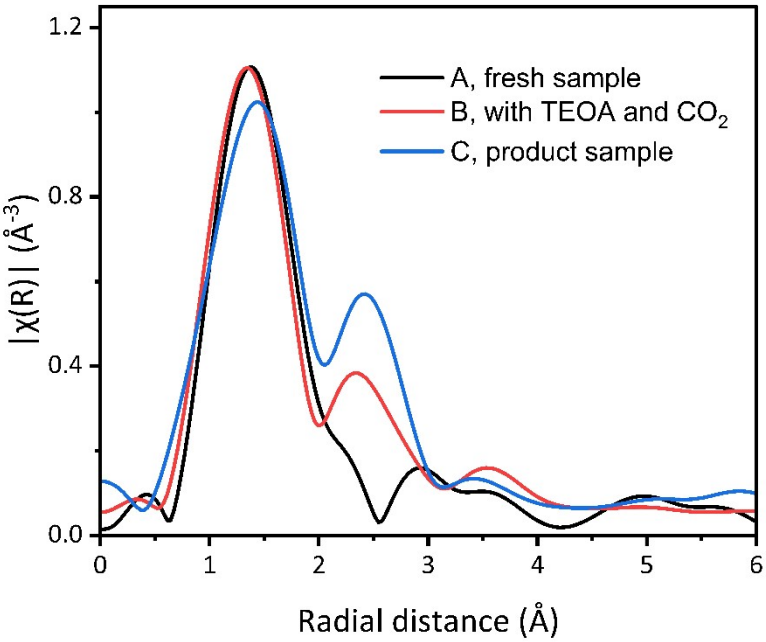

Figure S7. R-space spectra of the fresh sample (A), the sample with CO_2 purge and TEOA (B), the sample at the end of the reaction, the sample at the end of the reaction (the product sample, C). The k-range used in Fourier transforms was from 2 to 7 $Å^{-1}$.

Note S2. EXAFS fitting

The EXAFS spectra were analysed using the Artemis program from the IFEFFIT data analysis package.[2] Before fitting the spectra of the sample at the end of the reaction, the Co foil measured at the same beamline was analysed to obtain the passive electron reduction factor S_0^2 (0.61). The Co-O path from the Co_3O_4 structure was used to fit the first shell. As for the second shell, firstly, the Pd in the $[M(CO)_4][Sb_2F_{11}]_2$ (M = Pd, Pt) was replaced by Co. Then, the Co-C-O forward double scattering path was used to fit the second shell. The Co, C, and O are in the linear position. The fitting results of the product sample are shown in Table S2 and Fig. S8. However, the obtained coordination number is overestimated, most probably due to the limitations of the data quality that allowed for only limited theoretical information to be included for the second shell fitting.

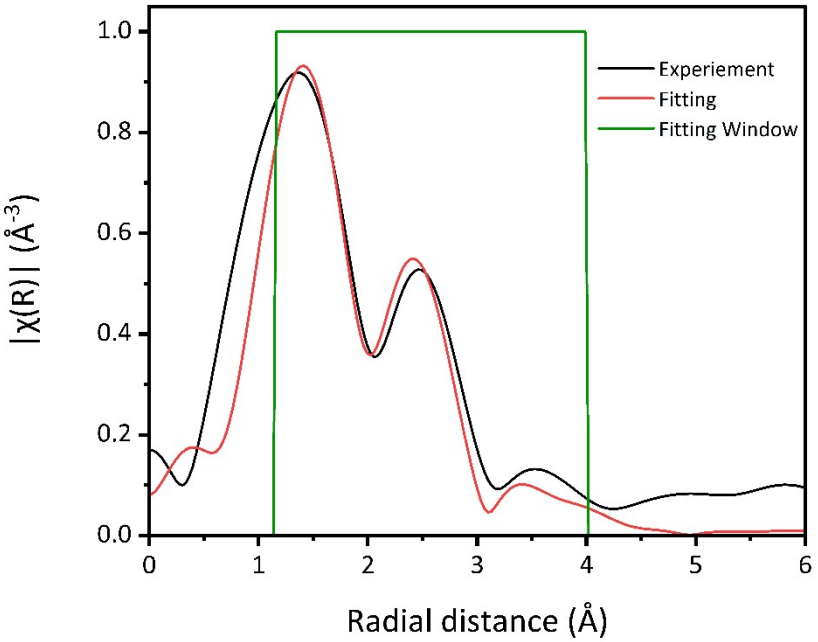

Figure S8. Fourier transform magnitudes of k^2-weighted EXAFS spectra of the sample at the end of the reaction, the product sample (black), and its theoretical fitting (red). The k-range used in Fourier transforms was from 2 to 7 $Å^{-1}$. The fitting range was from 1.15 to 4 Å.

Table S2. EXAFS fitting results.

	Contribution	CN	R(Å)	$\boldsymbol{\sigma}^2$ ($Å^2$)	ΔE_0(eV)
First Shell	Co-O	7.1(4.3)	1.95(6)	0.0127(97)	-6.7(6.5)
Second Shell	Co-CO	5.2(2.2)	3.31(7)	0.0127(97)	8.0(2.9)

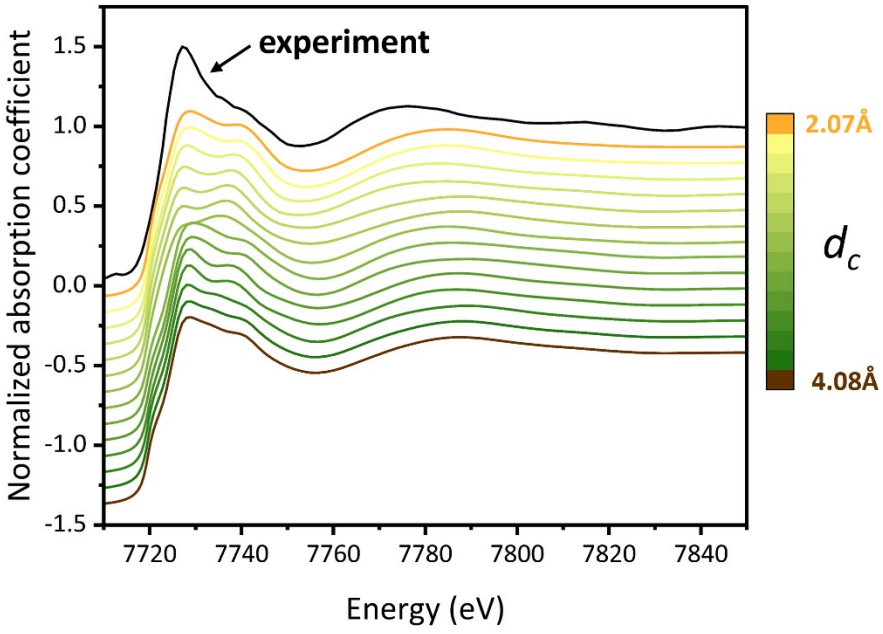

Figure S9. Theoretical XANES spectra calculated using FEFF code showing the sensitivity of d_C in the range of 2.07 Å to 4.08 Å.

Note S3. Details of FEFF parameters used in XANES simulation

FEFF 9.9.1 was used in this study to calculate the theoretical XANES spectra. The reference compound, Cobalt oxide (CoO), was chosen as a standard to optimize the non-structural parameters. S_0^2 of 0.8, where the core-hole was treated with random phase approximation (RPA), was used for the Co K-edge XANES simulation. The radius of the cluster was 7 Å, the maximum number of iterations was 100, the convergence accelerator factor was 3. These were the optimized parameters for FEFF's automated self-consistent potential calculations (*SCF* card). Hedin-Lundquist exchange correlation potential was used. The maximum k value of 7 $Å^{-1}$ and the grid of 0.07 $Å^{-1}$ were used for the XANES card in FEFF calculation.

Note S4. Neural network architecture.

To train, validate and test the model, around 5000 theoretical XANES spectra with two descriptors (d_C and d_O, shown in Fig. 7) changing were calculated by FEFF. Before the neural network training, all the theoretical XANES spectra were interpolated on the same energy mesh from 7711 eV and 7760 eV. Each spectrum was labelled with its corresponding simulated value of d_C and d_O. For each spectrum, its normalized absorption coefficient is stored in a list with these

two labels. Within the whole dataset, the lists of all the theoretical XANES spectra were shuffled to reduce bias[3] and achieve faster convergence and good speedup[4]. 80% of the data were used for training and validation, and 20% of the data were used for testing. For the neural network training, validation, and testing, Tensorflow 2.0 framework was utilized in this study. The number of neurons in the first layer was equal to the number of features. The number of neurons in the output layer was fixed to two since two descriptors needed to be predicted. By tuning hyperparameters (using random search method), such as the number of the hidden layers, the number of neurons in each hidden layer, activation function, the optimizer, with backpropagation[5, 6] the optimized parameters of network with lowest Mean Square Loss (MSE) on validation dataset were selected. In the final neural network architecture, there were two hidden layers with 101 neurons in the first hidden layer and 20 neurons in the second hidden layer. In a neural network, the activation function is responsible for transforming the summed weighted input from one neuron into the non-linear output from the neuron. The activation function for the hidden layers was the rectified linear unit activation function (ReLU)[7]. Adam optimizer[8] was used to optimize the model weights. For the learning process of the model, batch size of 64, epochs of 200, validation split of 0.1, validation batch size of 10 were used. (Validation split of 0.1 means that 10% of the training and validation data were used for validation) With the hyperparameters mentioned above, the neural network was trained and validated. The mean squared error (MSE) of training and validation as the function of epochs are shown in Fig. S11. Based on Fig. S11, the MSE from both training and validation data reached around 3.4×10^{-4}, indicating the model was well-trained. The testing results are shown in Fig. 9, and the corresponding error distributions are shown in Fig. S13 and Fig. S14. The R^2 score calculated from the testing data is 0.99. The equation of R^2 is shown as following:

$$R^2(y,\hat{y}) = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}, where\ \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

$\hat{y}_i$ is the predicted value of the i-th sample, y_i is the corresponding true value. R^2 score is an indication of the goodness of fit and a measure of how well the unseen samples are likely to be predicted by the model. The best possible score is 1.0.

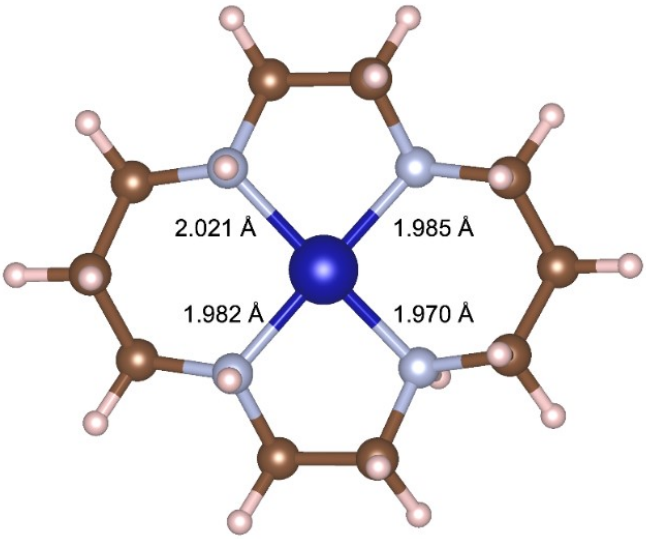

Figure S10. DFT result of Co-cyclam complex reported by Huang P. *et al*.[1] The four bond lengths between Co and the four N atoms from the cyclam according to the DFT results. (Additional details can be collected from N. Aaron Deskins)

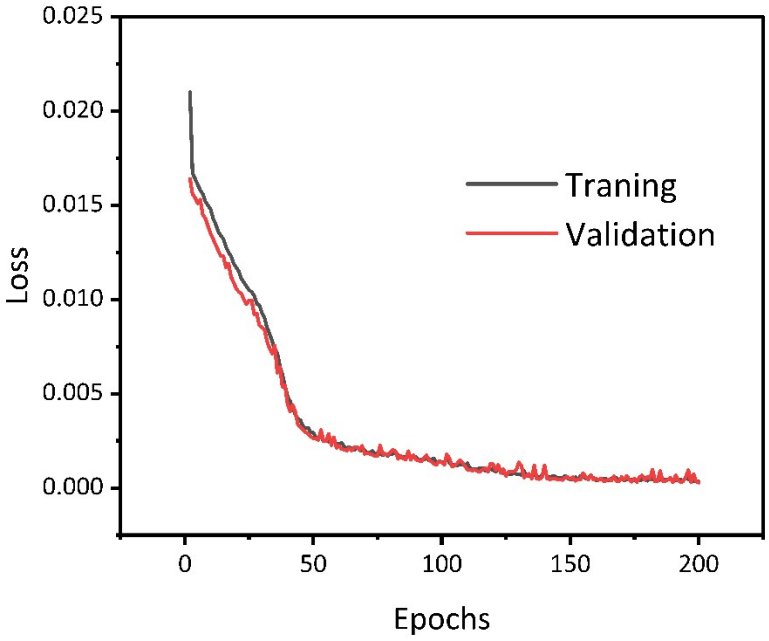

Figure S11. MSE loss during training and validation as the function of epochs.

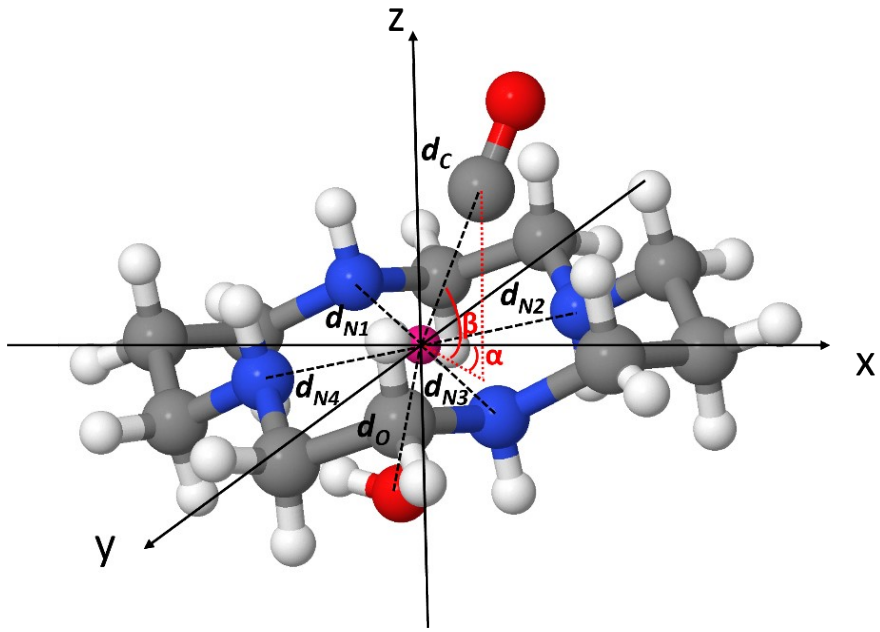

Figure S12. A schematic of Co-cyclam complex showing eight descriptors used for theoretical structure generation, including six bond lengths (d_{N1}, d_{N2}, d_{N3}, d_{N4}, d_C, and d_O) and two bond angles (α and β).

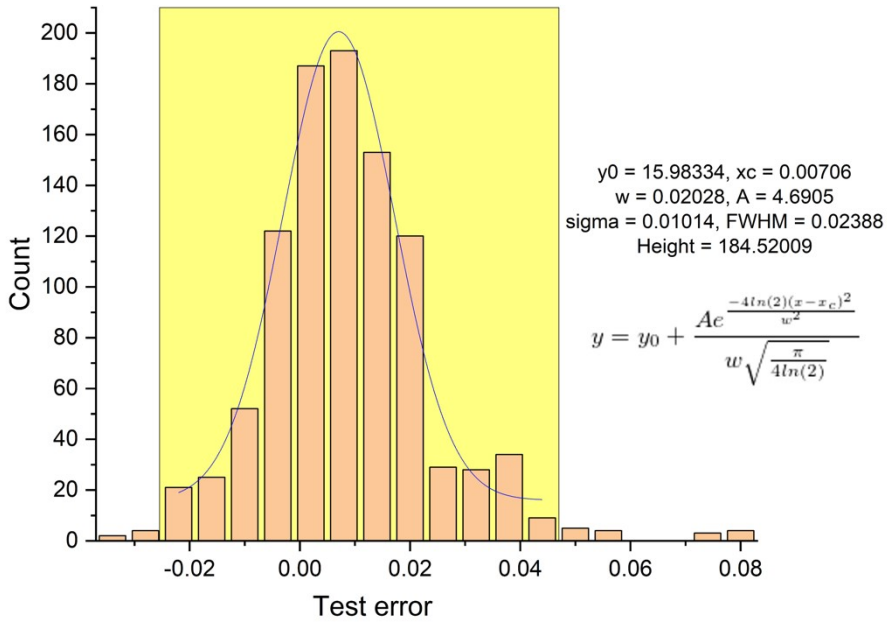

Figure S13. Test error distribution of d_C. The test error values represent the difference between true values and the predicted values from NN. A Gaussian function (the equation is shown in the figure) was used to fit the distribution and determine the standard deviation of test error to be 0.01.

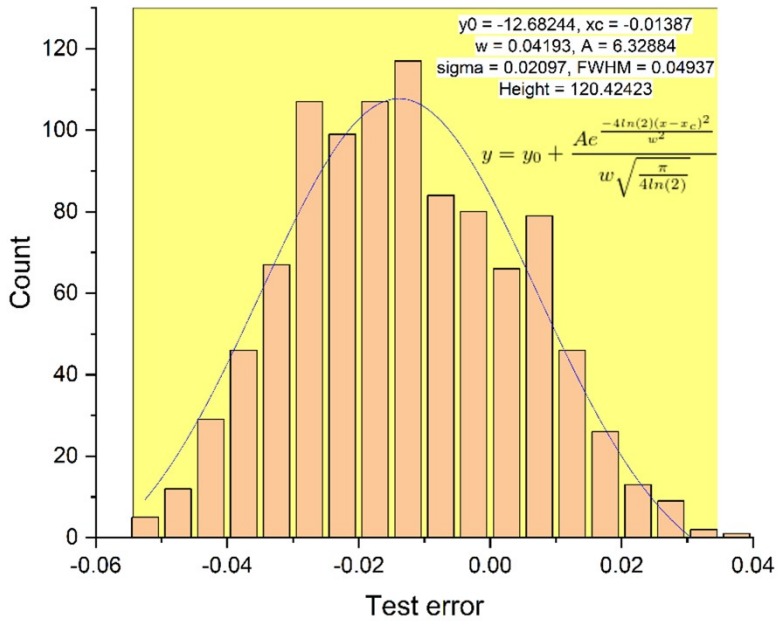

Figure S14. Test error distribution of d_O. The test error values represent the difference between true values and the predicted values from NN. A Gaussian function (the equation shown in the figure) was used to fit the distribution and determine the standard deviation of test error to be 0.02.

References:

1. P. P. Huang, S. A. Pantovich, N. O. Okolie, N. A. Deskins and G. H. Li, *Chemphotochem*, 2020, **4**, 420-426.
2. B. Ravel and M. Newville, *J. Synchrotron Radiat.*, 2005, **12**, 537-541.
3. M. A. Montemurro, R. Senatore and S. Panzeri, *Neural Comput*, 2007, **19**, 2913-2957.
4. Q. Meng, W. Chen, Y. Wang, Z.-M. Ma and T.-Y. Liu, *Neurocomputing*, 2019, **337**, 46-57.
5. D. E. Rumelhart, G. E. Hinton and R. J. Williams, *Nature*, 1986, **323**, 533-536.
6. H. Azami, M.-R. Mosavi and S. Sanei, *Wireless personal communications*, 2013, **71**, 789-803.
7. A. Agarap, *arXiv preprint*, 2019, **1803**, 08372.
8. D. P. Kingma and J. Ba, Adam: A Method for Stochastic Optimization. *Proceedings of the 3rd International Conference on Learning Representations* (ICLR 2015), 2015.