Supporting Information for: Comprehensive sampling of coverage effects in catalysis by leveraging generalization in neural network models

Daniel Schwalbe-Koda*[,1,2,3], Nitish Govindarajan*[,1,2,4], and Joel B. Varley[1,2]

[1] *Materials Science Division, Lawrence Livermore National Laboratory, Livermore, CA 94550, United States*

[2] *Laboratory for Energy Applications for the Future (LEAF), Lawrence Livermore National Laboratory, Livermore, CA 94550, United States*

[3] *Department of Materials Science and Engineering, University of California, Los Angeles, Los Angeles, CA 90095, United States*

[4] *School of Chemistry, Chemical Engineering, and Biotechnology, Nanyang Technological University, 21 Nanyang Link, Singapore 637371, Singapore*

*E-mail: dskoda@ucla.edu; nitish.govindarajan@ntu.edu.sg

Supplementary Text

Simplified SOAP + lateral interaction ML model

An overview of the simplified lateral interaction model is shown on Fig. S10. The interaction of the adsorbate with the surface and with other adsorbates are represented by an on-site descriptor (x_i) and adsorbate distances (d_{ij}), respectively (Fig. S10a). Such an approach

isolates the effects of adsorbate-surface and adsorbate-adsorbate interactions when predicting the binding energies for the systems of interest, and does not rely on learnable descriptors from graph neural networks (GNNs). Instead, radial basis functions (RBF) are used to map d_{ij} into vectors (r_{ij}) that are used along with the on-site descriptors x_i, x_j as inputs to a neural network (NN) that predicts the interaction energies between adsorbates (E_{ij}). A separate neural network using the on-site descriptor is used to predict the adsorbate-surface interaction energy (E_i) that is independent of the coverage. Finally, the total binding energy is predicted by combining the two NNs as the sum of these interaction energies (Fig. S10b).

The results regarding this simplified model for the dataset of CO on Cu are summarized in Fig. 3 and discussed in the main text. To implement this model, we used the Smooth Overlap of Atomic Positions (SOAP) as descriptor x_i for the binding sites. For each binding site in the structure, the power spectrum was computed using 7 radial basis functions, 6 angular basis functions, smearing of 0.3 Å, and a cutoff of 5.0 Å. The descriptors were computed using the package `dscribe` and only considered the C-Cu pairs. This led to vectors x_i of length 343 floating point numbers. The distance r_{ij} between adsorption sites was represented using a Gaussian basis function, with 8 trainable basis functions of varying widths and initially centered at zero. A cutoff of 5.0 Å was used for representing r_{ij}. The NNs used to predict E_i and E_{ij} were simple, feedforward NNs with 3 hidden layers of size 800 each. Because E_i is predicted directly from x_i, the input size of the first NN was 343. On the other hand, an input size of 694 was used to predict E_{ij}. The Mish activation function was used for the NNs. No batch normalization was used.

The code for the model was written in PyTorch, and the model was trained on 4 NVIDIA V100 GPUs in the Lassen supercomputer, with training and parallelization routines implemented within PyTorch Lightning. The AdamW optimizer with initial learning rate of 10^{-3} and weight decay of 10^{-2} was used to minimize the mean squared error between predicted and true binding energies. A learning rate scheduler was used to decrease the learning rate by a factor of 0.5 if the validation loss plateaus for 20 consecutive epochs. A stochastic

weight averaging (SWA) callback was also used after epoch 700, with a starting learning rate of 3×10^{-4} and cosine annealing function every 10 epochs. The simplified model was trained for a minimum of 1000 epochs and a maximum of 1200 epochs, and the best chosen model was the one that minimizes the validation loss. The final prediction of binding energies was performed using an ensemble of 6 simplified models with different intialization, but the same train/validation splits.

Supplementary Figures

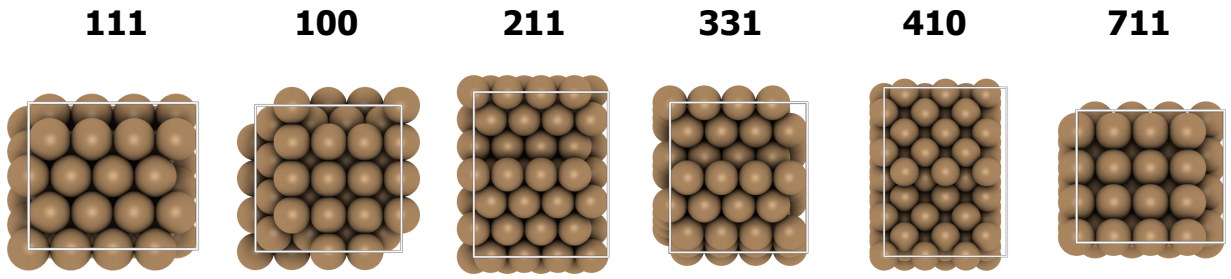

Figure S1: Visualization of the six Cu surfaces (111, 100, 211, 331, 410 and 711) considered in this study.

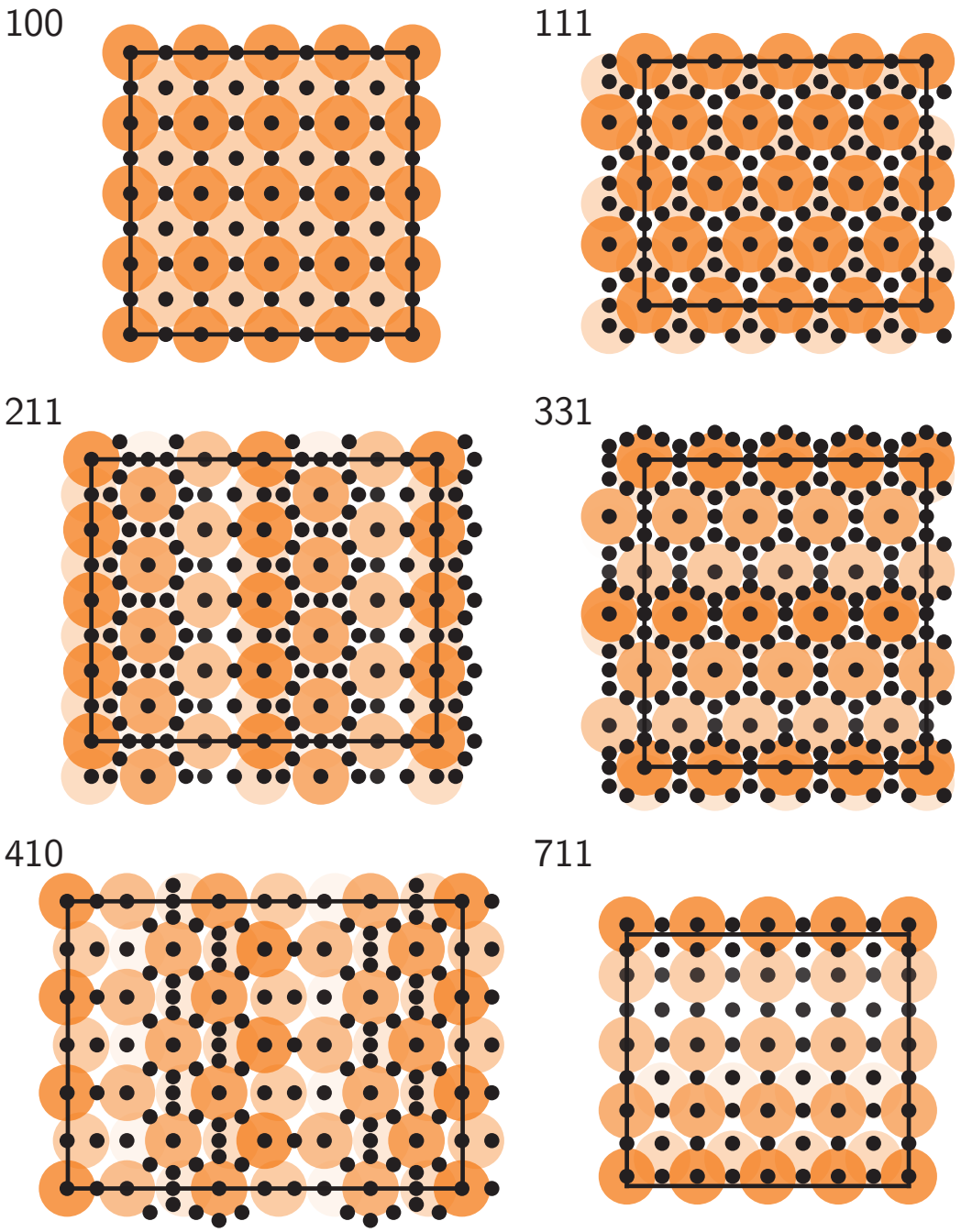

Figure S2: Visualization of binding sites from the six Cu surfaces (100, 111, 211, 331, 410 and 711) considered in this study. Binding sites are shown in black, and copper atoms are shown in orange. Transparency is used to indicate depth, with more opaque atoms closer to the surface.

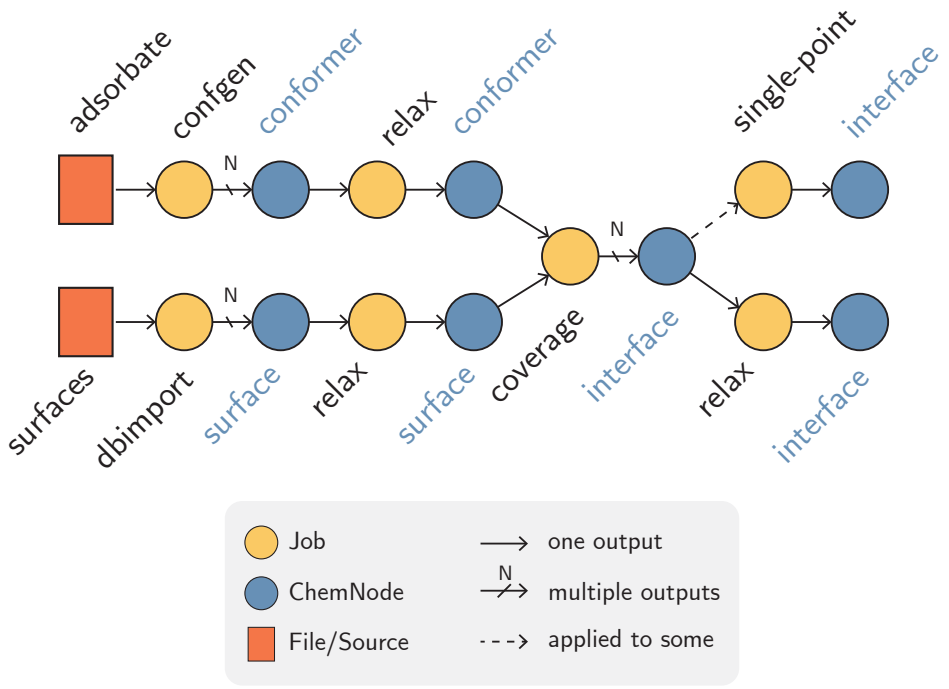

Figure S3: Schematic of the complete simulation workflow performed in this study. The pipeline was implemented using mkite.

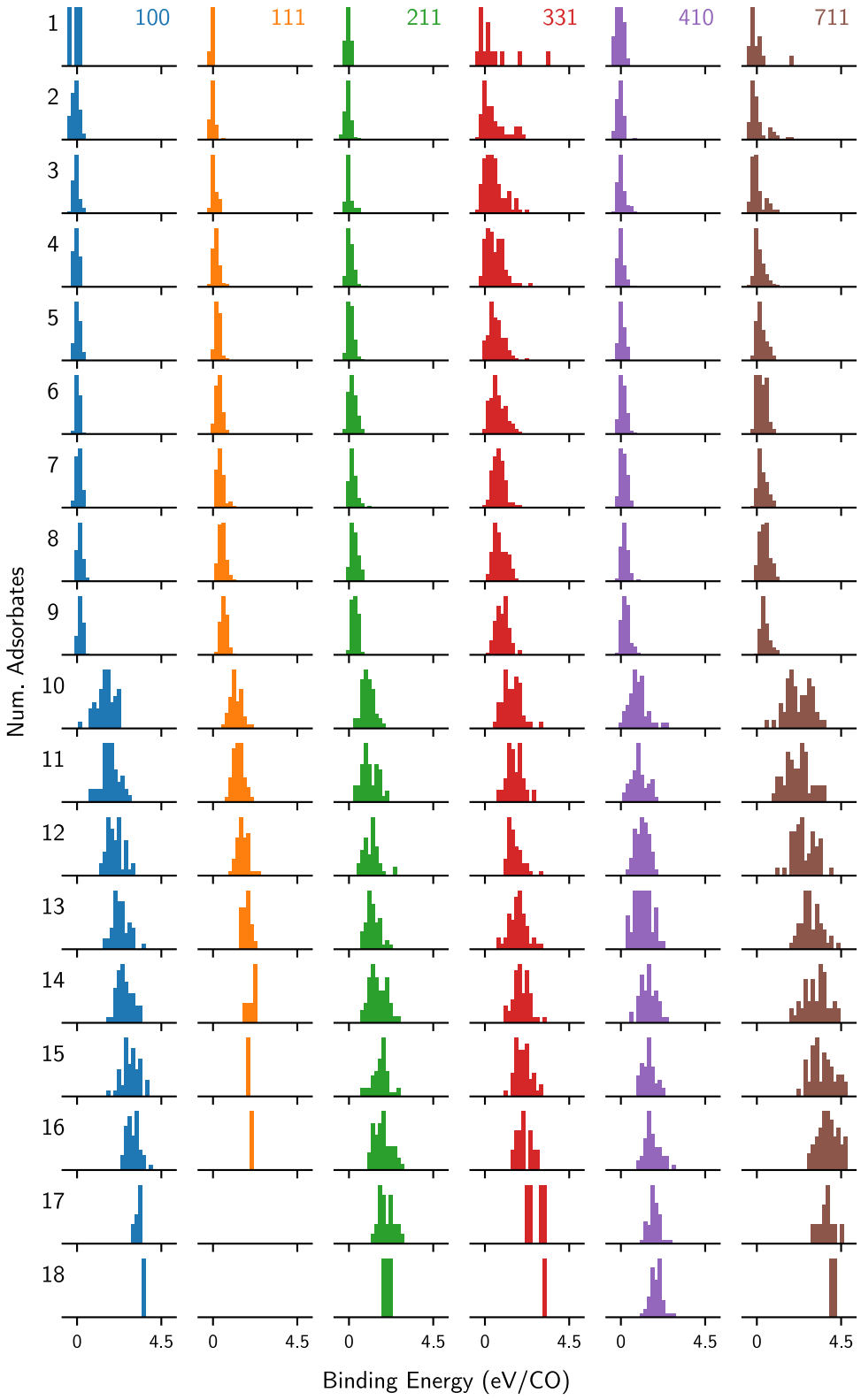

Figure S4: Distributions of the original dataset energies computed with DFT separated by facet and coverage. Configurations were sampled randomly (see Methods). The energies shown in these distributions are for unrelaxed systems, i.e., single-point DFT calculations on the energies from initial structures

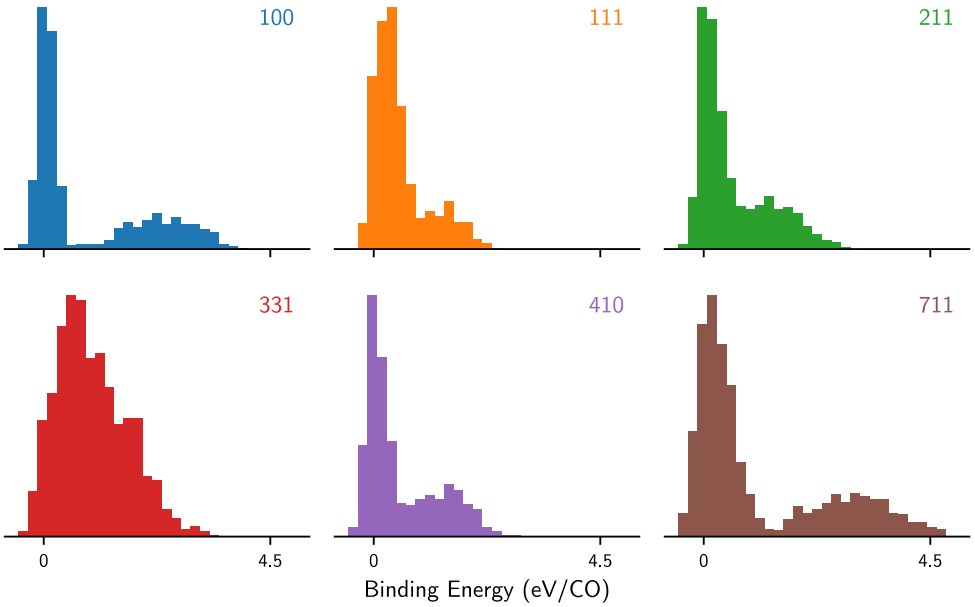

Figure S5: Distributions of the original (unrelaxed) dataset energies computed with DFT separated by facet. These distributions are obtained by consolidating the distributions in Fig. S4.

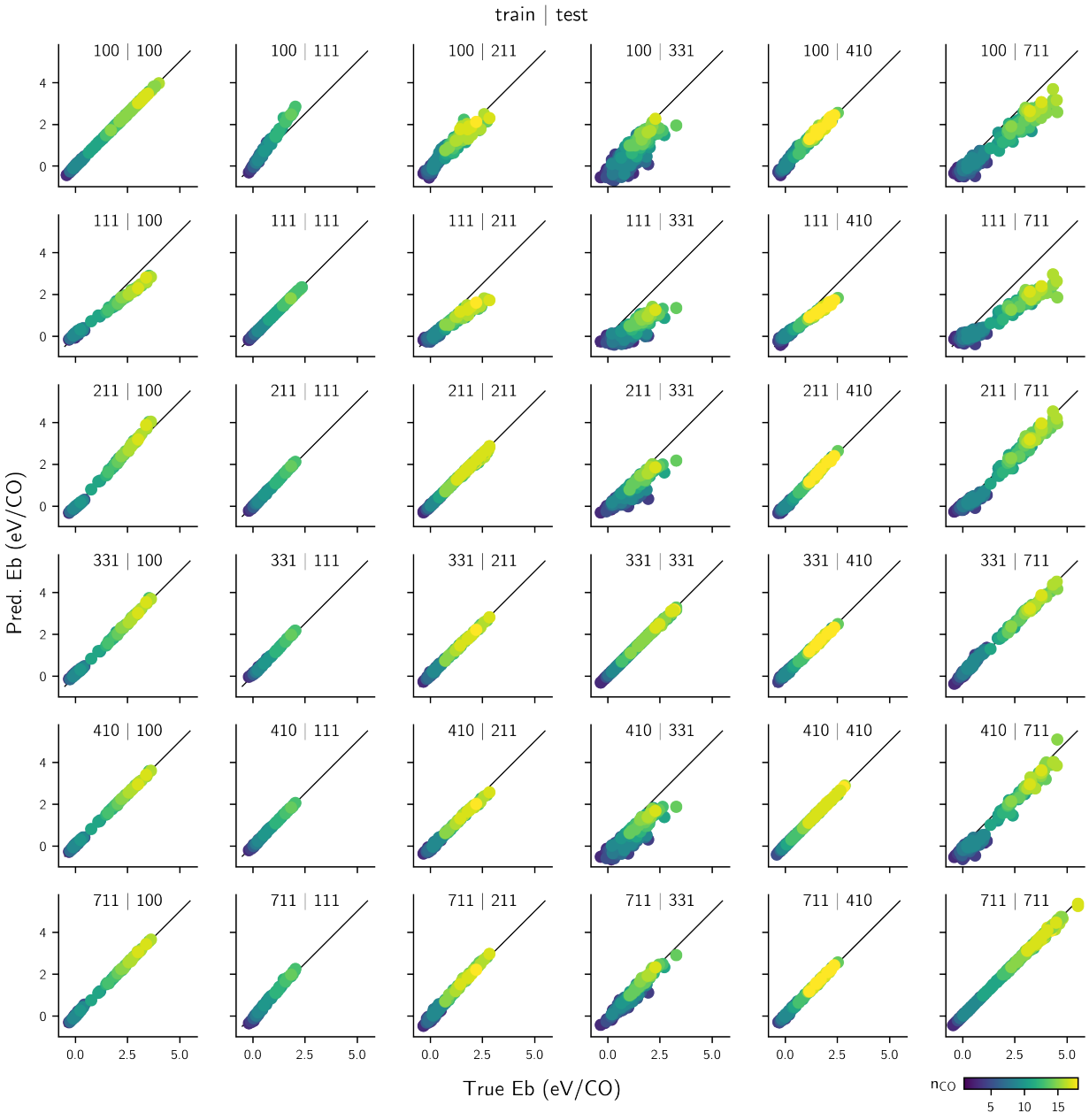

Figure S6: Parity plots for the MACE models trained in this study on different facets. Each row represents a model trained on a single facet, and each column represents the facet used for evaluation of that model. The plots are annotated with the (train | test) labels for clarity. The color depicts the coverage, with higher coverages corresponding to brighter colors. The black line represents the perfect prediction. For diagonal terms, only points in the test set are plotted.

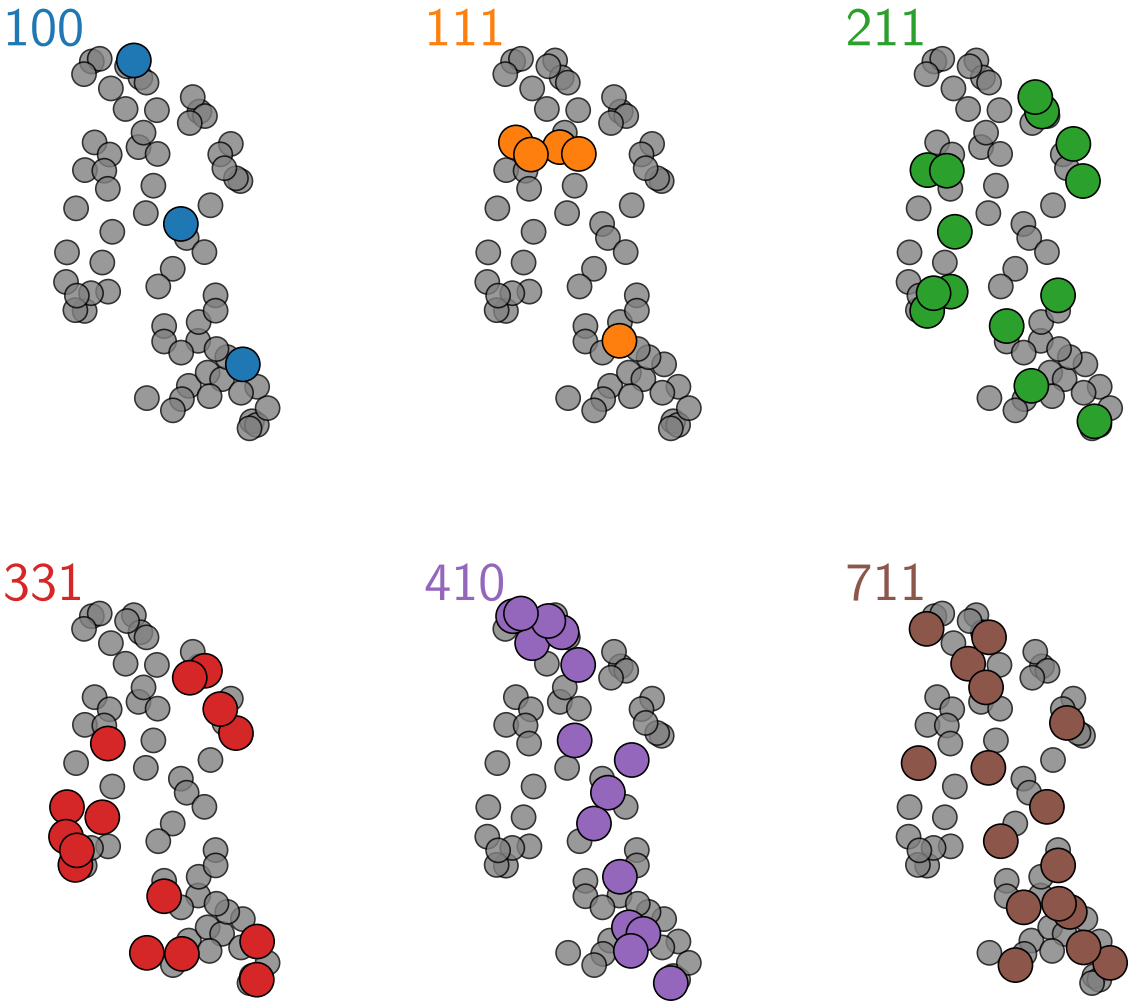

Figure S7: Two-dimensional visualization of unique binding sites in Cu facets, obtained from the cosine distance between SOAP fingerprints and UMAP. Binding sites for all facets are shown in grey, and colored dots are binding sites of each facet.

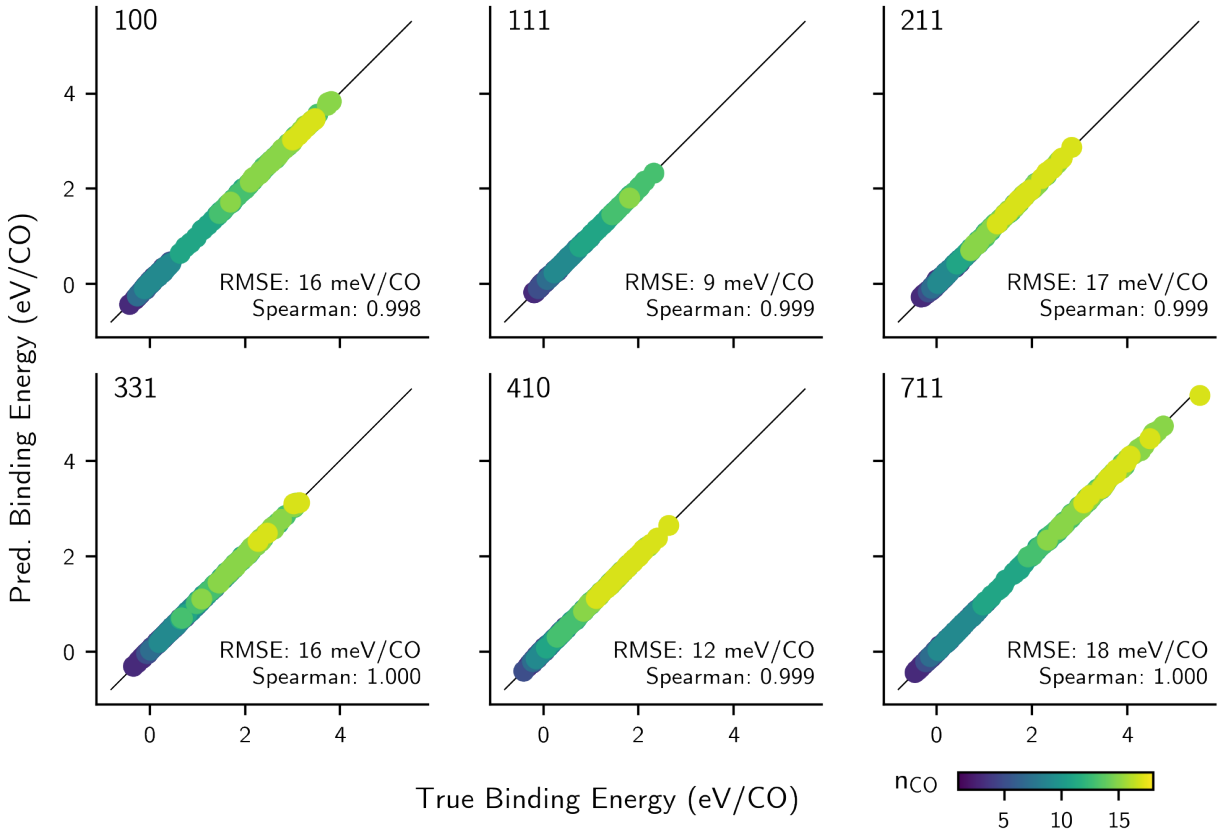

Figure S8: Parity plots for the MACE models trained on coverages with even number of CO adsorbates and tested on coverages with odd number of CO adsorbates, on a per-facet basis. All configurations with a single CO were included in the training set.

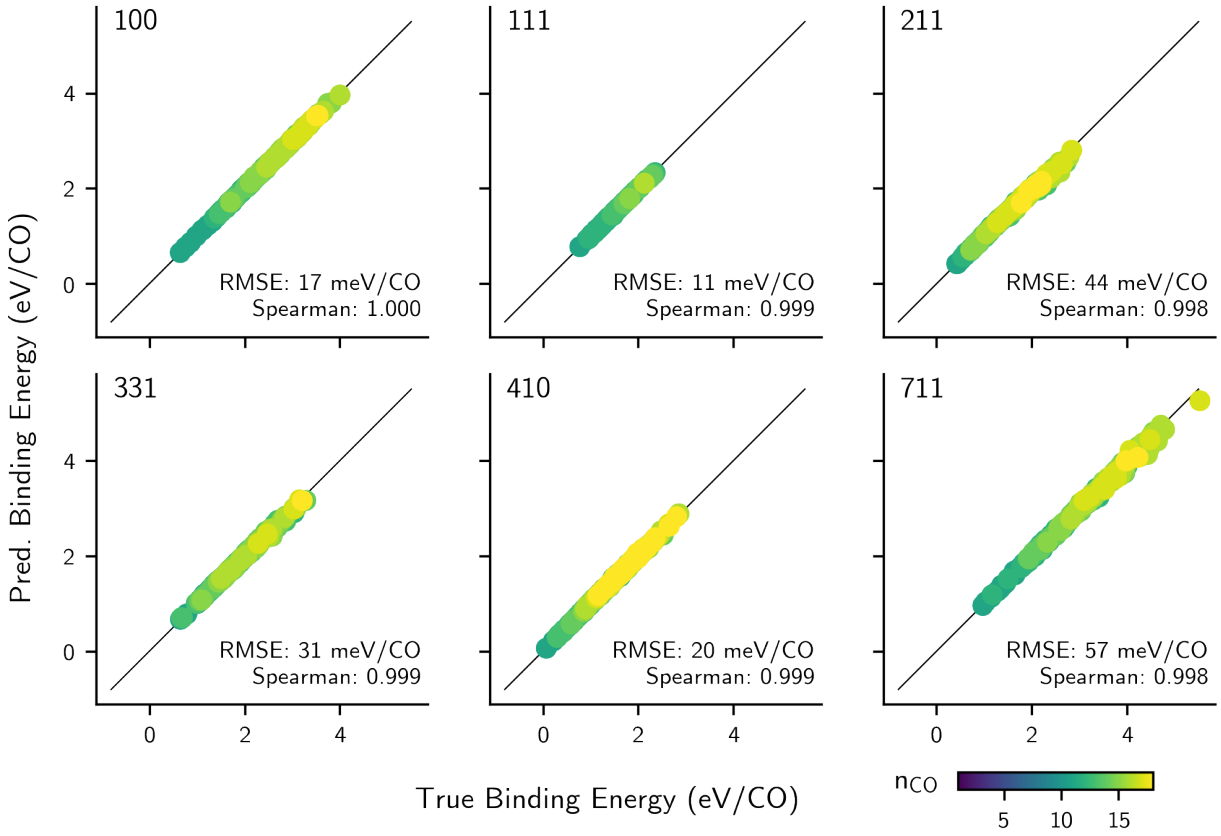

Figure S9: Parity plots for the MACE models trained on lower CO coverages ($n < 10$ CO) and tested on higher CO coverages ($n \geq 10$ CO) for each facet.

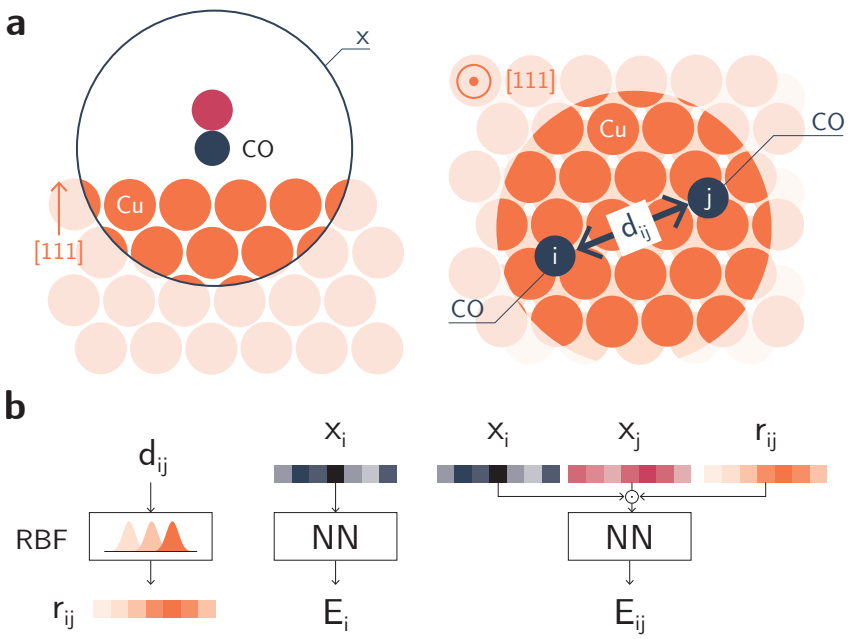

Figure S10: ML architecture for a simpler lateral interaction model. **a**. Interactions between adsorbates and surfaces are represented by an on-site descriptor (adsorption site) and an adsorbate-adsorbate distance (d_{ij}), **b**. The distance between the adsorbates is mapped into a vector using a Gaussian basis function and, along with the descriptors, is used to predict interaction energies. The descriptors alone are also used to predict an on-site interaction energy that does not depend on the coverage.

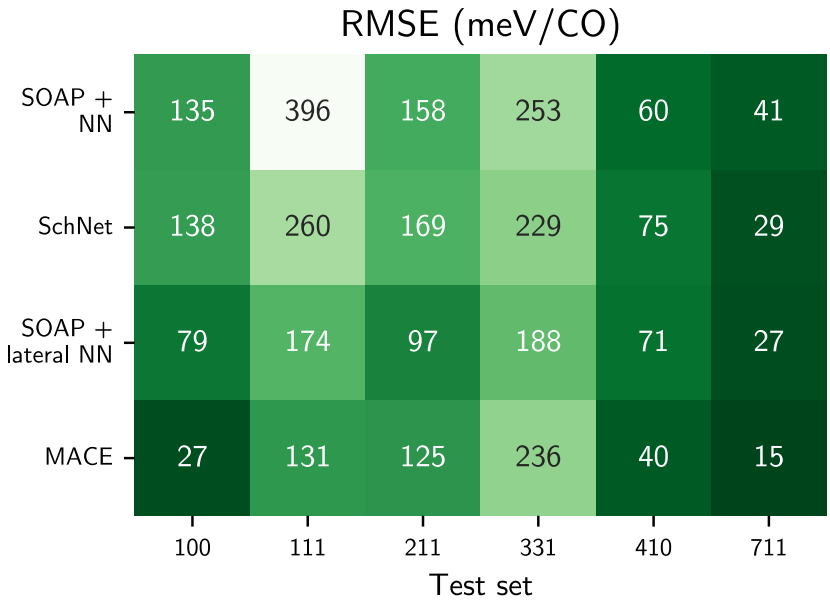

Figure S11: Errors for the models in Fig. 3d trained on Cu(711) configurations. The average RMSE for facets other than Cu(711) corresponds to the "other facets" label in Fig. 3d.

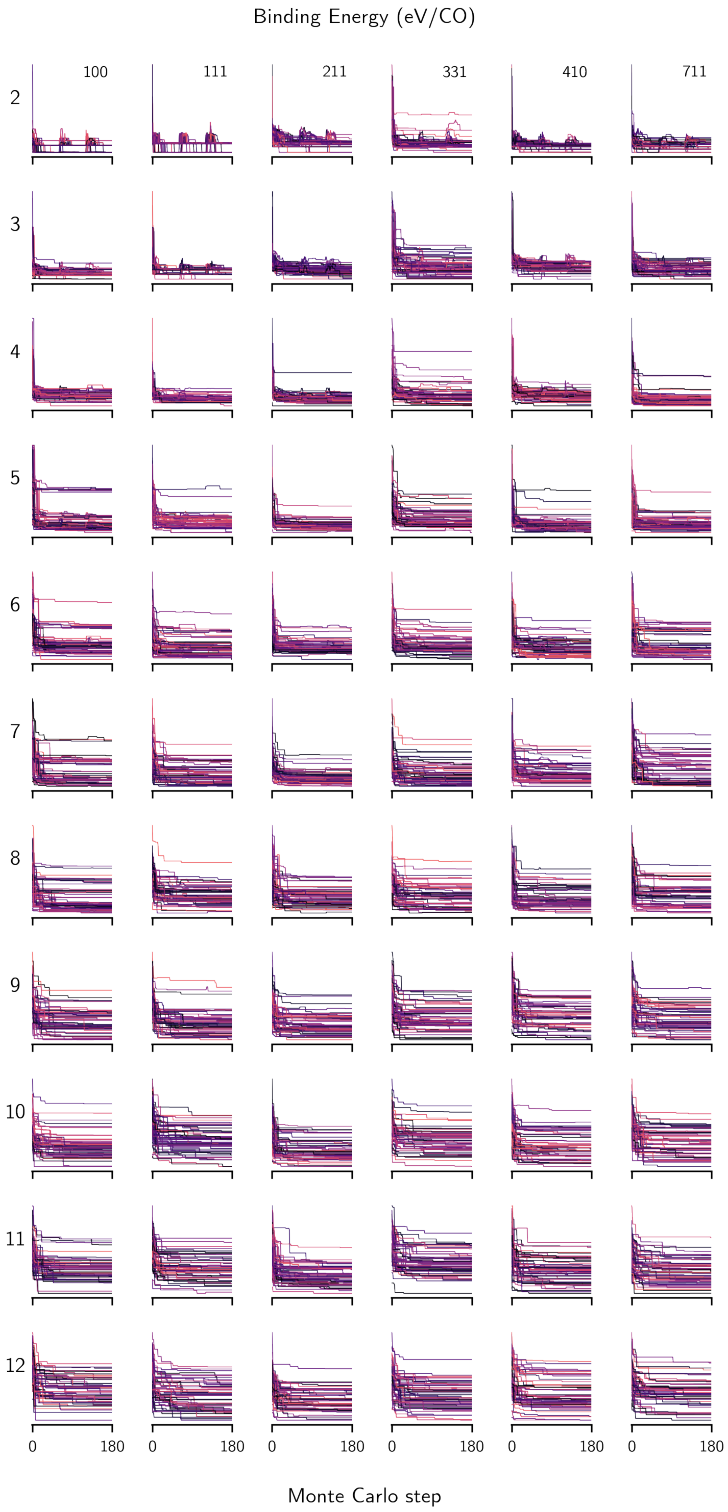

Figure S12: (part 1/2) Binding energy profiles for Cu facets (columns) with different CO coverage (rows) sampled with simulated annealing method with Metropolis algorithm. For clarity, only one every 15 sampling trajectories are shown. Colors are chosen randomly for each trajectory to aid visualization.

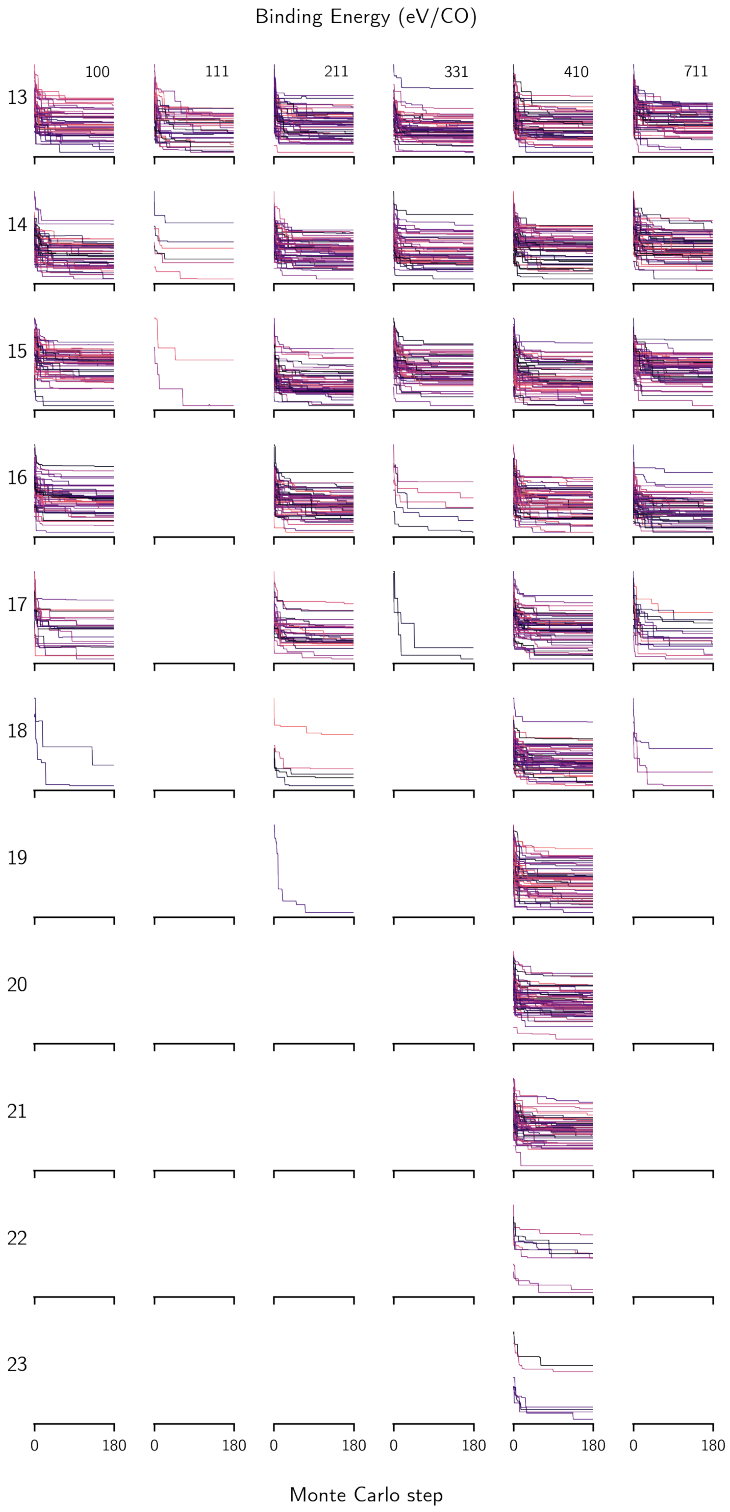

Figure S13: (part 2/2) Binding energy profiles for Cu facets (columns) with different CO coverage (rows) sampled with simulated annealing method with Metropolis algorithm. For clarity, only one every 15 sampling trajectories are shown. Colors are chosen randomly for each trajectory to aid visualization.

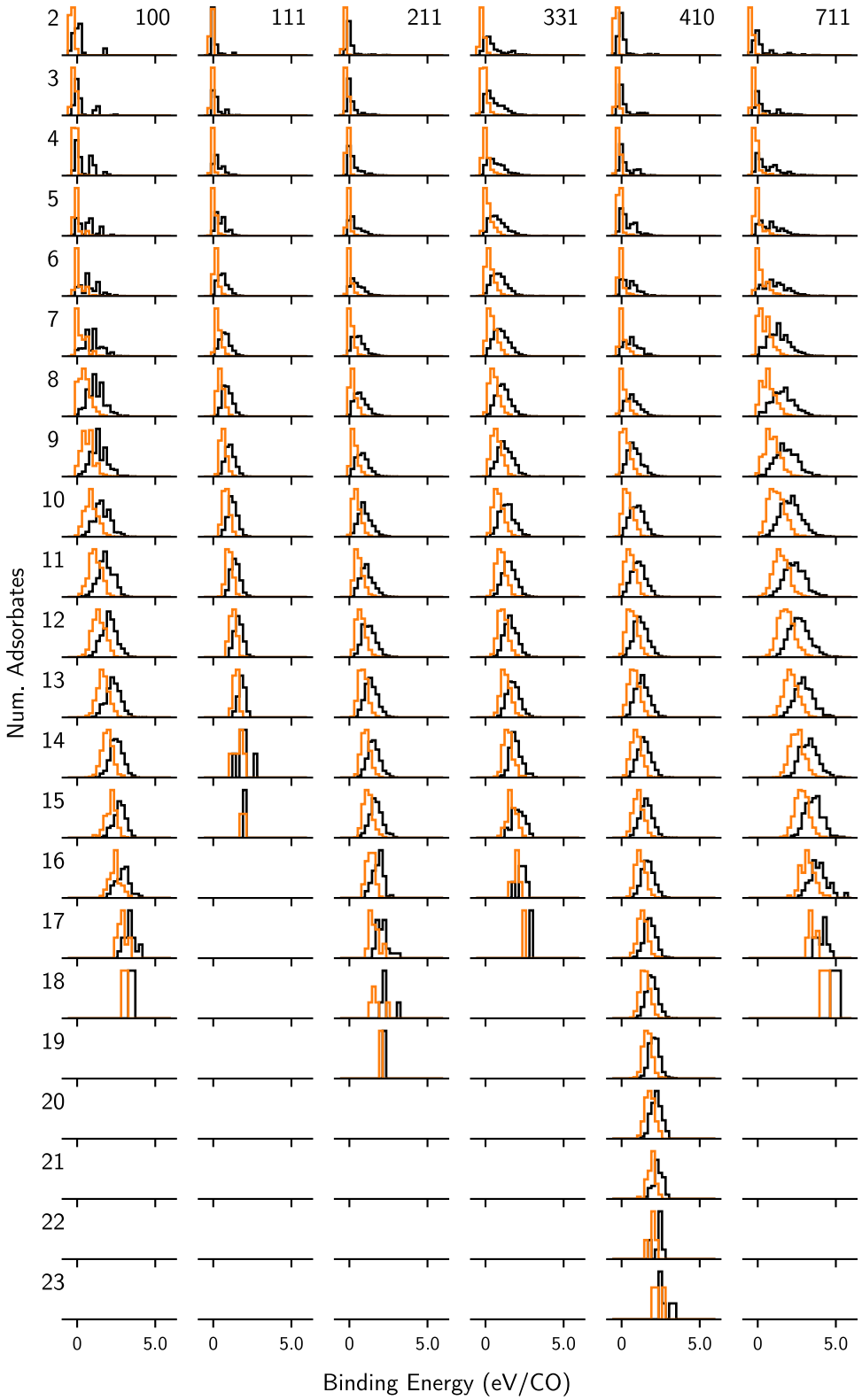

Figure S14: Distributions of energies of randomly sampled structures (Fig. S4, black) and lowest energy replicas from the MCMC sampling (orange).

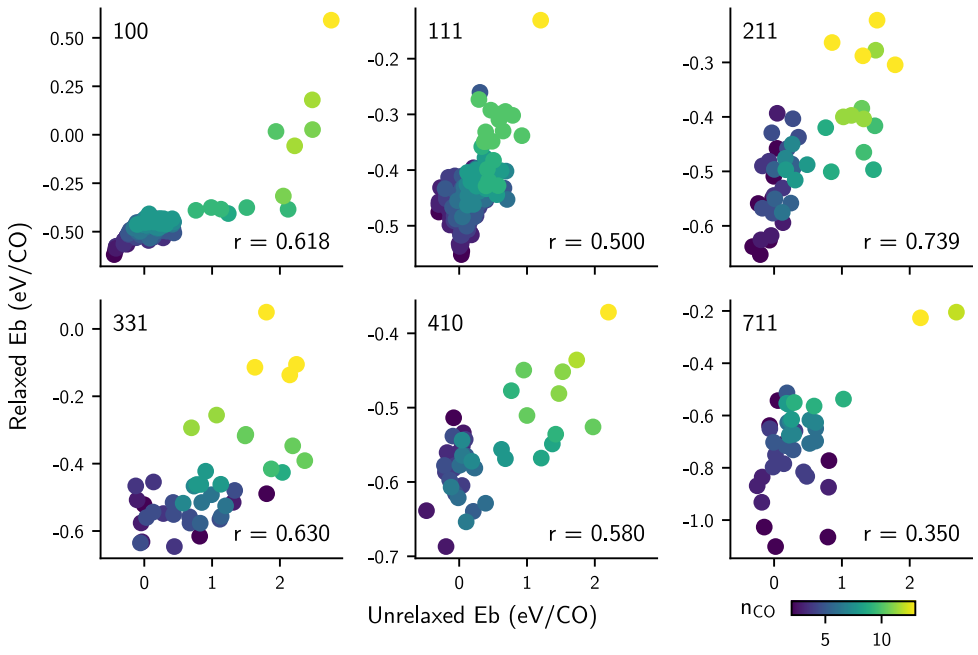

Figure S15: Correlation between binding energies obtained with DFT from relaxed and unrelaxed systems of CO on Cu facets from the original dataset. The value r corresponds to the Spearman's correlation coefficient between (relaxed, unrelaxed) energy pairs.

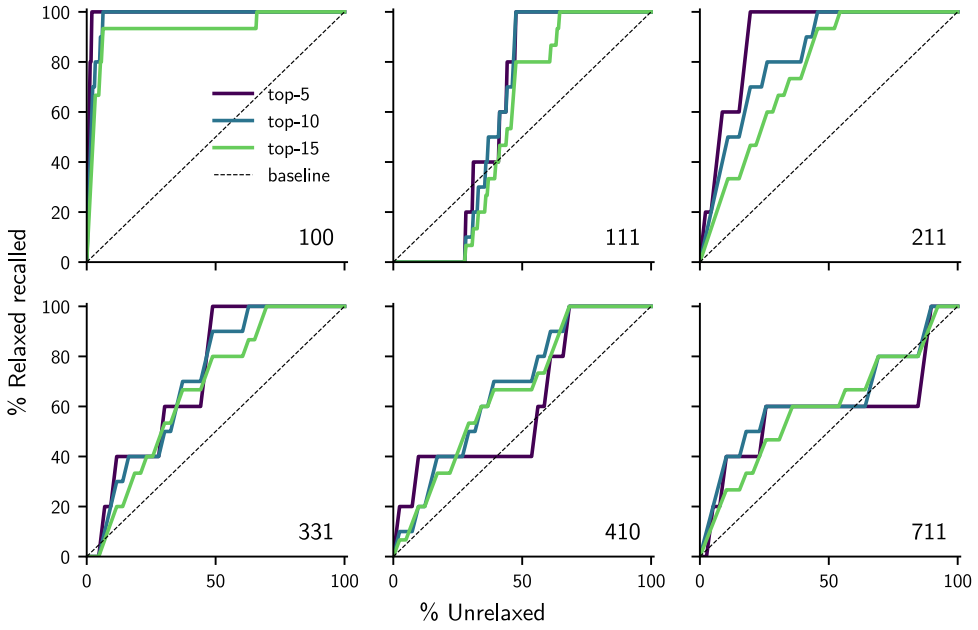

Figure S16: Recall curves of top-N relaxed configurations from unrelaxed binding energies. A higher recall is obtained from systems with higher rank-correlation between unrelaxed and relaxed binding energies (Fig. S19). Three scenarios are shown: top-5, top-10, and top-15 recall. The baseline recall curve is shown in dashed lines.

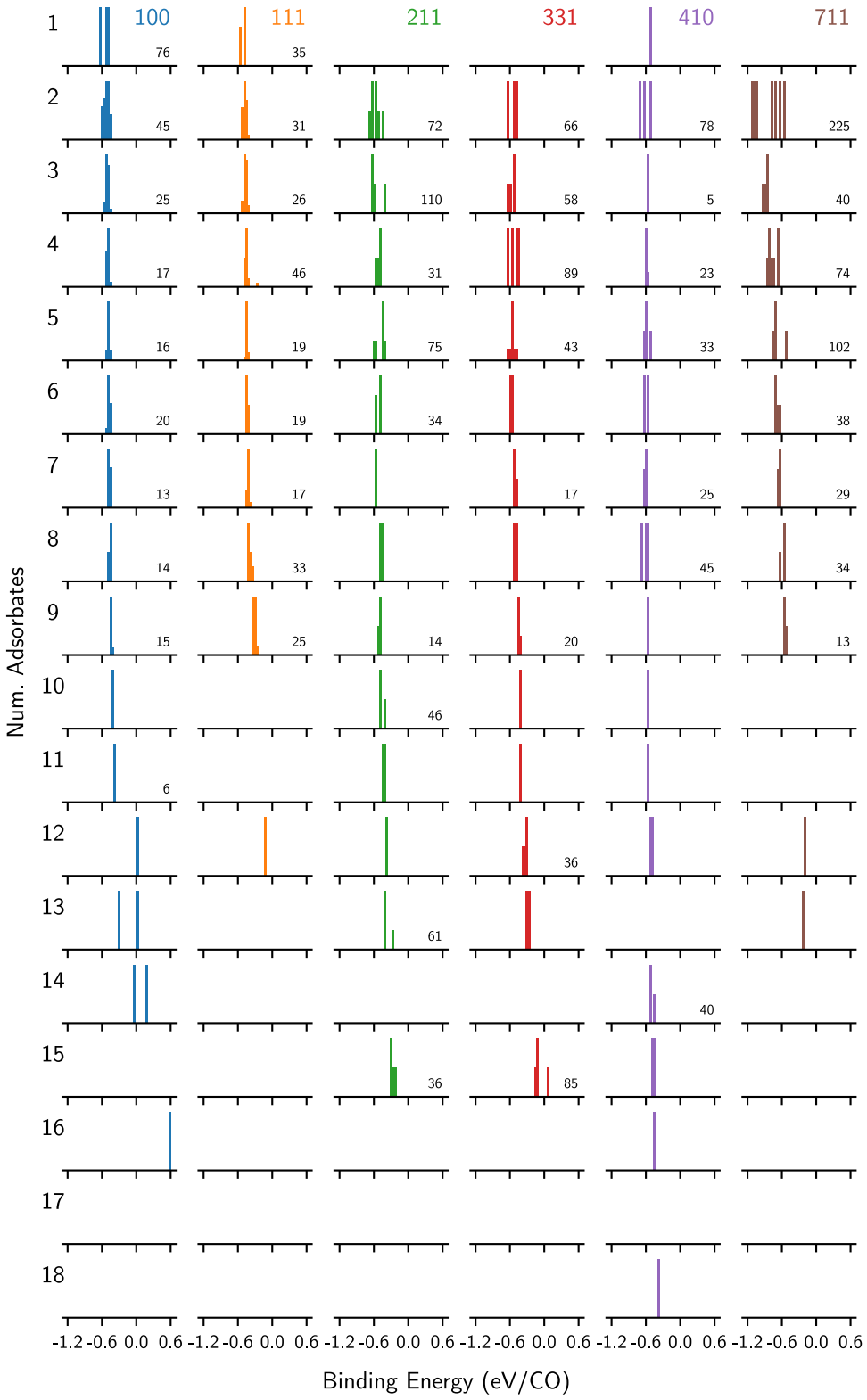

Figure S17: Distributions of binding energies of DFT-relaxed systems from the original dataset separated by facet (column) and coverage (row). Configurations were selected randomly among the unrelaxed configurations in the initial dataset. The number on the bottom right corresponds to the standard deviation (in meV/CO) of each distribution when more than two data points are available.

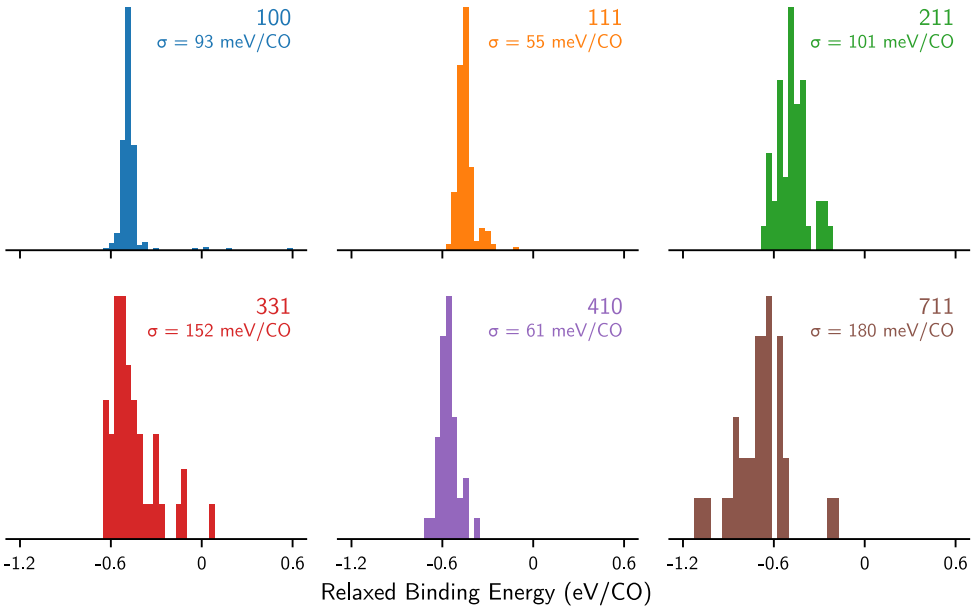

Figure S18: Distributions of binding energies of DFT-relaxed systems from the original dataset separated by facet (see also Fig. S17). σ corresponds to the standard deviation of each distribution.

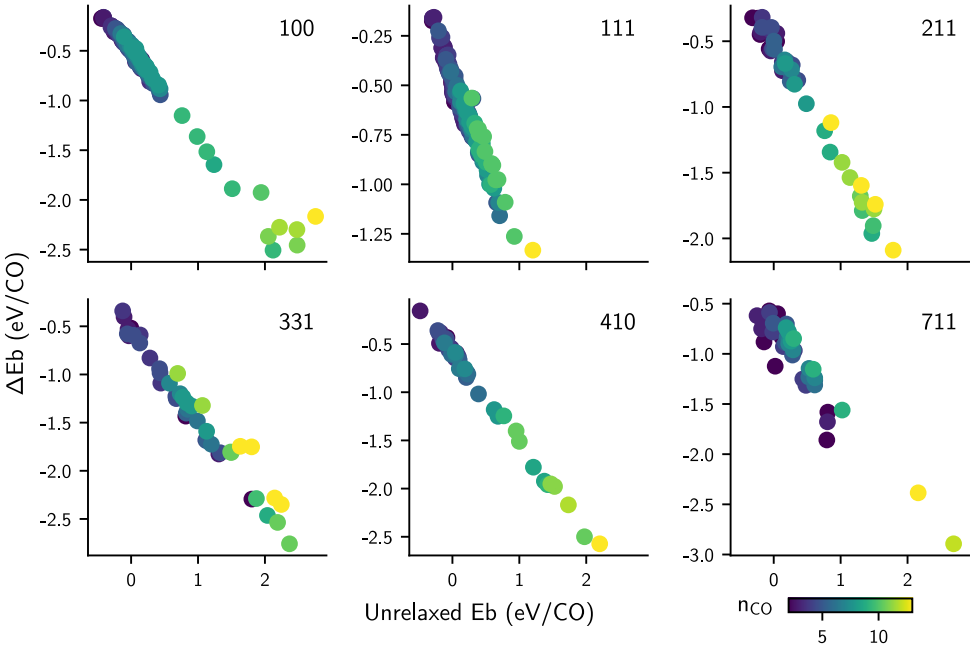

Figure S19: Correlation between the difference between relaxed and unrelaxed binding energies (ΔE_b) and unrelaxed binding energies (E_b) for the original dataset of CO on Cu facets. The near-linear relationship between ΔE_b and the unrelaxed E_b illustrates that the correlations are dominated by the unrelaxed energies, as they exhibit much higher variance than relaxed energies (Fig. S17).

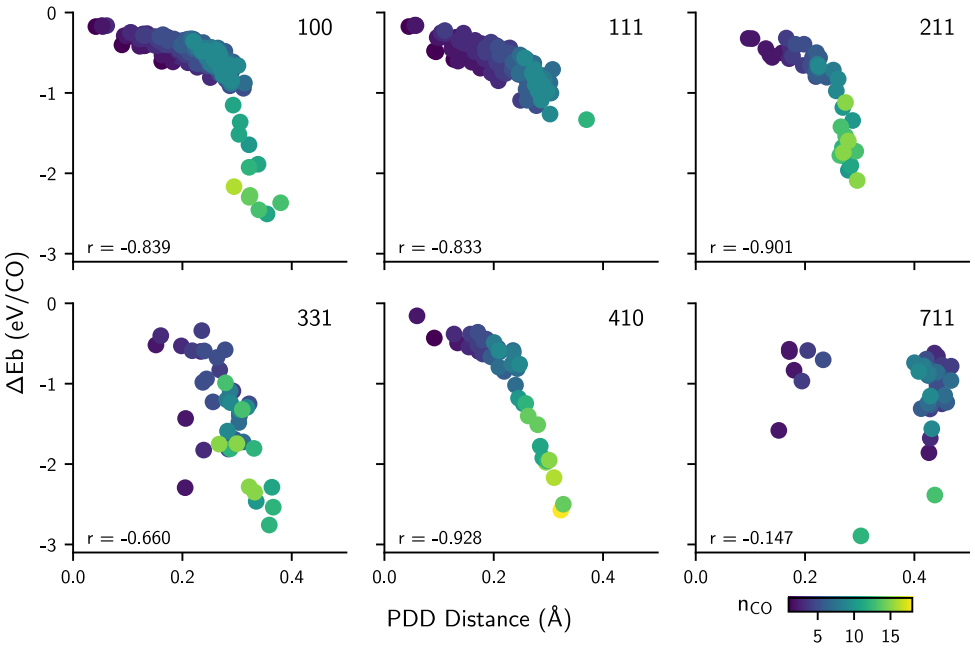

Figure S20: Correlation between relaxation energy differences ($\Delta E_b = E_b^{(\text{relax})} - E_b^{(\text{unrelax})}$) and distances between relaxed and unrelaxed structures, as computed by the pointwise distance distribution (PDD). The structures are from the original dataset of CO on Cu facets. The strong correlations show that a higher ΔE_b (and, thus a higher unrelaxed energy, as shown in Fig. S19) is connected to higher structural changes of the adsorbates and surfaces.

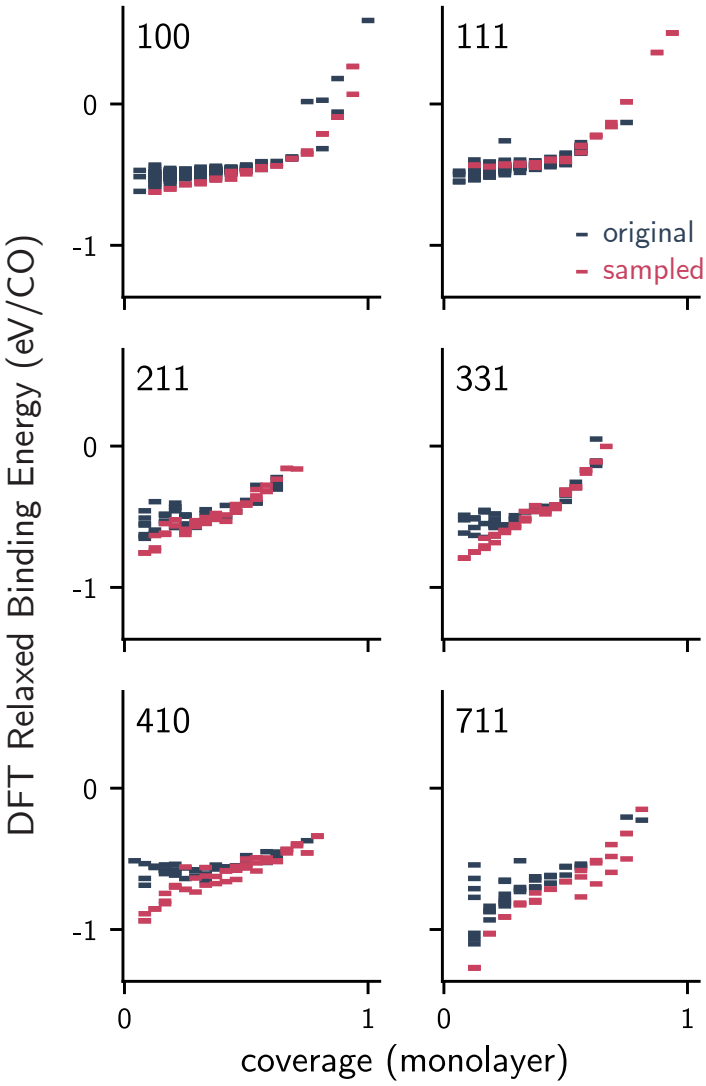

Figure S21: Dependence between relaxed binding energies and surface coverage of CO on Cu facets. Blue lines represent relaxed energies in the original dataset. Red lines represent relaxed energies of systems obtained from ML-accelerated MCMC sampling.

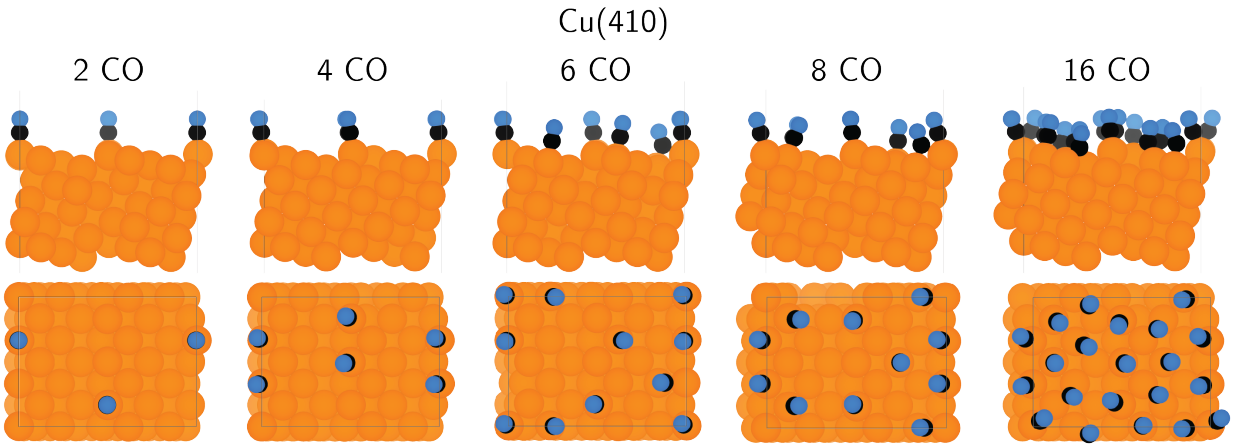

Figure S22: Side and top view of the lowest energy structures among the sampled dataset for *CO on Cu(410) for five different coverages. Carbon, oxygen, and copper atoms are depicted in black, blue, and orange, respectively.

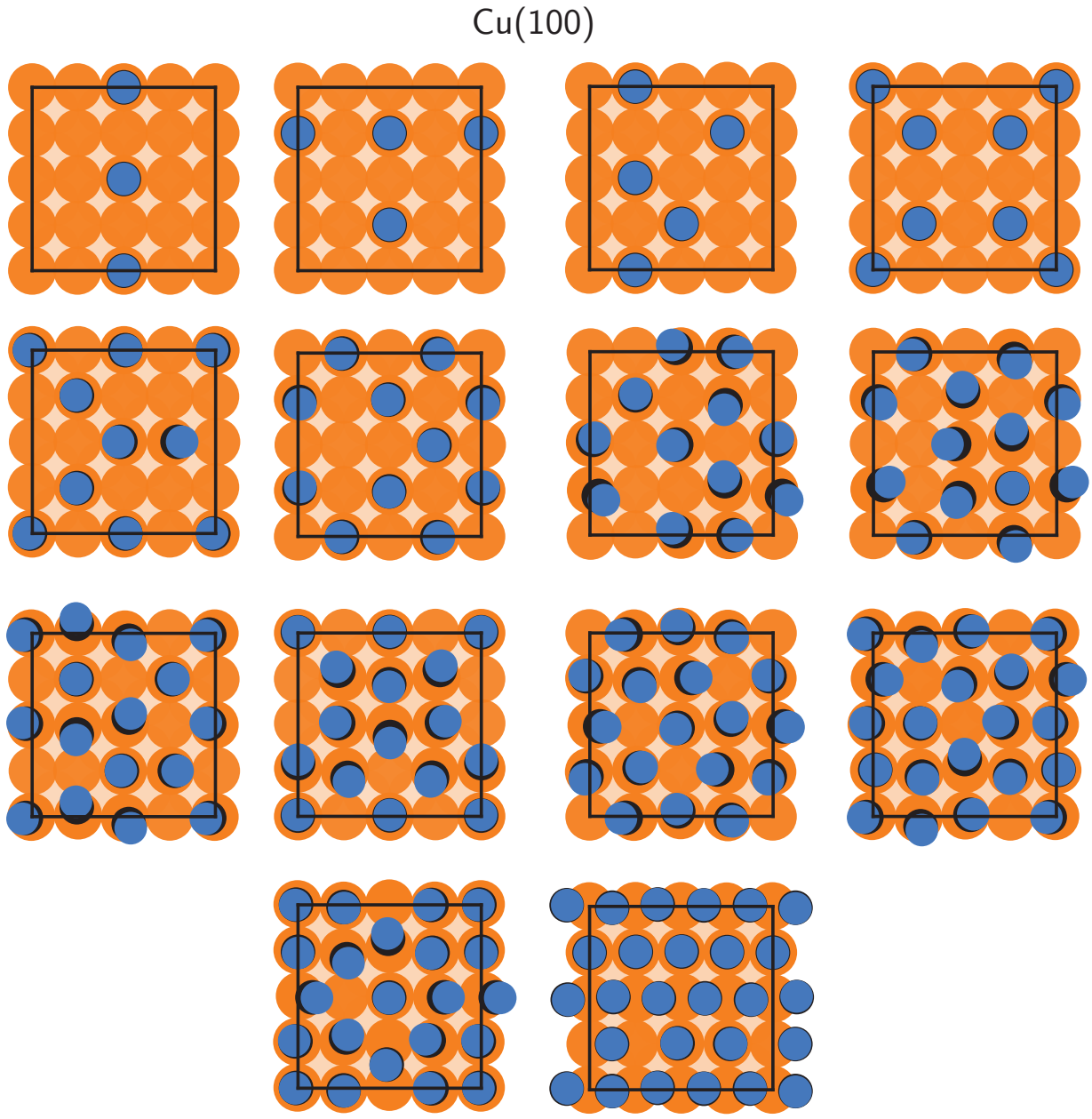

Figure S23: Visualization of lowest-energy structures for Cu(100) among the sampled configurations of *CO coverage. Copper, oxygen, and carbon atoms are depicted with orange, blue, and black circles. Color fading in copper atoms depict distance, with more opaque atoms closer to the surface.

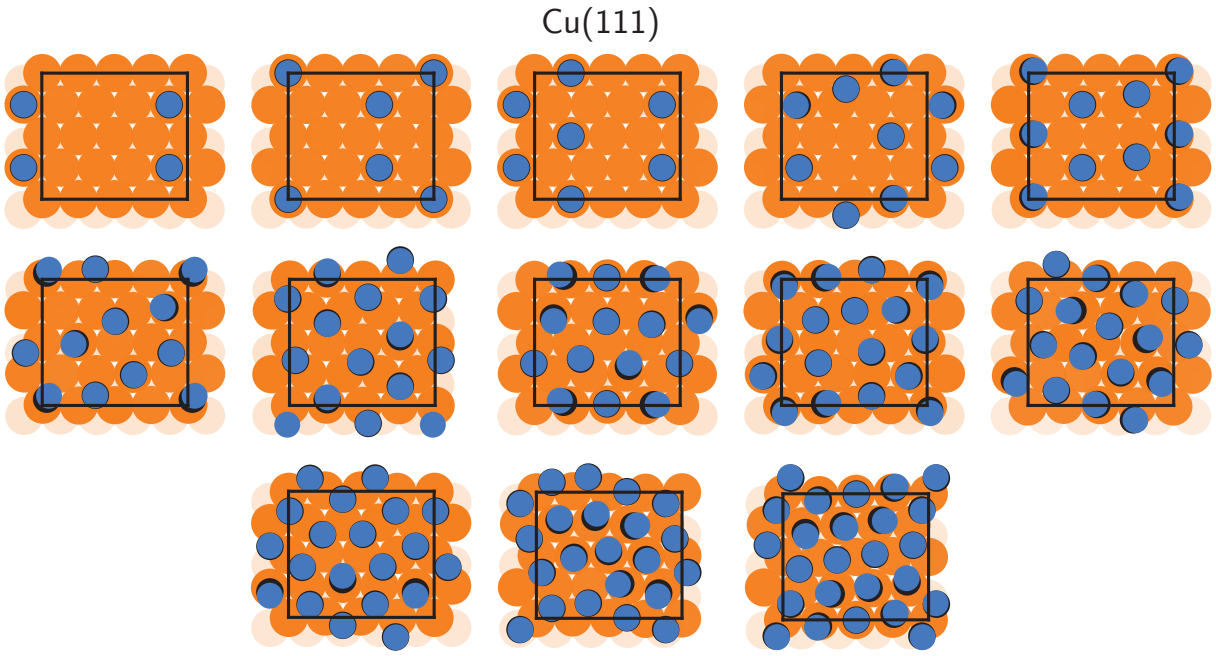

Figure S24: Visualization of lowest-energy structures for Cu(111) among the sampled configurations of *CO coverage. Copper, oxygen, and carbon atoms are depicted with orange, blue, and black circles. Color fading in copper atoms depict distance, with more opaque atoms closer to the surface. As discussed in the main text, sampled low-coverage configurations for Cu(111) do not correspond to the global energy minima for these systems.

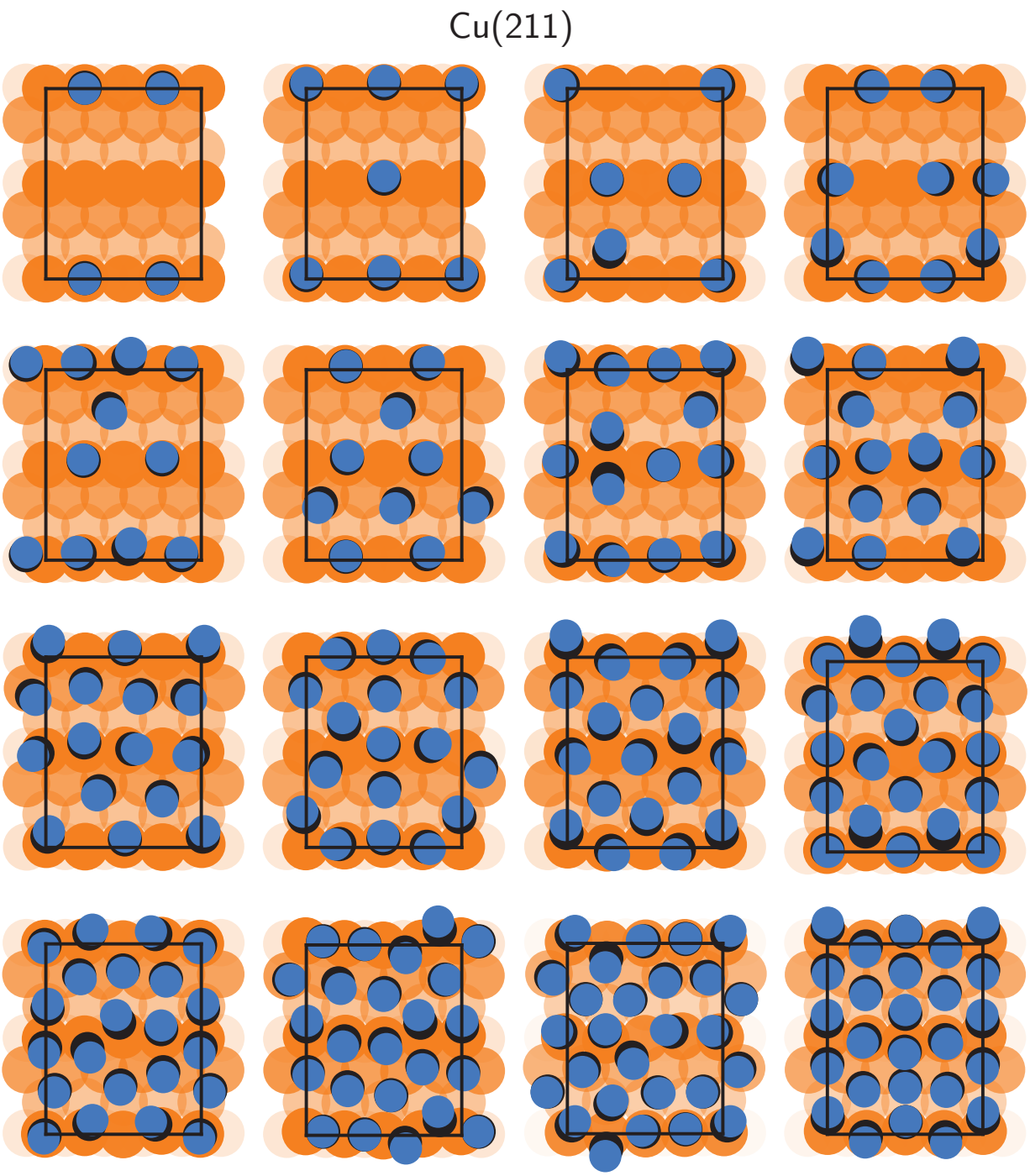

Figure S25: Visualization of lowest-energy structures for Cu(211) among the sampled configurations of *CO coverage. Copper, oxygen, and carbon atoms are depicted with orange, blue, and black circles. Color fading in copper atoms depict distance, with more opaque atoms closer to the surface.

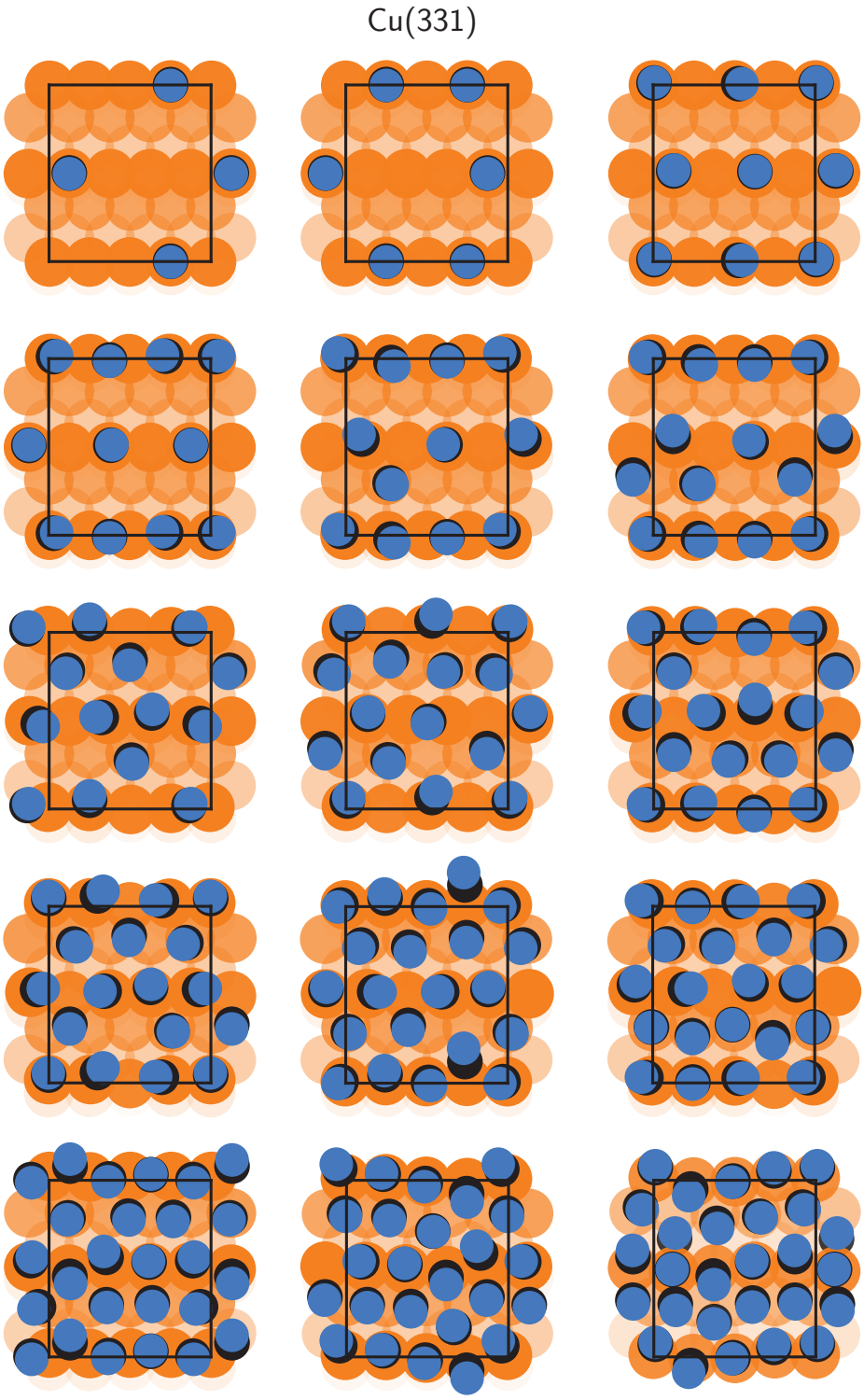

Figure S26: Visualization of lowest-energy structures for Cu(331) among the sampled configurations of *CO coverage. Copper, oxygen, and carbon atoms are depicted with orange, blue, and black circles. Color fading in copper atoms depict distance, with more opaque atoms closer to the surface.

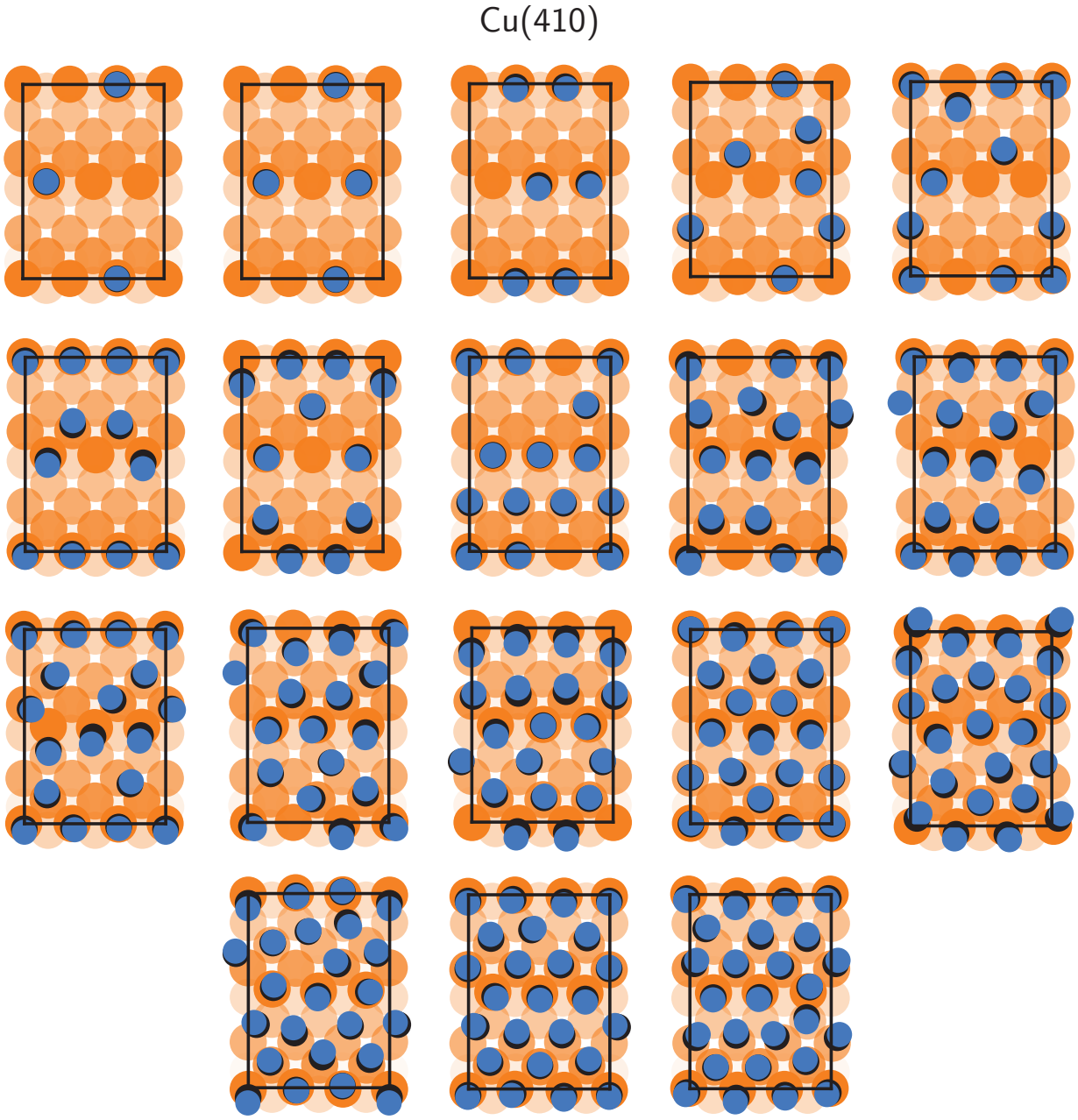

Figure S27: Visualization of lowest-energy structures for Cu(410) among the sampled configurations of *CO coverage. Copper, oxygen, and carbon atoms are depicted with orange, blue, and black circles. Color fading in copper atoms depict distance, with more opaque atoms closer to the surface.

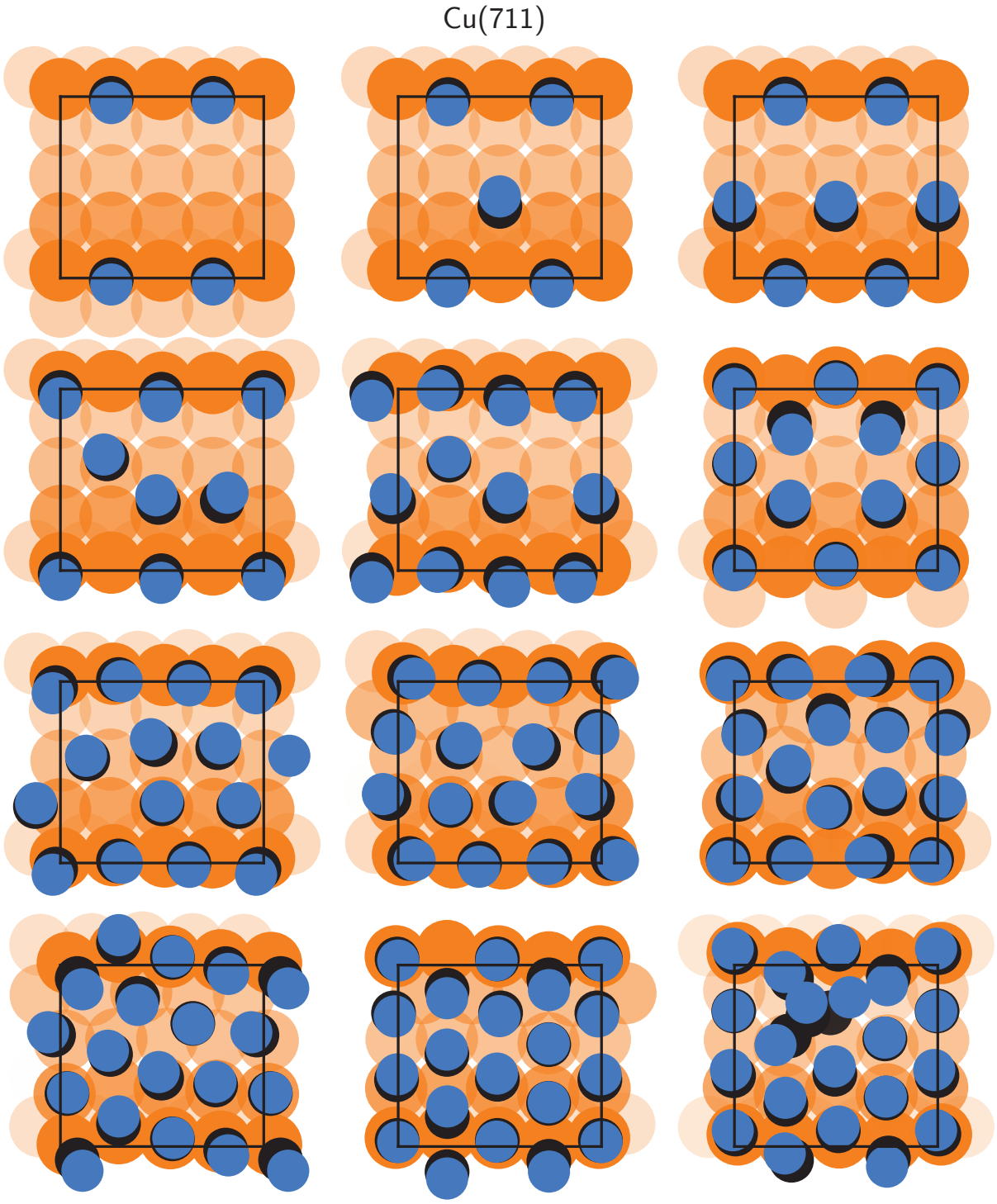

Figure S28: Visualization of lowest-energy structures for Cu(711) among the sampled configurations of *CO coverage. Copper, oxygen, and carbon atoms are depicted with orange, blue, and black circles. Color fading in copper atoms depict distance, with more opaque atoms closer to the surface.

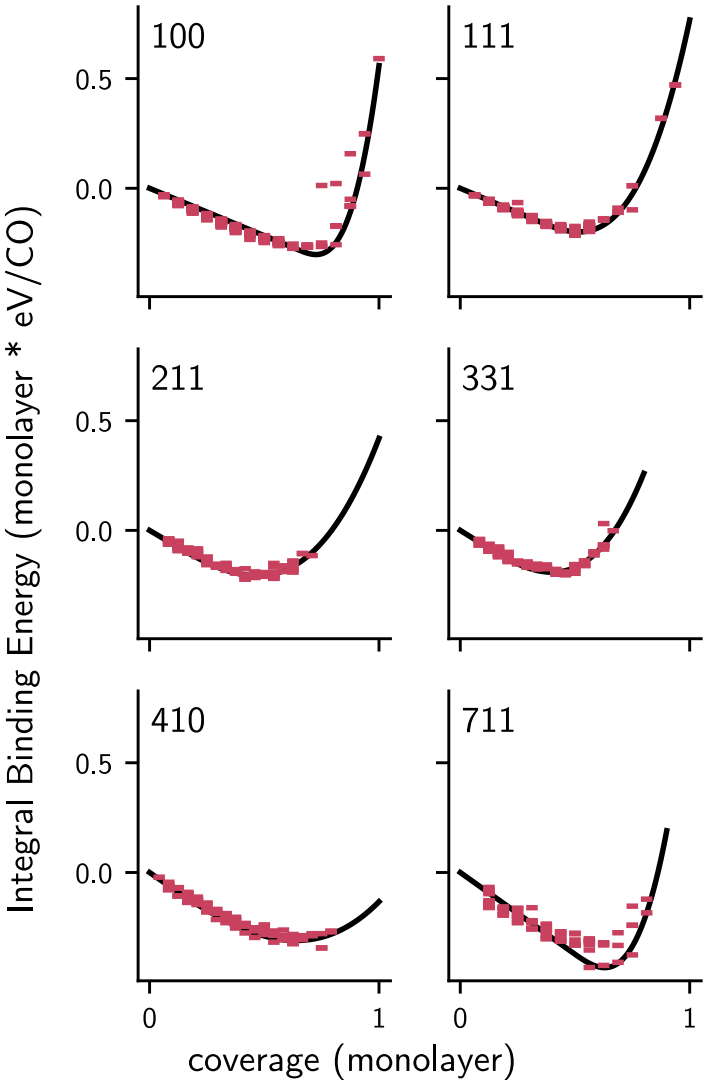

Figure S29: Relaxed binding energies (red) and fitted integral binding energy curve of CO on Cu facets. The curves were fitted with a polynomial that is linear at low coverages and cubic at high coverages. The fits were performed by first taking the Boltzmann average at 298 K of binding energies for each facet and coverage regime, then fitting the function parameters using a non-linear least squares.

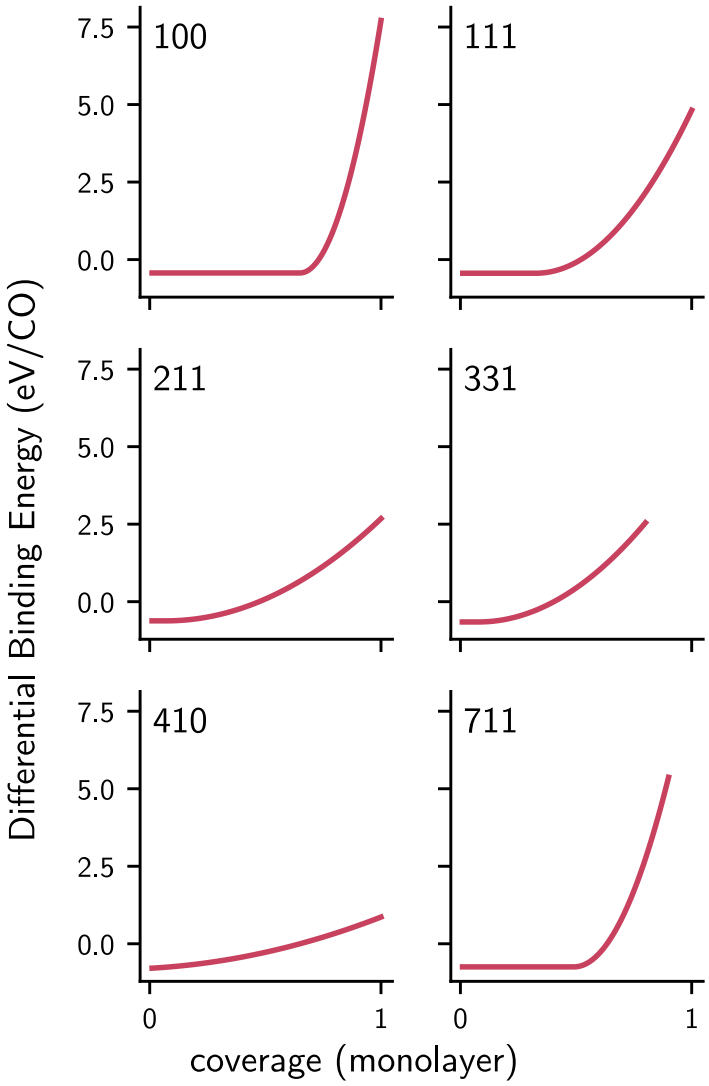

Figure S30: Differential binding energy curve of CO on Cu facets. The curves were obtained by taking the derivative of fitted integral binding energy curves with respect to the coverage, then combined together to form Fig. 6.

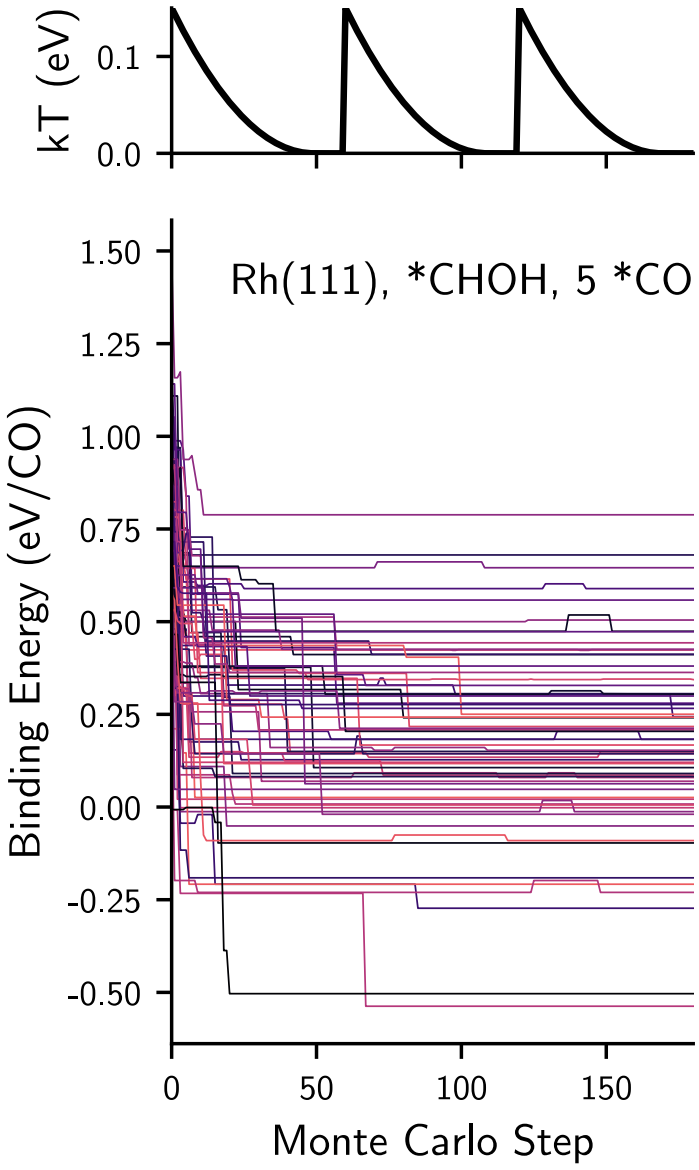

Figure S31: Example of temperature (top) and binding energy (bottom) profiles during MCMC sampling of CO configurations on the *CHOH-Rh(111) system. Only one every 15 trajectories is shown for clarity. Colors are chosen randomly to enable visualization of different sampling trajectories.

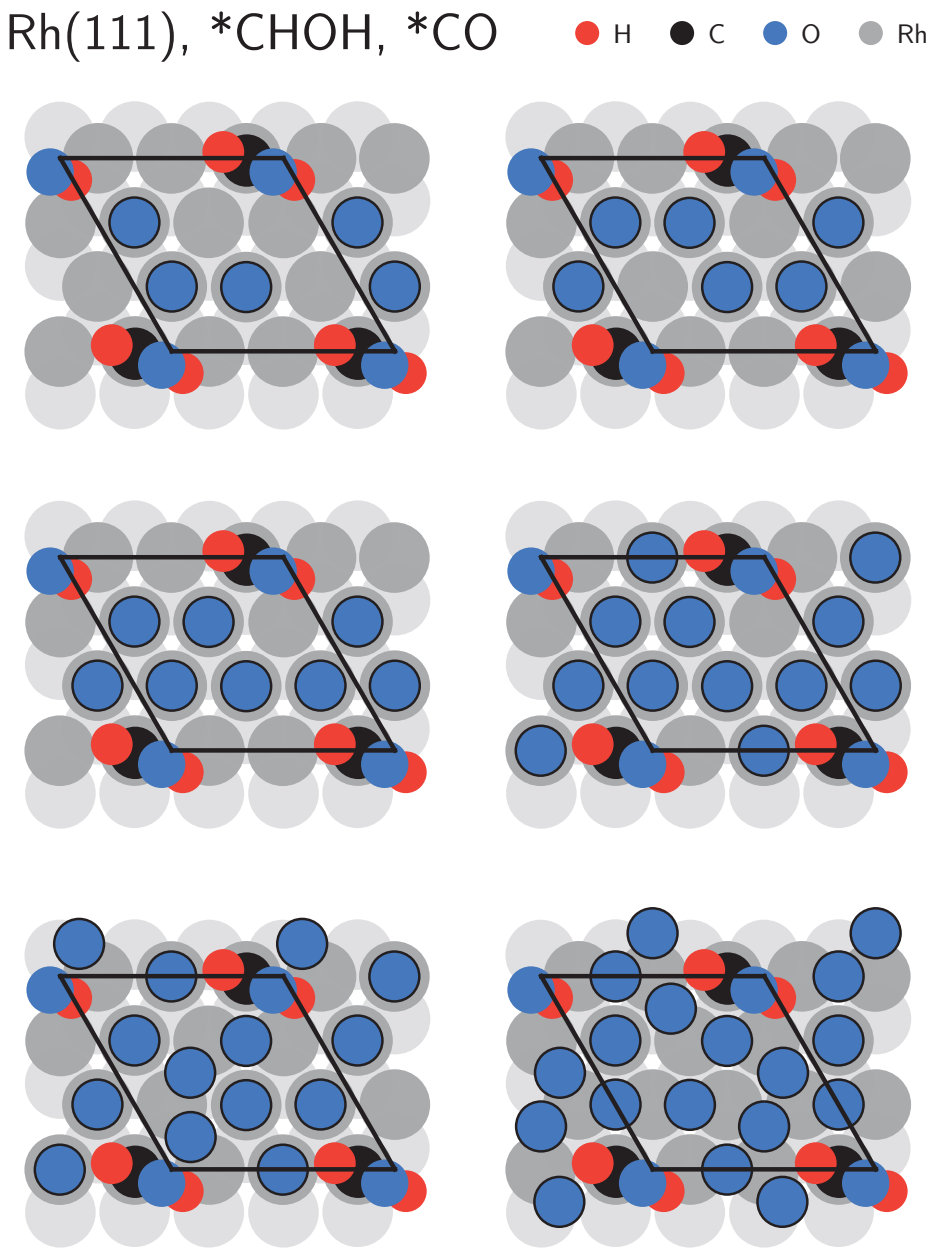

Figure S32: Lowest unrelaxed energy configurations of *CHOH (top site) and *CO on Rh(111) sampled with the ML-accelerated MCMC approach for increasing CO coverages. Rhodium, carbon, oxygen, and hydrogen atoms are depicted with gray, black, blue, and red colors, respectively.

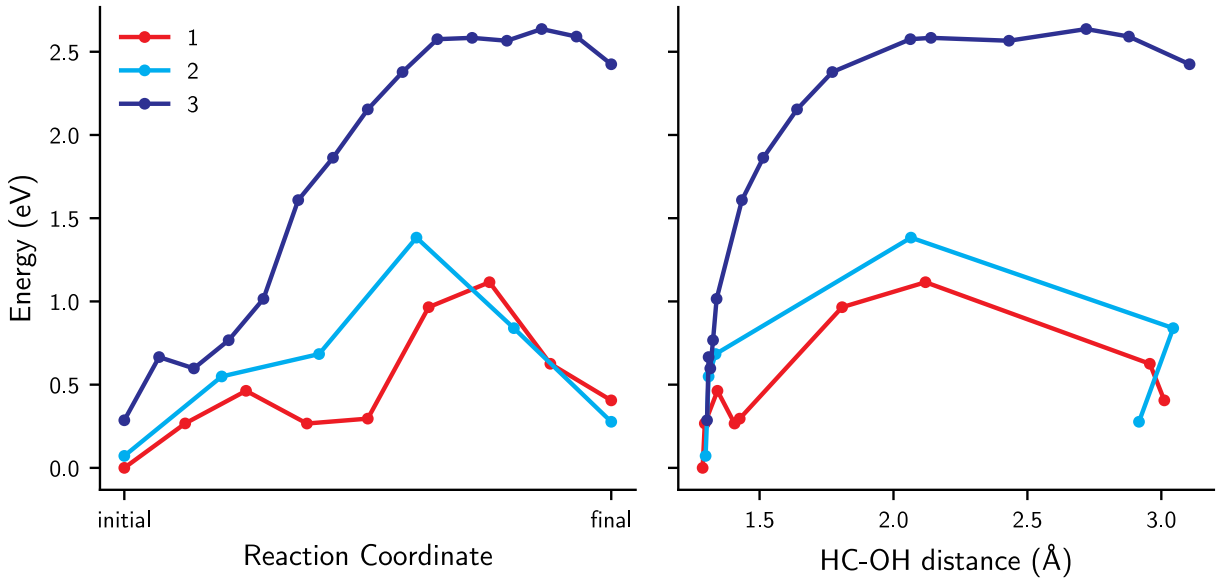

Figure S33: NEB trajectories for all three systems shown in Fig. 7c of the main text. The trajectories are shown according to equally spaced images (left) and to the distance between C-O atoms in the CHOH adsorbate (right). In all three cases, the transition state of CHOH bond scission is found for a C-O bond length around 2.1 Å. Energy differences for restructuring are further demonstrated with the unchanging distance between C-O atoms in the CHOH adsorbate (right panel).

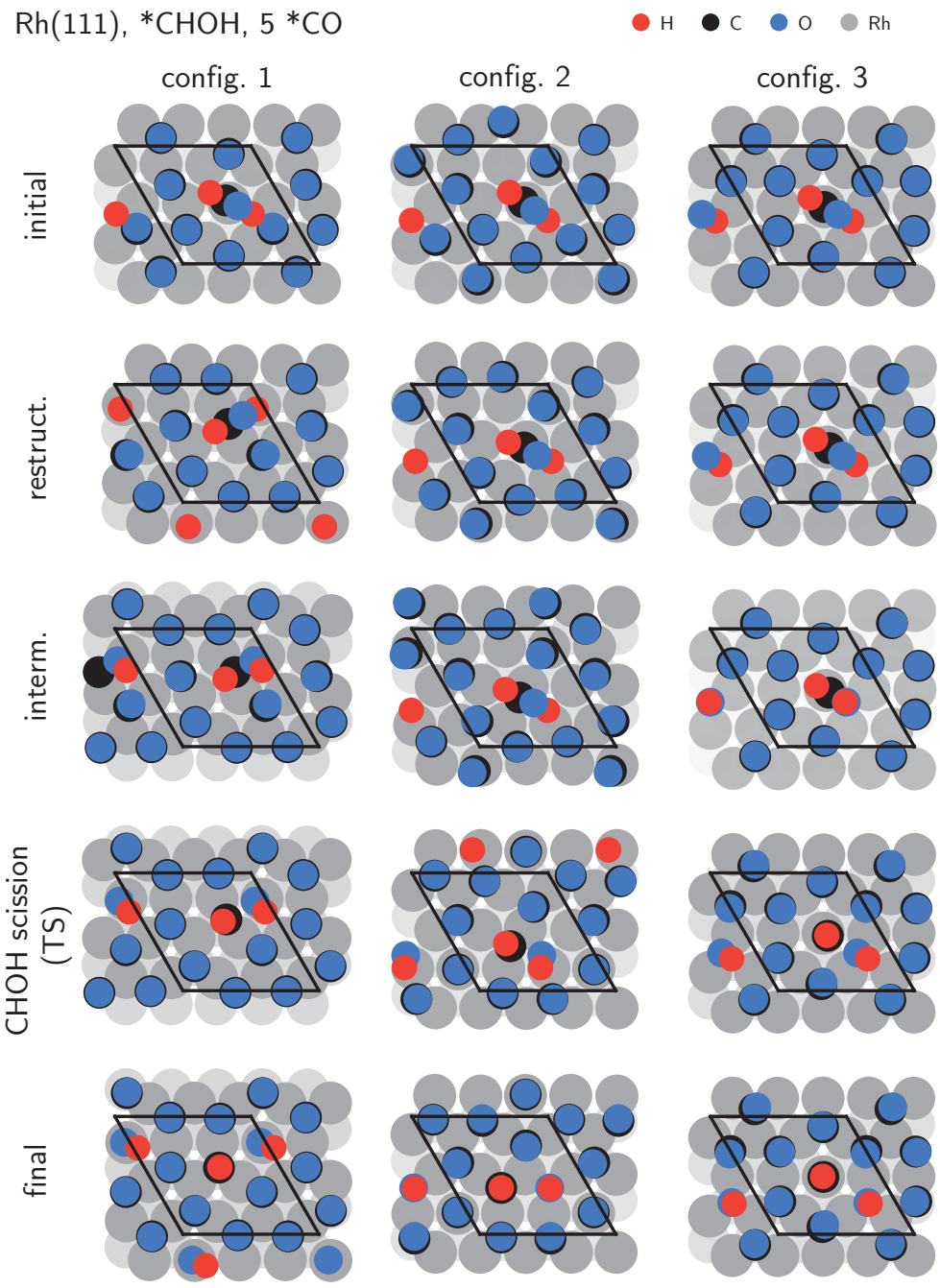

Figure S34: Schematic of NEB trajectories *CHOH bond scission for the three configurations shown in Fig. 7 of the main text. Rhodium, carbon, oxygen, and hydrogen atoms are depicted with gray, black, blue, and red colors, respectively.

Supplementary Tables

Table S1: Statistics on the Cu surface models used in this work. The lateral lattice vectors **a** and **b** are orthogonal to each other (Fig. S1), with magnitudes given in columns a, b. The surface depth is the cutoff value used to label atoms as "surface atoms". With this cutoff, different number of surface atoms and adsorption sites are found for each facet. Because high-coverage regimes are of interest, all adsorption sites are reported instead of only considering unique ones.

Facet	a (Å)	b (Å)	Surface Depth (Å)	Num. Surface Atoms	Num. Adsorption Sites
100	10.39	10.39	1.0	16	64
111	10.39	9.00	1.0	16	96
211	10.39	12.73	1.4	24	128
331	10.39	11.32	1.6	24	144
410	11.02	15.15	1.2	24	120
711	10.39	9.28	1.8	16	64

Table S2: Simplified estimates of the number of configurations of n adsorbates on each of the facets. Each number was obtained by choosing n groups of 4 adsorption sites among all adsorption sites shown in Table S1. The groups simplify the estimate by assuming that each adsorption site has roughly 3 nearest neighbors that cannot be simultaneously selected due to small distances, leading to a p choose n problem, where p is the number of adsorption sites in Table S1 divided by 4. These estimates do not account for symmetrically equivalent distributions of adsorbates on the facet given the periodic boundary conditions and surface geometry.

n	100	111	211	331	410	711
1	16	24	32	36	30	16
2	120	276	496	630	435	120
3	560	2024	4960	7140	4060	560
4	1820	1.06e+04	3.60e+04	5.89e+04	2.74e+04	1820
5	4368	4.25e+04	2.01e+05	3.77e+05	1.43e+05	4368
6	8008	1.35e+05	9.06e+05	1.95e+06	5.94e+05	8008
7	1.14e+04	3.46e+05	3.37e+06	8.35e+06	2.04e+06	1.14e+04
8	1.29e+04	7.35e+05	1.05e+07	3.03e+07	5.85e+06	1.29e+04
9	1.14e+04	1.31e+06	2.80e+07	9.41e+07	1.43e+07	1.14e+04
10	8008	1.96e+06	6.45e+07	2.54e+08	3.00e+07	8008
11	4368	2.50e+06	1.29e+08	6.01e+08	5.46e+07	4368
12	1820	2.70e+06	2.26e+08	1.25e+09	8.65e+07	1820
13	560	2.50e+06	3.47e+08	2.31e+09	1.20e+08	560
14	120	1.96e+06	4.71e+08	3.80e+09	1.45e+08	120
15	16	1.31e+06	5.66e+08	5.57e+09	1.55e+08	16
16	1	7.35e+05	6.01e+08	7.31e+09	1.45e+08	1
17	0	3.46e+05	5.66e+08	8.60e+09	1.20e+08	0
18	0	1.35e+05	4.71e+08	9.08e+09	8.65e+07	0
19	0	4.25e+04	3.47e+08	8.60e+09	5.46e+07	0
20	0	1.06e+04	2.26e+08	7.31e+09	3.00e+07	0
21	0	2024	1.29e+08	5.57e+09	1.43e+07	0
22	0	276	6.45e+07	3.80e+09	5.85e+06	0
23	0	24	2.80e+07	2.31e+09	2.04e+06	0
24	0	1	1.05e+07	1.25e+09	5.94e+05	0

Table S3: Number of data points sampled on a per-facet, per-coverage basis for the original dataset of *CO on copper. Binding energies for all these randomly sampled configurations were computed using single-point DFT calculations. The final number of configurations depends on the maximum number of structures allowed to be sampled and the symmetry equivalence between crystal structures.

	Facet					
n_{CO}	100	111	211	331	410	711
1	3	5	28	15	29	16
2	28	66	95	95	95	87
3	82	99	100	100	100	100
4	95	100	100	100	100	100
5	100	100	100	100	100	100
6	100	100	100	100	100	100
7	100	100	100	100	100	100
8	100	100	100	100	100	100
9	100	100	100	100	100	100
10	50	50	50	50	50	50
11	50	50	50	50	50	50
12	50	50	50	50	50	50
13	50	21	50	50	50	50
14	50	6	50	50	50	50
15	50	1	50	50	50	50
16	50	1	50	12	50	50
17	11	0	26	4	50	12
18	2	0	6	1	50	2

Table S4: Energies of CH-OH bond scission pathways for configurations 1, 2, and 3 shown in Fig. 7 of the main text. All energies use the energy of initial configuration 1 as reference. Energies are expressed in eV.

Config.	Initial	Restructuring	Intermediate	CH-OH scission	Final
1	0.00	0.46	0.27	1.12	0.41
2	0.07	0.55	0.68	1.38	0.28
3	0.29	0.67	0.60	2.58	2.42