

Supplementary Materials

Machine-learning prediction of facet-dependent CO coverage on Cu electrocatalysts

Shanglin Wu¹, Shisheng Zheng^{2,*}, Wentao Zhang¹, Mingzheng Zhang¹, Shunning Li^{1,*}, Feng Pan^{1,*}

¹School of Advanced Materials, Peking University, Shenzhen Graduate School, Shenzhen 518055, Guangdong, China.

²College of Energy, Xiamen University, Xiamen 361000, Fujian, China.

***Correspondence to:** Prof. Shunning Li, Prof. Feng Pan, School of Advanced Materials, Peking University, Shenzhen Graduate School, No. 2199 Lishui Road, Shenzhen 518055, Guangdong, China. E-mail: lisn@pku.edu.cn, panfeng@pkusz.edu.cn; Prof. Shisheng Zheng, College of Energy, Xiamen University, No. 4221, Xiang'an South Road, Xiamen 361000, Fujian, China. E-mail: zhengss@pku.edu.cn

Additional Method Details

Model Construction and Theoretical Calculation Methods

All calculations were performed using the DFT framework in the VASP software, utilizing PAW pseudopotentials. The exchange-correlation term was corrected using the RP functional (revised Perdew–Burke–Ernzerhof functionals). The surface slabs in all calculations had a minimum thickness of 8 Å in the z-direction and a minimum size of 7 Å in the x/y directions; the vacuum layer was 15 Å thick; the number of Cu atoms did not exceed 100; Cu atoms located more than 4 Å from the surface in the z-direction were fixed. During the optimization of the Cu bulk phase structure, the cut-off energy was 840 eV; for all other calculations, it was 400 eV. The convergence criteria for forces and energy were set to -0.03 eV/Å and 10^{-5} eV, respectively. The vaspkit tool package was used to generate the KPOINTS file necessary for structural optimization, employing the Gamma point with a K-Mesh sampling precision of 0.05; the Fermi level was broadened by 0.1 eV. Considering the significant impact of CO molecule interactions on adsorption energy calculations, all computational systems used the DFT-D3 dispersion correction. In this work, the coverage definition is the number of CO molecules per unit projected area on the xoy plane of the Slab, facilitating comparisons between different crystal surface adsorption configurations at high coverages; for comparisons within the same crystal surface, the phrase "number of CO molecules adsorbed on the Slab surface" is also used.

The average adsorption energy of CO on the Cu surface is defined as:

$$\Delta E_{\text{ads}} = \frac{E_{\text{ads-Slab}} - E_{\text{clean-Slab}} - n_{\text{CO}} E_{\text{CO(g)}}}{n_{\text{CO}}}$$

where $E_{\text{ads-Slab}}$, $E_{\text{clean-Slab}}$ and E_{CO} are the energies after structural optimization, and n_{CO} is the number of CO molecules adsorbed on the Slab.

Method to symbolize different sites and their combinations

We employ graph theory tools to identify and classify distinct surface sites in the adsorption structure. The procedure is as follows:

1. Graph Representation of the Adsorption Structure: The entire adsorption structure is abstracted as a graph, where atoms are represented as vertices and bonds are

represented as edges. Specifically, the C-O covalent bond, C-Cu bonding, and Cu-Cu metallic bonding are abstracted as edges with graph distances of 0, 1, and 2, respectively. This distance definition captures the direct bonding relationships between the atoms.

2. Extraction of Local Environment Subgraphs: For each probe C atom, a local environment subgraph is extracted with a graph distance of 3. This distance allows us to capture the second-order coordination of the surface atoms surrounding the C atom, which helps distinguish different site types. Including the local environment is crucial because atoms at the same coordination number (e.g., 3-coordinated sites on a Cu(111) surface) can have distinct local environments, such as the differences between the hcp-hollow and fcc-hollow sites.

3. Isomorphism Comparison of Subgraphs: We use the `is_isomorphic()` function from the NetworkX package (based on the VF2 algorithm) to compare the isomorphism of the local subgraphs. This allows us to determine the distinct independent subgraphs and identify independent site types on the crystal surface. Using this method, we have identified 68 different independent site types on the 8 Cu surface configurations studied.

4. Denotation of site types: We apply a combined letter naming scheme to distinguish between different site types on the surface. For example, "Bb" represents a site type where "B" indicates the coordination number ($n=2$), and "b" denotes that this is the second distinct graph structure of sites with the same coordination number. Similarly, other site types are named by following this scheme, where the first letter corresponds to the coordination number and the second letter differentiates the graph structure among sites with the same coordination number. To clarify with a specific example, consider the site sequence "Aa2Ab1Da1Dc2." Here, "A" represents a site with coordination number 1; "a" indicates that this is the first distinct graph structure of coordination number 1; "B" represents a site with coordination number 2; "b" indicates that this is the second distinct graph structure of coordination number 2, and so on.

Principle of Configuration Deduplication

Crystal surface structures, including the distribution patterns of adsorbed atoms or molecules, form a two-dimensional periodic pattern described by two-dimensional space groups. Additionally, projections of three-dimensional crystals in certain directions, as well as periodic patterns on textiles or wallpaper, are considered two-dimensional periodic patterns. There are 17 two-dimensional space groups, also known as crystallographic plane groups, distributed among five classes of plane lattices: oblique lattice points (mp): p1, p2; simple rectangular lattice points (op): pm, pg, cm; centered rectangular lattice points (op): p2mm, p2mg, p2gg, c2mm; square lattice points (tp): p4, p4mm, p4gm; hexagonal lattice points (hp): p3, p3m1, p31m, p6, p6mm. Three types of symmetry operations can be performed on two-dimensional periodic images: translation operations that describe the image's periodicity, point operations including pure rotations and pure mirror images, and compound operations, which are combinations of reflection relative to a line and translation half the period along that line. The plane groups corresponding to the 8 Cu surfaces studied in this work are shown in Table S3 and Table S4.

In the limited clean Cu-Slab, generating adsorption structures by filling with CO ignores the actual surface periodicity, leading to many equivalent adsorption structures, termed filling equivalent structures. Equivalent adsorption structures can be identified through corresponding two-dimensional symmetry operations. In essence, for a given adsorption configuration, applying symmetry operations unique to the substrate Slab can cut out different adsorbate atom distribution patterns within the diamond-shaped area enclosed by the Slab lattice vectors while maintaining the post-transformation Slab coincident with the original Slab. Symmetry operations do not alter the adsorption configuration itself; hence, the configurations before and after transformation are equivalent. The essence of our deduplication is to determine how many unique symmetry operations can extract different adsorbate atom distribution patterns within the diamond-shaped area for a given adsorbate CO atom lattice on a specific surface morphology and size of the substrate Slab, based solely on the clean Slab's expansion multiples in the x and y directions and its independent point operation types.

For adsorption configurations generated by Slabs with lattice vectors \vec{l}_1 , \vec{l}_2 , and expansion multiples A, B, if the point group of the plane has a set of point operations S (including the identity operation E) containing the adsorbate lattice P (with the same lattice vectors as the Slab), then any operation Q on the adsorbate atoms P can be represented as:

$$QP = S_k P + \frac{1}{AB} (\vec{l}_1 \quad \vec{l}_2) \begin{pmatrix} Ba \\ Ab \end{pmatrix},$$

$$S_k \in S; a \in [0, A - 1], b \in [0, B - 1], a, b \in Z$$

Since the set of point operations S is finite and the values of a, b are limited, the number of different adsorbate lattice patterns that can be extracted within the diamond-shaped area is at most $A \times B \times (\text{number of elements in } S)$, termed the characteristic transformation operation set for the Slab on the adsorbate lattice. Data for different surface Slabs are presented in Table S4.

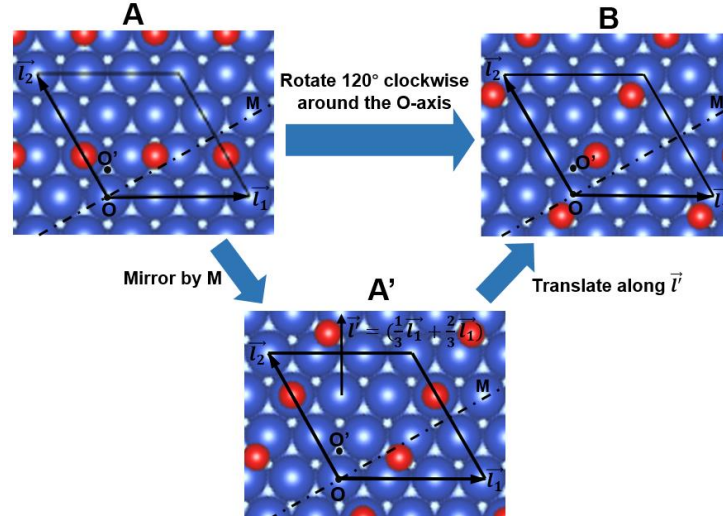
For a specified Cu surface (with its characteristic transformation operation set for the adsorbate lattice denoted as S_m), for two adsorption configurations C1 and C2 under a specified CO coverage: if the sets of adsorption configurations generated by applying all operations in S_m to both C1 and C2 are equivalent, then C1 and C2 are considered equivalent adsorption configurations.

For example, consider two configurations, A and B, corresponding to the adsorption of two CO molecules on the Cu(111)-Slab. The origin O of the slab unit cell is associated with both A and B, and \vec{l}_1 and \vec{l}_2 are the lattice vectors of the Cu(111) slab. In the POSCAR files for configurations A and B, the CO coordinates do not match exactly, meaning they would be initially considered different configurations. In other words, the CO molecules are placed in different regions within the quadrilateral formed by the lattice vectors. However, this arrangement does not take into account the periodicity and symmetry of the slab surface.

By fixing the lattice vectors, we can apply symmetry operations to transform configuration A into configuration B. Specifically, we can perform two operations:

1. A 120° clockwise rotation of A around a fixed axis O'.

2. A mirror reflection of A through a plane passing through point O, producing A', followed by a translation of A' along the vector $\vec{l}' = (\frac{1}{3}\vec{l}_1 + \frac{2}{3}\vec{l}_1)$, resulting in B.
- 3.



Supplementary Figure 1. Symmetry operations for transformation between configurations A and B.

These transformations can be written as:

$$QP = RP = TMP = MP + \left(\frac{1}{3}\vec{l}_1 + \frac{2}{3}\vec{l}_1\right)$$

where Q represents the combined operator, P represents the atomic coordinates of the adsorption configuration (usually in fractional coordinates), R represents the clockwise 120° rotation, M represents the mirror operation through point O, and T represents the translation operation.

This means that if there is a symmetry operation Q that transforms A into B, then A and B are equivalent configurations, or duplicate configurations. Furthermore, the transformation Q is not unique. We can prove that if configuration A is equivalent to configuration B, the corresponding transformation operations are finite and only depend on the inherent symmetry of the slab and the supercell expansion factors. For example, with a Cu(111) slab expanded by a factor of 3 in both directions, the total number of equivalent adsorption configurations generated by symmetry operations is finite and can be expressed as:

$$QP = S_k P + \frac{1}{AB} \begin{pmatrix} \vec{l}_1 & \vec{l}_2 \end{pmatrix} \begin{pmatrix} Ba \\ Ab \end{pmatrix},$$

$$S_k \in S; a \in [0, A - 1], b \in [0, B - 1], a, b \in Z$$

where S_k represents the set of symmetry operations associated with the slab (e.g., identity operation E, rotations R1, R2, and mirror operations M1, M2 and M3 for Cu(111)), A and B are the supercell expansion factors (both equal to 3 in this case). The total number of equivalent configurations is $A \times B \times$ number of symmetry operations, yielding a maximum of 54 equivalent configurations for each adsorption configuration.

In the algorithm, we generate all equivalent configurations of a given adsorption structure A and quickly check whether another configuration B is equivalent. The use of matrix methods allows this process to be done efficiently.

Definition of Average Minimum C-C Distance

For a given coverage, we define the set of configurations with predicted average adsorption energy less than the lowest adsorption energy + 0.01 eV as the most stable adsorption configuration set at that coverage, using the optimized force field to optimize this set. The average minimum C-C distance (MMCD) at this coverage is defined as:

$$\text{MMCD} = \frac{1}{N n_{CO}} \sum_{j=1}^N \sum_{i=1}^{n_{CO}} MCD_i$$

where MCD_i represents the distance between the C atom in the i-th CO and its nearest neighboring C atom, and N is the number of configurations in the most stable adsorption configuration set.

Evaluate the computational overhead of different methods

The time comparison results in Figure 4 were obtained using a CPU cluster provided by Baode Technology Group. Each CPU node in the cluster consists of two CPUs, with models including the Intel® Xeon® CPU E5-2650 v3 and Intel® Xeon® Platinum 9242. The former has 10 cores per CPU, while the latter has 48 cores per CPU.

To obtain 1592 DFT data points, continuous calculations on a 20-core node took 94 days. In contrast, using the trained deep learning force field to perform geometry

optimization on 186,764 structures on a 96-core node took 66 days. The training of the deep learning potential itself took 20 hours on a 96-core node. Training and inference of the graph neural network were performed on a single GPU, and the computational cost was relatively negligible.

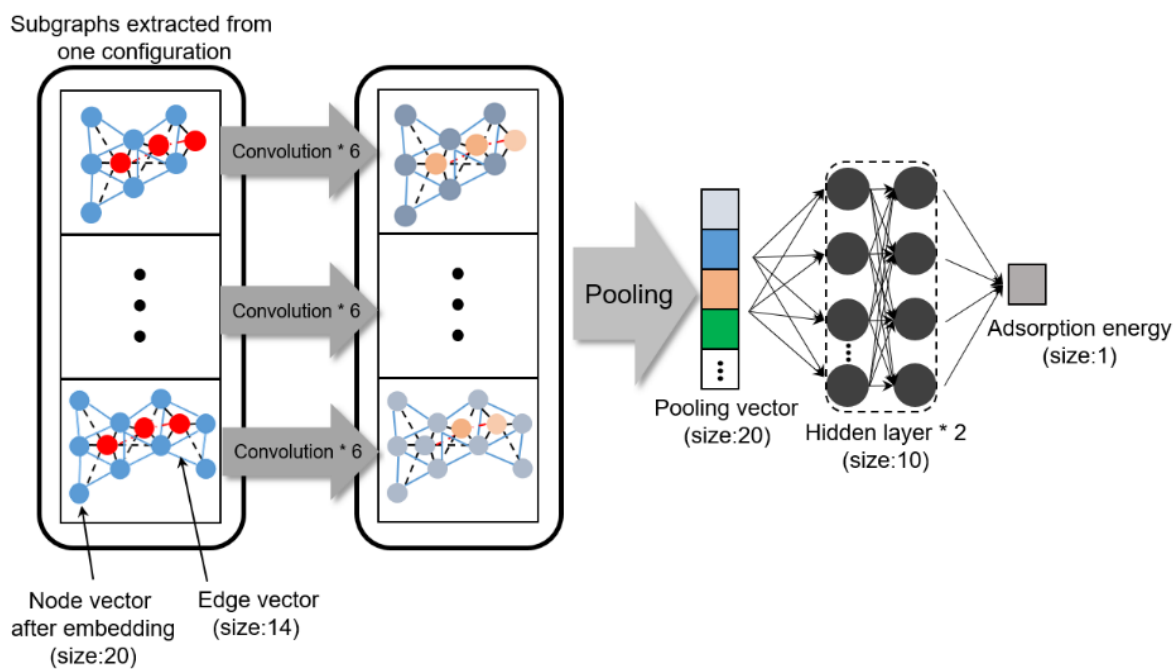
Hence, the computational costs for the three strategies are roughly as follows:

$$\text{DFT: } 94 \times 24 \times 20 \times 6957456 / 1592 = 197,186,190 \text{ (CPU hours)}$$

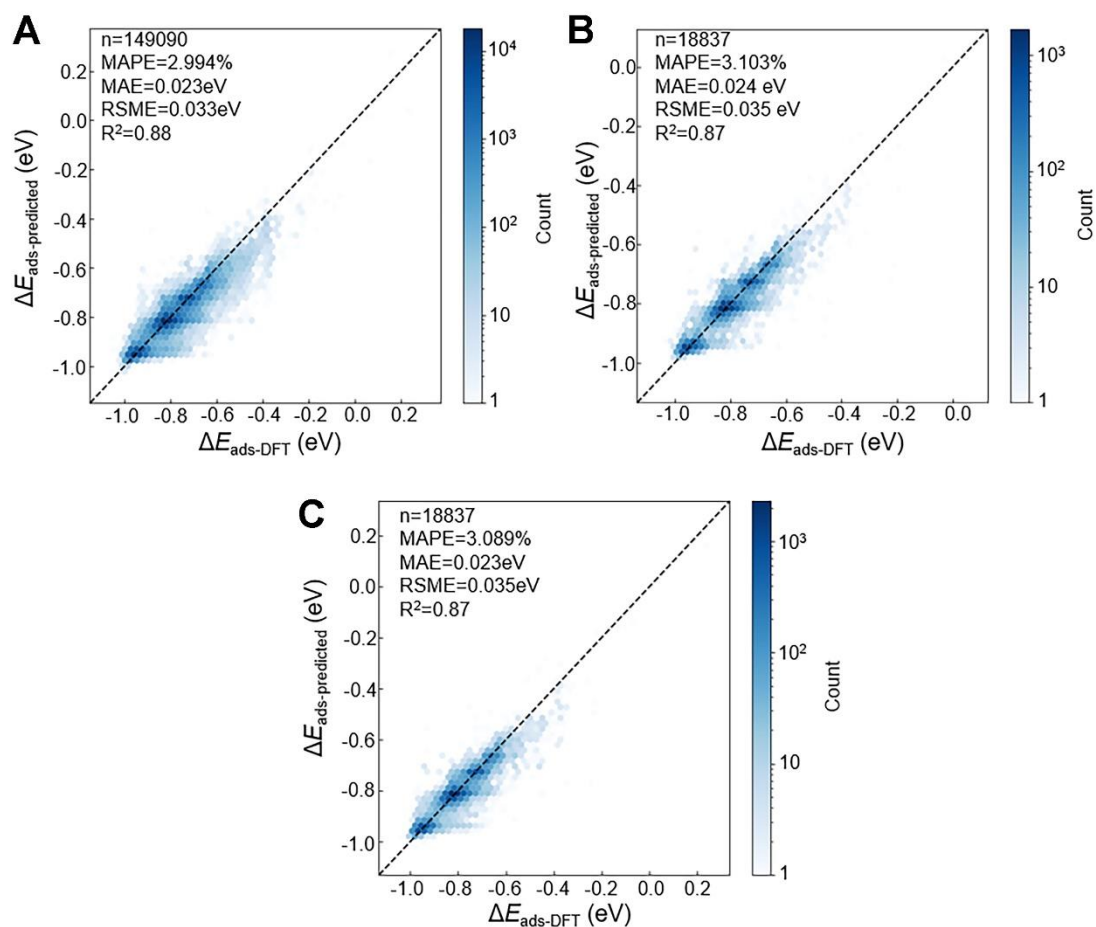
$$\text{DFT+MLFF: } 94 \times 24 \times 20 + 20 \times 96 + 66 \times 24 \times 96 \times (6957456 - 1592) / 186764 = 5,710,532 \text{ (CPU hours)}$$

$$\text{DFT+MLFF+GNN: } 94 \times 24 \times 20 + 20 \times 96 + 66 \times 24 \times 96 = 199,104 \text{ (CPU hours)}$$

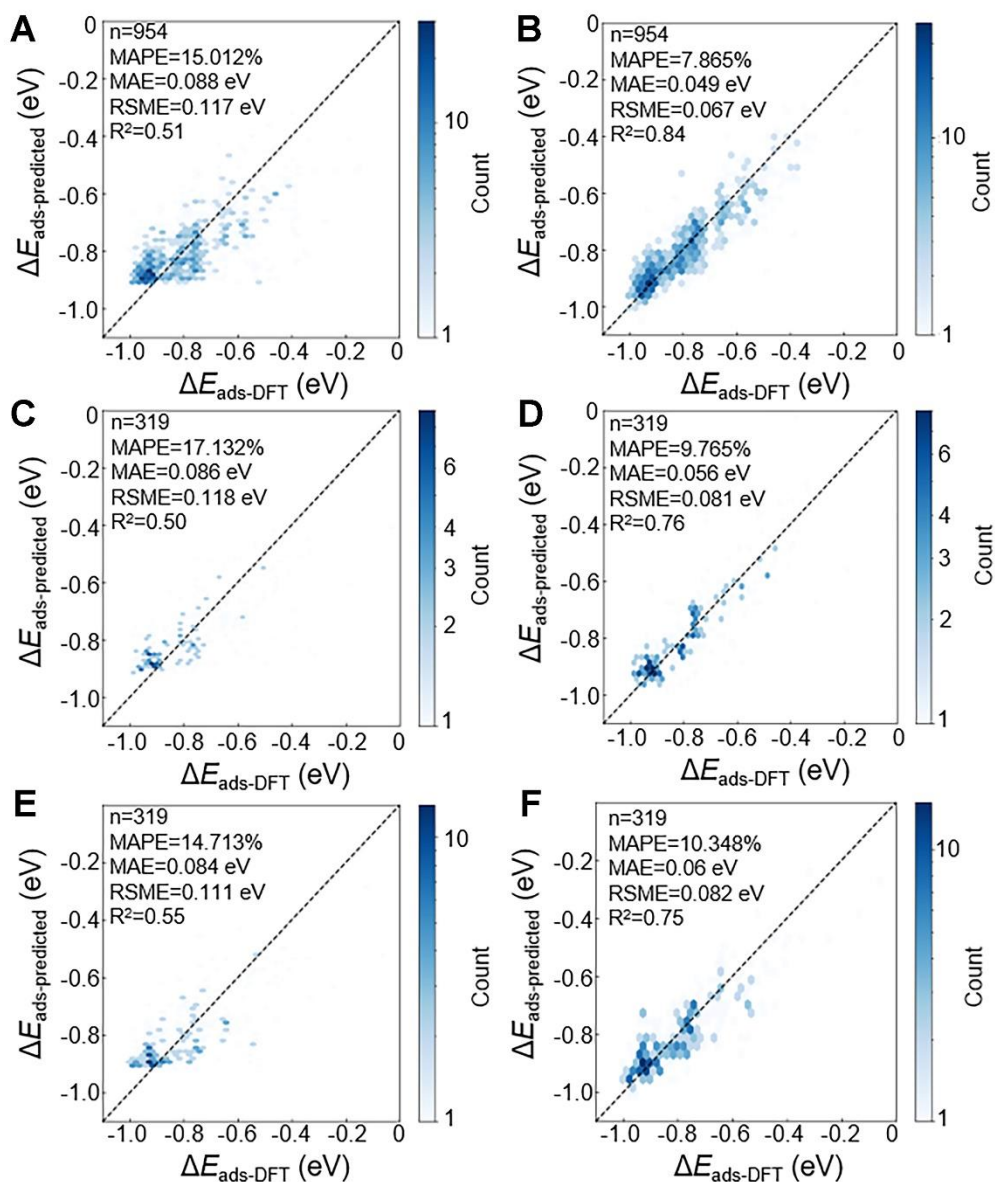
From these calculations, it is clear that the computational cost of the DFT+MLFF+GNN approach is approximately 1/1000 of that of full DFT calculations. It is important to note that this comparison is made for predicting a large number of different adsorption configurations (around 7 million), rather than accelerating the simulation of dynamics for a single configuration. This represents a significant improvement in prediction efficiency.



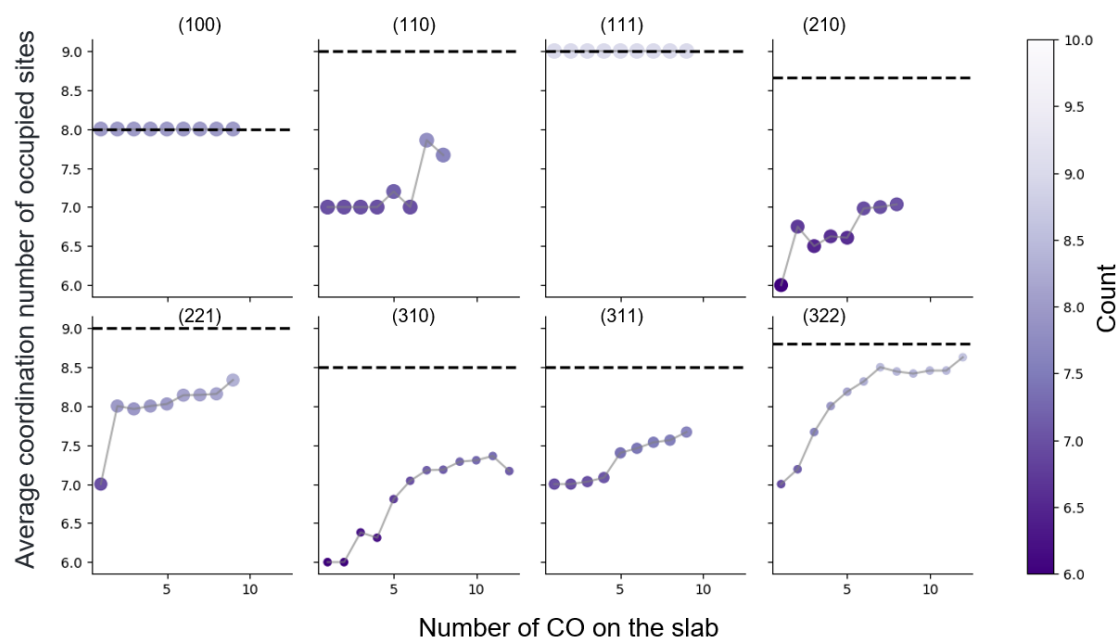
Supplementary Figure 2. Model architecture of GNN.



Supplementary Figure 3. Model performance on DFT+MLFF optimized configurations. (A-C) corresponding to training set, validation set and test set, respectively.



Supplementary Figure 4. Comparison of model performance trained with different feature extraction methods. (A), (C), and (E) corresponding to the training, validation, and testing results of models without considering non-bonded interactions; (B), (D), and (F) corresponding to the training, validation, and testing results of models considering non-bonded interactions.



Supplementary Figure 5. Line graph showing the variation of the average coordination number of Cu atoms occupied by CO as a function of CO coverage. The dashed line represents the average coordination number of Cu atoms on the corresponding clean surface.

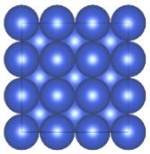
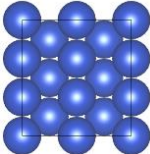
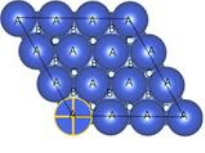
Supplementary Table 1. Configurations counting before deduplication

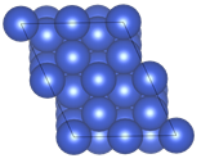
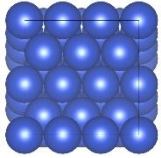
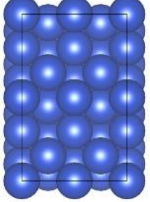
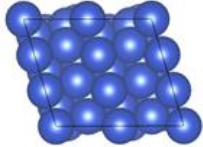
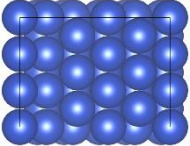
Number of COs on the slab	100	110	111	210	221	310	311	322
1	36	48	54	68	72	76	48	84
2	486	762	972	1662	1785	2242	870	2778
3	3108	4792	7083	17980	19122	33024	7528	46768
4	9837	11667	20196	87717	91155	264209	32463	434217
5	14940	9108	19044	176724	177996	1178728	67212	2264544
6	10596	2212	4824	165816	114869	2936752	62753	6541261
7	3420	24	432	74512	21039	4008556	21564	10090668
8	486	3	54	14256	2430	2848410	4464	7890852
9	28	0	6		317	953856	850	2903258
10						145920	96	492171
11						9728	12	53535
12						256	0	3846
13								159
14								15

Supplementary Table 2. Configurations counting after deduplication

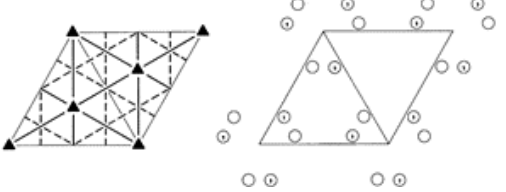
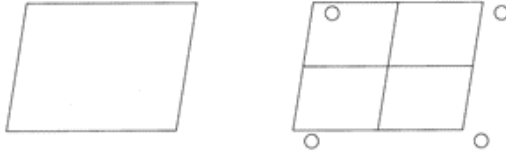
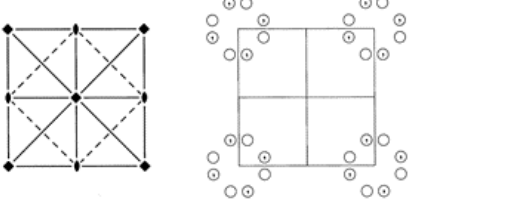
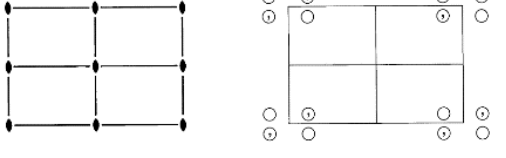
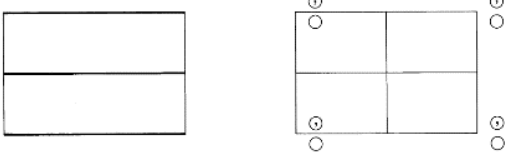
Number of COs on the slab	100	110	111	210	221	310	311	322
1	3	6	4	17	19	14	8	23
2	17	61	29	441	339	339	149	537
3	75	279	170	4495	3347	4276	1260	8182
4	199	642	442	22133	15557	33663	5436	73710
5	292	481	424	44181	30206	148017	11202	380847
6	230	165	145	41711	19678	368910	10518	1096669
7	90	3	23	18628	3704	501985	3594	1690397
8	22	1	4	3657	452	357710	749	1323291
9	7		4		73	119580	149	489281
10						18575	17	84197
11						1240	2	9550
12						46		833
13								39
14								5

Supplementary Table 3. Plane groups and plane point groups corresponding to different Miller indices

Cu facet	Plane group	Plane point group	Number of point operations	Slab expansion factor (A,B)	Slab top view
100	p4mm	4mm	8	(3, 3)	
110	p2mm	2mm	4	(2, 3)	
111	p3m1	3m	6	(3, 3)	

210	p1	1	1	(2, 2)	
221	plm1	m	2	(3, 1)	
310	plm1	m	2	(2, 2)	
311	p1	1	1	(2, 3)	
322	plm1	m	2	(1, 3)	

Supplementary Table 4. Symmetrical pattern of Plane group

Plane group	Symmetrical pattern
p3m1	
p1	
p4mm	
p2mm	
p1m1	

Supplementary Table 5. The number of configurations which were calculated by DFT calculation for training MLFF

Number of COs on the slab	100	110	111	210	221	310	311	322
1	3	5	4	5	5	5	4	4
2	5	13	6	15	15	15	10	10
3	15	56	34	25	25	25	14	14
4	40	129	89	30	27	30	15	14
5	59	97	85	22	28	28	15	14
6	46	33	29	22	28	28	15	14
7	18	3	5	14	24	28	7	14
8	5	1	4	14	16	28	6	14
9	5		4		12	10	6	14
10						10	3	14
11						8	1	14
12						4		14
13								2
14								1

Supplementary Table 6. Configurations for training the adsorption energy prediction model of GNN

Number of COs on the slab	100	110	111	210	221	310	311	322
1	3	6	4	12	16	14	7	15
2	17	61	29	78	136	105	28	120
3	75	279	170	364	695	560	84	606
4	199	642	442	1154	2307	2166	210	2159
5	292	481	424	1583	4835	5202	442	5867
6	230	165	145	1567	4748	8997	518	12943
7	90	3	23	752	1295	11334	280	23630
8	22	1	4	259	165	9743	103	32206

9	7		4		49	4496	40	26416
10						1140	7	10626
11						175	2	2250
12						15		399
13								22
14								4

In the machine learning force field section, we used the DPMD deep learning potential framework. Ensuring consistency in crystal facet and coverage, all trajectories of adsorption configurations optimized by DFT were thoroughly mixed, from which training and validation sets were divided in a 4:1 ratio.

Supplementary Table 7. MLFF training parameters

Item	Content	Value
Descriptor	"type"	"se_e2_a"
	"sel"	"auto"
	"rcut_smth"	0.5
	"rcut"	10
	"neuron"	[25, 50, 100]
	"resnet_dt"	False
	"axis_neuron"	16
Fitting net	"neuron"	[240, 240, 240]
	"resnet_dt"	True
Learning rate	"type"	"exp"
	"decay_steps"	2500
	"start_lr"	0.001
	"stop_lr"	3.51e-08
Loss	"type"	"ener"
	"start_pref_e"	0.02
	"limit_pref_e"	1
	"start_pref_f"	1000
	"limit_pref_f"	1
	"start_pref_v"	0
	"limit_pref_v"	0
Training	"batch_size"	20
	"numb_btch"	6
	"numb_steps"	500000

In the adsorption energy prediction model section, we employed a graph embedding network model based on the local environment of the adsorbate, a technique for feature engineering on graph-structured data, consisting of embedding layers, convolutional layers, and pooling layers, with the final output mapped to label data through hidden layers. During the dataset preparation phase: the dataset was divided into training, validation, and test sets in a 3:1:1 ratio; feature selection included

elemental features and structural features, with elemental features covering electronegativity, atomic radius, valence electron number, first ionization energy, etc., and structural features including bond types and bond lengths. These features were processed through one-hot encoding to obtain initial node feature vectors and edge feature vectors.

Supplementary Table 8. Attributes selection

Attribute	Range	Interval
Pauling electronegativity	0.5-4.0	10
Atom radius (pm)	25-250	10
Valence electrons	1-12	12
First ionization energy (eV)	1.3-3.3	10
Bond length (Angstrom)	0.5-4.5	10
Bond type	0,1,2,3	

Supplementary Table 9. Training parameters

Hyperparameters	Content
optimizer	Adam
momentum	0.9
weight_decay	0
log_learning_rate	-7
lr_milestones	[100,]
batch_size	128
epochs	200
dropout	0.1