

Supporting Information

Identifying Active Sites for CO₂ Reduction on Dealloyed Gold Surfaces by Combining Machine Learning with Multiscale Simulations

Yalu Chen[†], Yufeng Huang[†], Tao Cheng^{†§}, William A. Goddard, III^{*†}

[†] Materials and Process Simulation Center (MSC) and Joint Center for Artificial Photosynthesis (JCAP), California Institute of Technology, Pasadena, California 91125, United States

[§] Institute of Functional Nano & Soft Materials (FUNSOM), Jiangsu Key Laboratory for Carbon-Based Functional Materials & Devices, Joint International Research Laboratory of Carbon-Based Functional Materials and Devices, Soochow University, 199 Renai Road, Suzhou, 215123, Jiangsu, PR China

E-mail: wag@caltech.edu

S1. Gold Nanoparticle Synthesis using ReaxFF Reactive Molecular Dynamics

Using the same approach as in our earlier work¹, the 10nm gold nanoparticle investigated here is computationally synthesized by carrying out ReaxFF reactive molecular dynamics, where a multiwall carbon nanotube (CNT) is used as the catalysis support. We use the Embedded-atom-model (EAM)² to describe the interactions between gold atoms, and the Lennard-Jones (LJ) potential to model the interaction between gold atoms and CNT. To best mimic experimental chemical vapor deposition (CVD), gold atoms are added to the simulation box with the deposition rate of 3.0 Å/ns for 35 ns (the CVD experiment deposition rate is 2 Å/s for 50 s) and simulated annealing with the peak temperature at 1164K is carried out to heal the defect afterwards. 20 ps reactive molecular dynamics³ at 300K is then carried out to refine the final structure after 63 annealing cycles. The predicted XRD spectra and TEM images show that the simulated structure agrees with experiment. All the simulations are carried out in LAMMPS. The predicted TEM image and atomic structure of synthesized AuNP on CNT are also showed in the earlier work¹.

S2. Surface Atoms Extraction using the Surface Vector Method

Surface and bulk atoms are distinguished by the following procedure: for each gold atom, we first cut a sphere of 8 Å radius around the selected atom(cAu) from whole nanoparticle, as shown in **Figure S1**. Then we sum up all the cAu-nAu vectors pointing from the center atom(cAu) to other atoms(nAu) in the nanoclusters and take the negative norm of this summed vector as surface vector (red vector in Figure S1). This selected Au atom is defined as a surface atom if all the angles between cAu-nAu vector and surface vector are larger than the given threshold, which is optimized to 30 degree in our system, otherwise it's defined as a bulk atom. The same 8 Å nanocluster model will be used later for DFT calculations and feature extraction for neural network(NN) inputs.

S3. Datasets Generation from DFT Calculations

All datasets presented in this work are generated from density functional theory(DFT) calculations in VASP package⁴, using the PBE functional including the D3 London Dispersion correction and the projector augmented wave (PAW) method⁵. We set the kinetic energy cutoff at 400 eV and the Methfessel-Paxton smearing of second order is used with the width of 0.2 eV. The convergence criteria are 10⁻⁵ eV differences for electronic energy and all geometries are relaxed until the force converged to 10⁻² eV/Å. Only gamma point is considered in these calculations. We use the 8 Å nanocluster as described above for DFT calculations. The cluster with different adsorbates (CO or HOCO) is put into a 20 Å cubic box with all Au atoms fixed. The two physical descriptors we use here for evaluating the activity of a site are ΔE_{CO} and ΔE_{HOCO} as described in **Section 2.1**.

S4. Modeling Dealloyed Gold Surfaces

To model the dealloyed gold surfaces as in earlier work⁶, we first cut an 10nm gold sphere from gold single crystal and randomly remove 25% of Au atoms on the surface. Simulated annealing with the peak temperature at 1164K is carried out to equilibrate the structure afterwards. 100 ps ReaxFF reactive molecular dynamics at 300K is then carried out to refine the final structure after 10 annealing cycles. All calculations are carried out in LAMMPS.

S5. Neural Network Based Machine Learning Model

We use the neural network based machine learning algorithm in this work. The overall structure of model topology is shown schematically in **Figure 2**. For a nanocluster with N atoms taking the cAu atom as center, the Cartesian coordinates of all atoms in the cluster are given. We first calculate the N-1 interatomic distances (R_{ij}), and transform them to a set of symmetry function values, which we call input features. Two main parts of features are two-body terms(C_2) and three-body terms(C_3). As shown in **Figure S2(B)**, two-body terms are constructed by summing over all nAu atoms in the systems and three-body terms are iterated through all triangles taking cAu atom in one corner. The mathematical representation of two-body terms and three-body terms are also shown in **Equation 1 and 2**, where f is the symmetry function. This type of feature representation method is derived from the work by Behler and Parrinello in 2007⁷. We use localized piecewise cosine function as symmetry functions instead of Gaussian functions, as in **Equation 3**. These piecewise cosine functions are more localized than Gaussian functions with the value of zero outside their cutoff distance^{8,9}. For the three-body term, we no longer need the angle of dependence with three sides treated equally, which is also one of the reasons why we are using this simplified version of Gaussian functions. In our model, we use 12 symmetry functions for two-body terms and 3 symmetry functions for three-body terms, leading to a total number of 39 input features. We consider this gives the best balance of dataset size and model complexity. Having defined a set of features, a fully connected two-layer neural network with 30 nodes in the first layer and 50 nodes in the second layer are followed to fit two selected physical descriptors: ΔE_{CO} and ΔE_{HOCO} . The total number of model parameters is 2801. The main idea is to represent the physical descriptors as a function of two-body terms and three-body terms with the weight parameters w and bias parameters b , as in **Equation 4**.

$$\text{Two-Body Term: } C_{2m,i} = \sum_j f_m(R_{ij}), (1)$$

$$\text{Three-Body Term: } C_{3mnl,i} = \sum_{jk} f_m(R_{ij})f_n(R_{ik})f_l(R_{jk}), (2)$$

Basis Function (Localized Cosine Piecewise Function):

$$f_m(R_{ij}) = \begin{cases} \frac{1}{2} \cos\left(\frac{R_{ij}-d_m}{r} \pi\right) + \frac{1}{2}, & |R_{ij} - d_m| < r; \\ 0, & \text{Otherwise} \end{cases}, (3)$$

$$E_i = F_{NN}(C_{2m,i}, C_{3mnl,i}; w, b), (4)$$

Where,

C_2 , C_3 are the symmetry functions of two-body terms and three-body terms, and m, n, l are indices of these symmetry functions;

i is the index of surface atom, and j and k are the indices of nAu atom (**Figure S1**);

d_m and r is the center and width of symmetry function as shown in **Figure S2(A)**;

F_{NN} is the neural network function taking the two-body terms and three-body terms as variables and in parameter of weights and bias.

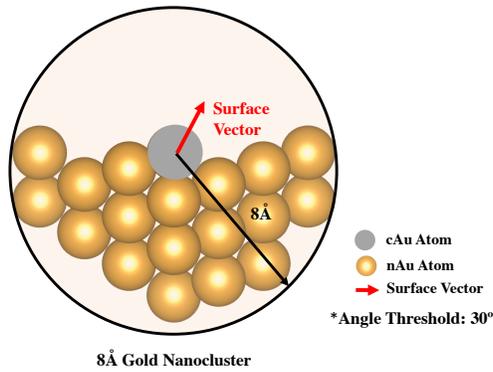


Figure S1. 8 Å nanocluster model, where we use surface vector method to distinguish whether the cAu atom is a surface atom or a bulk atom. Same model will be used later for DFT calculations and feature extraction for neural network inputs. cAu Atom: center atom; nAu Atom: other atoms in the nanocluster.

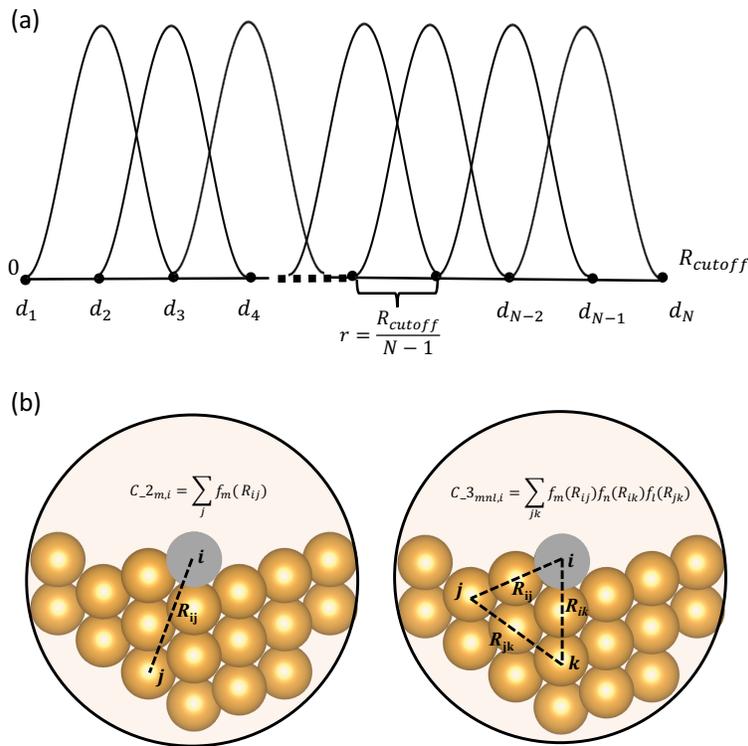


Figure S2. Mapping geometric features to symmetry functions. (a) Localized piecewise cosine symmetry functions. d_m and r are the center and width of symmetry function. In our model, R_{cutoff} is set to 8 Å; (b) Illustration of Two-Body Term and Three-Body Term. Two-body terms are constructed by summing over all nAu atoms and three-body terms are iterated through all triangles taking cAu atom in one corner. We use 12 symmetry functions for two-body terms and 3 symmetry functions for three-body terms, which leads to a total number of 39 input features to neural network.

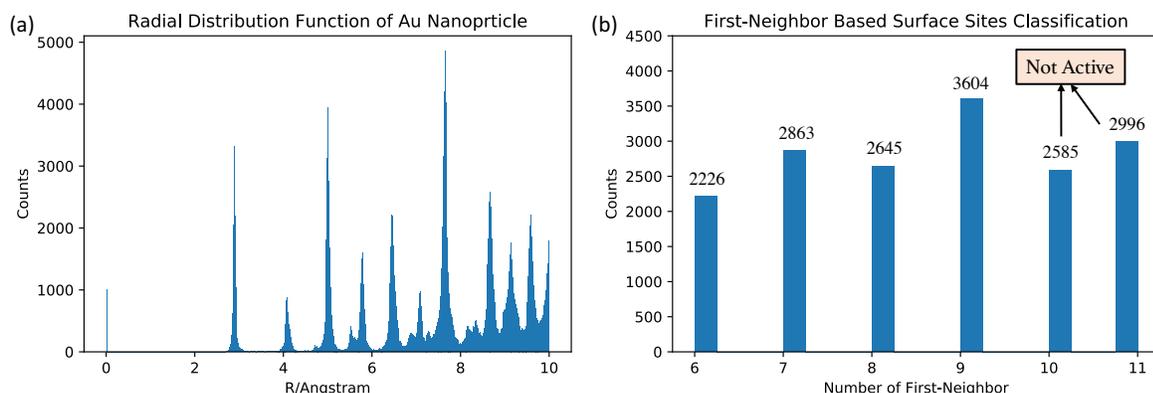


Figure S3. (a) Au-Au radial distribution function for Au nanoparticles synthesized by reactive molecular dynamics. The first peak appears at 2.80 Å and the second peak appears at 4.20 Å. We choose first-neighbor cutoff at 3.30 Å here; (b) Surface sites classification based on first-neighbor. We classify all 16919 surface sites into six groups based on their number of first-neighbor (coordination). Sites with 10 or 11 first-neighbor are not active for CO₂RR since they cannot adsorb HOCO or show very high HOCO formation energy.

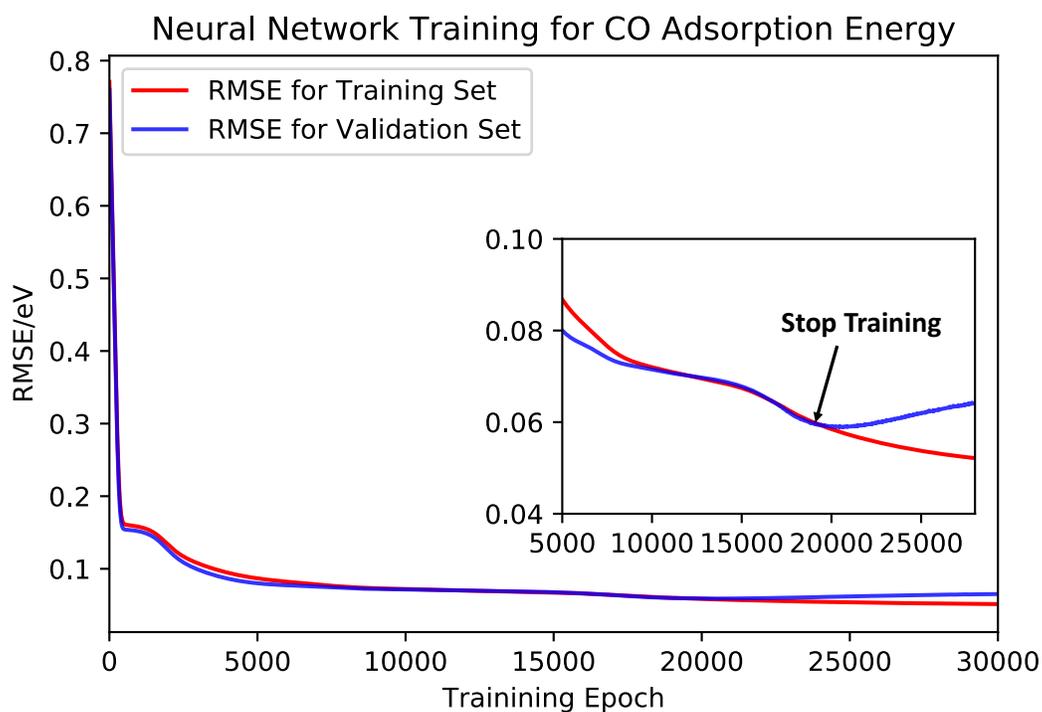


Figure S4. Training log for CO adsorption energy, ΔE_{CO} . Validation set is used for preventing overfitting (early-stop). At epoch 19000, the RMSE of validation set reached minimum at 0.0591 eV with the RMSE of training set at 0.0563 eV, where we stop the training.

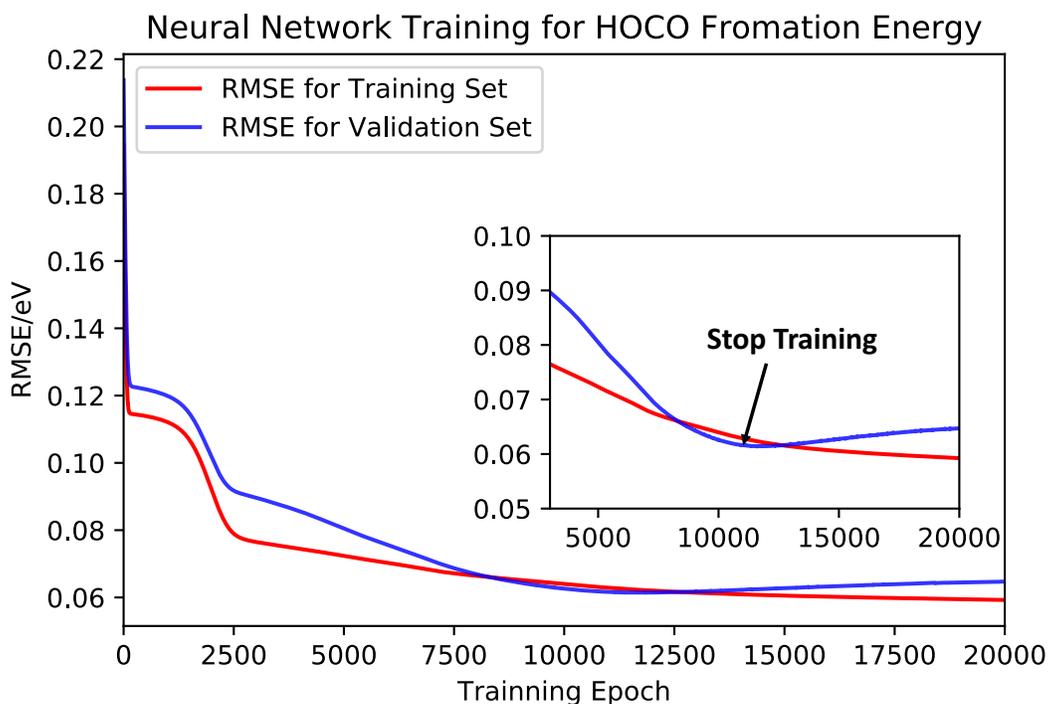


Figure S5. Training log for COOH formation energy, ΔE_{HOCO} . Validation set is used for early-stop and prevent overfitting. At epoch 11000, the RMSE of validation set reaches minimum at 0.0593 eV with the RMSE of training set at 0.0616 eV, where we stop the training.

Data Sets	Coordination				Data Set Size	Final RMSE/eV
	6	7	8	9		
Training Set	276	276	276	276	1104	0.0563
Validation Set	35	35	35	35	140	0.0591
Testing Set	35	35	35	35	140	0.0521

Table S1. Partition of data sets and final RMSE for CO adsorption energy training, ΔE_{CO} . To have a complete training set, we constrain the ratio of sites from each coordination group to be equal within each set. All surface sites within each group are selected randomly and all three sets are totally independent.

Data Sets	Coordination				Data Set Size	Final RMSE/eV
	6	7	8	9		
Training Set	224	214	209	212	859	0.0616
Validation Set	25	25	25	25	100	0.0593
Testing Set	25	25	25	25	100	0.0614

Table S2. Partition of data sets and final RMSE for HOCO formation energy, ΔE_{HOCO} . To have a complete training set, we constrain the ratio of sites from each coordination group to be equal within each group. All surface sites within each group are selected randomly and all three sets are totally independent.

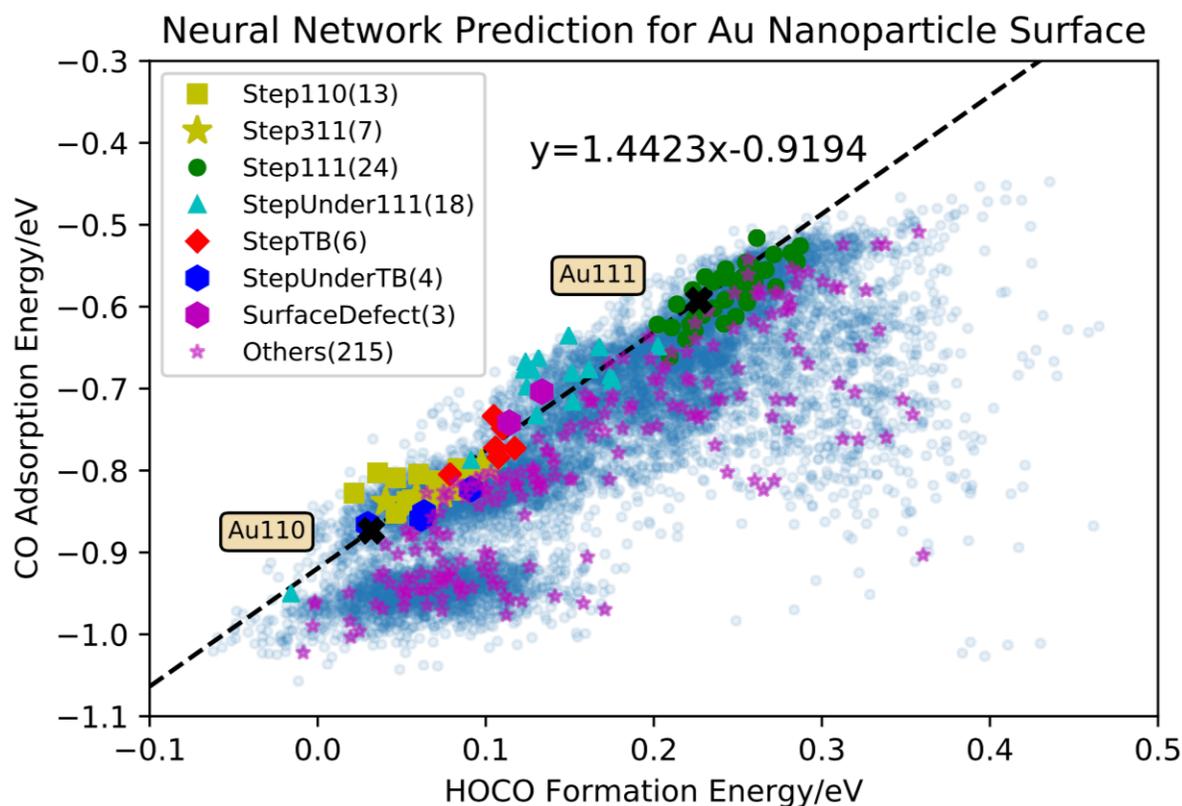


Figure S6. Identification results of 300 sites randomly selected from all surface sites with the number of sites shown in the bracket next to the markers. The majority of randomly selected 300 sites are not from seven active groups, indicating that the seven groups are only concentrated above the straight line with a value at 0.9194.

Type of Sites	Top 300 Sites	Random 300 Sites
Step110	66	13
Step311	24	7
Step111	47	24
StepUnder111	115	18
StepTB	34	6
StepUnderTB	11	4
SurfaceDefect	3	3
Others	0	215

Table S3. Comparison of top 300 sites and random 300 sites. The majority of random 300 sites are not from seven active groups (as marked as star in Figure S6), which implies that seven active groups mainly concentrate above and around the straight line.

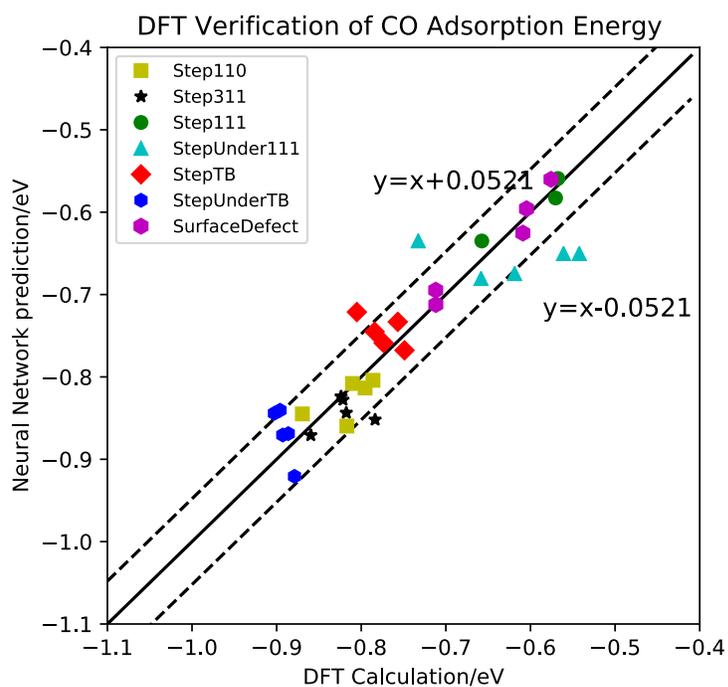


Figure S7. DFT verification of CO adsorption energy for seven active groups. RMSE of machine learning model for CO adsorption energy is 0.0521 eV (dashed line is the error bound). We randomly selected 5 sites from each group and as we could see most sites lie within the error bound, which support our model is accurate and on the other hand validate the seven groups are the sites with better CO₂RR performance.

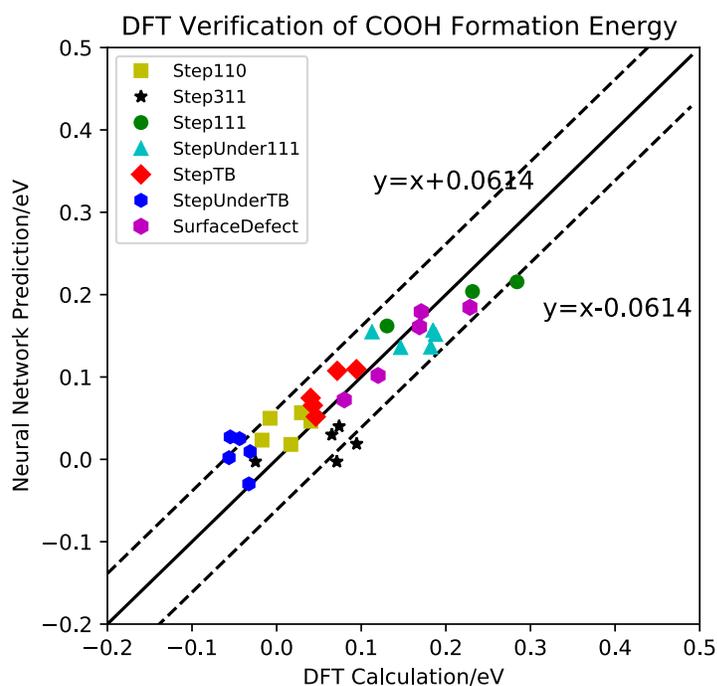


Figure S8. DFT verification of HOCO formation energy for seven active groups. RMSE of machine learning model for HOCO formation energy is 0.0614 eV (dashed line is the error bound). We randomly selected 5 sites from each group and as we could see most sites lie within the error bound, which support our model is accurate and on the other hand validate the seven groups are the sites with better CO₂RR performance.

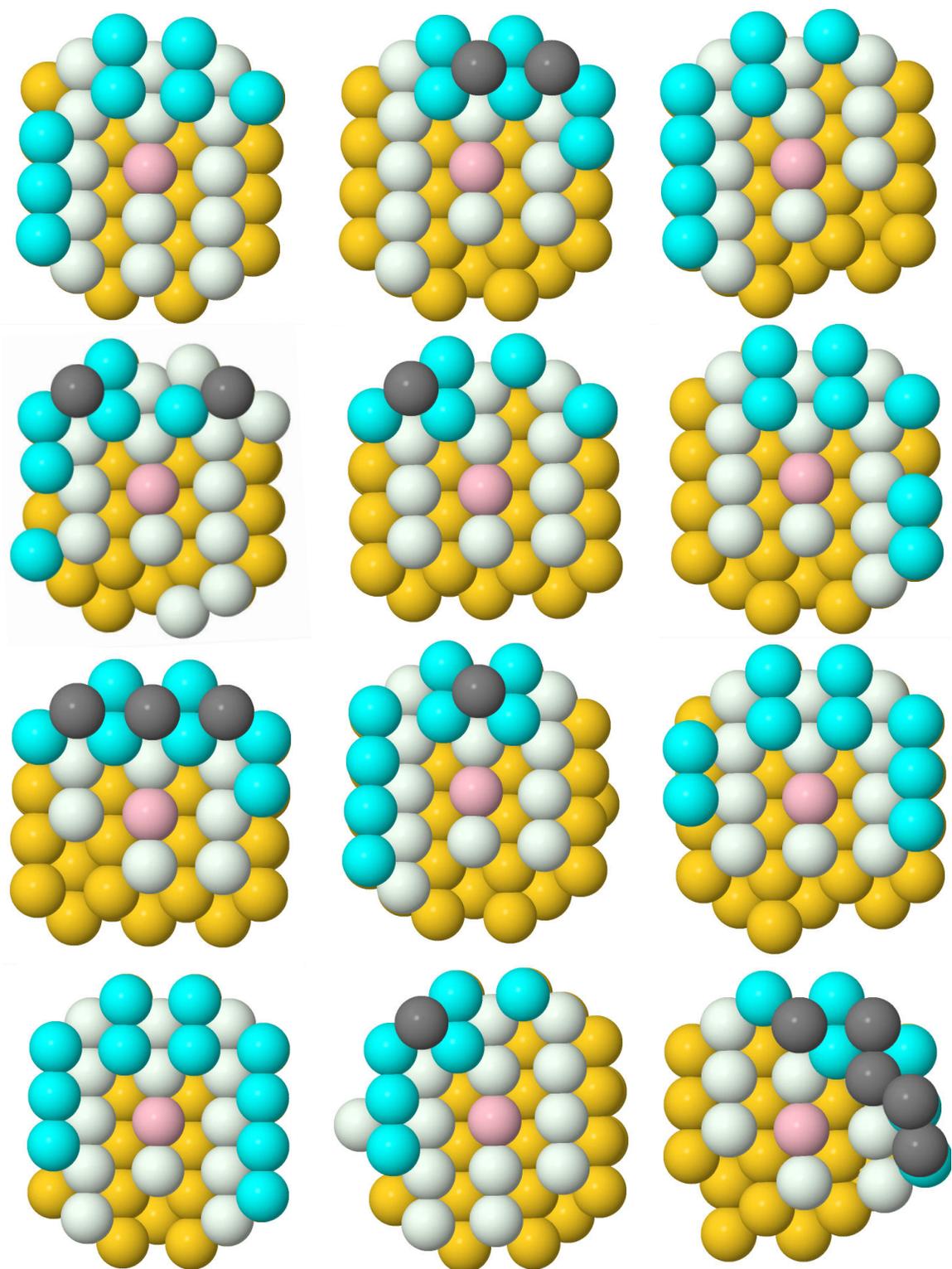


Figure S9. 12 structures from Group of Step110. The center atom is pink, while atoms at the same layer are white. Atoms in the layer below white atoms are gold, while atoms one layer above center atom are cyan. Atoms above cyan atoms are gray and twin boundaries are dashed line.

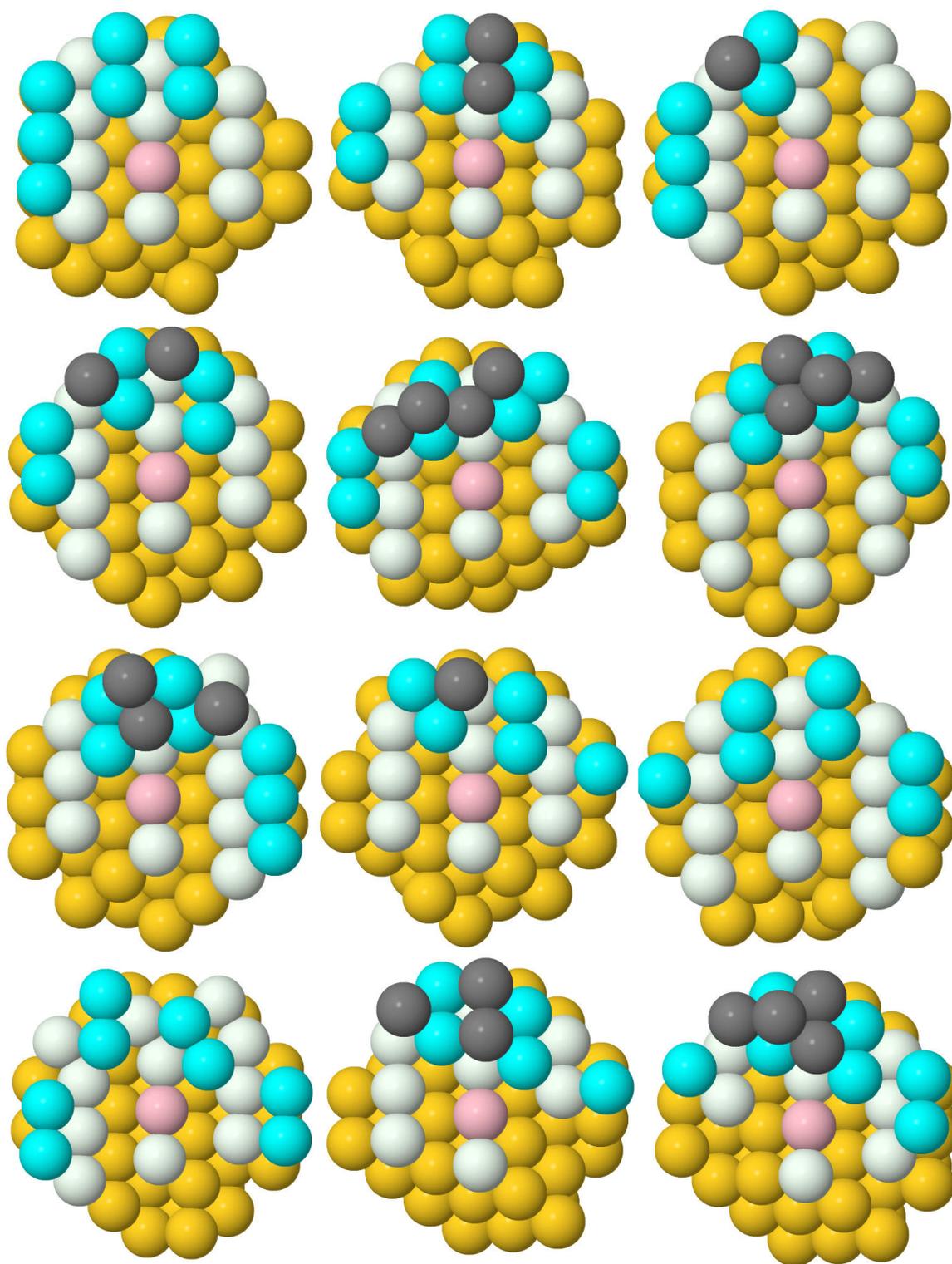


Figure S10. 12 Structures from Group of Step311. The center atom is pink, while atoms at the same layer are white. Atoms in the layer below white atoms are gold, while atoms one layer above center atom are cyan. Atoms above cyan atoms are gray and twin boundaries are dashed line.

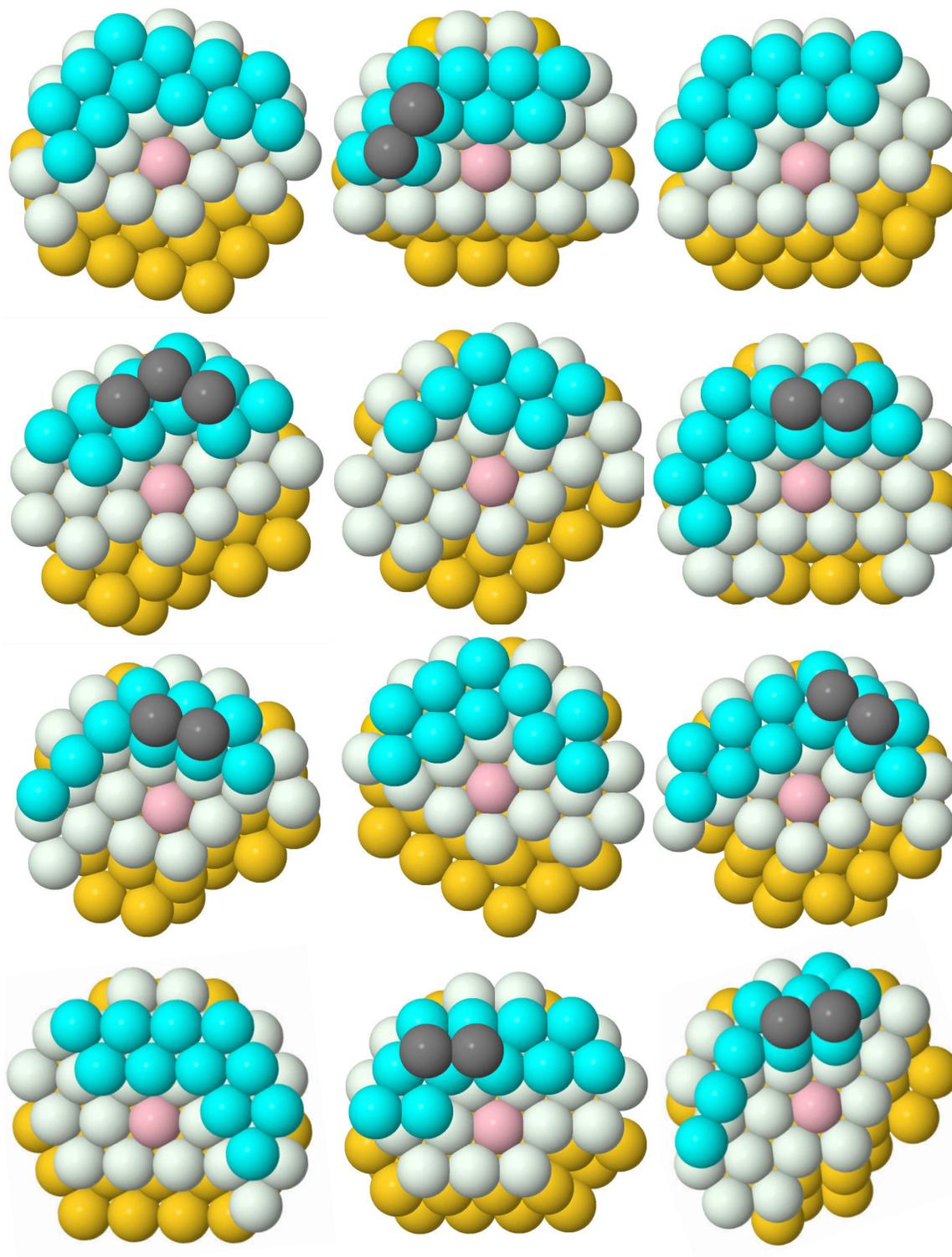


Figure S11. 12 Structures from Group of Step111. The center atom is pink, while atoms at the same layer are white. Atoms in the layer below white atoms are gold, while atoms one layer above center atom are cyan. Atoms above cyan atoms are gray and twin boundaries are dashed line.

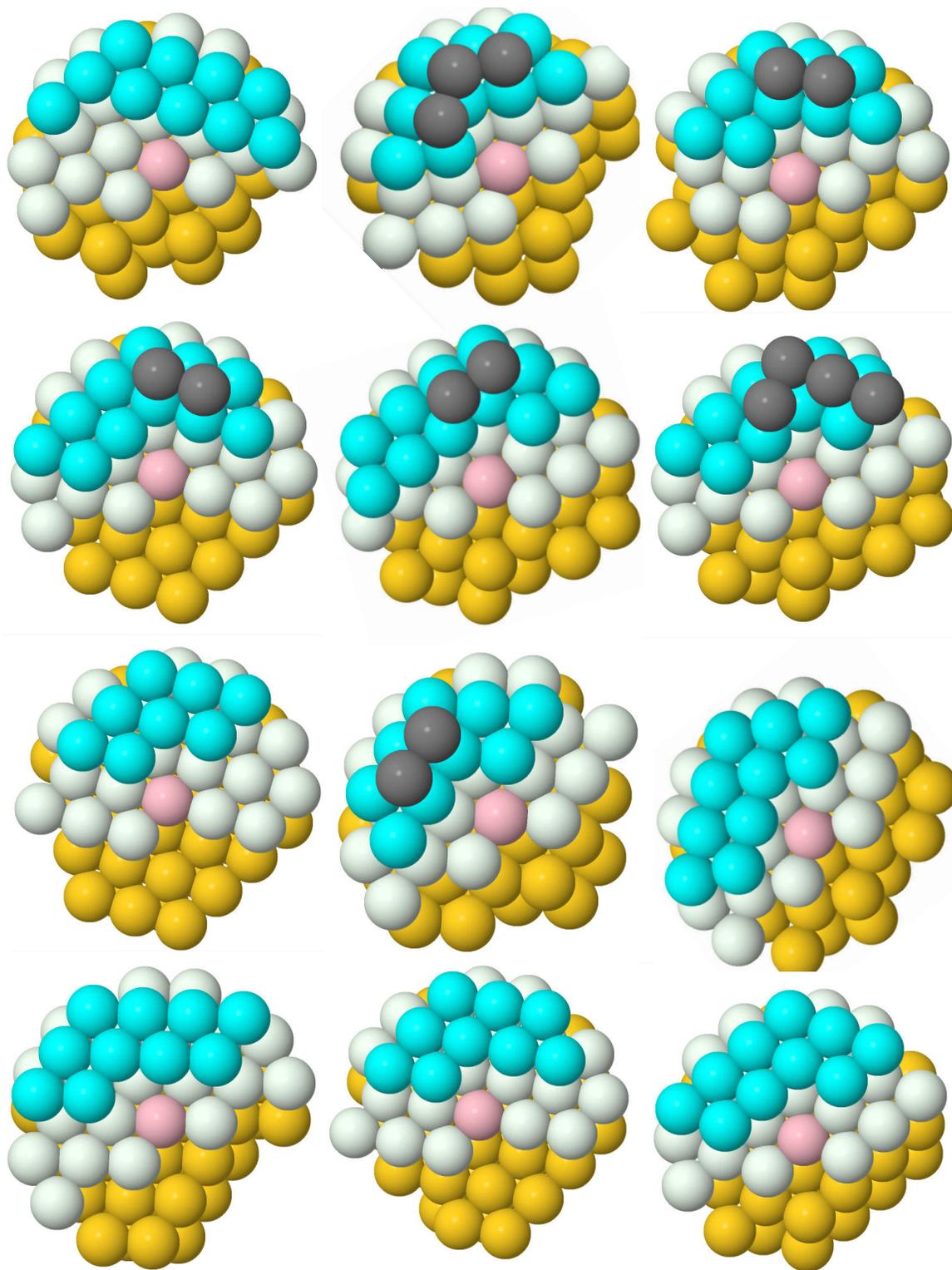


Figure S12. 12 Structures from Group of StepUnder111. The center atom is pink, while atoms at the same layer are white. Atoms in the layer below white atoms are gold, while atoms one layer above center atom are cyan. Atoms above cyan atoms are gray and twin boundaries are dashed line.

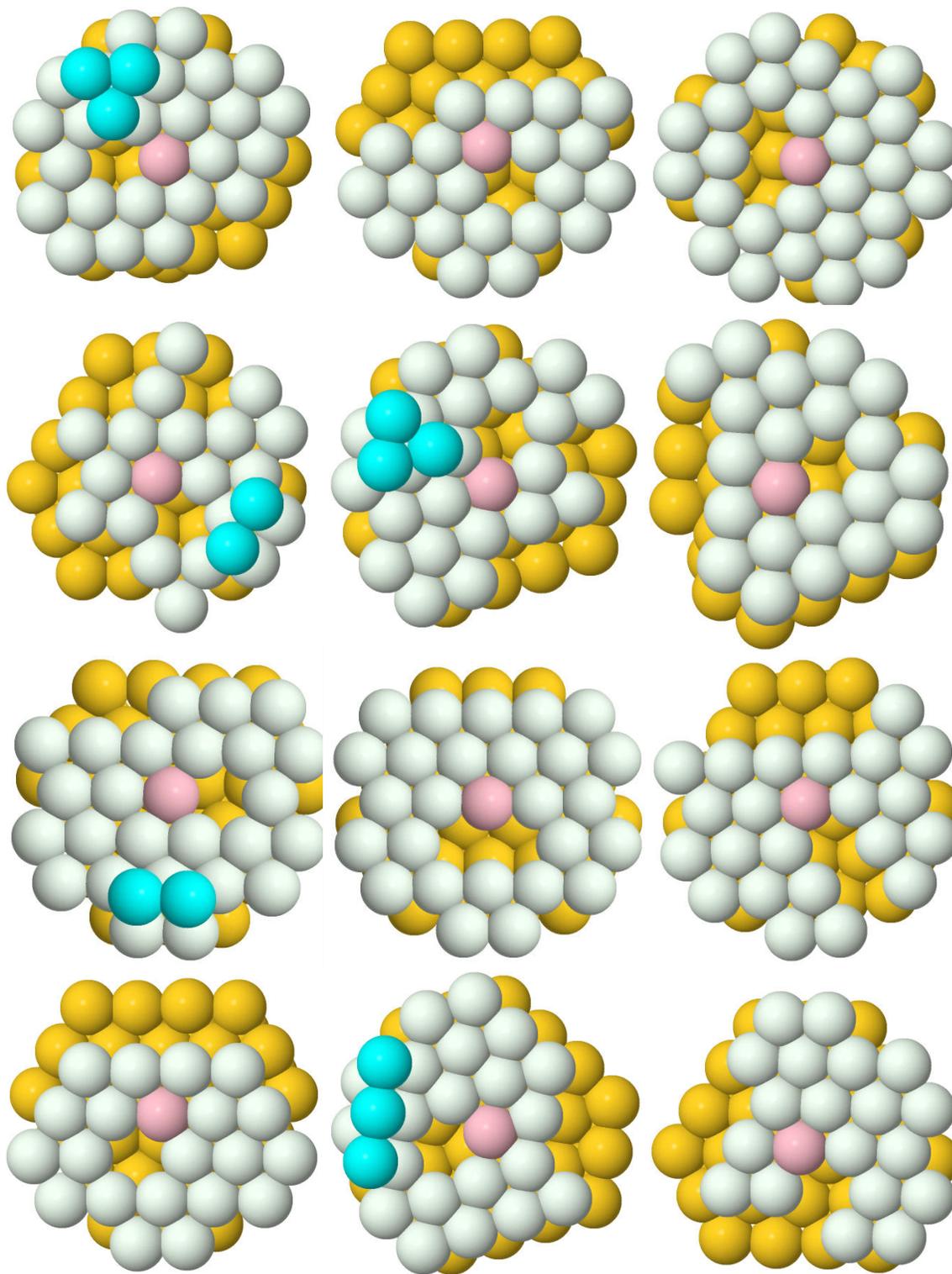


Figure S13. 12 Structures from Group of SurfaceDefect. The center atom is pink, while atoms at the same layer are white. Atoms in the layer below white atoms are gold, while atoms one layer above center atom are cyan. Atoms above cyan atoms are gray and twin boundaries are dashed line.

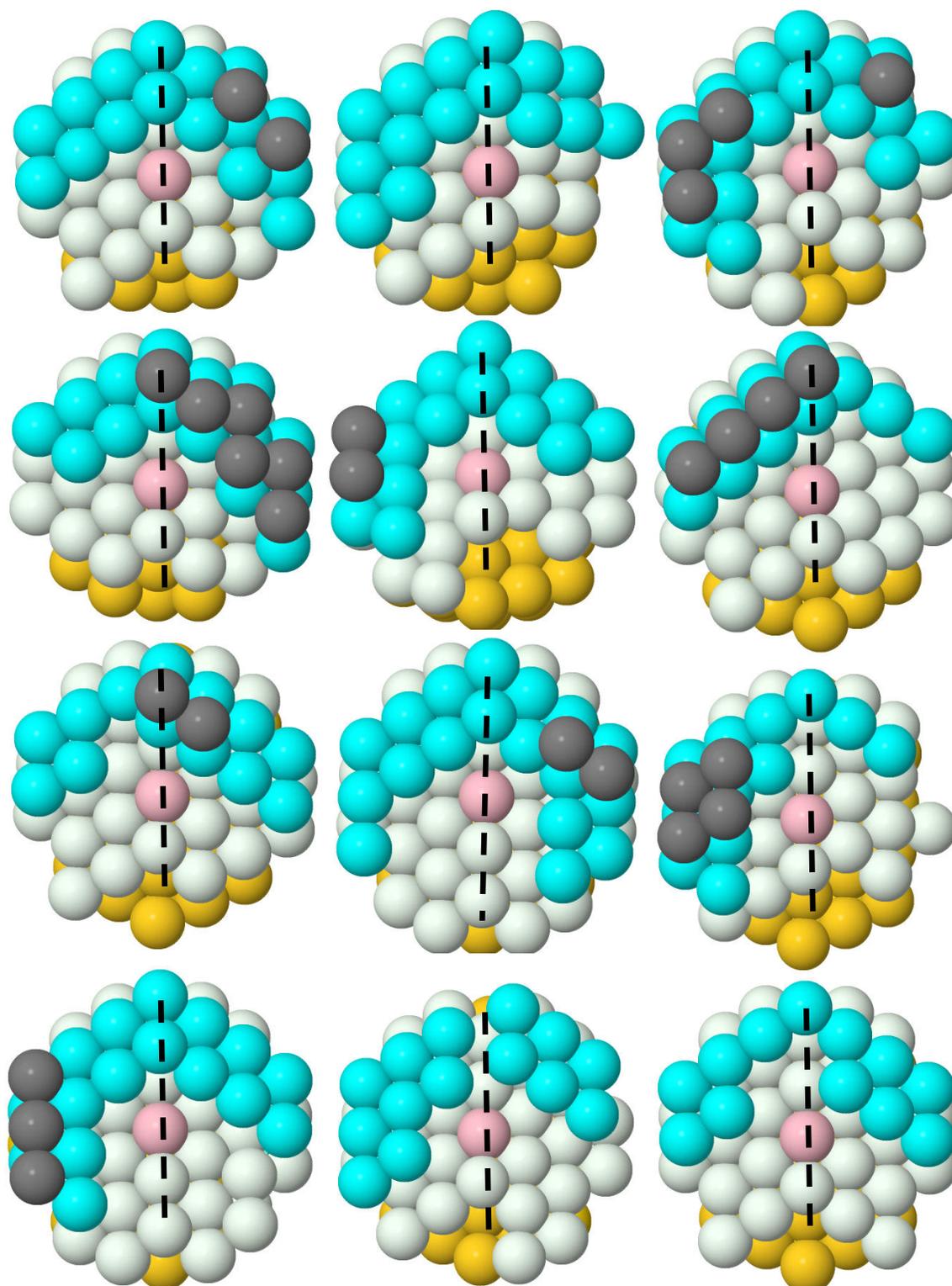


Figure S14. 12 Structures from Group of StepTB. The center atom is pink, while atoms at the same layer are white. Atoms in the layer below white atoms are gold, while atoms one layer above center atom are cyan. Atoms above cyan atoms are gray and twin boundaries are dashed line.

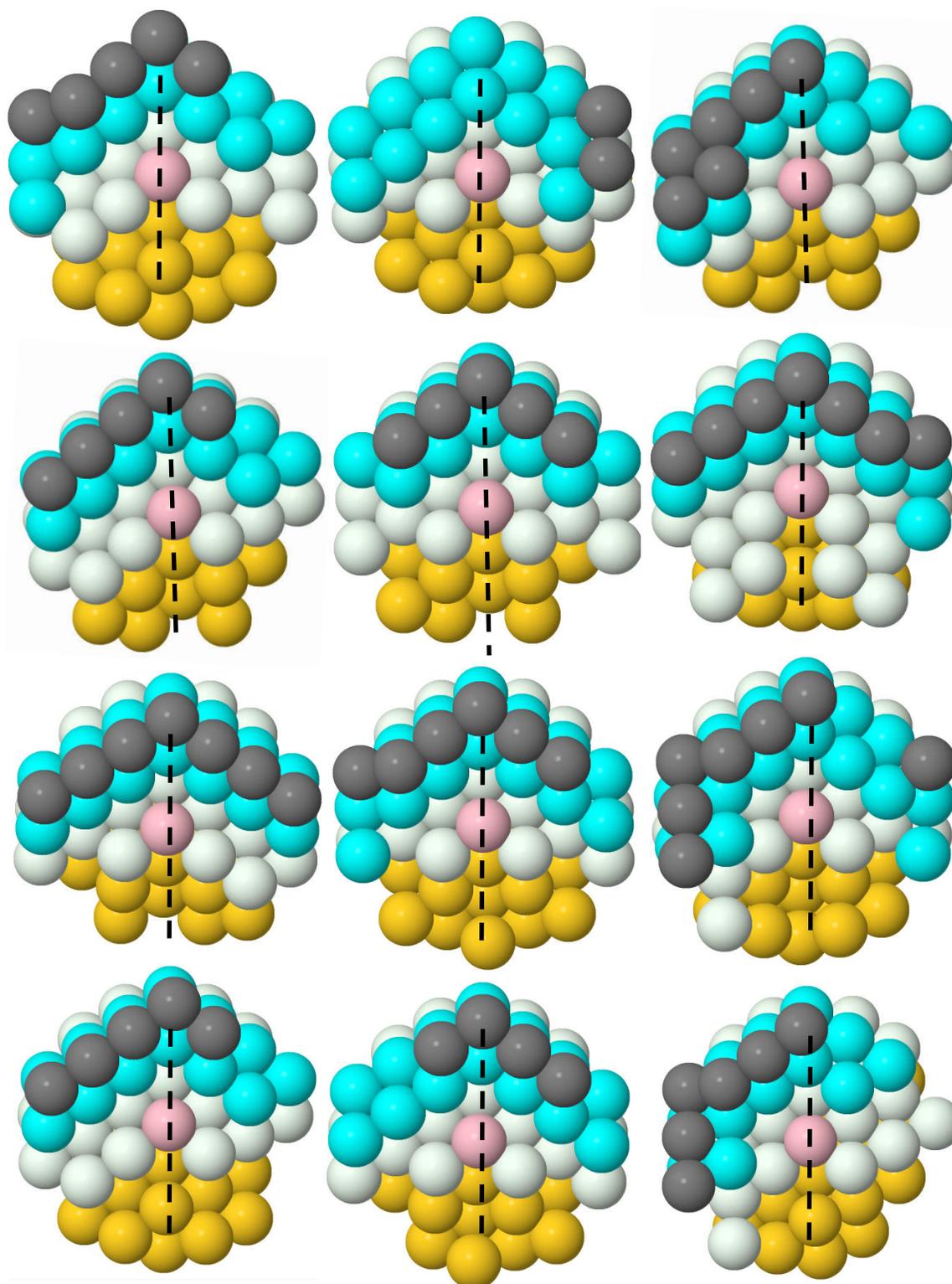


Figure S15. 12 Structures from Group of StepUnderTB. The center atom is pink, while atoms at the same layer are white. Atoms in the layer below white atoms are gold, while atoms one layer above center atom are cyan. Atoms above cyan atoms are gray and twin boundaries are dashed line.

REFERENCE

- (1) Cheng, T.; Huang, Y.; Xiao, H.; Goddard, W. A. Predicted Structures of the Active Sites Responsible for the Improved Reduction of Carbon Dioxide by Gold Nanoparticles. *The Journal of Physical Chemistry Letters* **2017**, *8* (14), 3317–3320. <https://doi.org/10.1021/acs.jpcclett.7b01335>.
- (2) Foiles, S. M.; Baskes, M. I.; Daw, M. S. Embedded-Atom-Method Functions for the Fcc Metals Cu, Ag, Au, Ni, Pd, Pt, and Their Alloys. *Phys. Rev. B* **1986**, *33* (12), 7983–7991. <https://doi.org/10.1103/PhysRevB.33.7983>.
- (3) van Duin, A. C. T.; Dasgupta, S.; Lorant, F.; Goddard, W. A. ReaxFF: A Reactive Force Field for Hydrocarbons. *The Journal of Physical Chemistry A* **2001**, *105* (41), 9396–9409. <https://doi.org/10.1021/jp004368u>.
- (4) Kresse, G.; Hafner, J. *Ab Initio* Molecular-Dynamics Simulation of the Liquid-Metal–Amorphous-Semiconductor Transition in Germanium. *Physical Review B* **1994**, *49* (20), 14251–14269. <https://doi.org/10.1103/PhysRevB.49.14251>.
- (5) Kresse, G.; Joubert, D. From Ultrasoft Pseudopotentials to the Projector Augmented-Wave Method. *Physical Review B* **1999**, *59* (3), 1758–1775. <https://doi.org/10.1103/PhysRevB.59.1758>.
- (6) Sun, K.; Cheng, T.; Wu, L.; Hu, Y.; Zhou, J.; Maclennan, A.; Jiang, Z.; Gao, Y.; Goddard, W. A.; Wang, Z. Ultrahigh Mass Activity for Carbon Dioxide Reduction Enabled by Gold–Iron Core–Shell Nanoparticles. *Journal of the American Chemical Society* **2017**, *139* (44), 15608–15611. <https://doi.org/10.1021/jacs.7b09251>.
- (7) Behler, J.; Parrinello, M. Generalized Neural-Network Representation of High-Dimensional Potential-Energy Surfaces. *Physical Review Letters* **2007**, *98* (14). <https://doi.org/10.1103/PhysRevLett.98.146401>.
- (8) Huang, Y.; Chen, Y.; Cheng, T.; Wang, L.-W.; Goddard, W. A. Identification of the Selective Sites for Electrochemical Reduction of CO to C₂₊ Products on Copper Nanoparticles by Combining Reactive Force Fields, Density Functional Theory, and Machine Learning. *ACS Energy Letters* **2018**, 2983–2988. <https://doi.org/10.1021/acsenenergylett.8b01933>.
- (9) Huang, Y.; Kang, J.; Goddard, W. A.; Wang, L.-W. Density Functional Theory Based Neural Network Force Fields from Energy Decompositions. *Physical Review B* **2019**, *99* (6), 064103. <https://doi.org/10.1103/PhysRevB.99.064103>.