Supporting Information

Machine Learning Assisted Stability and CO_2 Reduction Reaction Activity Prediction of Single Atom Alloys

Qiuyan Zhu, Dongxu Tian*, Xiaru Du[b]

School of Chemistry, Dalian University of Technology, No.2 Linggong Road, Dalian City, Liaoning Province, P. R. China, 116024

[b]Dalian Kaiteli Catalytic Engineering Technology Co. Ltd., No.19 Huangpu Road, Dalian City, Liaoning Province, P. R. China, 116024

Table S1. Comparison of XGBoost Model Performance in Retaining and Removing Outliers.

	Test_R^2	Test_RMSE	Test_MAE
Retain Outliers	0.85	0.45	0.27
Remove outliers	0.92	0.31	0.21

Table S2. Features selected for formation energy prediction model.

	Name	Symbol	Unit
Target variables	Formation energy	E_f	eV
Element Properties	Atomic number	Z_a、Z_b	
	Atomic radius	r_a、r_b、Δr	pm
	Work function	W_a、W_b、ΔW	eV
	Ionization energy	IE_a、IE_b、ΔIE	MJ/mol
	Electronegativity	χ_a、χ_b、Δχ	
	Electron affinity	EA_a、EA_b、ΔEA	eV
	Number of valence electrons	VE_a、VE_b、ΔVE	
	Period	Period_a、Period_b	
	d-band filling of the host metal	f	
	d-band center of the host metal	ε_d	eV
	Mass	M_a、M_b、ΔM	
	Density	ρ_a、ρ_b、Δρ	g/cm^3
	Specific heat capacity	C_p_a、C_p_b、ΔC_p	J/Kg*K
	Cohesion energy	CEbulk_a、CEbulk_b、ΔCEbulk	eV
	Cohesive energy/coordination number of a single atom	CEbulk_a/CN、CEbulk_b/CN、ΔCEbulk/CN	eV
Structural characteristics	Coordination number of a single atom	CN	
	Miller index	MI	

	Bond length between dopant and host metal	l_1、l_2、Δl_1、Δl_2	Å
	Lattice constant	a_a、a_b、b_a、b_b、c_a、c_b、Δa、Δb、Δc	Å

‘a’ denotes the host metal and ‘b’ denotes the dopant atom.

Table S3. The cohesive energy of 46 kinds of metals.

Atom	CEbulk (eV/mol)	Atom	CEbulk (eV/mol)
Sc	-4.53	Os	-9.05
Ti	-6.09	Ir	-8.30
V	-5.82	Pt	-6.30
Cr	-9.87	Au	-3.65
Mn	-3.90	Be	-3.70
Fe	-5.35	Mg	-1.73
Co	-5.76	Ca	-2.06
Ni	-5.52	Sr	-1.76
Cu	-3.97	Ba	-2.01
Zn	-1.42	B	-6.49
Y	-4.63	Al	-3.66
Zr	-7.03	Ga	-2.85
Nb	-7.57	In	-2.53
Mo	-8.06	Tl	-2.24
Ru	-8.55	Si	-4.71
Rh	-6.63	Ge	-3.86
Pd	-4.35	Sn	-3.30
Ag	-2.99	Pb	-3.11
Cd	-1.04	As	-3.24
Hf	-7.00	Sb	-2.99
Ta	-8.96	Bi	-2.86
W	-9.45	Se	-2.87
Re	-8.38	Te	-2.70

Table S4. R^2 and MAE for each ML algorithm with optimal hyperparameters.

ML algorithm	CV MAE (eV)	CV R^2	Train MAE (eV)	Train R^2	Test MAE (eV)	Test R^2	Tuned hyperparameter value
Linear regression	0.58	0.53	0.56	0.58	0.54	0.57	Non-parametric
Ridge regression	0.57	0.53	0.55	0.57	0.54	0.57	alpha=1.0
Lasso	0.58	0.53	0.56	0.58	0.54	0.57	alpha=0.0001
ElasticNet	0.58	0.53	0.56	0.58	0.54	0.57	alpha=0.0001326, l1_ratio=0.4, max_iter=5000
Kneighbors	0.38	0.74	0	1	0.32	0.83	algorithm='auto', n_neighbors=5, p=1, weights='distance'
Decision Tree	0.48	0.57	0	1	0.41	0.73	max_depth=None, min_samples_split=2
Support vector	0.22	0.88	0.08	0.95	0.20	0.91	C=10.0, epsilon=0.01, gamma='auto', kernel='rbf'
RandomForest	0.33	0.80	0.12	0.97	0.26	0.88	n_estimators=100, max_depth=none, min_samples_split=2
GradiebtBoosting	0.27	0.85	0.0024	1	0.23	0.91	learning_rate=0.1, max_depth=7, max_features='sqrt', min_samples_leaf=1, min_samples_split=5, n_estimators=300
ExtraTree	0.27	0.85	0	1	0.20	0.93	n_estimators=100, max_depth=None, min_samples_split=2
Adaboost	0.474	0.67	0.42	0.76	0.45	0.73	learning_rate=1.0, loss='square', n_estimators=200
XGBoost	0.27	0.85	0.02	0.99	0.21	0.92	colsample_bytree=0.8, gamma=0, learning_rate=0.1, max_depth=5, n_estimators=300, reg_alpha=0, reg_lambda=1, subsample=0.8

Table S5. Performance of the XGBoost Model on Random and Extrapolation Test Sets.

	Test_R^2	Test_RMSE	Test_MAE
Random test set	0.93	0.28	0.21
Extrapolation test set	0.87	0.32	0.26

Table S6. The relationship between Δχ and E_f with the similar Δr.

Number	SAAs	E_f (eV)	Δχ	Δr (pm)
1	Rh_1V	-1.66	-0.65	0
2	Ru_1V	-1.16	-0.57	0
3	Pd_1Re	-0.77	-0.3	0
4	Nb_1Ta	-0.27	-0.1	0
5	Rh_1Ru	-0.17	-0.08	0
6	V_1Zn	1.25	0.02	0
7	Cr_1Cu	0.53	0.24	0
8	Re_1Pd	0.14	0.3	0
9	Ga_1Os	-0.76	0.39	0
10	V_1Ru	-1.19	0.57	0
11	Zn_1Rh	-1.43	0.63	0
12	V_1Rh	-1.48	0.65	0

Table S7. With the similar ΔCEbulk and Δχ, the relationship between Δr and E_f.

Number	SAAs	E_f (eV)	ΔCEbulk (eV)	Δχ	Δr (pm)
1	Bi_1W	-2.09	-9.34	-0.2	-15.7
2	Sn_1W	-1.65	-9.53	-0.26	-12
3	Te_1Re	-1.27	-9.59	-0.2	-5
4	Cu_1W	0.39	-9.35	-0.2	11
5	Sb_1Ta	-1.88	-7.98	-0.55	1
6	Te_1Nb	-1.80	-7.37	-0.5	4
7	Ge_1Ta	-1.42	-7.72	-0.51	18
8	Ga_1Hf	-1.40	-7.24	-0.51	24
9	Be1Mo	-0.22	-7.40	0.59	27.7

Table S8. Features selection for ΔE_{CO} prediction model.

	Name	Symbol	Unit
Target variable	CO adsorption energy	ΔE_{CO}	eV
Element Properties	Atomic number	Z_a、Z_b	
	Atomic radius	r_a、r_b、Δr	pm
	Work function	W_a、W_b、ΔW	eV
	Ionization energy	IE_a、IE_b、ΔIE	MJ • mol^{-1}
	Electronegativity	χ_a、χ_b、Δχ	
	Electron affinity	EA_a、EA_b、ΔEA	eV
	Valence electron number	VE_a、VE_b、ΔVE	
	Period	Period_a、Period_b	
	The d-band filling of the main metal	f	
	The center of the d-band of the main metal	ε_d	eV
	Mass	M_a、M_b、ΔM	
	Density	ρ_a、ρ_b、Δρ	g • cm^{-3}
	Specific heat capacity	C_p_a、C_p_b、ΔC_p	J • (Kg • K)$^{-1}$
	Cohesive energy	CEbulk_a、CEbulk_b、ΔCEbulk	eV
	Surface energy of the host metal	E_{surf}	eV
Structural characteristics	Coordination number of a single atom	CN	
	Miller indices	MI	
	The bond length between C and a single atom	l_{C-N}	Å
	C-O bond length	l_{C-O}	Å
	Lattice constant	a_a、a_b、b_a、b_b、c_a、c_b、Δa、Δb、Δc	Å

'a' denotes the host metal and 'b' denotes the dopant atom.

Table S9. Optimized Model Hyperparameters and Performance on Cross-Validation, Training and Test Sets.

ML algorithm	CV MAE (eV)	CV R^2	Train MAE (eV)	Train R^2	Test MAE (eV)	Test R^2	Hyperparameter value
Linear regression	0.32	-7.52	0.20	0.85	0.26	0.73	Non-parametric
Ridge regression	0.33	-0.85	0.28	0.74	0.29	0.74	alpha=100
Lasso	0.45	0.32	0.44	0.44	0.45	0.43	alpha=0.27, max_iter=1000, selection='random'
ElasticNet	0.48	0.31	0.47	0.35	0.50	0.34	alpha=0.36, l1_ratio=0.9
Kneighbors	0.29	0.69	0.23	0.80	0.30	0.68	algorithm='auto', n_neighbors=5
Decision Tree	0.14	0.88	0.041	0.99	0.18	0.77	max_depth=9, min_samples_split=5
Support vector	0.12	0.88	0.050	0.95	0.088	0.98	C=1000, epsilon=0.01, gamma=0.001, kernel='rbf'
RandomForest	0.12	0.90	0.044	0.99	0.088	0.97	max_features=0.5, n_estimators=300
Gradiebt Boosting	0.14	0.90	0.06	0.98	0.11	0.96	learning_rate=0.05, max_depth=3, max_features='sqrt', n_estimators=300, min_samples_leaf=5, min_samples_split=10
ExtraTree	0.12	0.92	0.03	0.99	0.09	0.97	n_estimators=100, min_samples_split=2, min_samples_leaf=2, min_samples_split=3
Adaboost	0.18	0.87	0.16	0.92	0.20	0.86	learning_rate=1.0, loss='square', n_estimators=200
XGBoost	0.13	0.90	0.04	0.99	0.09	0.97	colsample_bytree=0.9, gamma=0, learning_rate=0.05, max_depth=7, n_estimators=300, reg_alpha=1.0, reg_lambda=2.0, subsample=0.8

Table S10. Ten-fold cross validation results of RFR model.

Fold	R^2	RMSE (eV)	MAE (eV)
1	0.94	0.18	0.10
2	0.94	0.19	0.12
3	0.93	0.19	0.12
4	0.91	0.20	0.12
5	0.97	0.11	0.09
6	0.97	0.12	0.09
7	0.96	0.15	0.09
8	0.93	0.18	0.11
9	0.95	0.18	0.11
10	0.97	0.13	0.10
Mean	0.95	0.16	0.11
Std	0.02	0.03	0.01

Table S11. Performance of the RFR Model on Random and Extrapolation Test Sets.

	Test_R^2	Test_RMSE	Test_MAE
Random test set	0.97	0.12	0.09
Extrapolation test set	0.88	0.08	0.06

Table S12. Relationship between ΔE_{CO} and CEbulk_b for Cu-based SAAs.

Number	SAAs	ΔE_{CO} (eV)	CEbulk_b (eV)	l_{C-O} (Å)	l_{C-N} (Å)
1	Rh_1Cu	-2.18	-6.63	1.17	1.85
2	Pt_1Cu	-1.65	-6.30	1.16	1.87
3	Ti_1Cu	-1.40	-6.09	1.16	2.10
4	Sc_1Cu	-0.88	-4.53	1.15	2.33
5	Al_1Cu	-0.67	-3.66	1.16	1.99
6	Ag_1Cu	-0.46	-2.99	1.15	2.11
7	Zn_1Cu	-0.41	-1.42	1.15	2.02

Table S13. Threshold sensitivity analysis of ΔE_{CO} with fixed ΔE_H screening ($\Delta E_H \notin$ [-0.47, -0.07] eV).

ΔE_{CO} range (eV)	Candidates	Retention of original 26
[-0.77, -0.57]	26	100%
[-0.72, -0.62]	8	31%
[-0.72, -0.57]	17	65%
[-0.77, -0.62]	17	65%
[-0.75, -0.59]	23	88%
[-0.75, -0.57]	23	88%
[-0.77, -0.59]	22	85%

Table S14. Threshold sensitivity analysis of ΔE_H with fixed ΔE_{CO} screening ($\Delta E_{CO} \in$ [-0.77, -0.57] eV).

ΔE_H excluded range (eV)	Candidates	Retention of original 26
[-0.47, -0.07]	26	100%
[-0.52, -0.02]	26	100%
[-0.57, 0.03]	26	100%
[-0.62, 0.08]	26	100%
[-0.67, 0.13]	26	100%
[-0.72, 0.18]	26	100%
[-0.77, 0.23]	26	100%

Table S15. Gibbs free energy changes for each elementary step of CO_2RR toward CH_4 on Cu (111).

Reaction mechanism	ΔG (eV)
$CO_2+*\rightarrow CO_2*$	0.31
$CO_2*+H^++e^-\rightarrow *COOH$	0.46
$*COOH+H^++e^-\rightarrow *CO+H_2O$	-0.48
$*CO+H^++e^-\rightarrow *CHO$	0.67
$*CHO+H^++e^-\rightarrow *OCH_2$	-0.11
$*OCH_2+H^++e^-\rightarrow *OCH_3$	-0.79
$*OCH_3+H^++e^-\rightarrow *O+CH_4$	-0.49
$*O+H^++e^-\rightarrow *OH$	-0.52
$*OH+H^++e^-\rightarrow *H_2O$	-0.04

Table S16. Gibbs free energy changes for each elementary step of CO_2RR toward CH_3OH on Cu (111).

Reaction mechanism	ΔG (eV)
$CO_2+*\rightarrow CO_2*$	0.31
$CO_2*+H^++e^-\rightarrow *COOH$	0.46
$*COOH+H^++e^-\rightarrow *CO+H_2O$	-0.48
$*CO+H^++e^-\rightarrow *CHO$	0.67
$*CHO+H^++e^-\rightarrow *OCH_2$	-0.11
$*OCH_2+H^++e^-\rightarrow *OCH_3$	-0.79
$*OCH_3+H^++e^-\rightarrow *OHCH_3$	0.05
$*OHCH_3\rightarrow CH_3OH+*$	0.05

Table S17. Gibbs free energy changes for each elementary step of CO_2RR toward CH_4 on Al_1/Cu (111).

Reaction mechanism	ΔG (eV) (top Al)	ΔG (eV) (top Cu)
$CO_2+*\rightarrow CO_2*$	0.43	
$CO_2*+H^++e^-\rightarrow *COOH$	0.03	
$*COOH+H^++e^-\rightarrow *CO+H_2O$	0.14	-0.16
$*CO+H^++e^-\rightarrow *CHO$	0.08	0.39
$*CHO+H^++e^-\rightarrow *OCH_2$	-0.32	
$*OCH_2+H^++e^-\rightarrow *OCH_3$	-0.76	
$*OCH_3+H^++e^-\rightarrow *O+CH_4$	-0.78	
$*O+H^++e^-\rightarrow *OH$	-0.21	
$*OH+H^++e^-\rightarrow *H_2O$	0.28	
$*H_2O\rightarrow H_2O+*$	0.10	

Table S18. Gibbs free energy changes for each elementary step of CO_2RR toward CH_3OH on Al_1/Cu (111).

Reaction mechanism	ΔG (eV) (top Al)	ΔG (eV) (top Cu)
$CO_2+*\rightarrow CO_2*$	0.43	
$CO_2*+H^++e^-\rightarrow *COOH$	0.03	
$*COOH+H^++e^-\rightarrow *CO+H_2O$	0.14	-0.16
$*CO+H^++e^-\rightarrow *CHO$	0.08	0.39
$*CHO+H^++e^-\rightarrow *OCH_2$	-0.32	
$*OCH_2+H^++e^-\rightarrow *OCH_3$	-0.76	
$*OCH_3+H^++e^-\rightarrow *OHCH_3$	0.35	
$*OHCH_3\rightarrow CH_3OH+*$	0.21	

Table S19. The Gibbs free energy changes for each elementary step of the CO_2RR towards CH_4 on the Cu (111) and Au_1/Pd (111).

Reaction mechanism	ΔG (eV)	
	Cu (111)	Au_1/Pd (111)
$CO_2+*\rightarrow CO_2*$	0.31	0.31
$CO_2*+H^++e^-\rightarrow *COOH$	0.46	-0.08
$*COOH+H^++e^-\rightarrow *CO+H_2O$	-0.48	0.38
$*CO+H^++e^-\rightarrow *CHO$	0.67	-0.35
$*CHO+H^++e^-\rightarrow *OCH_2$	-0.11	0.27
$*OCH_2+H^++e^-\rightarrow *OCH_3$	-0.79	0.13
$*OCH_3+H^++e^-\rightarrow *O+CH_4$	-0.49	-0.56
$*O+H^++e^-\rightarrow *OH$	-0.52	-0.38
$*OH+H^++e^-\rightarrow *H_2O$	-0.04	-0.72

Table S20. The Gibbs free energy changes for each elementary step of the CO_2RR towards CH_3OH on the Cu (111) and Au_1/Pd (111).

Reaction mechanism	ΔG (eV)	
	Cu (111)	Au_1/Pd (111)
$CO_2+*\rightarrow CO_2*$	0.31	0.31
$CO_2*+H^++e^-\rightarrow *COOH$	0.46	-0.08
$*COOH+H^++e^-\rightarrow *CO+H_2O$	-0.48	0.38
$*CO+H^++e^-\rightarrow *CHO$	0.67	-0.35
$*CHO+H^++e^-\rightarrow *OCH_2$	-0.11	0.27
$*OCH_2+H^++e^-\rightarrow *OCH_3$	-0.79	0.13
$*OCH_3+H^++e^-\rightarrow *OHCH_3$	0.05	-0.52

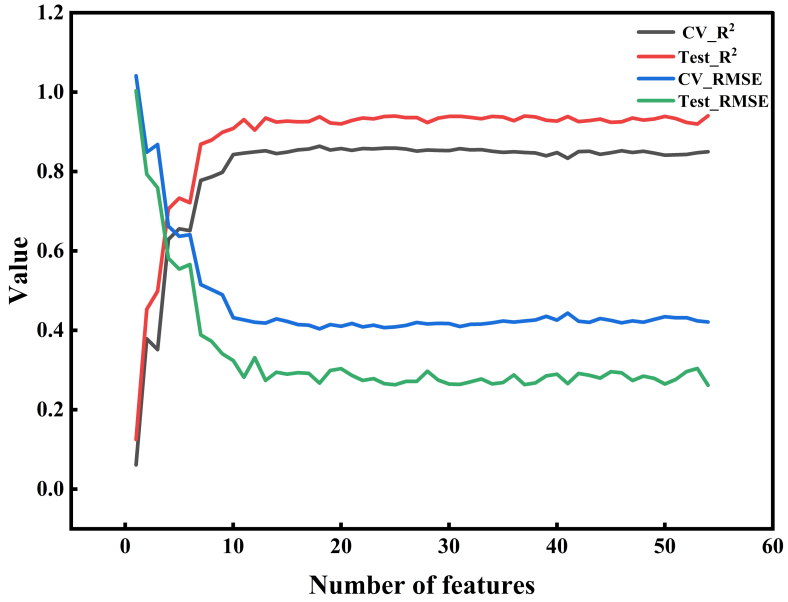

Figure S1. RFE learning curve for XGBoost model.

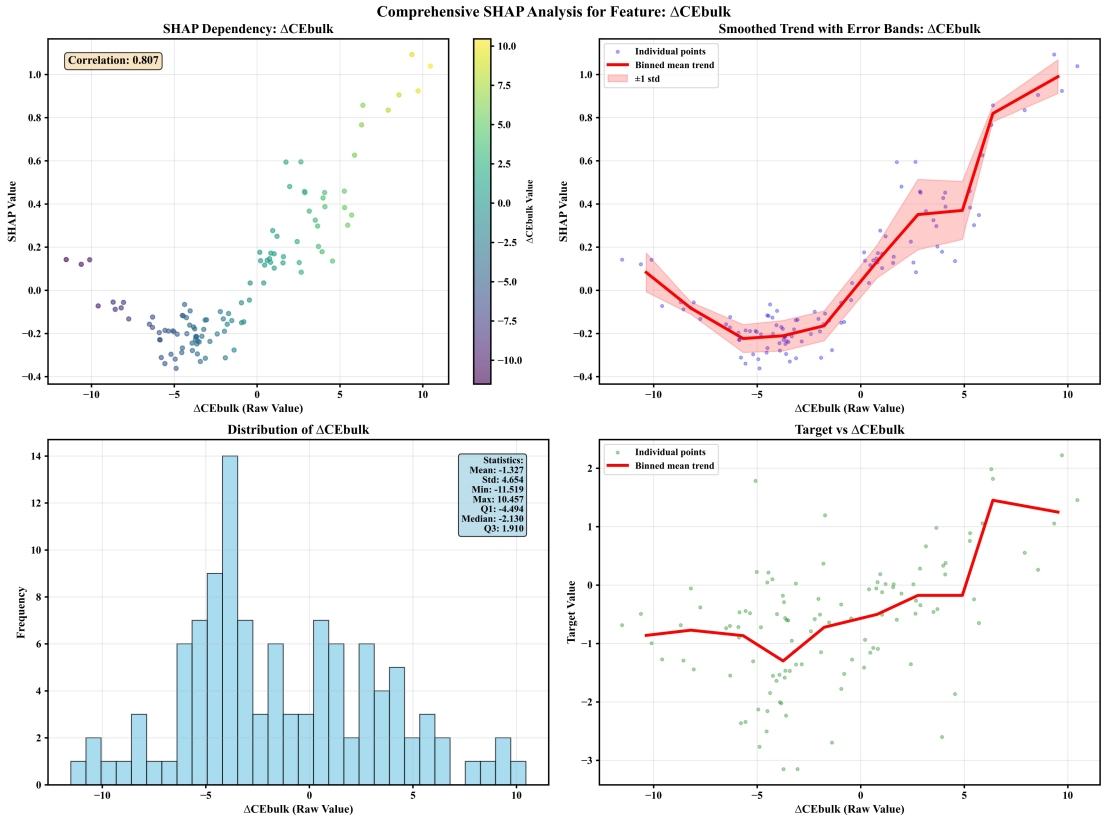

Figure S2. Comprehensive SHAP analysis for feature ΔCEbulk.

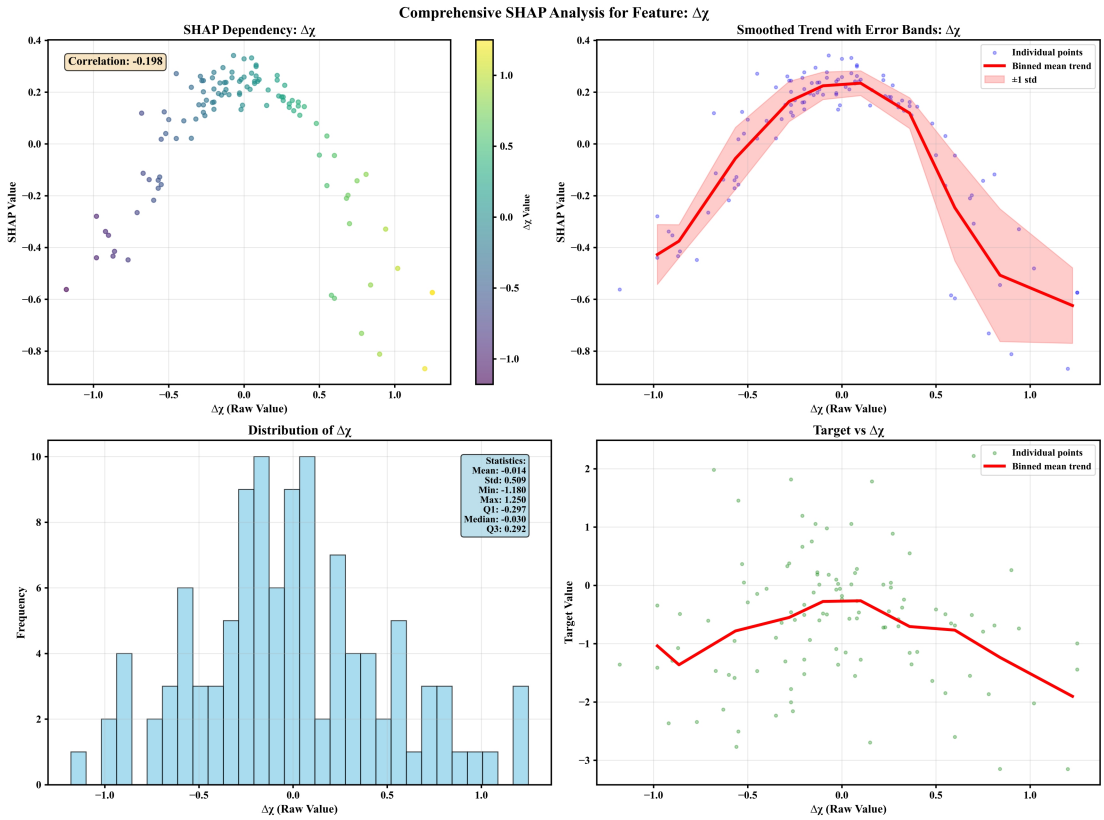

Figure S3. Comprehensive SHAP analysis for feature Δχ.

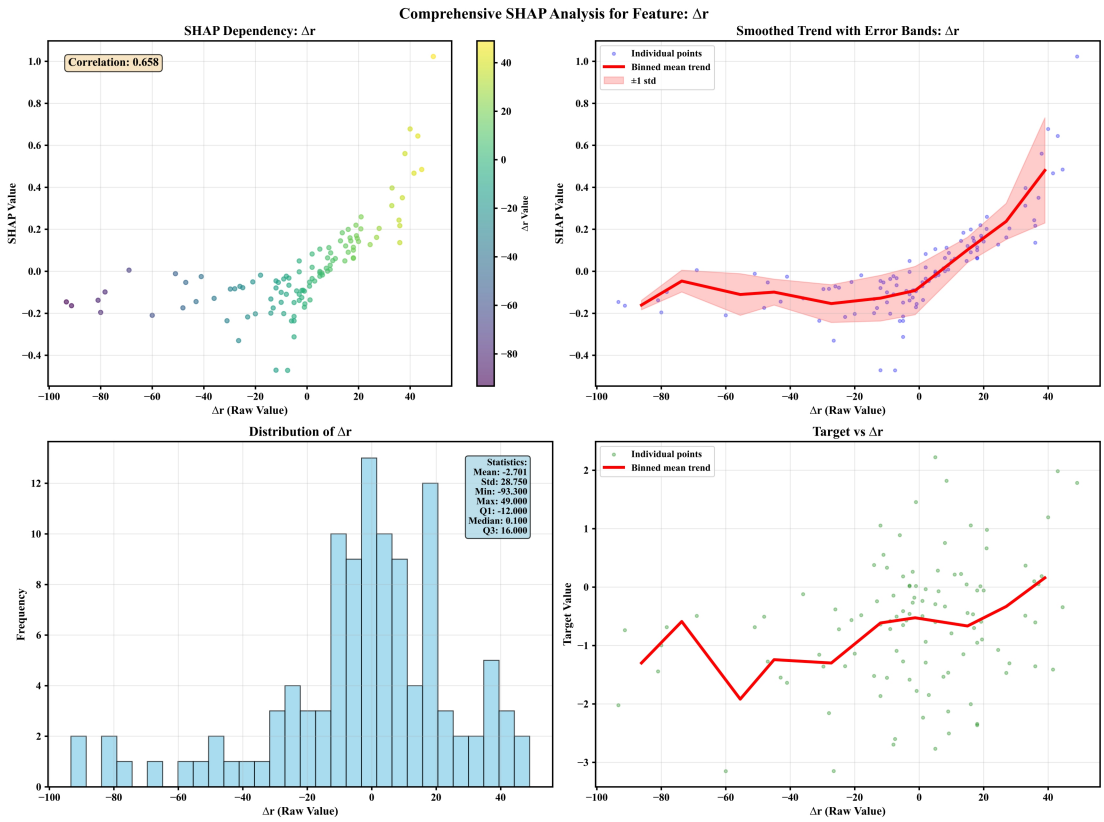

Figure S4. Comprehensive SHAP analysis for feature Δr.

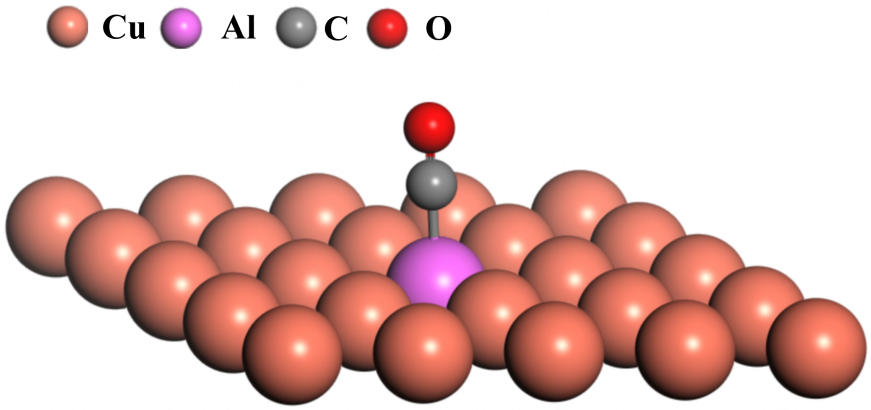

Figure S5. Schematic diagram of CO adsorption state.

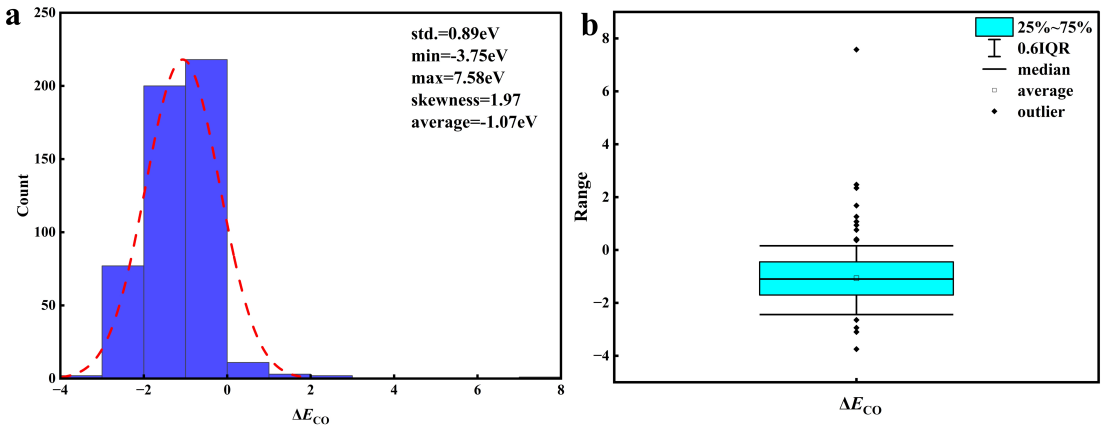

Figure S6. (a) Histogram with the standard deviation of 1.28eV and the average value of -0.64eV. (b) Box plot analysis with the upper bound of 2.27 eV and the lower bound of -3.70 eV.

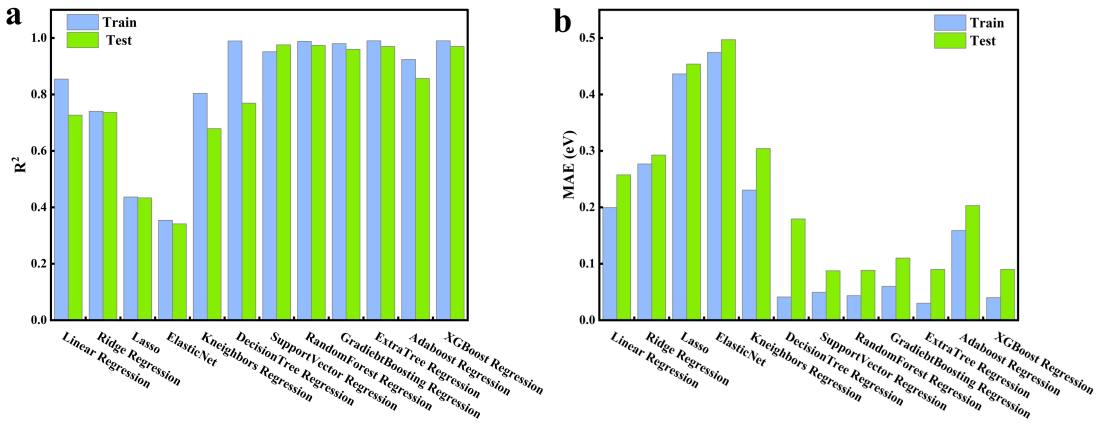

Figure S7. (a) R^2 and (b) MAE (eV) of different machine learning algorithms.

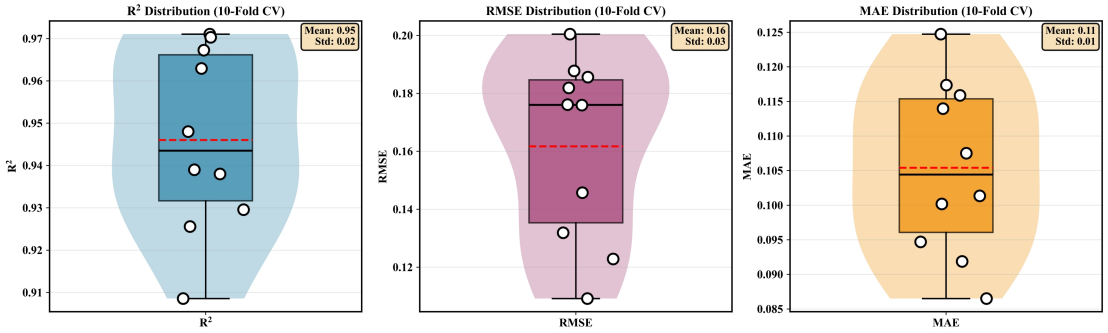

Figure S8. Distribution diagram of RFR model 10-fold cross validation.

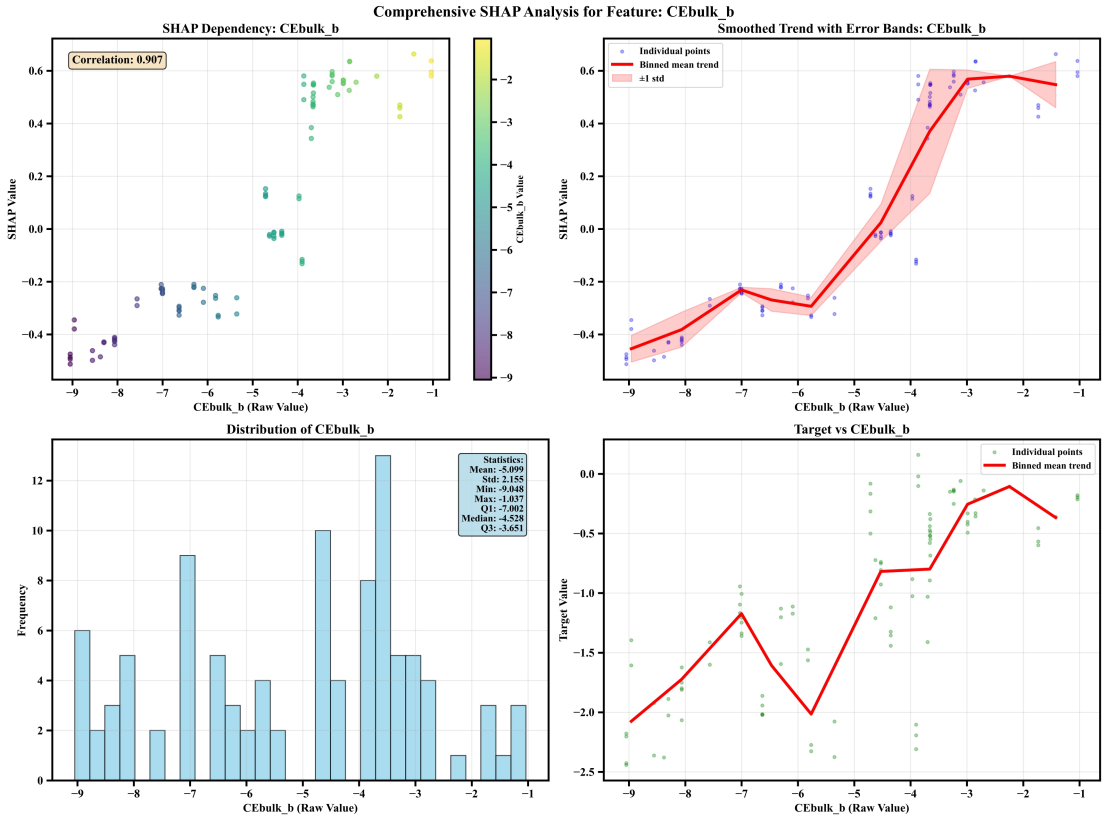

Figure S9. Partial Dependence Plots (PDPs) analysis of CEbulk_b.

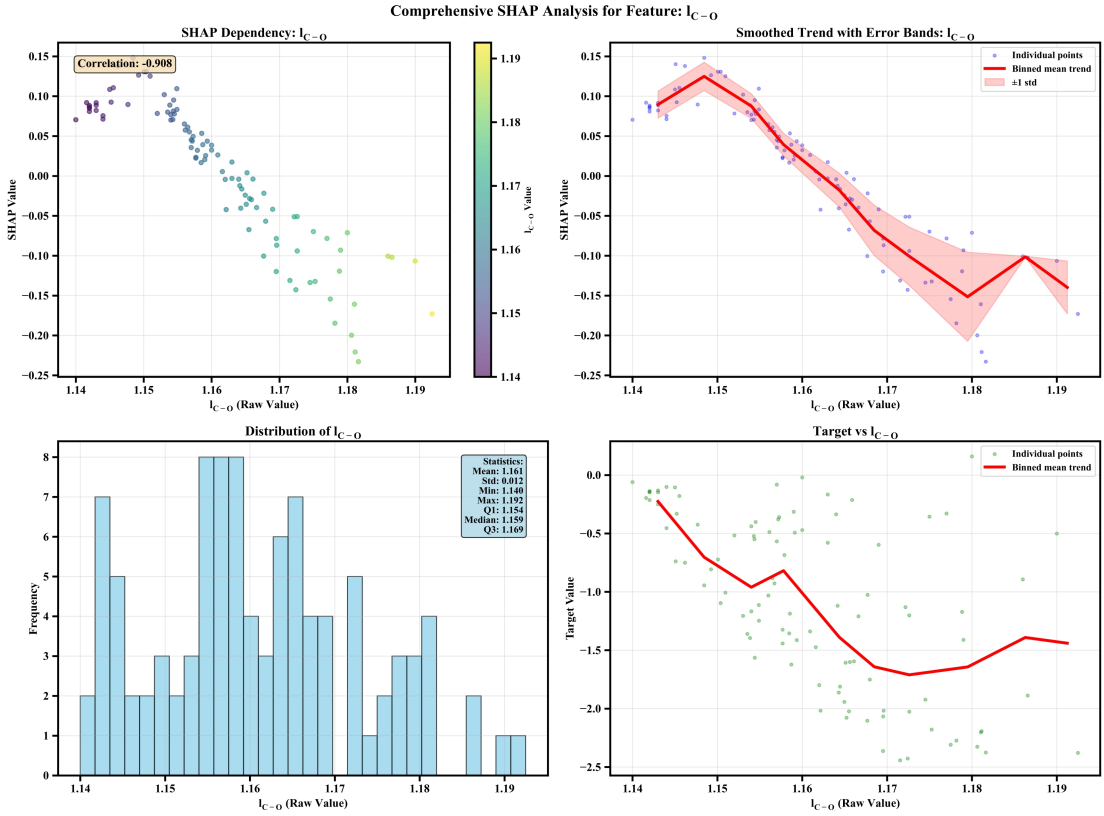

Figure S10. Partial Dependence Plots (PDPs) analysis of l_{C-O}.

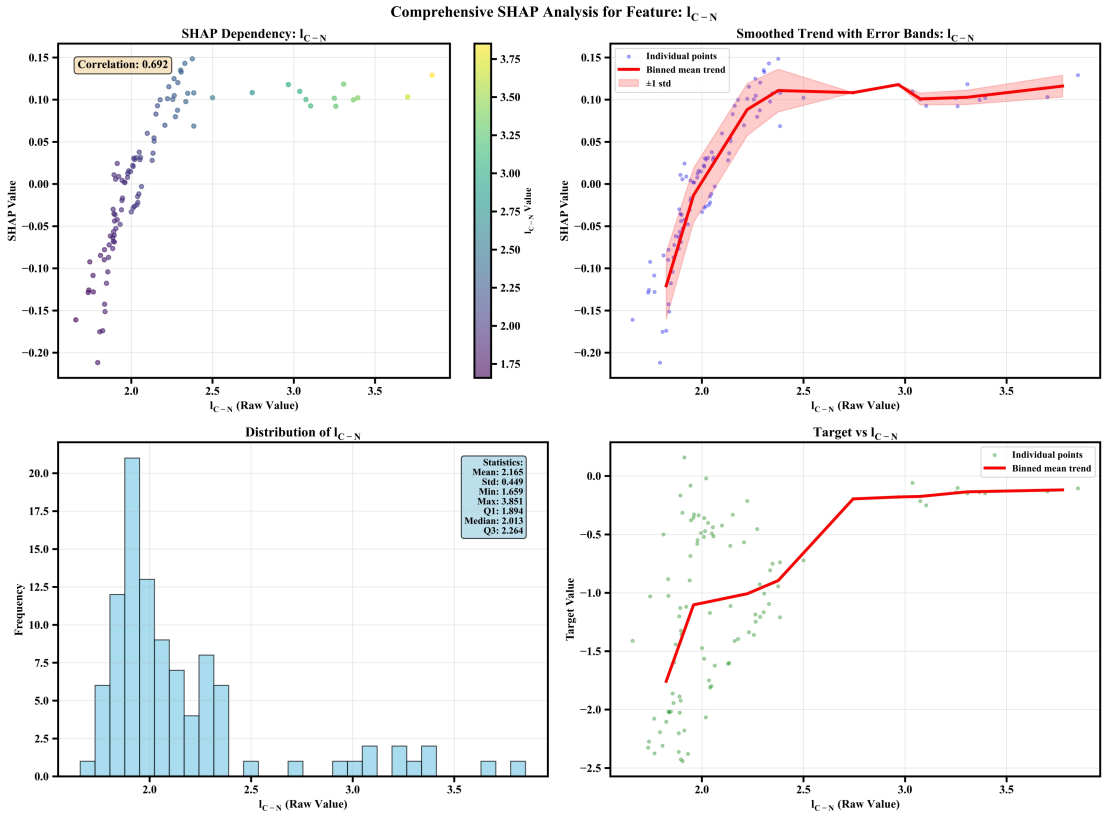

Figure S11. Partial Dependence Plots (PDPs) analysis of $l_{C\text{-}N}$.

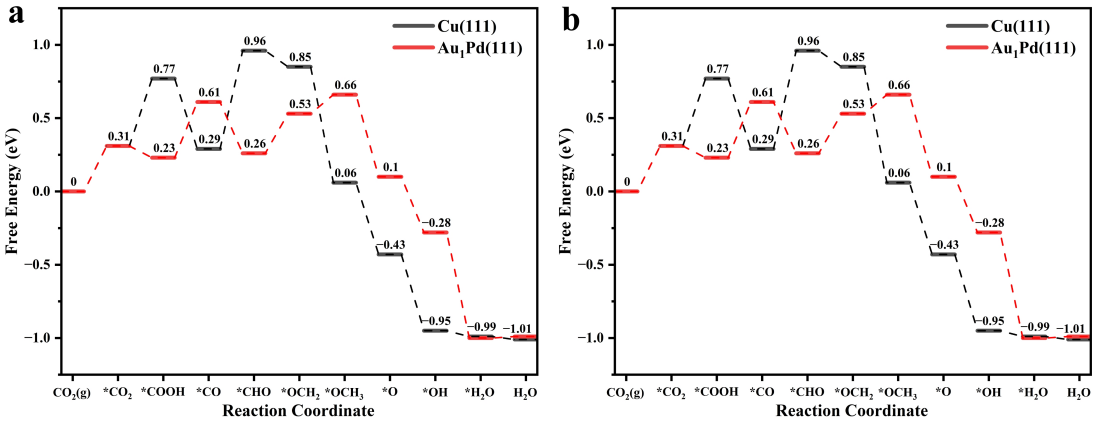

Figure S12. Gibbs Free Energy Diagram for CO_2RR to CH_4 (a), and to CH_3OH (b) on Cu (111) and Au_1/Pd (111).