Supporting Information

Data-Guided Design of Double-Atom Catalysts for Enhanced Electrocatalytic Performance

Chenyang Wei,[a] Wenbo Mu,*[,b] Hongyuan Zhang,[a] Zhenghui Liu*[,c] and Tiancheng Mu*[,a,d]

[a] *School of Chemistry and Life Resources, Renmin University of China, Beijing 100872, P.R. China. E-mail: tcmu@ruc.edu.cn (TM)*

[b] *Department of Computer Science and Engineering, University of California San Diego, La Jolla, CA, 92093-0404, USA. Email: wmu@ucsd.edu (WM)*

[c] *School of Pharmaceutical and Chemical Engineering, Taizhou University, Taizhou 318000, Zhejiang, China. Email: liuzhenghui@iccas.ac.cn (ZL)*

[d] *Key Laboratory of Green Chemical Media and Reactions, Ministry of Education, School of Chemistry and Chemical Engineering, Henan Normal University, Xinxiang, Henan 453007, P. R. China.*

**Corresponding author Email: tcmu@ruc.edu.cn (TM)*

Computational details

The binding energy of adsorbate ad^*, denoted as ΔE_{ad^*} was determined by the equation:

$$\Delta E_{ad^*} = E_{DACs^*} - E_{DACs} - E_{ad}$$

In this equation, E_{DACs^*} represents the total energy of DACs bound with the adsorbate ad^*, while E_{DACs} and E_{ad} correspond to the individual energies of the DACs and ad, respectively.

The free energy change for each reaction step was computed utilizing the computational hydrogen electrode (CHE) model, described as:

$$\Delta G = \Delta E + \Delta ZPE - T\Delta S + \Delta G_U + \Delta G_{pH}$$

where ΔE is the total energy variation derived from DFT calculations, ΔZPE indicates the zero-point energy difference, and ΔS accounts for the entropy change. The term ΔG_{pH} is evaluated using the formula $\Delta G_{pH} = -k_B T ln[H^+] = 0.0592pH$, with k_B representing the Boltzmann constant and T the absolute temperature. Under standard reactions (electrode potential $U = 0\ eV$, temperature $T = 298.15\ K$, pressure $P = 1\ bar$ and $pH = 0$), the free energy change simplifies to:

$$\Delta G = \Delta E + \Delta ZPE - T\Delta S$$

Consequently, the Gibbs free energy change for the adsorption of the adsorbate ad^*, ΔG_{ad^*} can be elucidated as:

$$\Delta G_{ad^*} = \Delta E_{ad^*} + \Delta ZPE - T\Delta S$$

In the present study, the CO_2RR proceeds through the steps outlined below:

(1) $Slab + CO_2(g) + 2H^+ + e^- \rightarrow COOH^* + H^+$ ΔG_1

(2) $COOH^* + H^+ + e^- \rightarrow CO^* + H_2O$ ΔG_2

(3) $CO^* + H_2O \rightarrow Slab + CO(g) + H_2O$ ΔG_3

Correspondingly, the limiting potential (U_L) is given by:

$$U_L = -\frac{\Delta G_{max}}{e} = -max(\Delta G_1, \Delta G_2, \Delta G_3)/e$$

Moreover, the OER is characterized by the following sequential reactions:

(1) $Slab + 2H_2O \rightarrow OH^* + H_2O + H^+ + e^-$ ΔG_1

(2) $OH^* + H_2O + H^+ + e^- \rightarrow O^* + H_2O + 2H^+ + 2e^-$ ΔG_2

(3) $O^* + H_2O + 2H^+ + 2e^- \rightarrow OOH^* + 3H^+ + 3e^-$ ΔG_3

(4) $OOH^* + 3H^+ + 3e^- \rightarrow Slab + O_2(g) + 4H^+ + 4e^-$ ΔG_4

From this, the overpotential (η^{OER}) can be deduced by calculating the greatest free energy

change across these steps, expressed as:

$$\eta^{OER} = \left[\max\left(\Delta G_1, \Delta G_2, \Delta G_3, \Delta G_4\right) / e\right] - 1.23\ eV$$

while HER proceeds through:

(1) $2H^+ + 2e^- \rightarrow H^+ + e^- + H^*$

(2) $H^+ + e^- + H^* \rightarrow H_2$

To discern the thermodynamic and electrochemical stability of DACs, definitions for the binding energy (ΔE_{bind}) and dissolution potential (U_{diss}) are as follows:

$$\Delta E_{bind} = E_{DACs} - E_{Substrate} - n_{M1}E_{M1} - n_{M2}E_{M2}$$

$$U_{diss} = U_{diss}^{0}(bulk) - \frac{E_{form}}{ne}$$

Here, E_{DACs} symbolizes the total energy of DACs, whereas $E_{Substrate}$ indicates the energy of substrate. The terms n_{M1} and n_{M2} signify the counts of metal 1 and metal 2 within DACs respectively, with $n_{M1} = n_{M2} = 1$ in the case of DACs. The energies of single metal 1 (E_{M1}) and metal 2 (E_{M2}) atom within their bulk configurations are designated as such. $U_{diss}^{0}(bulk)$ represents the standard dissolution potential of the bulk metal, while n is the number of electrons partaking in the dissolution process. Lastly, the formation energy (E_{form}) of DACs can be computed by dividing the sum difference of the DACs energies, substrate energy, and the energies of the bulk metals by the total count of metals, which is also half of the binding energy:

$$E_{form} = \frac{E_{DACs} - E_{Substrate} - n_{M1}E_{M1} - n_{M2}E_{M2}}{n_{M1} + n_{M2}} = \frac{\Delta E_{bind}}{2}$$

Table S1 Summary of feature names and corresponding abbreviations in machine learning (ML) models applying for data analysis

Machine learning (ML) details

All ML algorithms are implemented using Python 3 in the Jupyter Notebook environment. For model training and evaluation, we use Optuna for hyperparameter optimization, XGBoost for regression modeling, and PyTorch for training the diffusion model. The data preprocessing steps are performed using pandas and scikit-learn libraries. The dataset is divided into training and test sets using a 7:3 split. Optuna are applied to optimize the regression models. For each trial, XGBoost Regressor is trained with randomly optimized hyperparameters, including learning rate, subsample rate, and the number of estimators etc.

The diffusion model is trained in parallel with the regression model. The model architecture is a simple fully connected neural network with ReLU activations, and it is trained using Adam optimizer to minimize the loss function. Generated samples from the diffusion model are used to augment the training set, with the final model being evaluated on both the original and augmented datasets.

Cross-validation is employed during hyperparameter optimization, and model performance is assessed using R^2 on the test set. Additionally, t-SNE is utilized for visualizing the low-dimensional projection of the original and generated data to visualized the similarity of both type of data. Moreover, three different quantifying methods are applied for the evaluation of the generated data as follow:

MDD (Maximum Mean Discrepancy)

MMD measures the difference between the distributions of two datasets. It is defined as:

$$MMD^2 = \| \mu_X - \mu_Y \|^2$$

where μ_X and μ_Y are the mean embeddings of the two datasets in a reproducing kernel Hilbert space. MDD quantifies the global similarity between generated and real data, with a smaller MDD indicating higher distribution alignment.

Average Cosine Similarity (ACS)

ACS measures the directionality between two sets of vectors, defined as:

$$ACS = \frac{1}{N}\sum_{i=1}^{N} \frac{A_i . B_i}{\| A_i \| \| B_i \|}$$

where A_i and B_i are the feature vectors of the generated and real data. ACS assesses the directional similarity of the datasets, with higher values indicating closer alignment in feature space.

Nearest-Neighbor Consistency (NNC)

NNC evaluates local similarity by calculating the nearest neighbor distance:

$$NNC = \frac{1}{N}\sum_{i=1}^{N} \min_j \| X_i - Y_j \|$$

where X_i and Y_j are points in the generated and real datasets. Smaller NNC values indicate that the generated data points are closer to real data points, reflecting local consistency.

These metrics are used to quantify the reliability of the generated data, ensuring its global and local consistency with real data. Lower values in MDD, ACS, and NNC indicate that the generated data closely resemble the original data, confirming its reasonableness for use in ML tasks.

Table S1 Summary of feature names and corresponding abbreviations in ML models applying for data analysis

Atomic features	Structure features
Atomic number (N)	**The number of carbons on the substrate (number of C)**
Atomic mass (M)	**The number of nitrogens on the substrate (number of N)**
Atomic radius (r)	**The ratio of carbons to nitrogens on the substrate (C:N)**
Electronegativity (χ)	**The distance between transition metal 1 and transition metal 2 (M1-M2)**
Electron affinity (EA)	**The distance between the central transition metals and the atoms of substrates surrounding them (e.g., M1-N1, N2-6)**
First ionization energy (EI)	**The distance between the central transition metals and the atoms of adsorbates (e.g., H-M1, H-M2)**
The number of d-electrons (θd)	**The distance of the adsorbate's atoms (e.g., C-O, H-O)**
The number of s-electrons (θs)	**-**
The number of outermost electron (Ne)	**-**

Table S2 Summary of details of features using for describing elements of double-atom catalysts (DACs)

Element	N	M	R(Å)	χ	EA (eV)	EI (eV)	θd	θs	Ne
Sc	21	44.96	1.64	1.36	0.19	6.56	1	2	3
Ti	22	47.87	1.47	1.54	0.09	6.83	2	2	4
V	23	50.94	1.35	1.63	0.53	6.75	3	2	5
Cr	24	52	1.25	1.66	0.68	6.77	5	1	6
Mn	25	54.94	1.37	1.55	0.97	7.43	5	2	7
Fe	26	55.85	1.26	1.83	0.15	7.90	6	2	8
Co	27	58.93	1.25	1.88	0.66	7.88	7	2	9
Ni	28	58.69	1.25	1.91	1.16	7.64	8	2	10
Cu	29	63.55	1.28	1.9	1.24	7.73	10	1	11
Zn	30	65.39	1.37	1.65	0.09	9.39	10	2	12
Y	39	88.91	1.82	1.22	0.31	6.22	1	2	3
Zr	40	91.22	1.6	1.33	0.43	6.63	2	2	4
Nb	41	92.91	1.43	1.6	0.89	6.76	4	1	5
Mo	42	95.96	1.4	2.16	0.75	7.09	5	1	6
Ru	44	101.07	1.34	2.2	1.05	7.36	7	1	8
Rh	45	102.91	1.34	2.28	1.14	7.46	8	1	9
Pd	46	106.42	1.37	2.2	0.54	8.34	10	0	10
Ag	47	107.87	1.44	1.93	1.30	7.58	10	1	11
Cd	48	112.41	1.49	1.69	0.27	8.99	10	2	12
Hf	72	178.49	1.56	1.3	0.63	6.83	2	2	4
Ta	73	180.95	1.43	1.5	0.32	7.55	3	2	5
W	74	183.85	1.37	2.36	0.82	7.86	4	2	6
Re	75	186.21	1.37	1.9	0.38	7.89	5	2	7
Os	76	190.23	1.35	2.2	1.08	8.7	6	2	8
Ir	77	192.22	1.36	2.2	1.56	9.1	7	2	9
Pt	78	195.08	1.39	2.28	2.13	9	9	1	10
Au	79	196.97	1.44	2.54	2.31	9.23	10	1	11

Table S3 The details (performance and hyperparameters) of each ML model for data analysis. All models were based on Random Forests algorithm, while all hyperparameters were obtained by 10-fold cross validation (CV) through Grid Search method. Hyperparameters not mentioned were kept at their default values

ML models	Hyperparameters	Accuracy	Grid Search range
ΔE_{bind}	**n_estimators = 25** **max_depth = 5** **max_features = 0.7**	**R^2: 0.92**	
ΔG_{CO^*}	**n_estimators = 5** **max_depth = 3** **max_features = 1**	**R^2: 0.83**	**n_estimators = [5,10,15, 20,25, 50]** **max_depth = [3,5,7]** **max_features = [0.6, 0.7, 1]**
ΔG_{OH^*}	**n_estimators = 5** **max_depth = 7** **max_features = 1**	**R^2: 0.80**	
ΔG_{H^*}	**n_estimators = 10** **max_depth = 7** **max_features = 1**	**R^2: 0.88**	

Table S4 Bader charge analysis (e) of key atoms (transition metal 1 and 2, adsorbates) on DACs discussed in this manuscript

Adsorbate	Bader charge transfer e											
CO		**FeSc**		**RhCu**		**VZr**		**PtNi**		**PtPt**		**WMo**
		CN		**CN**		**CN**		**C_2N**		**C_2N**		**C_2N**
	Fe	-0.542	**Rh**	-1.11	**V**	-1.23	**Pt**	-0.498	**Pt**	-0.864	**W**	-1.20
	Sc	-1.37	**Cu**	-0.225	**Zr**	-1.45	**Ni**	-0.647	**Pt**	-0.918	**Mo**	-1.05
	C	-0.533	**C**	-0.676	**C**	-0.174	**C**	-0.644	**C**	-0.371	**C**	-0.330
	O	+0.967	**O**	+0.950	**O**	+0.978	**O**	+0.899	**O**	+0.712	**O**	+0.684
		NbZr		**NiSc**		**NiW**		**PtMn**		**RhZn**		**VTi**
		g-C_3N_4		**g-C_3N_4**		**g-C_3N_4**		**N-C_3N_4**		**N-C_3N_4**		**N-C_3N_4**
	Nb	-1.12	**Ni**	-0.393	**Ni**	-0.229	**Pt**	-0.195	**Rh**	-0.296	**V**	-1.01
	Zr	-2.04	**Sc**	-1.58	**W**	-1.57	**Mn**	-0.947	**Zn**	-0.705	**Ti**	-1.25
	C	+0.290	**C**	-0.551	**C**	-0.283	**C**	-0.649	**C**	-0.548	**C**	-0.262
	O	+0.987	**O**	+0.857	**O**	+0.833	**O**	+0.866	**O**	+0.750	**O**	+0.925
H		**CoFe**		**CoPt**		**CoW**		**CoFe**		**CoNi**		**CoPt**
		CN		**CN**		**CN**		**C_2N**		**C_2N**		**C_2N**
	Co	-0.788	**Co**	-0.863	**Co**	-0.265	**Co**	-0.801	**Co**	-0.807	**Co**	-0.921
	Fe	-0.997	**Pt**	-0.474	**W**	-1.94	**Fe**	-1.04	**Ni**	-0.753	**Pt**	-0.973
	H	+0.370	**H**	+0.639	**H**	+0.589	**H**	+0.380	**H**	+0.337	**H**	+0.608
		FeMn		**MnMn**		**NiMn**		**NiCo**		**NiFe**		**NiMn**
		g-C_3N_4		**g-C_3N_4**		**g-C_3N_4**		**N-C_3N_4**		**N-C_3N_4**		**N-C_3N_4**
	Fe	-0.852	**Mn**	-0.892	**Ni**	-0.390	**Ni**	-0.618	**Ni**	-0.505	**Ni**	-0.571
	Mn	-1.04	**Mn**	-0.897	**Mn**	-0.997	**Co**	-0.953	**Fe**	-1.05	**Mn**	-1.17
	H	+0.633	**H**	+0.545	**H**	+0.570	**H**	+0.486	**H**	+0.471	**H**	+0.512
OH		**CoCo**		**CoPt**		**NiFe**		**NiMn**		**NiNi**		**WFe**
		N-C_3N_4		**N-C_3N_4**		**N-C_3N_4**		**N-C_3N_4**		**N-C_3N_4**		**N-C_3N_4**
	Co	-0.0859	**Co**	+0.623	**Ni**	+0.00524	**Ni**	+0.532	**Ni**	-0.109	**W**	-1.73
	Co	-0.848 e	**Pt**	-0.258	**Fe**	-1.10	**Mn**	-1.11	**Ni**	-0.559	**Fe**	-0.724
	O	+0.792	**O**	+0.0847	**O**	+0.719	**O**	+0.346	**O**	+1.07	**O**	+1.07
	H	-0.446	**H**	+0.104	**H**	-0.314	**H**	-0.0352	**H**	-0.566	**H**	-0.393

Table S5 Summary of d-bands center (ε_d) of DACs discussed in this manuscript. All ε_d were directly obtained from VASPKIT package (standard edition 1.4.1)[1]

Adsorbate	ε_d (eV) of DACs										
CO		**FeSc CN**		**RhCu CN**		**VZr CN**		**PtNi C_2N**		**PtPt C_2N**		**WMo C_2N**
	Fe	-0.612	**Rh**	-1.64	**V**	0.797	**Pt**	-1.96	**Pt**	-2.83	**W**	0.329
	Sc	1.89	**Cu**	-2.36	**Zr**	1.86	**Ni**	-1.33	**Pt**	-2.80	**Mo**	0.125
	Total	0.563	**Total**	-2.02	**Total**	1.26	**Total**	-1.64	**Total**	-2.81	**Total**	0.221
		NbZr g-C_3N_4		**NiSc g-C_3N_4**		**NiW g-C_3N_4**		**PtMn N-C_3N_4**		**RhZn N-C_3N_4**		**VTi_ N-C_3N_4**
	Nb	0.823	**Ni**	-2.04	**Ni**	-1.31	**Pt**	-3.41	**Rh**	-1.99	**V**	0.807
	Zr	1.53	**Sc**	1.16	**W**	0.140	**Mn**	-0.983	**Zn**	-6.58	**Ti**	0.861
	Total	1.18	**Total**	-0.531	**Total**	-0.675	**Total**	-2.17	**Total**	-4.33	**Total**	0.838
H		**CoFe CN**		**CoPt CN**		**CoW CN**		**CoFe C2N**		**CoNi C_2N**		**CoPt C_2N**
	Co	-0.897	**Co**	-0.688	**Co**	-1.04	**Co**	-1.03	**Co**	-0.791	**Co**	-0.756
	Fe	-1.03	**Pt**	-2.24	**W**	0.384	**Fe**	-1.10	**Ni**	-1.40	**Pt**	-2.35
	Total	-0.964	**Total**	-1.43	**Total**	-0.420	**Total**	-1.07	**Total**	-1.10	**Total**	-1.52
		FeMn g-C_3N_4		**MnMn g-C_3N_4**		**NiMn g-C_3N_4**		**NiCo N-C_3N_4**		**NiFe N-C_3N_4**		**NiMn N-C_3N_4**
	Fe	-0.774	**Mn**	-0.376	**Ni**	-1.40	**Ni**	-1.97	**Ni**	-1.99	**Ni**	-2.06
	Mn	-0.396	**Mn**	-0.477	**Mn**	-0.782	**Co**	-1.10	**Fe**	-1.34	**Mn**	-0.650
	Total	-0.586	**Total**	-0.427	**Total**	-1.09	**Total**	-1.55	**Total**	-1.68	**Total**	-1.37
OH		**CoCo N-C_3N_4**		**CoPt N-C_3N_4**		**NiFe N-C_3N_4**		**NiMn N-C_3N_4**		**NiNi N-C_3N_4**		**WFe N-C_3N_4**
	Co	-1.03	**Co**	-1.52	**Ni**	-2.12	**Ni**	-2.11	**Ni**	-1.67	**W**	0.133
	Co	-1.15	**Pt**	-1.61	**Fe**	-1.38	**Mn**	-0.472	**Ni**	-1.23	**Fe**	-0.797
	Total	-1.09	**Total**	-1.56	**Total**	-1.76	**Total**	-1.31	**Total**	-1.45	**Total**	-0.398

Table S6 Details of ML models for data prediction using Smooth Overlap of Atomic Positions (SOAP) or Coulomb Matrix (CM) as descriptors for DACs. Hyperparameters were optimized using 10-fold CV via the Optuna library

Prediction models	**Accuracy**	**Optuna hyperparameters tuning range**
ΔE_{bind}	**R^2: 0.863 (CM)** **R^2: 0.976 (SOAP)**	
U_{diss}^{MI}	**R^2: 0.898 (CM)** **R^2: 0.993 (SOAP)**	**params = {** **"n_estimators": trial.suggest_int("n_estimators", 100, 500),**
U_{diss}^{M2}	**R^2: 0.851 (CM)** **R^2: 0.987 (SOAP)**	**"max_depth": trial.suggest_int("max_depth", 3, 7),** **"learning_rate": trial.suggest_float("learning_rate", 0.005, 0.2, log=True),**
ΔG_{H^*}	**R^2: 0.796 (CM)** **R^2: 0.943 (SOAP)**	**"subsample": trial.suggest_float("subsample", 0.6, 1.0),** **"colsample_bytree": trial.suggest_float("colsample_bytree", 0.6, 1.0),** **"reg_alpha": trial.suggest_float("reg_alpha", 1e-5, 10.0, log=True),**
U_L	**R^2: 0.770 (CM)** **R^2: 0.928 (SOAP)**	**"reg_lambda": trial.suggest_float("reg_lambda", 1e-5, 10.0, log=True),** **}**
η^{OER}	**R^2: 0.772 (CM)** **R^2: 0.883 (SOAP)**	

Table S7 Model evaluation metrics before and after using the generative model

Model	R^2 after using generative model (test set)	R^2 before using generative model (test set)	MDD	ACS	NNC
ΔG_{H^*}	0.943	0.872	0.380	0.917	0.335
U_L	0.928	0.861	0.300	0.908	0.274
η^{OER}	0.883	0.840	0.258	0.888	0.263

Table S8 Comparison of theoretical and predicted values for catalytic performance of different materials

Property/System	**DFT calculation / literature value**	**ML model predicted value**	**Reference**
ΔG_{H^*} of NiZr_CN in HER	-0.35 eV	-0.34 eV	This work
ΔG_{H^*} of MnMn_g-C_3N_4 in HER	-0.86 eV	-0.83 eV	This work
U_L of FeMo@C_2N in CO_2RR	-1.8 eV[a]	-1.6 eV	10.1039/D4CP00213J
η^{OER} of NiCu@C_2N in OER	0.42 eV	0.45 eV	10.1007/s42823-024-00693-6
η^{OER} of Cu_2@C_2N in OER	0.38 eV	0.40 eV	

[a]Although this work only reports the binding energy of CO ($\Delta G_{ads}(CO)$), we have thoroughly discussed in the main text that when the adsorption strength of CO is excessively high (typically below about 1.05 eV), CO desorption is most likely to become the PDS in CO_2RR. In such cases, $\Delta G_{ads}(CO)$ serves as a reliable descriptor for the U_L in CO_2RR. Therefore, we adopt the reported $\Delta G_{ads}(CO)$ values from the literature as a surrogate for U_L in the corresponding catalytic systems.

Table S9 Summary of the designed scoring system used to evaluate the four substrates in this study[a]

Substrate_name	Count_meet_standard	Proportion_meet_standard	Average_Score
CN	4	0.154	6.08
C_2N	5	0.200	6.15
g-C_3N_4	4	0.160	6.37
N-C_3N_4	10	0.244	6.14

[a]Statistical data are based on ML model predictions for the test and training sets. The stability criterion is defined as $\Delta E_{bind} < 0\ ev$ and $U_{diss} > 0\ eV$. Since the data from each substrate vary, we also calculated the proportion of data that meet the stability standard, shown in the 'Proportion_meet_standard' column, which indicates the likelihood of each substrate satisfying the stability criterion in an electrochemical environment. Additionally, we assigned a score based on the specific values of three energies (ΔE_{bind}, U_{diss} for metals 1 and 2). The score is on a 10-point scale, with more negative ΔE_{bind} and more positive U_{diss} leading to a higher score. However, this scoring method is influenced by the data values and is intended for reference only. Therefore, the 'Proportion_meet_standard' column is the primary criterion for recommendation. From this table, we observe that N-C_3N_4 has both the highest count ('Count_meet_standard') and proportion ('Proportion_meet_standard') of stable configurations, making it the most stable substrate for recommendation.

1. $Proportion\_meet\_standard = Count\_meet\_standard / Total\_prediction\_counts$

   'Count_meet_dandard' represents the number of substrates that meet the stability criteria, while 'Total_prediction_count' represents the total number of predicted substrates.

2. $Average\_Score = \left(\Delta E_{bind} + U_{diss}^{MI} + U_{diss}^{MI}\right) / 3$

   Before substituting the values into the formula for calculation, all three parameters are converted to a 10-point scale.

Table S10 Summary of formulas generated by High-Performance Symbolic Regression in Python and Julia (PySR) along with their performance metrics[a]

Adsorption energy	Accuracy	Formula generated by PySR
ΔG_{H^*}	**R^2: 0.870**	(((0.8742166 ^ ***M2-N4***) ^ (((cos((((***N2*** + (***M2-N4*** - 1.710808)) - sin(***El1***)) * ***M2-N6***) ^ 0.8572206) - ((***M1-N2*** - ((***M1-N1*** -0.23210852) ^ ***M2-N6***)) * ***C:N***)) * ***M1-N2***) + ***C:N***)) - sin(***ϑs2*** * (sin(***M2-N5***) ^ ***M1-N2***))) / sin(***C:N*** / 1.2761757)
	R^2: 0.739	Simple version: sin((sin((***number of C*** +3.701664) + ((***EA1*** - sin(***M2***)) * -0.16440153)) / ***C:N***) * 1.961267) - sin(***ϑs2***)
ΔG_{CO^*}	**R^2: 0.924**	(sin((***C-M2*** - (***EA2*** - 0.21993148)) * (min(min(***M1-N2***, ((***M1-N2*** + ((((***C-M2*** - 0.65825975) - 1.6706436) * ***x2***) * ***EA2***)) / (***M2-N6*** -***ϑs1***)) / ***x2***), sin(***M2-N6***)) + (max(min(0.8547337, ***O-M1*** - 1.6706436) - ***EA2***, sin(sin(***M1***) - ***C-O***)) - ***ϑs2***))) * ***C-O***) / 0.79802406
	R^2: 0.699	Simple version: cos(exp(sin(exp(cos(***El1*** * (***El2*** / 0.91185826)))) ^ (***ϑs2*** + cos(***x1*** ^***ϑd2***))) + ***El2***) / cos(cos((***ϑs2*** / 0.0140918335) * ***El1***))
ΔG_{OH^*}	**R^2: 0.973**	((sin((***Ne2*** - ((***M2-N4*** - ***M1-N3***) - (cos(sin(***N2***) / sin(***H-M2***)) / 0.7027012))) * -0.45550263) - sin(sin(1.5241305 * sin(sin(***M2-N6*** * ***M2-N5***) / (***M2-N5*** ^ cos(sin(***M1-N3***) +***ϑd1***)))) + ***M2-N6***)) + 0.20597531) / (0.9472296 ^ (***M1-N3*** ^ ***EA1***))
	R^2: 0.863	Simple version: cos(***El2*** ^ 1.1826223) * (0.8775042 + sin(***El2*** ^ 1.7333285))

[a]In simple terms, the "simple version" excludes all structural features requiring DFT optimization, relying instead on element data from public databases and other non-DFT features to construct the formulas. By sacrificing some model accuracy, the aim is to highlight PySR's ability to construct highly interpretable formulas and predict adsorption energies using basic features.

Key hyperparameters of PySR:
niterations=200,
maxsize=25 for simple version while maxsize=50 for other models,
populations=200,
loss="loss(x, y) = (x - y)^2"
binary_operators=["+", "-", "*", "/", "^"]
unary_operators=["sin", "cos", "exp"] for other models while binary_operators=["+", "-", "*", "/", "^", "max", "min"] for ΔG_{co*} for better fitting result.

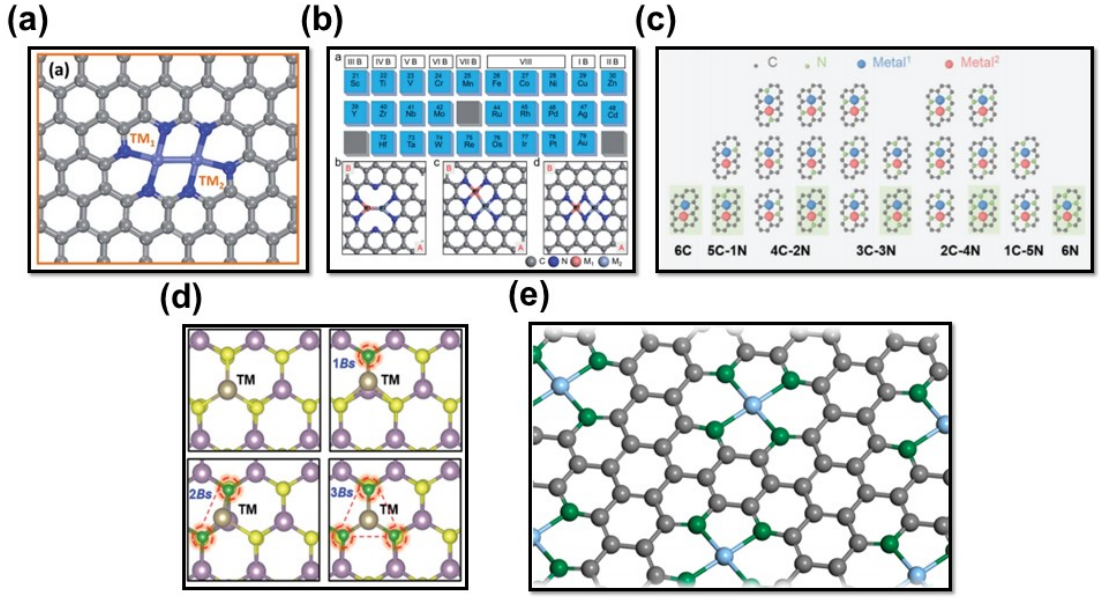

Fig. S1 Current trends in ML-led research of atomic catalysts (ACs): Panels (a) to (c) depict a selection of recent investigations into DACs facilitated by ML. Panel (a) focuses on studies using a single type of substrate, whereas panels (b) and (c) examine DACs on N-doped graphene substrates that undergo minor variations (distinct doping configurations) around the central double transition metals. Panels (d) and (e) illustrate research exploring single atom catalysts (SACs) utilizing ML, specifically examining SACs on MoS_2 substrates with varying B doping levels, as well as graphene doped with different elements (C, N, O, P, or S), respectively. Notably, the substrate structures in all illustrated examples remain largely unchanged except the atoms around the central transition metals. This figure has been adapted from the original illustrations cited as references 2 to 6 (a-e, respectively).

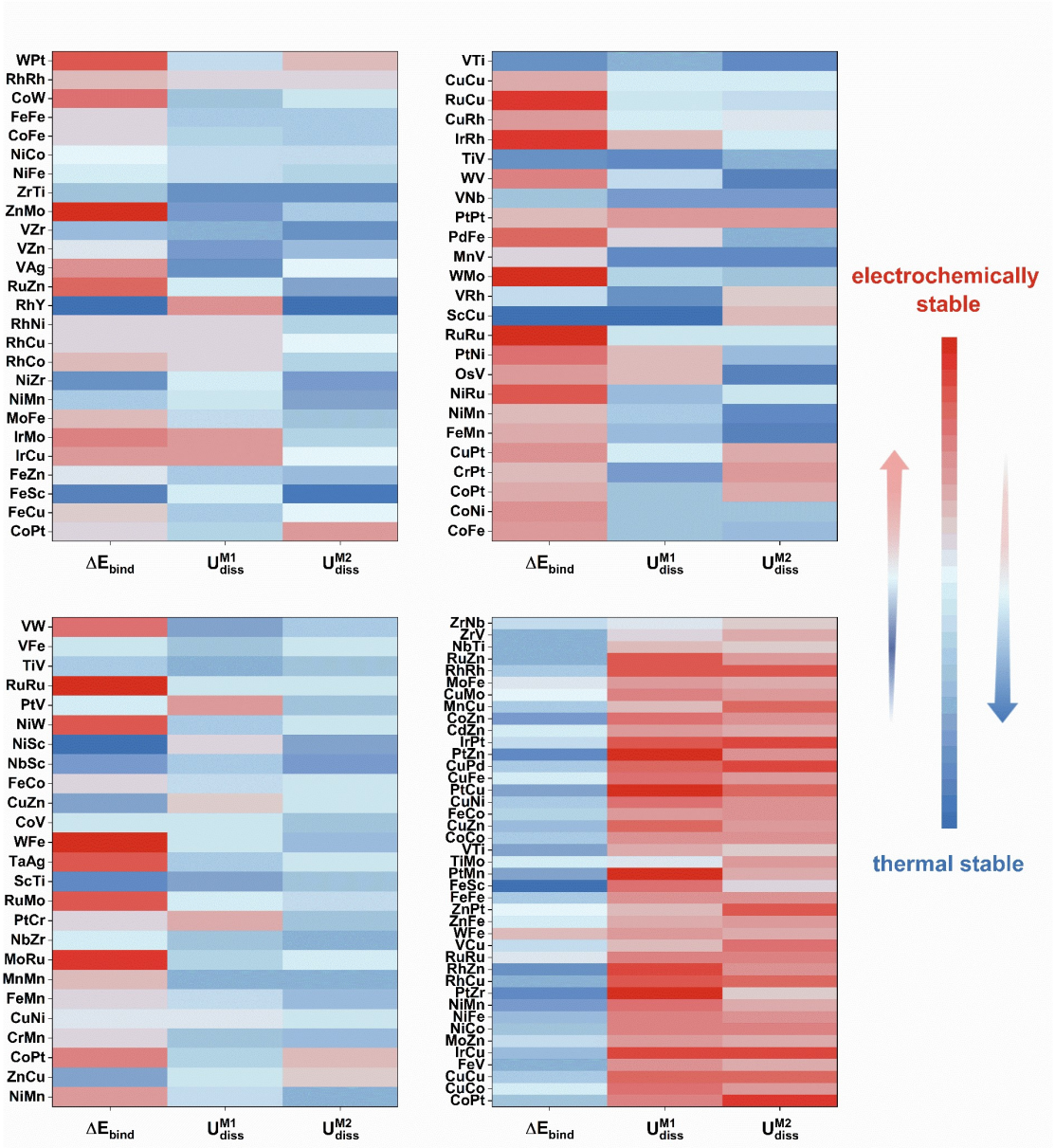

Fig. S2 Heatmaps of binding energy ($^{\Delta E_{bind}}$) and dissolution potential ($^{U_{diss}}$) across diverse DACs: This figure displays four distinct heatmaps representing the calculated $^{\Delta E_{bind}}$ and $^{U_{diss}}$ for DACs on different substrate geometries: CN (top left), C_2N (top right), g-C_3N_4 (bottom left), and N-C_3N_4 (bottom right). The heatmaps utilize a color gradient scale where red denotes positive energy differences and navy blue indicates negative ones. A negative $^{\Delta E_{bind}}$ suggests that the metal bind stably to the substrate, while a positive $^{U_{diss}}$ denotes a metal's resistance to dissolution during electrochemical reactions, both critical factors for the stability and longevity of DAC catalysts. The N-C_3N_4 substrate geometry, in particular, shows remarkable potential for steadfast dual-metal atom anchoring. Furthermore, a majority of the DACs with N-C_3N_4 substrate assessed display a positive $^{U_{diss}}$, reflecting their considerable electrochemical sturdiness.

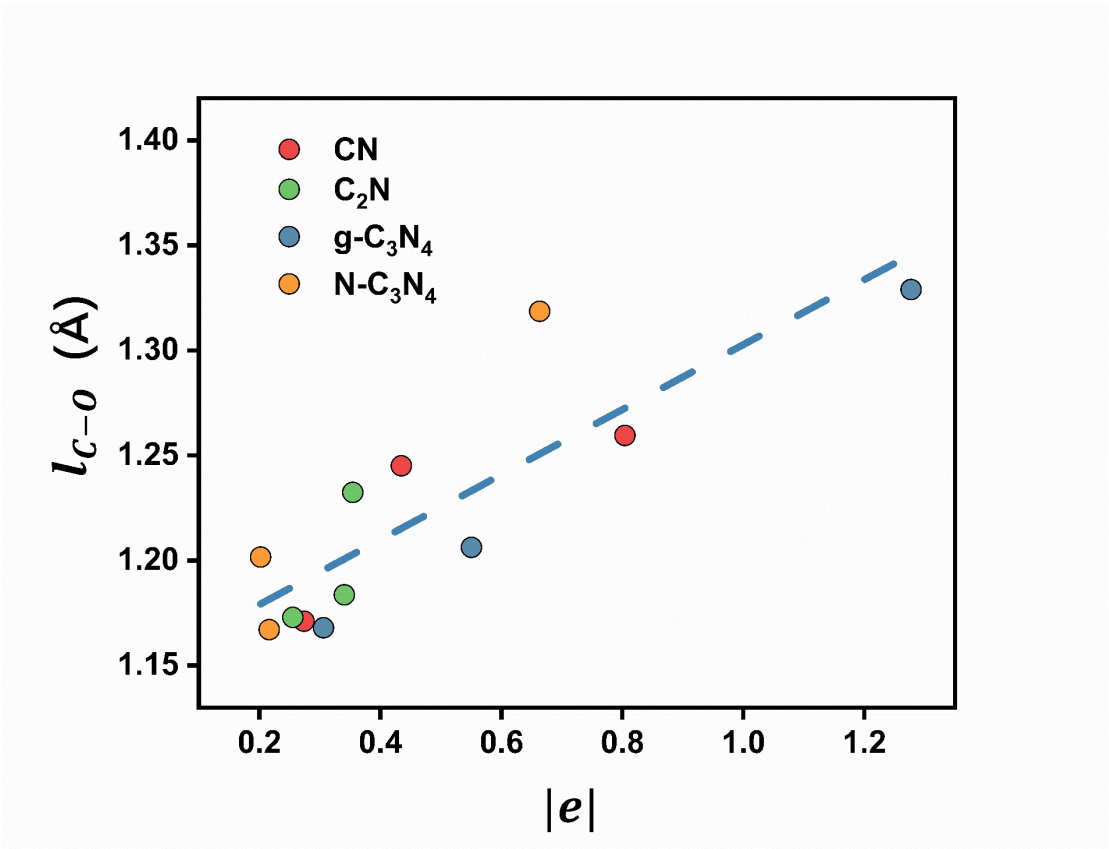

Fig. S3 Interrelational study of $|e|$ and C-O bond length (l_{C-O}) in DACs: This figure portrays a correlation analysis where the data points, color-coded as red, green, blue, and orange, represent DACs with the substrate geometries CN, C_2N, g-C_3N_4, and N-C_3N_4, respectively. These data points reflect the correlated behavior between both essential catalytic features ($|e|$ and l_{C-O}), offering insights into their roles within the catalytic efficiency and mechanisms.

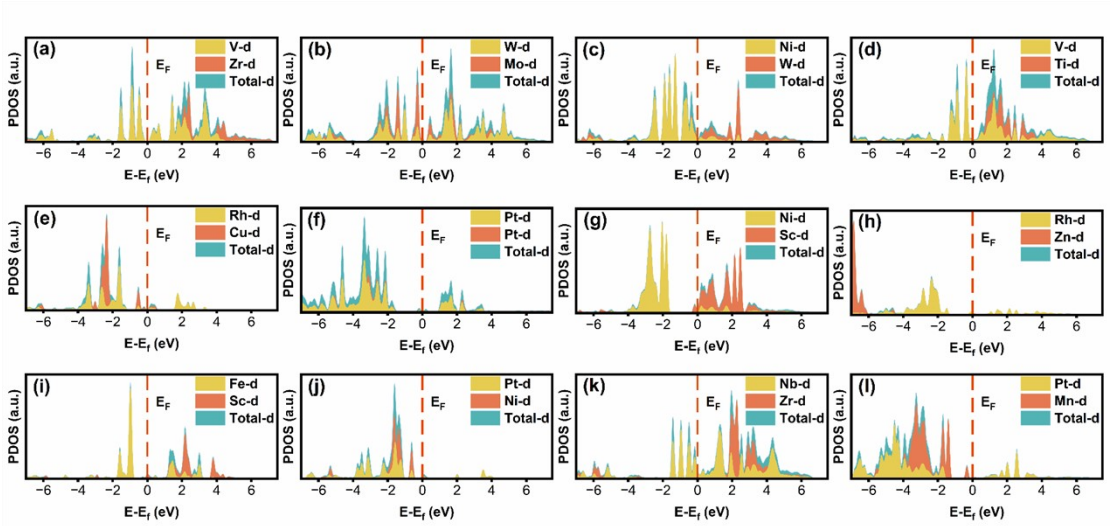

Fig. S4 Projected partial density of states (PDOS) for CO_2RR on DACs catalysts: This series depicts the PDOS of 12 DACs optimized for facilitating the CO_2RR. The Fermi level (E_F) is delineated by a red dashed line across each graph. Annotations in the upper right corners communicate the specific transition metals constituent of the DACs. Color-coding is used to differentiate the PDOS contributions: transition metal 1 (M1) is represented in yellow, transition metal 2 (M2) in orange, and the total d-orbital is conveyed in cyan. Groupings (a,e,i), (b,f,j), (c,g,k), and (d,h,l) correspond to DACs with underlying CN, C_2N, g-C_3N_4, and N-C_3N_4 substrate geometries, respectively.

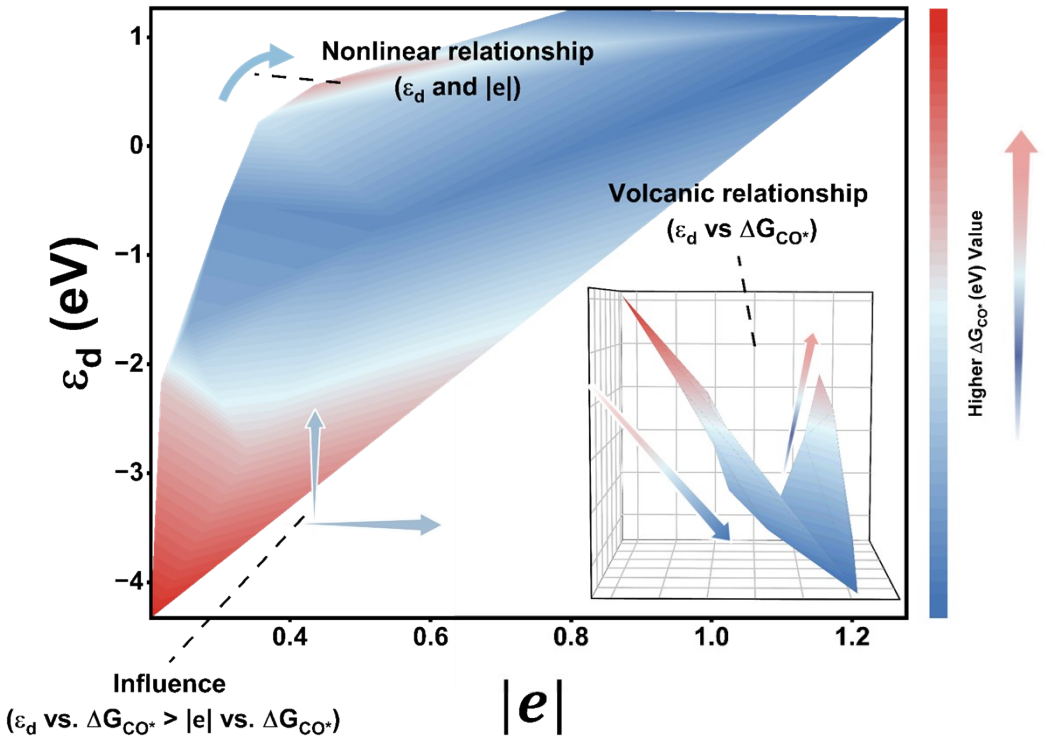

Fig. S5 The Overview of correlation analyses of ε_d, $|e|$, and ΔG_{CO^*}, with a color bar representing ΔG_{CO^*}. The lower right corner provides a front view of the image.

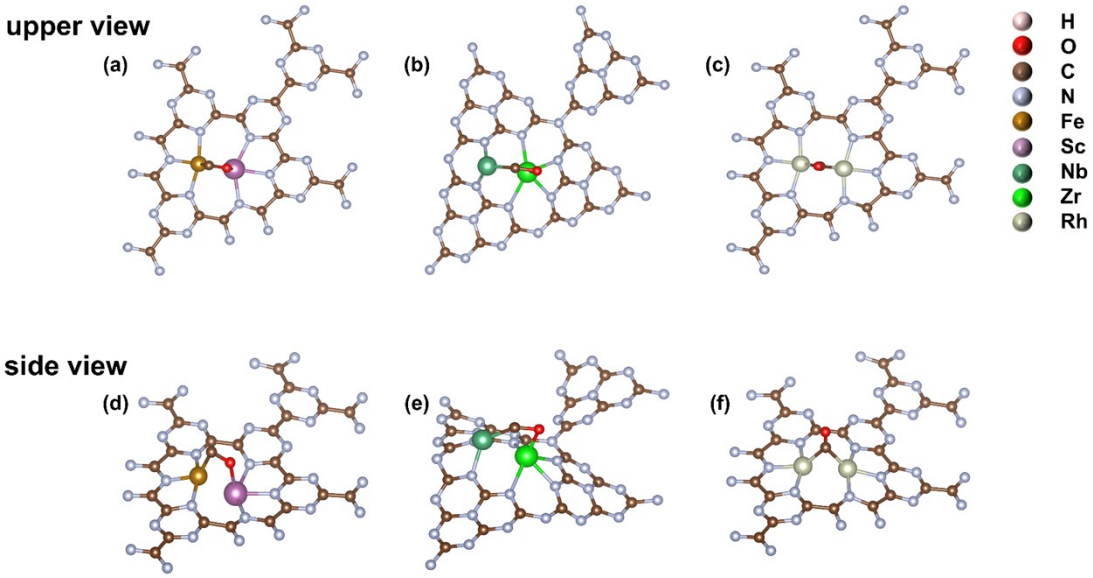

Fig. S6 Depiction of CO adsorption modes on optimized DACs structures: This figure delineates the optimized geometries of various DACs with CO adsorbed. Panels (a) and (d) feature the FeSc-CN catalyst, while panels (b) and (e) illustrate the NbZr_g-C_3N_4 catalyst, and both DACs are akin to a side-on adsorption mode for CO. Panels (c) and (f) present the RhRh_CN catalyst, which adopts an end-on adsorption configuration but with C as the terminal adsorption point. This last configuration is identified as the predominant CO adsorption mode (end-on with C as the terminal binding site) for the majority of the DACs investigated within this study.

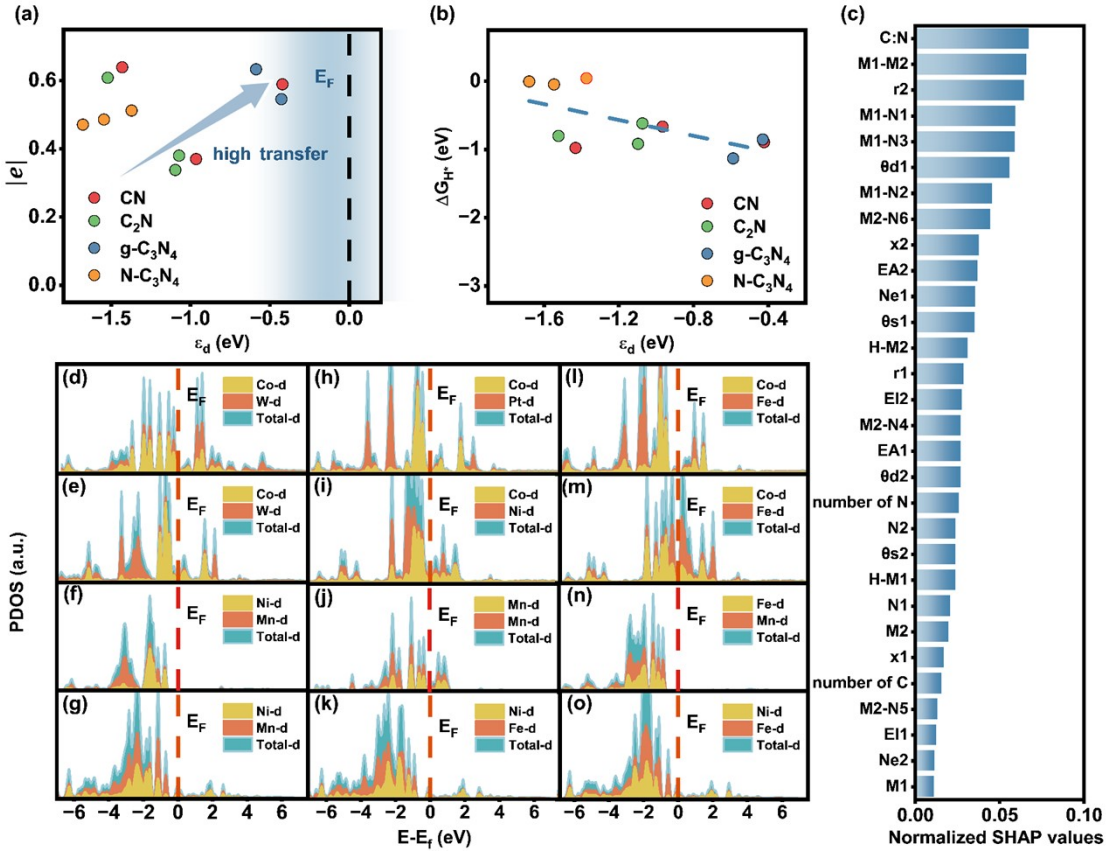

Fig. S7 Correlation of electronic descriptors and feature impact analysis of H adsorption on DACs: (a) & (b) plot the correlation between $|e|$ and ε_d as well as the Gibbs free energy change for hydrogen adsorption (ΔG_{H^*}) and ε_d, respectively. (c) presents the relative importance of various features as determined by SHAP (Shapley Additive exPlanations) values in the ML model that assesses the ΔG_{H^*} values of DACs. (d-o) display the PDOS of the optimized structures of 12 DACs that serve as hydrogen evolution reaction (HER) catalysts. The red dashed line marks E_F, with labels in the upper right corners indicating the associated transition metals (M1 and M2). The PDOS contributions of M1, M2, and the total d-orbital are colored yellow, orange, and cyan, respectively. Figures (d,h,l), (e,i,m), (f,j,n), and (g,k,o) represent DACs with CN, C_2N, g-C_3N_4, and N-C_3N_4 substrate geometries, correspondingly.

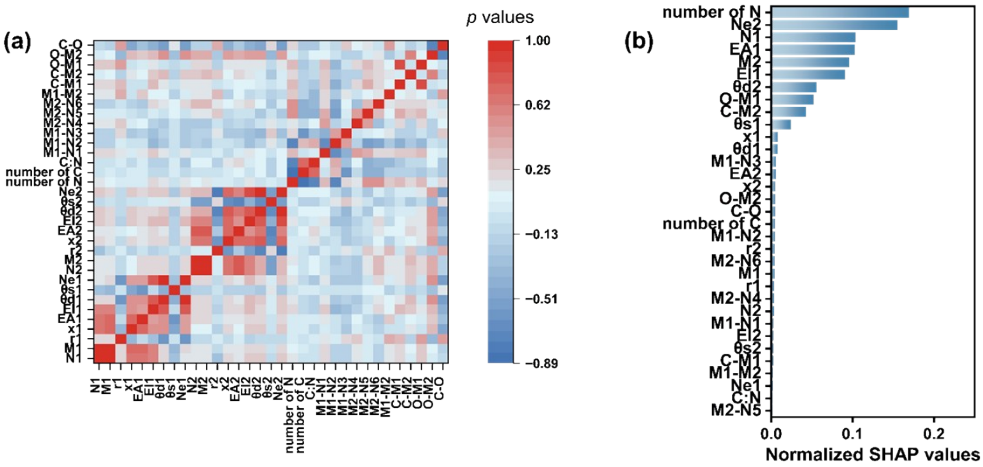

Fig. S8 Correlation heatmap and SHAP value analysis for CO adsorption on DACs: (a) Heatmap of Pearson's correlation coefficient (p) displaying the interdependence of full feature set: The range of colors from orange to dark purple indicates the values of p, with the most intense hues at both endpoints of color bar signifying larger absolute values of p and notable correlation between paired features. (b) Bar chart depicting the relative importance of features according to SHAP values within the ML model: This visualization prioritizes key descriptors influencing the Gibbs free energy change for CO adsorption (ΔG_{CO^*}) values of DACs, as identified by SHAP value assessment, establishing a hierarchy of feature significance in CO adsorption behavior.

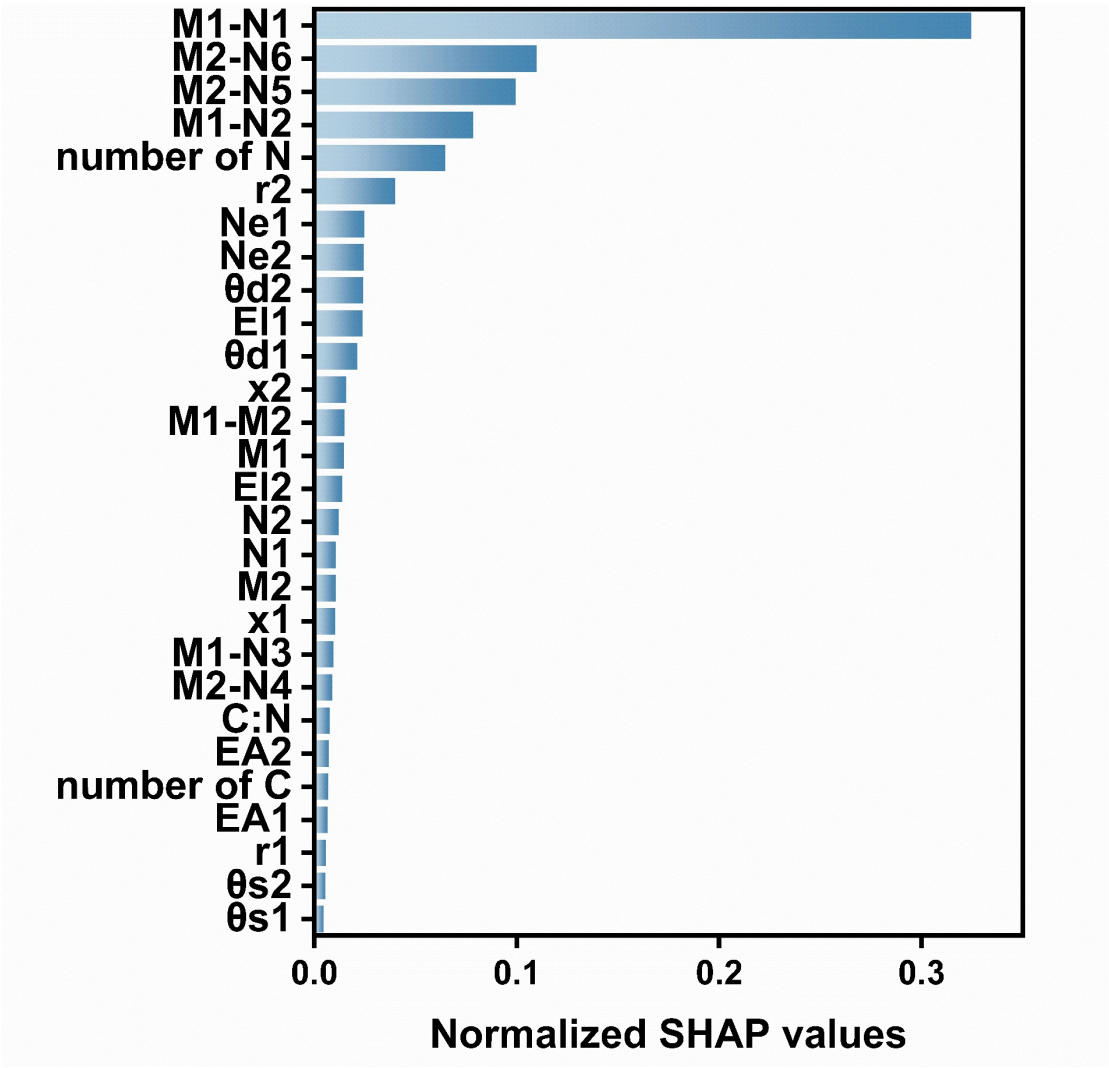

Fig. S9 Bar plot of SHAP values depicting feature importance for $^{\Delta E_{bind}}$ of DACs: This bar plot visualizes the relative importance of the features, as determined by SHAP values, within the ML model developed to analyze $^{\Delta E_{bind}}$ of DACs. Each bar represents the magnitude of impact that a particular feature contributes to the ML model's output, providing insights into the most influential factors for DAC stability.

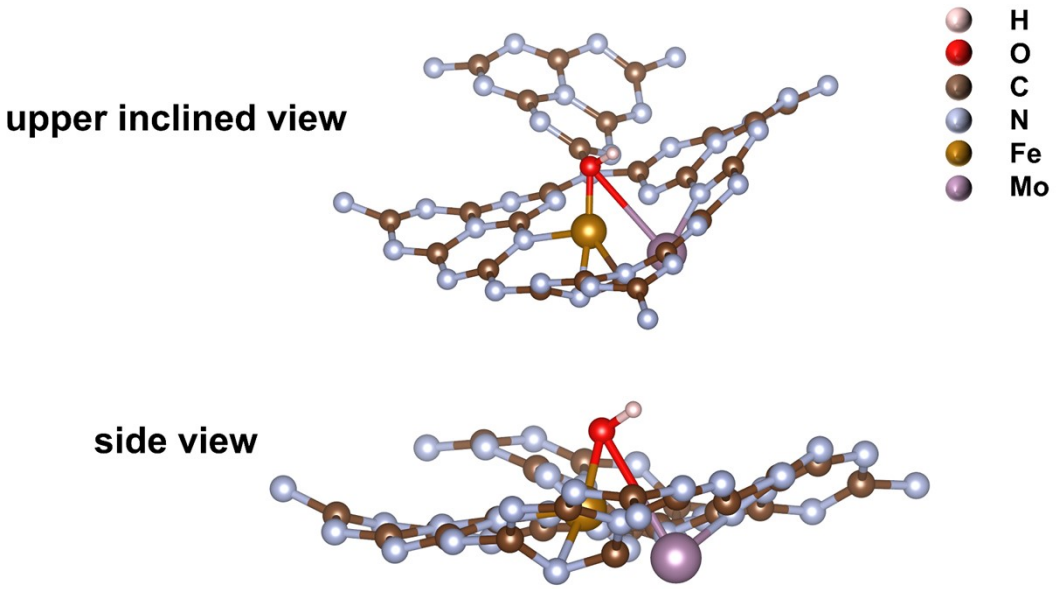

Fig. S10 Optimized configuration for MoFe DAC on N-C_3N_4 substrate: This figure illustrates the optimized molecular structure of the MoFe DAC interfaced with N-C_3N_4 substrate geometries. Quantitatively, the Fe-O bond measures 1.80 Å, which is notably shorter than the Mo-O bond at 3.25 Å. Such disparity suggests Fe as the predominant metal influence in our initial methodology, therefore, the d-electron count of MoFe N-C_3N_4 DAC was apportioned as 6, paralleled d-electron count of Fe. Subsequently, we refine our approach by incorporating bond lengths as weights to the respective d-electron counts, thereby deriving a weighted d-electron average, φ, described by the equation:

$$\phi = \frac{d_1 l_1 + d_2 l_2}{l_1 + l_2}$$

where d_1 and d_2 are the d-electron counts for transition metals 1 (M1) and 2 (M1), while l_1 and l_2 are the respective bond lengths of M1-O and M2-O. This refined quantification allows for a more precise tally of electron quantities within the bimetallic system of DACs.

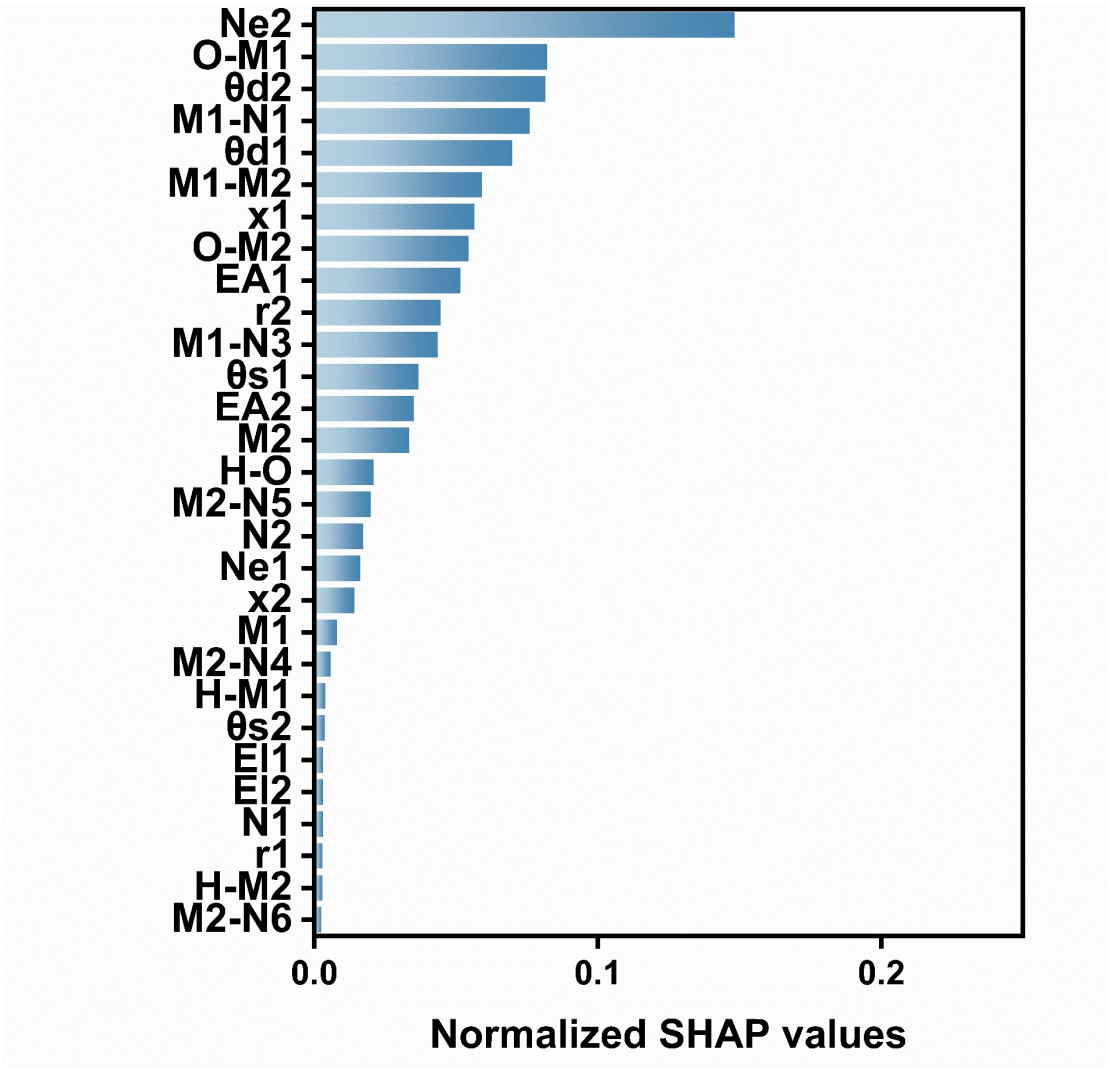

Fig. S11 SHAP values' insight on the Gibbs free energy change for OH adsorption (ΔG_{OH^*}) in DACs: The bar chart presents the SHAP value-derived relative importance of various features within the ML model employed to analyze ΔG_{OH^*} for the oxygen evolution reaction (OER) on DACs. This visualization elucidates the impact of individual features on the model's predictions, highlighting those that significantly influence OH adsorption during OER process.

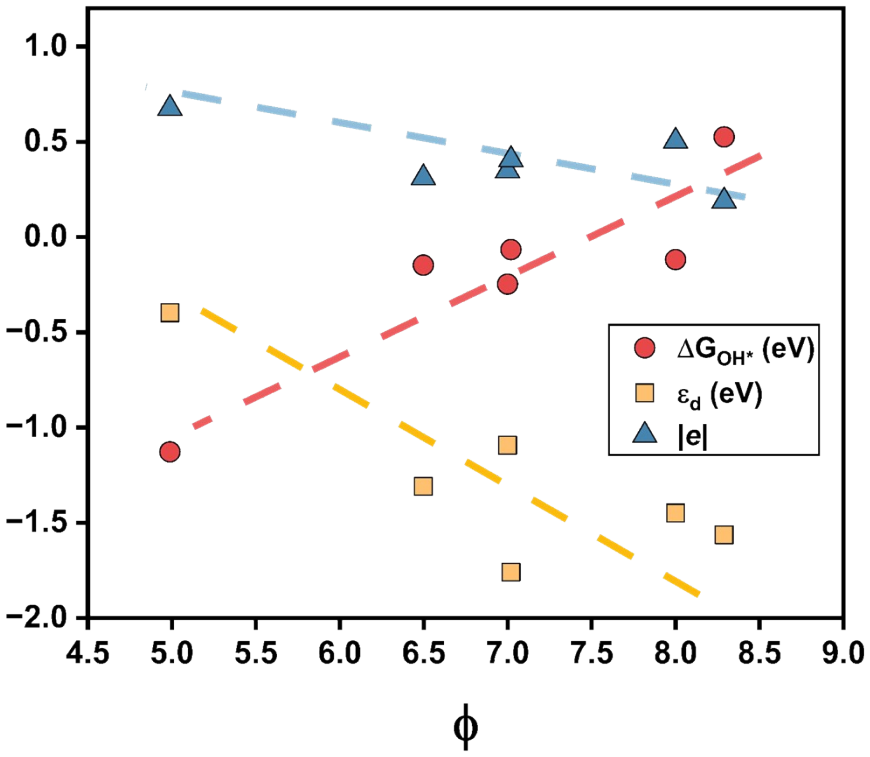

Fig. S12 Statistical correlation of φ with key parameters related to OER process in DACs: This figure delineates the statistical association of the parameter φ with ΔG_{OH^*}, ε_d, and $|e|$ for various DACs. Acknowledged studies[7,8] have previously confirmed the intrinsic connection of these parameters (ΔG_{OH^*}, ε_d and $|e|$) with the d-electron count in transition metals. The depicted correlations affirm that φ serves as a reasonable statistical approach for determining the effective d-electron numbers in complex systems featuring dual transition metals.

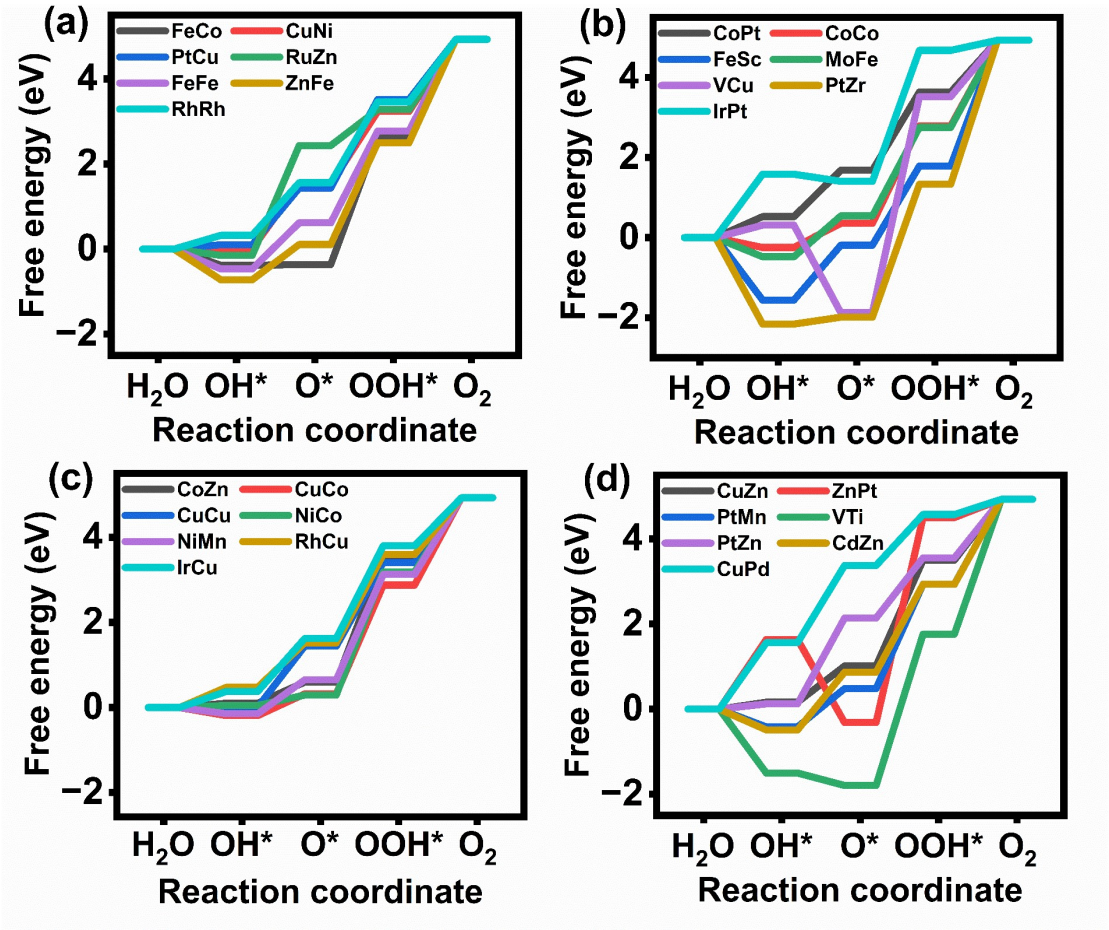

Fig. S13 Comparative free energy diagrams of OER for various DACs: Panels (a)-(d) exhibit the relative free energy profiles for the OER catalyzed by different DACs with N-C_3N_4 substrate. The diagrams present the free energy changes at each step of the OER process, facilitating an assessment of the catalytic efficiency and potential energy barriers across a range of DACs.

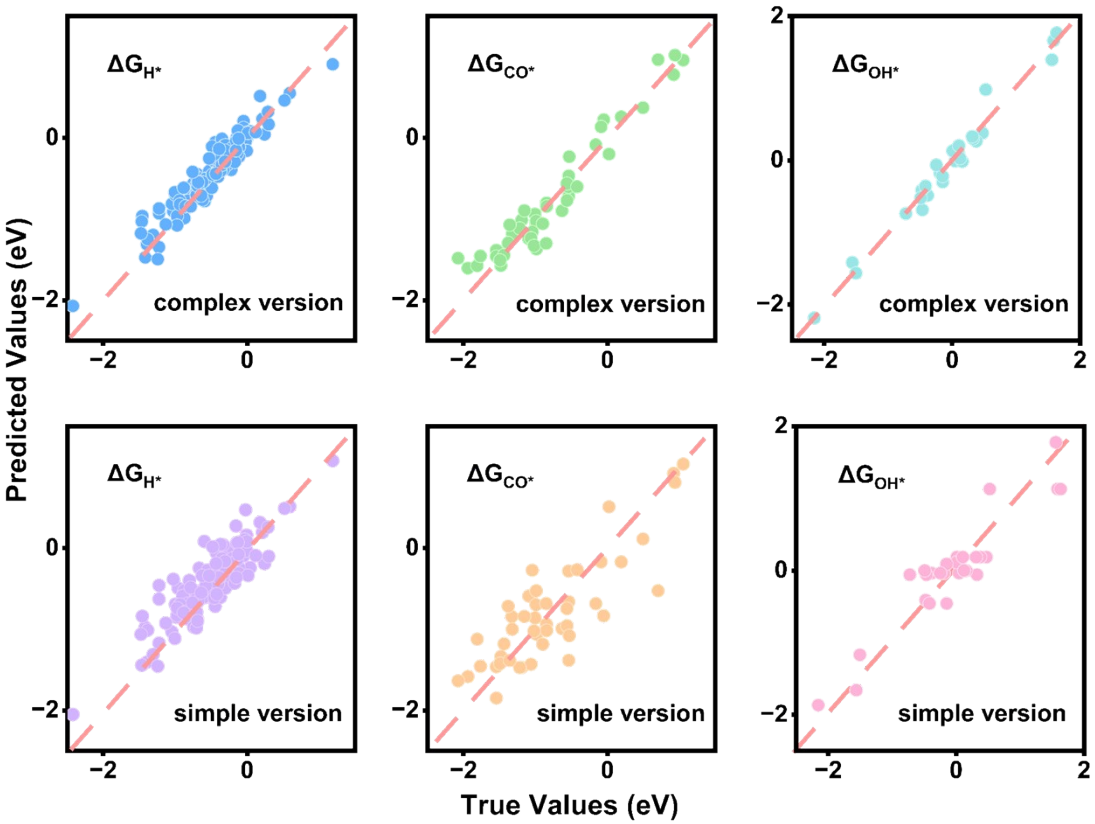

Fig. S14 Fitting results of PySR, with the corresponding key intermediate adsorption energy shown in the upper left corner. Without the use of structural descriptors, relying solely on elemental descriptors for fitting (simple version below) leads to varying degrees of accuracy deterioration.

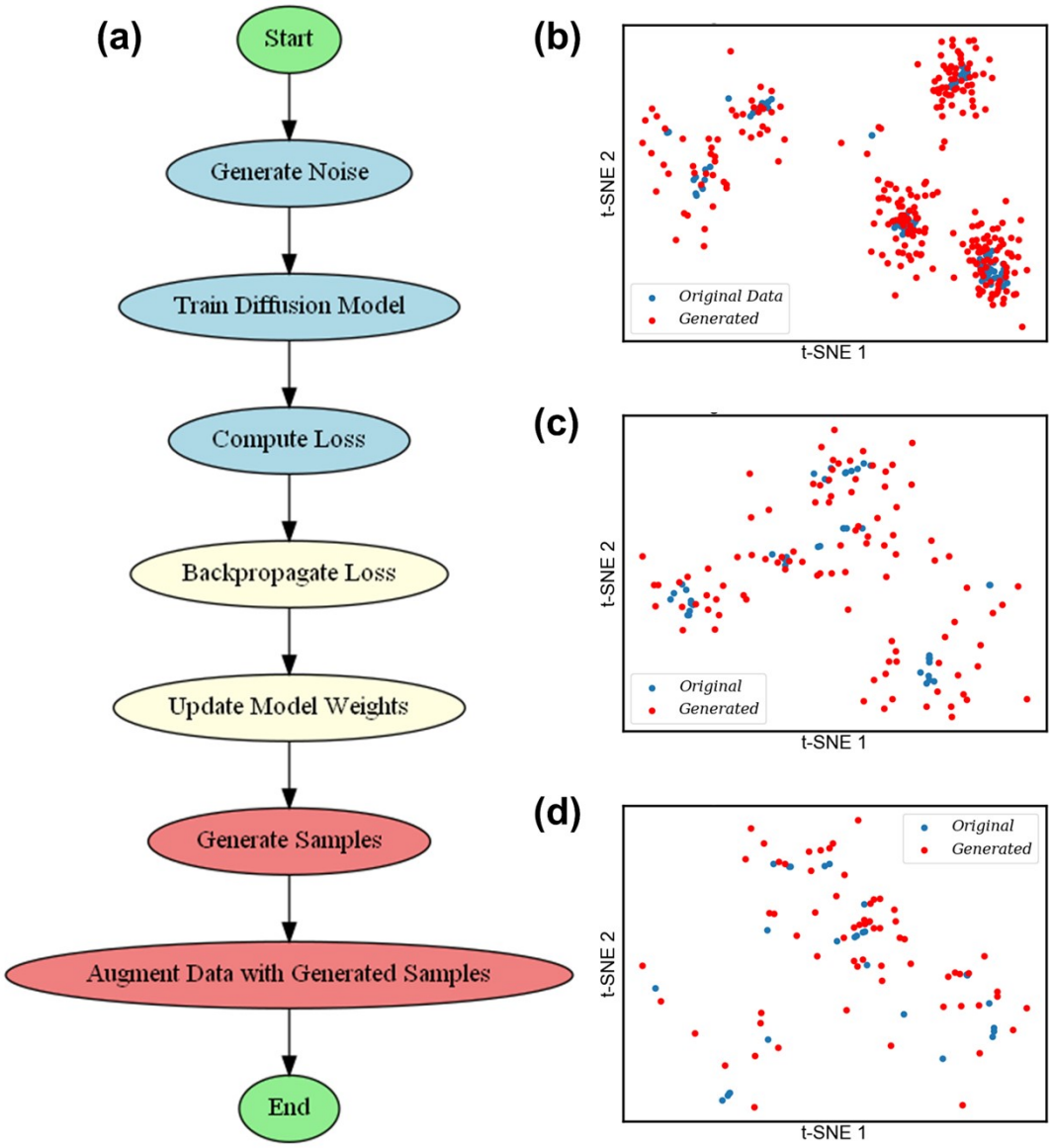

Fig. S15 (a) Flowchart of the diffusion model, designed to learn from the training set of DFT data and generate corresponding data to expand the dataset, ensuring effective model training. (b)-(d) t-SNE projections of the generated data (red dots) and raw data (blue dots), which are closely aligned, demonstrating the model's effective data generation.

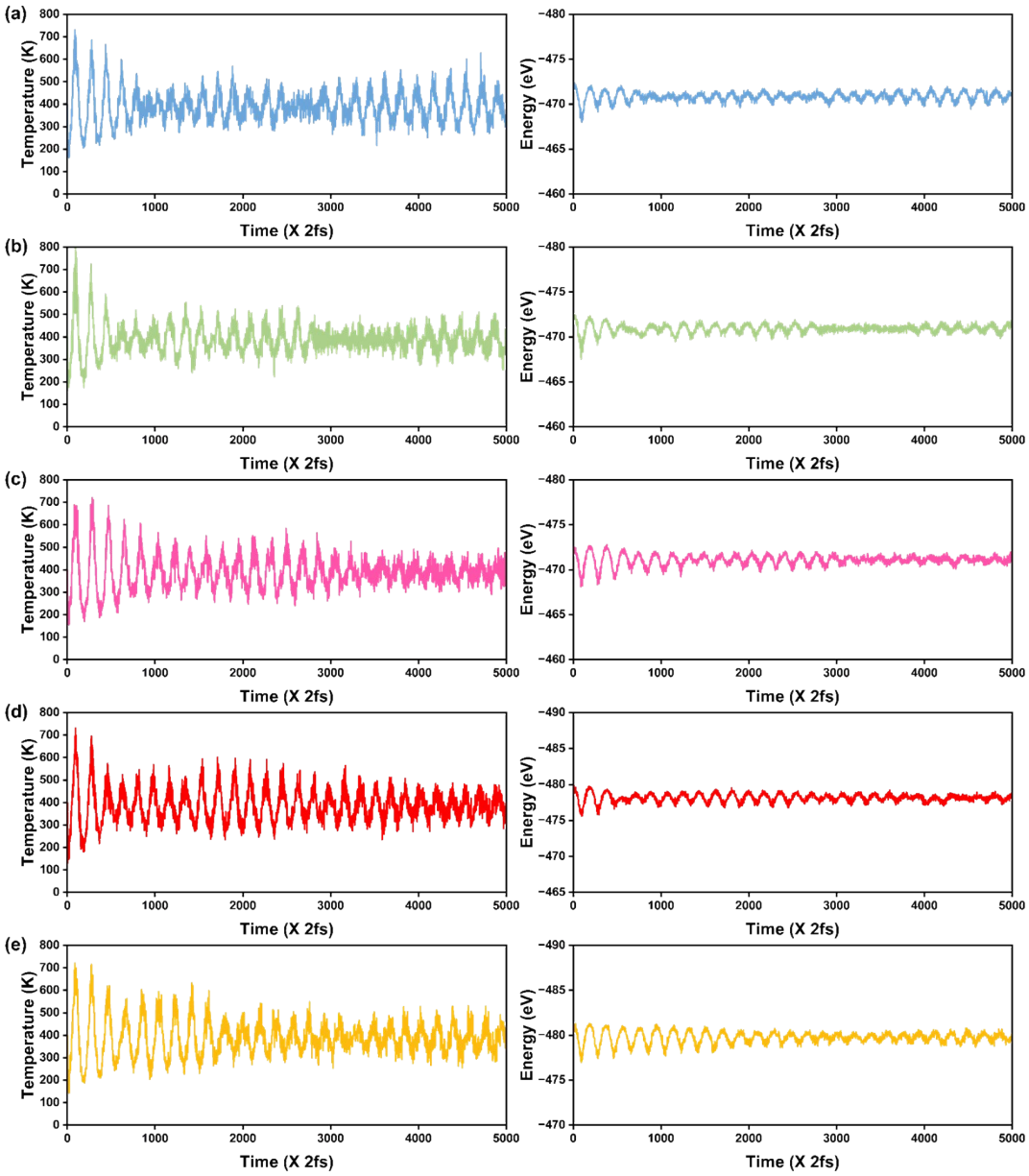

Fig. S16 Temperature (left) and energy (right) fluctuations during AIMD simulations performed at 400 K for 10 ps with a time step of 2 fs under the NVT ensemble for: (a) CuPd_N-C_3N_4, (b) PtZn_N-C_3N_4, (c) CuNi_ N-C_3N_4, (d) PtMn_N-C_3N_4, and (e) VTi_N-C_3N_4.

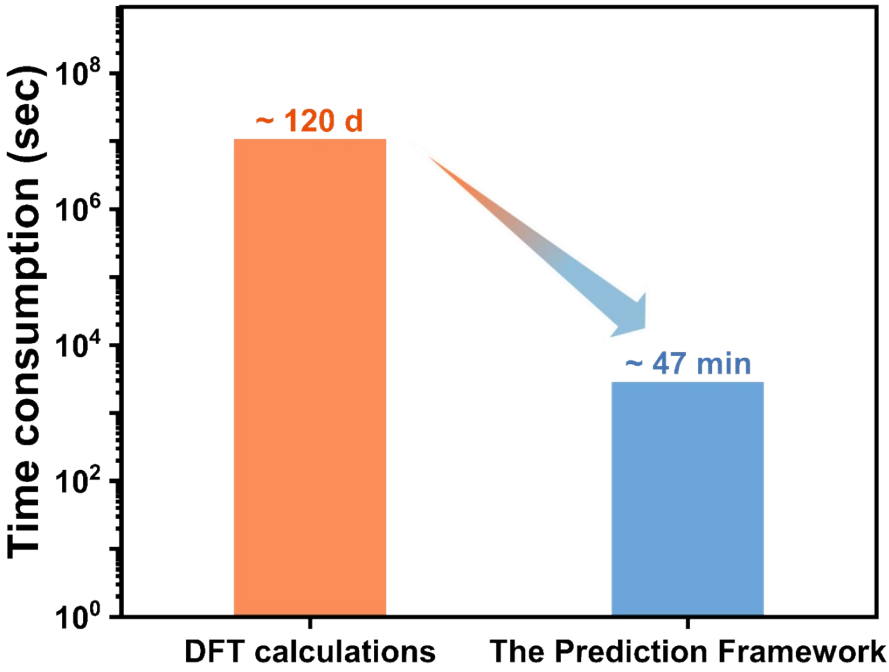

Fig. S17 Comparative analysis of computational time for molecular property predictions using density functional theory (DFT), traditional ML, and full process ML frameworks: This bar chart contrasts the computational time required for predicting molecular properties of all data through DFT calculations, a traditional ML framework, and an advanced full process ML framework. DFT calculations were executed on the Vienna Ab initio Simulation Package (VASP)[9] with an 88-core CPU, consuming approximately 10.6 million CPU-seconds. In stark contrast, the ML framework requires merely 47 minutes on a consumer-grade CPU, demonstrating efficiency that surpasses DFT calculations by over 3750 times. While the traditional ML framework and the full process ML framework exhibit comparable computational times, the latter significantly reduces manual effort and enhances prediction accuracy, indicating a substantial advancement in the field of computational material science.

References

1. V. Wang, N. Xu, J.-C. Liu, G. Tang and W.-T. Geng, *Comput. Phys. Commun.*, 2021, **267**, 108033.

2. C. Deng, Y. Su, F. Li, W. Shen, Z. Chen and Q. Tang, *J. Mater. Chem. A*, 2020, **8**, 24563-24571.

3. H. Feng, H. Ding, P. He, S. Wang, Z. Li, Z. Zheng, Y. Yang, M. Wei and X. Zhang, *J. Mater. Chem. A*, 2022, **10**, 18803-18811.

4. L. Wu, T. Guo and T. Li, *Adv. Funct. Mater.*, 2022, **32**, 2203439.

5. Z. Shu, H. Yan, H. Chen and Y. Cai, *J. Mater. Chem. A*, 2022, **10**, 5470-5478.

6. A. Chen, X. Zhang, L. Chen, S. Yao and Z. Zhou, *J. Phys. Chem. C*, 2020, **124**, 22471-22478.

7. T. Yan, X. Li, Z. Wang, Q. Cai and J. Zhao, *J. Colloid Interface Sci.*, 2023, **649**, 1-9.

8. X. Lu, J. Li, S. Cao, Y. Hu, C. Yang, Z. Chen, S. Wei, S. Liu and Z. Wang, *ChemSusChem*, 2023, **16**, e202300637.

9. A. Fonari and S. Stauffer, *vasp_raman.py*, https://github.com/raman-sc/VASP/, 2013.