Electronic Supplementary Material (ESI) for Journal of Materials Chemistry A.

Supporting Information

Data-driven design of double-atom catalysts with high H_2 evolution activity/CO_2 reduction selectivity based on simple features

Chenyang Wei[a], Dingyi Shi[b], Zhaohui Yang[a], Zhimin Xue*[,c], Shuzi Liu[a], Ruiqi Li*[,b] and Tiancheng Mu*[,a]

a Department of Chemistry, Renmin University of China, Beijing 100872, China. Email: tcmu@ruc.edu.cn

b College of Information Science and Technology, Beijing University of Chemical Technology, Beijing 100029, China. Email: lir@buct.edu.cn

c Beijing Key Laboratory of Lignocellulosic Chemistry, College of Materials Science and Technology, Beijing Forestry University, Beijing 100083, China. E-mail: zmxue@bjfu.edu.cn

Table S1 The key hyperparameters for each algorithm in our ML models applied for grid search (hyperparameters not mentioned were kept at their default values).

Algorithms	Hyperparameters
Support Vector Regression	C = [0.1, 0.2, 0.3, 1, 2, 3, 10, 20] gamma = [1, 0.1, 0.01, 0.001]
Random Forest Regression	n_estimators = [5, 10, 20, 50, 70, 100] max_depth = [5, 6, 7, 9, 10, 20] max_features = [0.6, 0.7, 1]
XGBoost Regression	n_estimators = [5, 10, 20, 50, 70, 100, 200] max_depth = [5, 6, 7, 8] max_delta_step = [1, 3, 5, 7]
Artificial Neural Network	-

Table S2 Summary of feature names and corresponding abbreviations.

Atomic features	Structure features
Atomic number (N)	The number of carbon atoms on doping sites (Nc)
Atomic mass (M)	The number of boron atoms on doping sites (Nb)
Atomic radius (r)	Doping position ($p_{1\text{-}6}$)
Electronegativity (χ)	The bonding length of boron and metal (B-M)
Electron affinity (EA)	The bonding length of carbon and metal (C-M)
First ionization energy (EI)	The bonding length of metal 1 and metal 2 (M1-M2)
The number of d-electron (θd)	The bonding length of hydrogen and metal for H absorption(H-M)
The number of s-electron (θs)	The bonding length of C, O and metal for CO absorption (CM and OM, respectively)
The number of outermost electron (Ne)	The bonding length of carbon and oxygen for CO absorption (CO)

Table S3 Summary of features for ML construction.

ML models	Features
Prediction of ΔG_{H^*}, ΔG_{CO^*}, $\Delta E_{binding}$	N, M, r, χ, EA, EI, θd, θs, Ne, Nb, Nc, p_{1-6}
Analysis of ΔG_{H^*}	N, M, r, χ, EA, EI, θd, θs, Ne, Nb, Nc, p_{1-6}, M_1-M_2, C_{1-6}-M_{1-2}, B_{1-6}-M_{1-2}, H-M_{1-2}
Analysis of ΔG_{CO^*}	N, M, r, χ, EA, EI, θd, θs, Ne, Nb, Nc, p_{1-6}, M_1-M_2, C_{1-6}-M_{1-2}, B_{1-6}-M_{1-2}, CM_{1-2}, OM_{1-2}, CO
Analysis of $\Delta E_{binding}$	N, M, r, χ, EA, EI, θd, θs, Ne, Nb, Nc, p_{1-6}, M_1-M_2, C_{1-6}-M_{1-2}, B_{1-6}-M_{1-2}

Table S4 Summary of features using for describing elements of double-atom catalysts (DACs).

Atom	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	52.00	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	0.56	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	7.83	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 S5 Summary of Bader charge analysis for key atoms for 6 B-doped graphene DACs discussed in the manuscript.

	FeZn_B		FeZn_2B		FeZn_3B		RhCu_B		RhCu_2B		RhCo_B
Fe	-0.690 e	**Fe**	-0.741 e	**Fe**	-0.751 e	**Rh**	-0.133 e	**Rh**	-0.106 e	**Rh**	-0.188 e
Zn	-0.232 e	**Zn**	-0.466 e	**Zn**	-0.538 e	**Cu**	-0.539 e	**Cu**	-0.446 e	**Co**	-0.544 e
B	-0.977 e	**B**	-0.765 e	**B**	-0.874 e	**B**	-1.47 e	**B**	-1.55 e	**B**	-1.52 e
C	+0.410 e	**B**	-0.576 e	**B**	-0.678 e	**C**	+0.257 e	**B**	-1.18 e	**C**	+0.280 e
C	+0.316 e	**C**	+0.189 e	**B**	-1.08 e	**C**	+0.117 e	**C**	+0.201 e	**C**	-0.0133 e
C	+0.125 e	**C**	+0.123 e	**C**	+0.384 e	**C**	+0.240 e	**C**	+0.363 e	**C**	+0.128 e
C	-0.0274 e	**C**	+0.214 e	**C**	+0.316 e	**C**	+0.174 e	**C**	+0.162 e	**C**	+0.170 e
C	+0.463 e	**C**	+0.326 e	**C**	+0.135 e	**C**	+0.0292 e	**C**	+0.221 e	**C**	+0.183 e
H	+0.417 e	**H**	+0.245 e	**H**	+0.255 e	**H**	-0.0626 e	**H**	-0.0431 e	**H**	-0.0356 e

Table S6 Different numbers of boron doping and the corresponding numbers of B-doped graphene DACs with $\Delta E_{binding} < 0$.

The number of boron doping	**The number of DACs with stable structure**
0	537
1	922
2	1747
3	1103
4	825
5	189
6	57

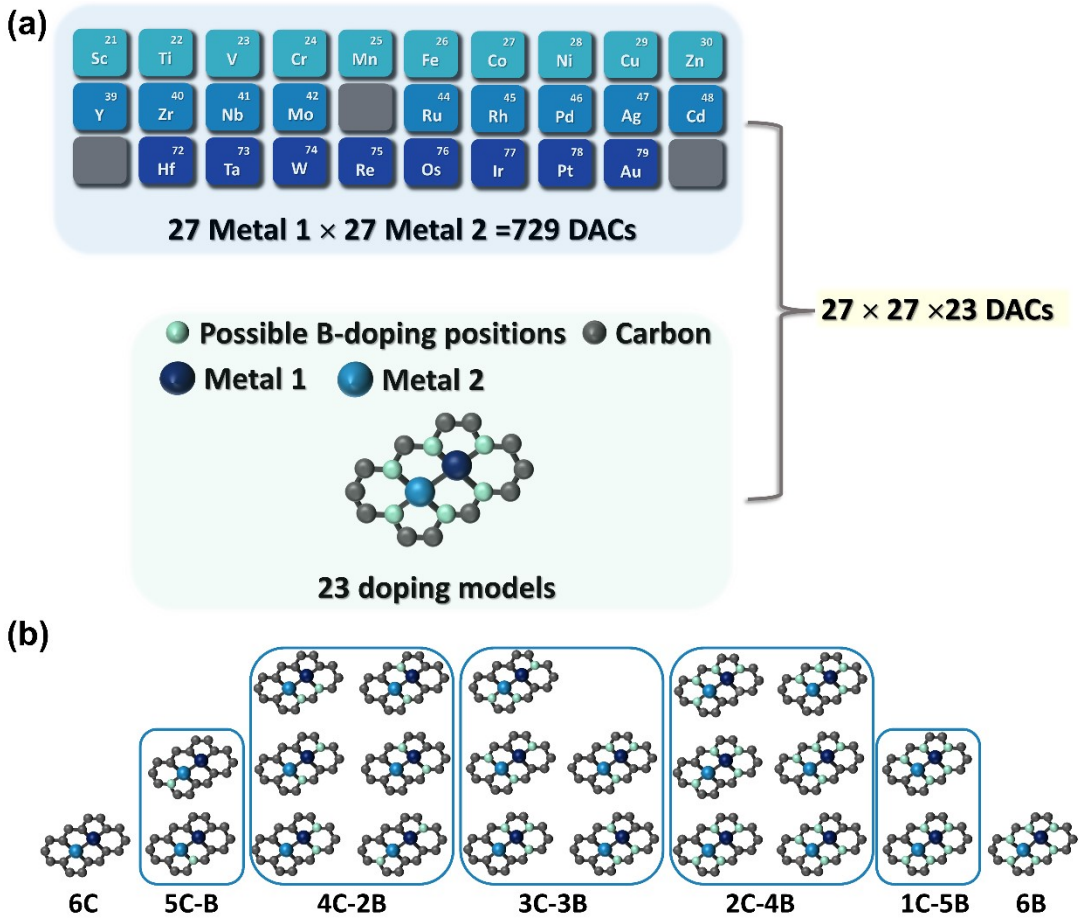

Figure S1 (a)The design spacing of 16,767 B-doped graphene DACs. (b) The details of 23 different doping models for B-doped graphene DACs.

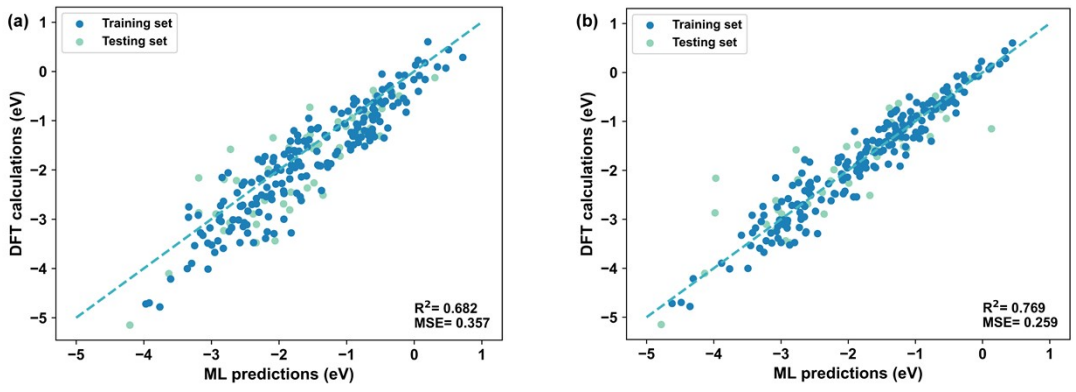

Figure S2 The comparison of ΔG_{H^*}. between the DFT calculations and GNN predictions by different graph: (a) graph G', (b) the key subgraph of G'

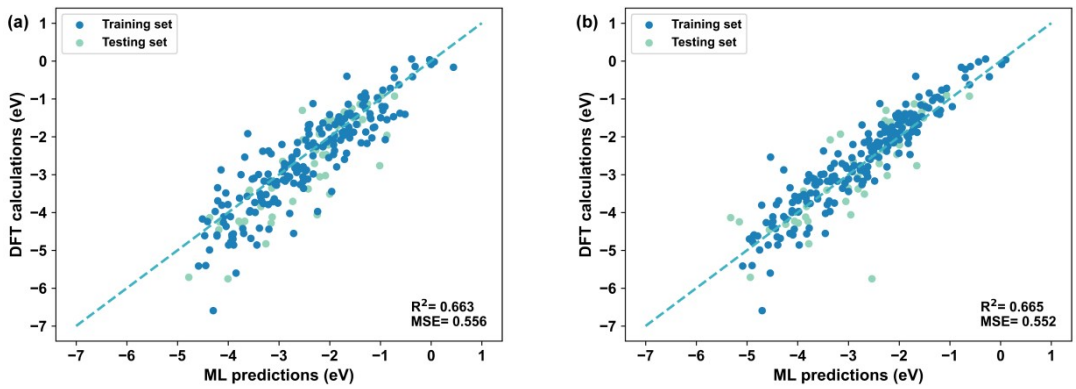

Figure S3 The comparison of ΔG_{CO^*} between the DFT calculations and GNN predictions by different graph: (a) graph G', (b) the key subgraph of G'

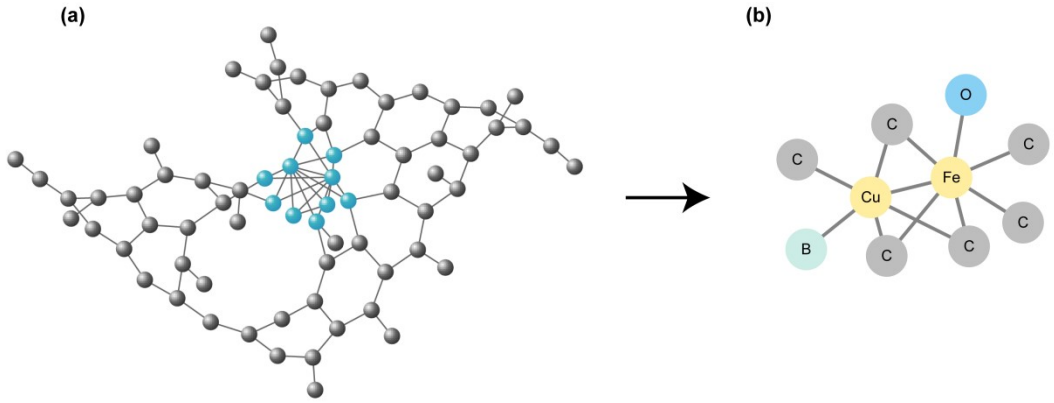

Figure S4 Using SubgraphX to find the key subgraphs that influences ΔG_{CO^*}.

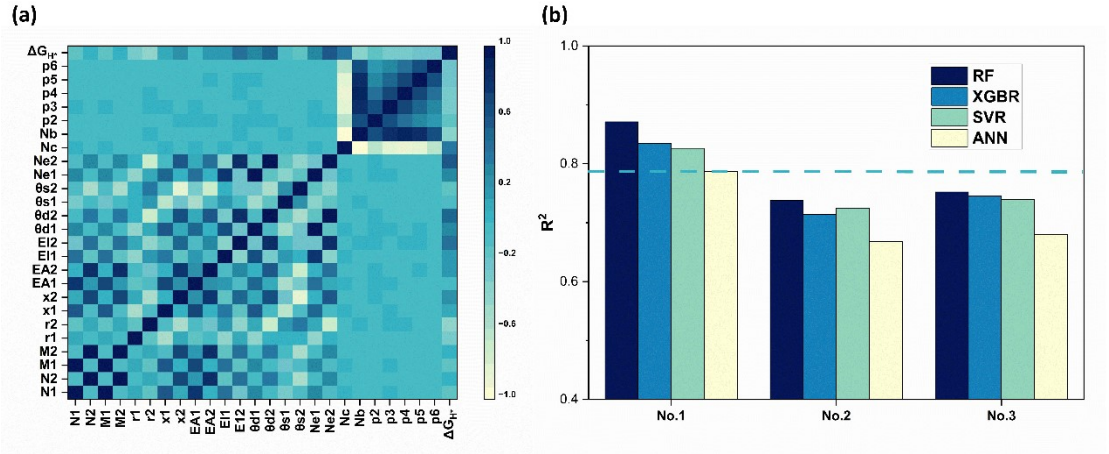

Figure S5 (a) The heatmap of feature analysis for different features. (b) The accuracy of various machine learning models (No.1: the initial feature set; No.2: feature set deleting feature N; No.3: feature set deleting feature M).

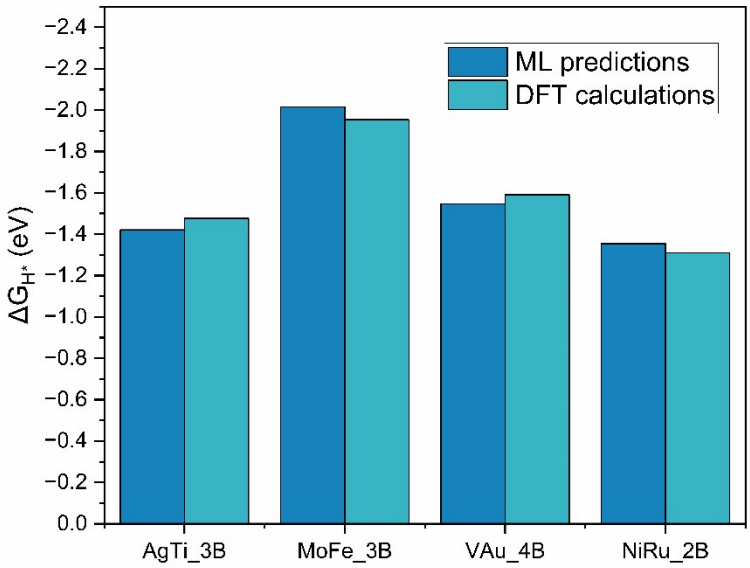

Figure S6 Comparation of DFT calculated ΔG_{H^*} and ML predicted ΔG_{H^*} of some examples of DACs with RF algorithm.

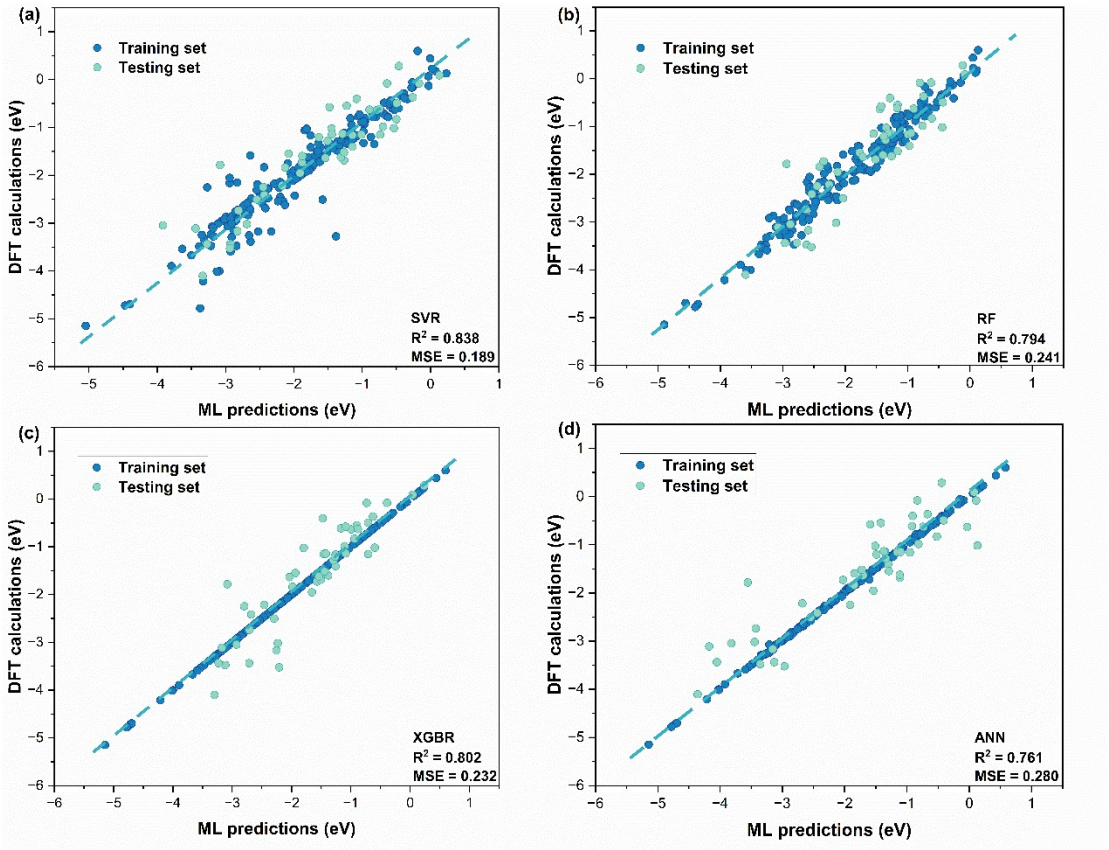

Figure S7 The performance for ML model used for analyzing the activity origin of ΔG_{H^*}.

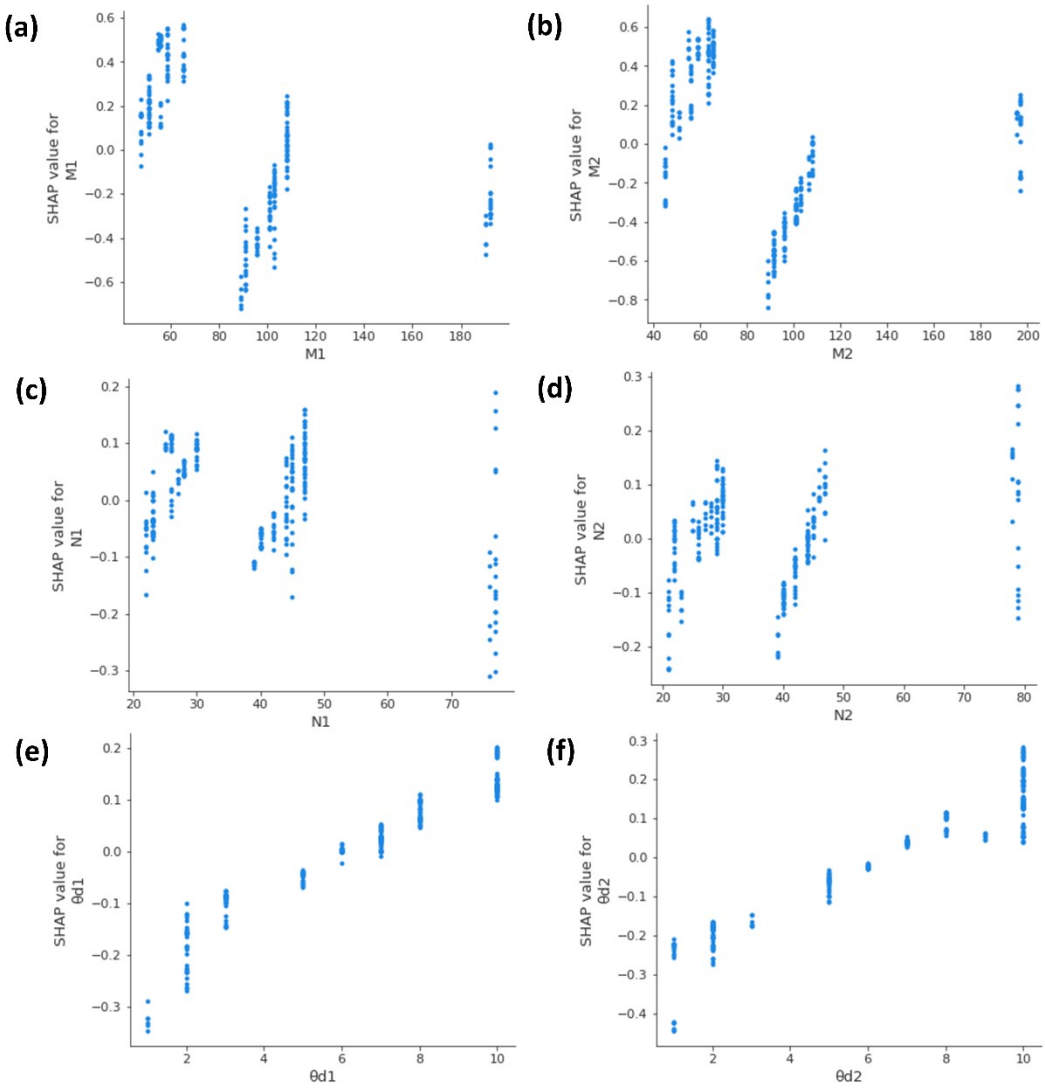

Figure S8 The partial dependence plots between features atomic mass (M1 and M2), atomic number (N1 and N2), the number of d-electrons (θd1 and θd2) and SHAP values.

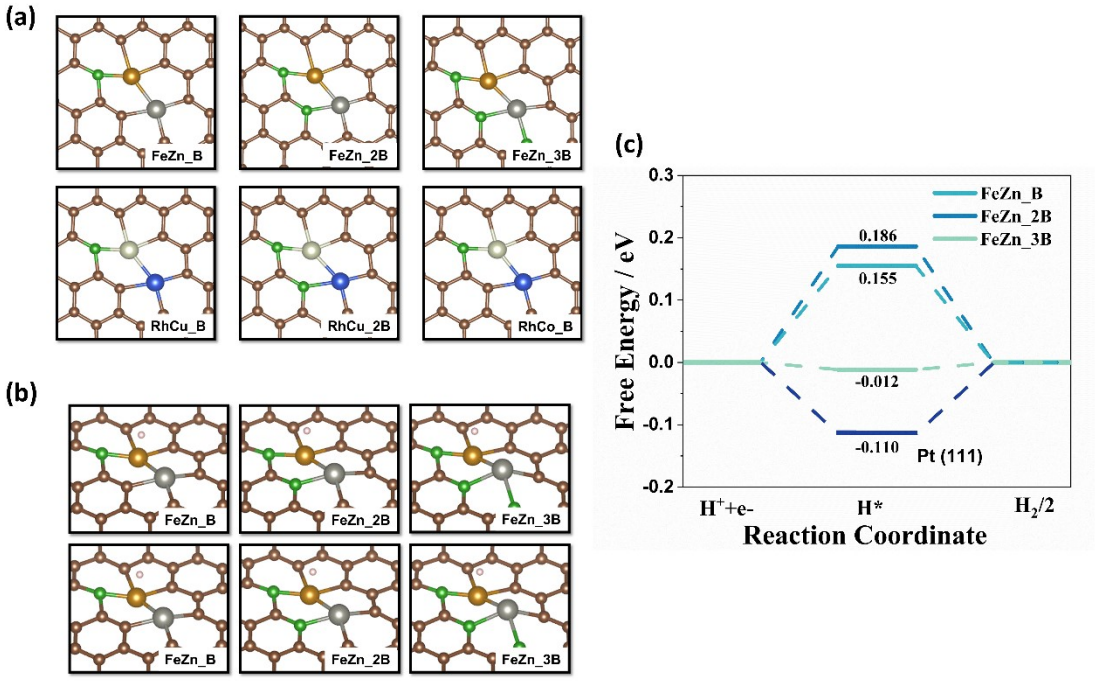

Figure S9 (a) The optimized structures of FeZn_B, FeZn_2B, FeZn_3B, RhCu_B, RhCu_2B and RhCo_3B without applying DFT-D3 dispersion correction. (b) The optimized structures of FeZn_B, FeZn_2B, FeZn_3B absorption hydrogen models with (above) and without (below) applying DFT-D3 dispersion correction. (c) The relative free energy profiles of the HER process for B-doped graphene DACs (FeZn_B, FeZn_2B and FeZn_3B) calculated without apply DFT-D3 dispersion correction.

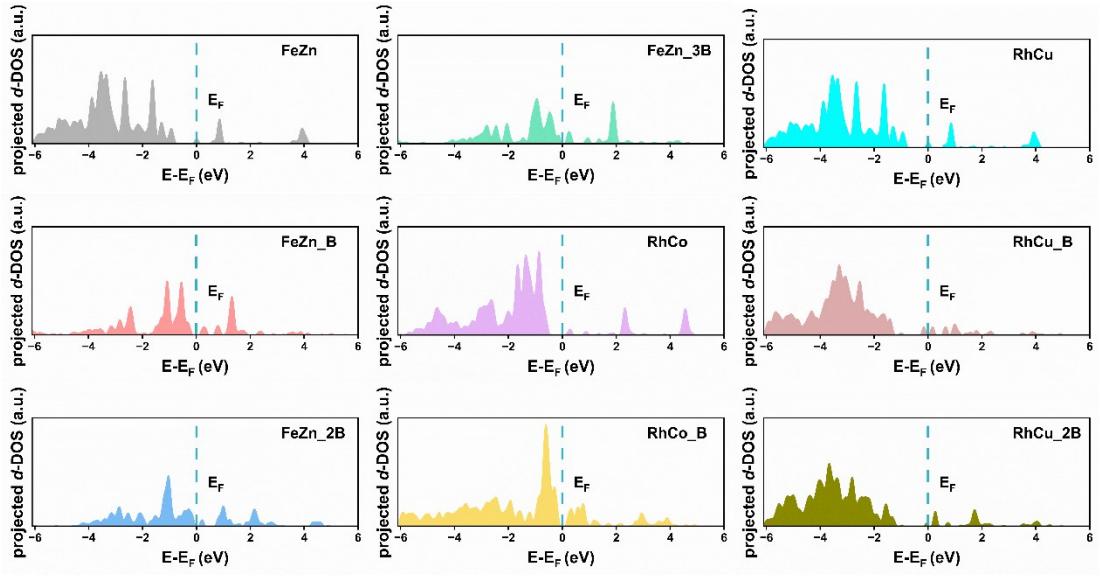

Figure S10 The projected density of states (PDOS) of the optimized structures of H-adsorbed B-doped graphene DACs.

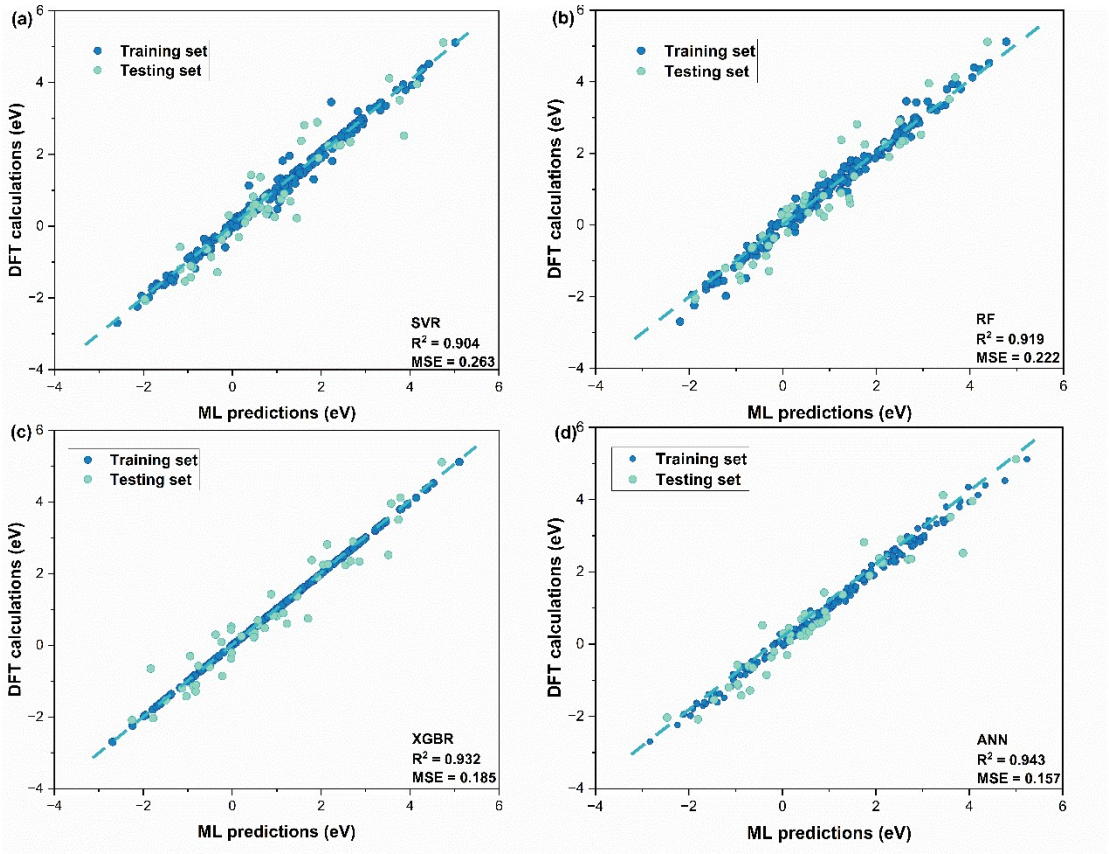

Figure S11 The comparison of $\Delta E_{binding}$ between the DFT calculations and ML predictions by different ML models: (a) SVR, (b) RF, (c) XGBR and (d) ANN.

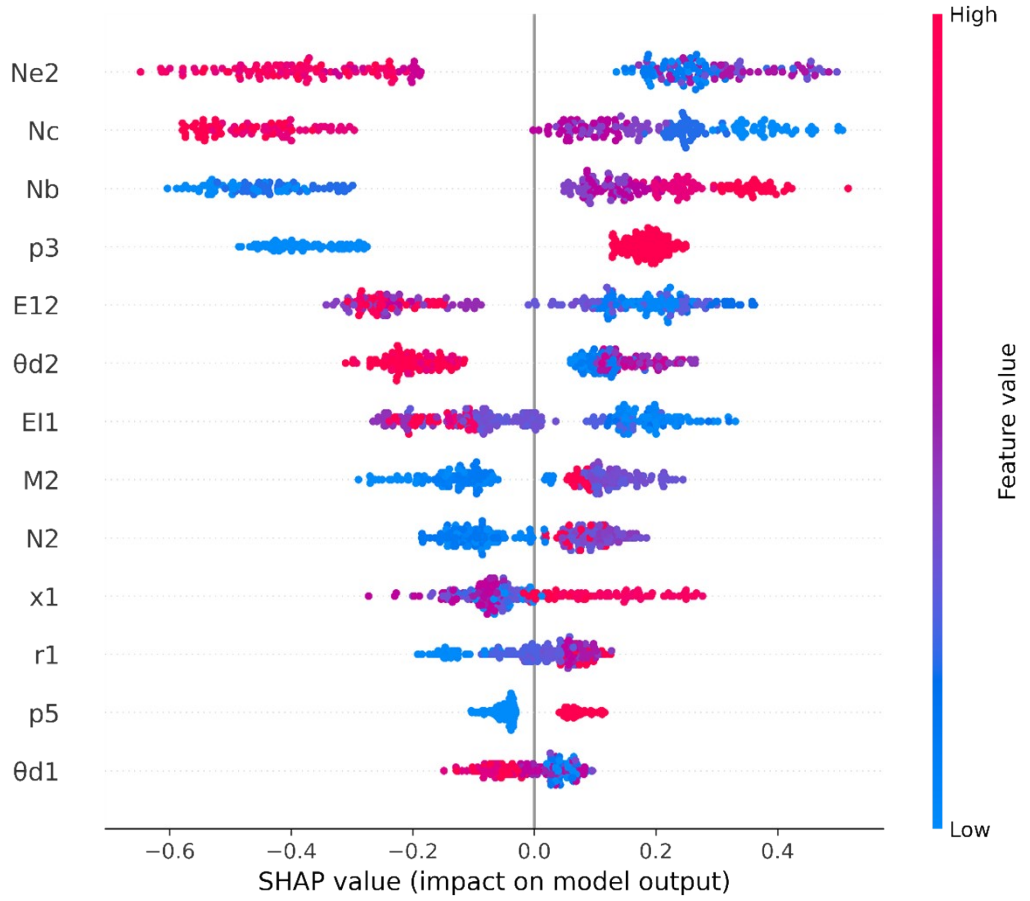

Figure S12 The SHAP values of ML model used for analyzing $\Delta E_{binding}$.

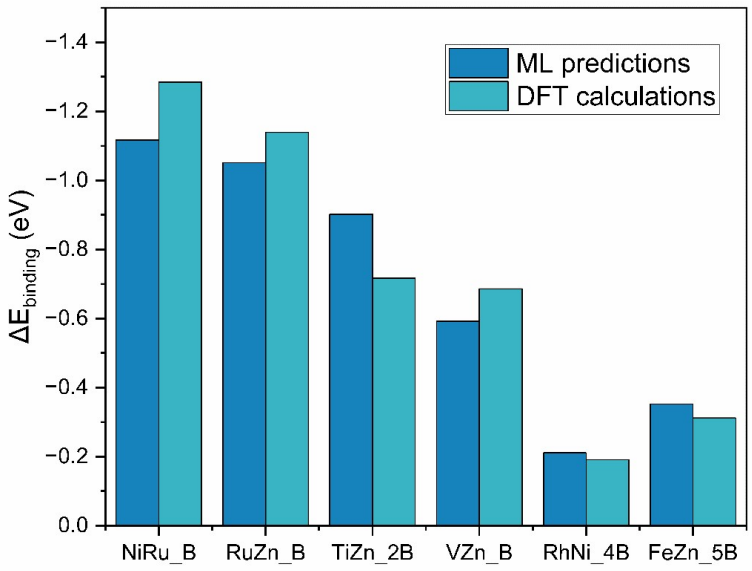

Figure S13 Comparation of DFT calculated $\Delta E_{binding}$ and ML predicted $\Delta E_{binding}$ of some examples

of DACs.

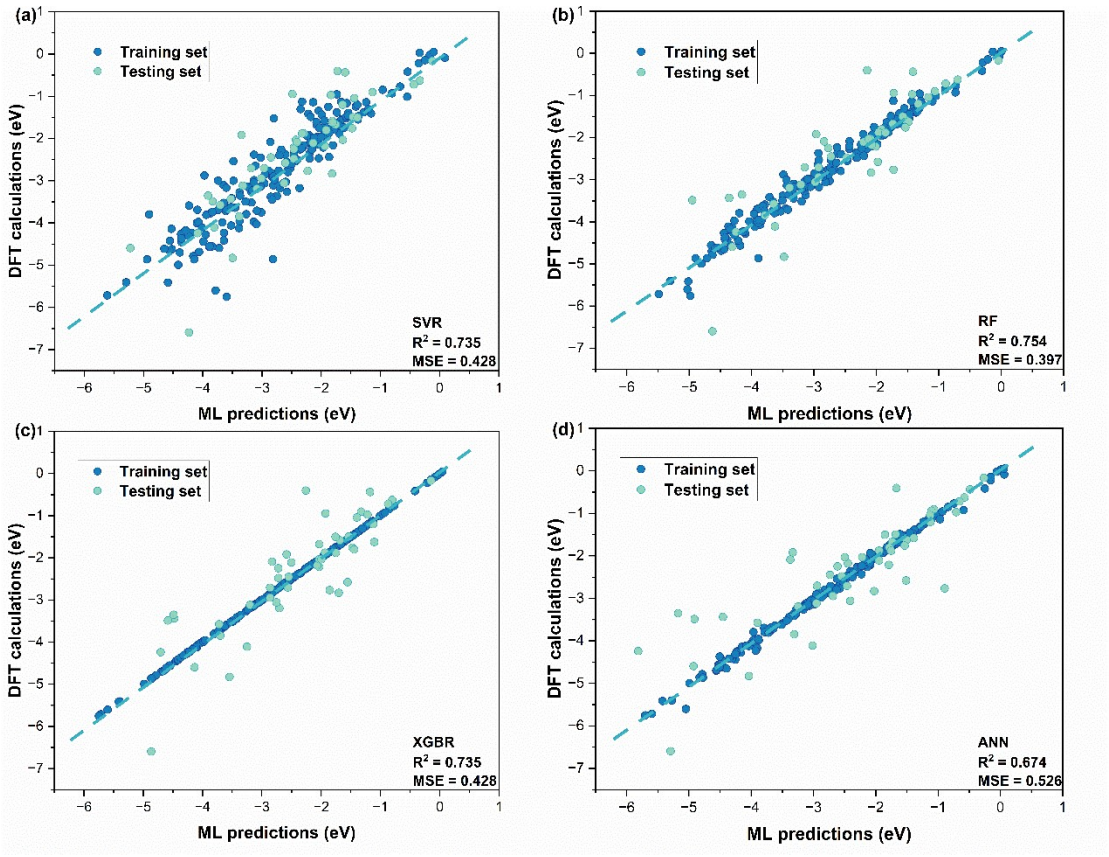

Figure S14 The comparison of ΔG_{CO^*} between the DFT calculations and ML predictions by different ML models: (a) SVR, (b) RF, (c) XGBR and (d) ANN.

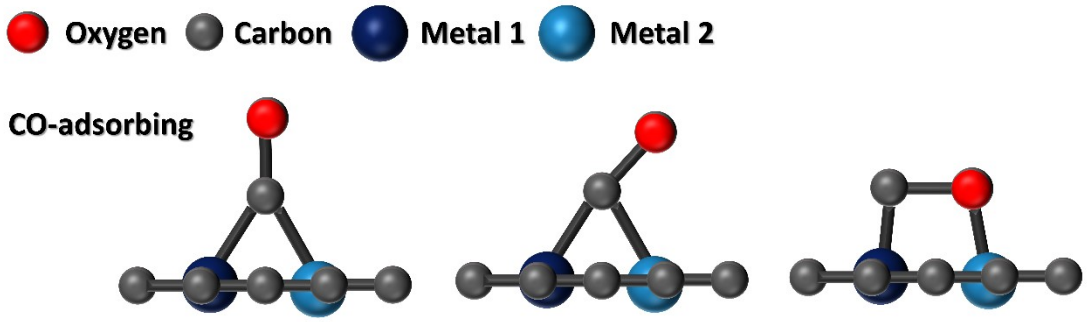

Figure S15 Examples of different adsorbing structures for CO_2 reduction reaction.

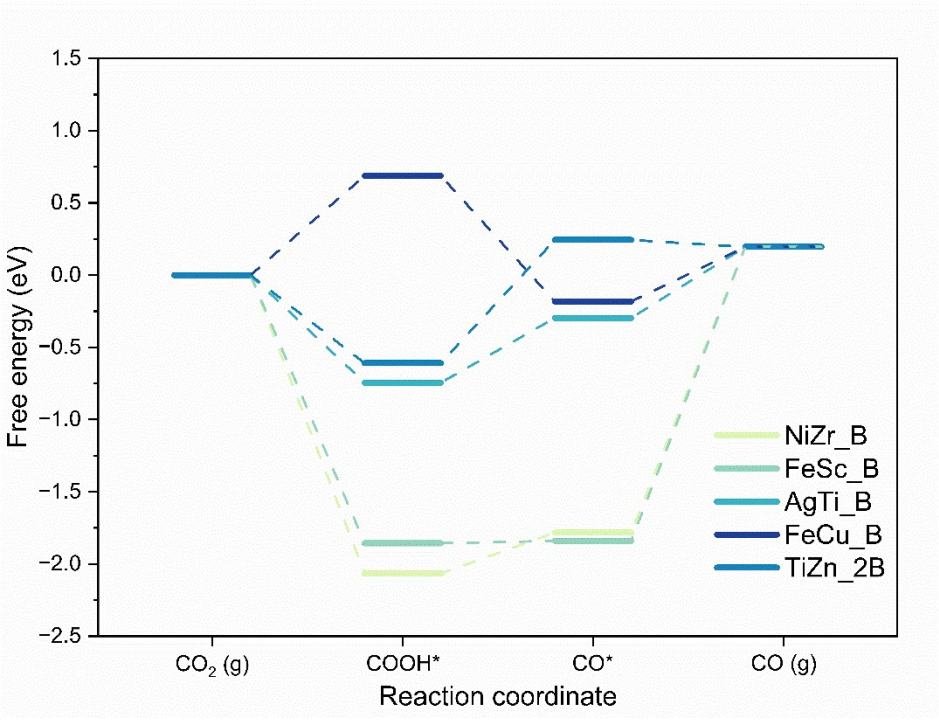

Figure S16 The relative free energy profiles of the CO_2RR process for 5 selected B-doped graphene DACs.

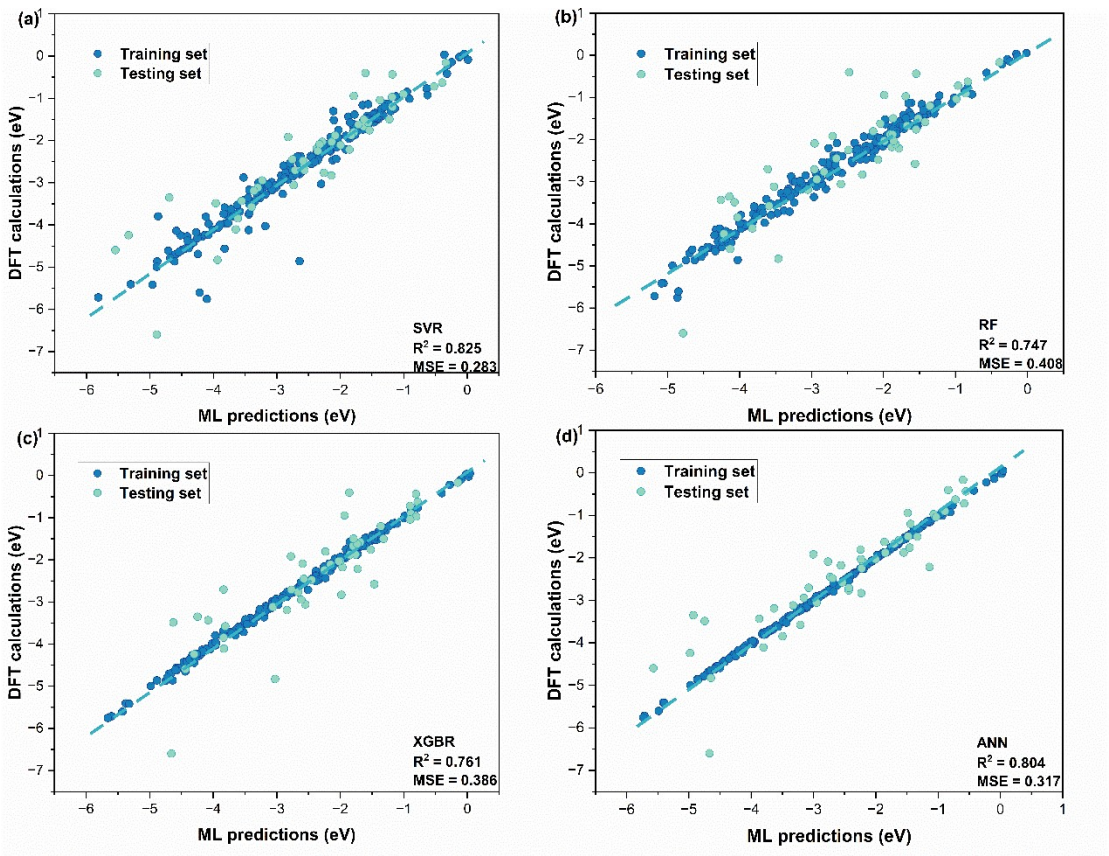

Figure S17 The performance for ML model used for analyzing the activity origin of ΔG_{CO^*}.

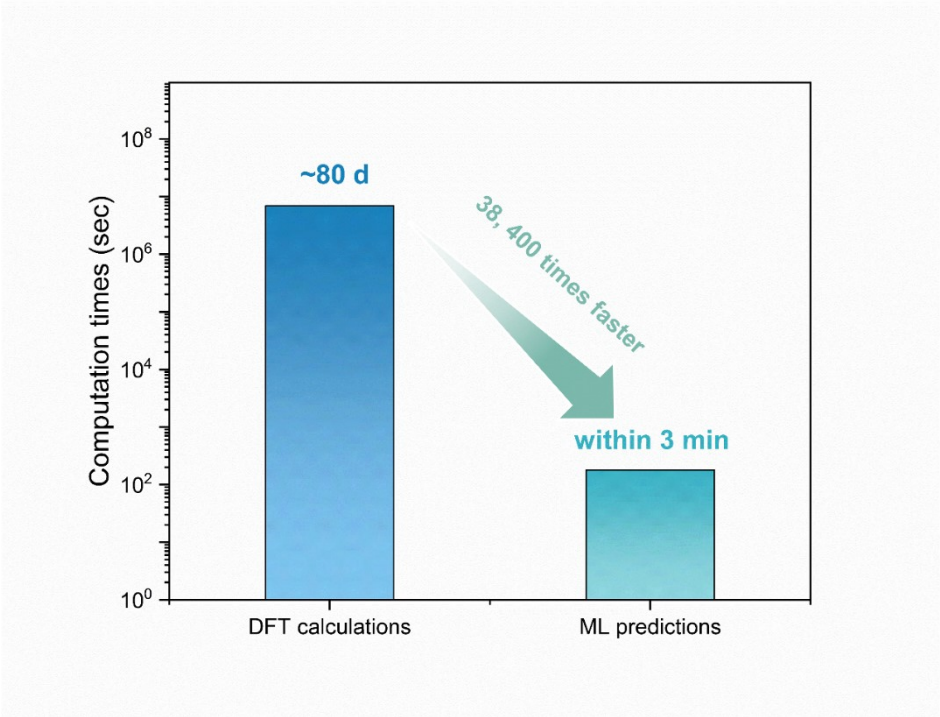

Figure S18 Comparation of computation cost for ML prediction and DFT calculations. For DFT calculations, we only estimate the time used for generating the dataset. For ML predictions, the whole process (starting from training the models to testing the models, and to predicting the ΔG_{H^*}, ΔG_{CO^*}, $\Delta E_{binding}$ of ~17000 DACs) was included in the computation cost.