Electronic Supporting Information for

Unravelling the Origin of Bifunctional OER/ORR Activity of Single-Atom Catalysts Supported on C_2N by DFT and Machine Learning

Yiran Ying [a], *Ke Fan* [a], *Xin Luo* [b]*, *Jinli Qiao* [c], *and Haitao Huang* [a]*

[a] Department of Applied Physics and Research Institute for Smart Energy, The Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong, P.R. China

[b] State Key Laboratory of Optoelectronic Materials and Technologies, Centre for Physical Mechanics and Biophysics, School of Physics, Sun Yat-sen University, Guangzhou, Guangdong Province, P.R. China, 510275

[c] State Key Laboratory for Modification of Chemical Fibers and Polymer Materials, College of Environmental Science and Engineering, Donghua University, 2999 Renmin North Road, Shanghai 201620, P. R. China

*E-mail: luox77@mail.sysu.edu.cn (X.L.); aphhuang@polyu.edu.hk (H.H.)

Electronic Supporting Information (ESI) Note 1. Additional computational details for DFT calculations

For each elementary step in OER and ORR, the Gibbs free energy changes ΔG is calculated as:

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

where ΔE, ΔZPE, TΔS, ΔG_{pH}, and ΔG_U represent changes in DFT-calculated total energy, zero-point energy (ZPE), entropic contribution, pH contribution, and applied electrode potential contribution, respectively. Adsorption energy for O_xH_y species involved in OER/ORR was calculated by:[1]

$$\Delta E_{*OxHy} = E_{*OxHy} - E_* - [xE_{H2O} - (2x-y)E_{H2}/2]$$

where * denotes active sites on the catalyst surface, and E_{H2O} and E_{H2} represent the energy of free H_2O and H_2 molecule, respectively. ZPE and entropic contribution terms were calculated by the vibrational frequencies of free molecules and adsorbates, and these values were listed in Table S1. ΔG_{pH} was calculated by $k_B T \ln 10 \times pH$, and the pH value was set as zero in this work; for non-zero pH (*e.g.* in alkaline condition), the reaction Gibbs free energy can be corrected by ΔG_{pH}, which will not be discussed in this work.[2, 3] ΔG_U was defined as -eU, where e and U represent the number of transferred electrons and applied electrode potential, respectively. Gibbs free energy for the H^+/e^- pair was calculated using Nørskov's computational hydrogen electrode (CHE) model:[2] $G(H^+ + e^-) = 0.5G(H_2)$. The solvent effect was considered by adding a -0.30 eV correction to all the DFT-calculated total energy of *OH and *OOH, because they can form hydrogen bonds with H_2O when solvated and thus are more stable.[4]

The protonation of *OOH in ORR can proceed through two different pathways: (1) $*OOH + H^+ + e^- \rightarrow *O + H_2O$ ($4e^-$ ORR); (2) $*OOH + H^+ + e^- \rightarrow * + H_2O_2$ ($2e^-$ ORR). Thus, the $4e^-/2e^-$

selectivity of ORR can be judged by comparing the potential barriers of reactions 1 and 2, that are, the potential barriers of *O formation and *H_2O_2 formation, respectively, with one H_2O molecule. If reaction 1 is more favorable than reaction 2 (*i.e.* $\Delta G_O + \Delta G(H_2O) < \Delta G(H_2O_2)$), that is, $\Delta G_O <$ 3.52 eV ($\Delta G(H_2O_2)$-$\Delta G(H_2O)$),[5] then the selectivity toward $4e^-$ ORR can be ensured. It should be noticed that our analysis is based on thermodynamics following the conventional model. Recently, Zhao *et al.* developed an advanced 'constant-potential hybrid-solvation dynamic model' to study the $2e^-/4e^-$ selectivity of ORR by considering the reaction kinetics at the solid-water interface.[6] Their results indicate that the ORR selectivity also roots in the atomic-scale reaction kinetics and is dependent on pH and potential (which is related to the different proton affinity to two O in *OOH).[6] However, this model needs extensive grand-canonical DFT and AIMD simulations considering the explicit water molecules and implicit solution modelling, which is not suitable for our screening due to the high time cost. As such, for ORR selectivity, we only apply the thermodynamic analysis in this work using the conventional model.

For the density of states (DOS) calculations, we have employed the DFT(PBE)+U scheme with Dudarev's rotational approach[7, 8] for localized 3*d* electrons in the fourth-period TM@C_2N following the computational scheme for SACs in ref. 4. The values of U-J for Sc-Zn are set as 2.11, 2.58, 2.72, 2.79, 3.06, 3.29, 3.42, 3.4, 3.87, and 4.12 eV.[4] A much denser 5×5×1 Monkhorst-Pack *k*-points mesh was used in the DOS calculations. Note that for adsorbates calculations we do not employ DFT+U, since several previous reports suggest that DFT+U results may contradict experimental results or result in unreliable results in adsorption energy,[9, 10] and PBE can effectively describe the catalytic activity of SACs.

ESI Note 2. Investigation of thermodynamic and electrochemical stability of TM@C_2N SACs

We define formation energy E_f as $E_f = E(TM@C_2N) - E(TM\_bulk) - E(C_2N)$, where $E(TM@C_2N)$, E(TM_bulk), and $E(C_2N)$ represent the DFT-calculated energies of TM@C_2N SAC, single TM atom in its most stable bulk structure, and C_2N monolayer, respectively. By calculating E_f, we can investigate the thermodynamic stability of TM@C_2N by comparing the binding and cohesive ability of single TM atoms on C_2N. To be more specific, if $E_f<0$, the binding of single TM atoms on C_2N is more favorable than their aggregating into clusters, so that the thermodynamic stability of TM@C_2N can be confirmed.

For the electrochemical stability, we use the dissolution potential (U_{diss}) versus standard hydrogen electrode (SHE) as an indicator. U_{diss} is defined as $U_{diss} = U^0_{diss} - E_f/eN_e$, where U^0_{diss} and N_e represent the standard dissolution potential of bulk metal and the number of electrons in the dissociation process, respectively. Positive U_{diss} values suggest electrochemical stable systems, where the dissolution of single TM atoms can be neglected in the electrochemical conditions.[5, 11]

All the results listed in Table S5 indicate that among 27 TM@C_2N SACs studied, 20 SACs (except for Mn, Mo, Ru, W, Os, Ir, and Pt@C_2N) exhibit thermodynamic stability; further screening using U_{diss} suggest that Ni, Cu, Rh, Pd, Ag, Cd, Hf, and Au@C_2N are stable in an electrochemical environment.

ESI Note 3. Details of machine learning

Machine learning (ML) was performed in the scikit-learn package,[12] and the random forest algorithm was chosen to train the ML model. To enlarge the chemical space for ML, except TM@C_2N with TM-N_2 coordination, we also consider the SACs with N_1C_1 and N_1S_1 coordination (Fig. S10), and the TM@C_2N dataset is amplified by a factor of 2, so the total number of data in the dataset becomes 108. 20% of the data were randomly chosen as the test set and the rest of them formed the training set. We have tried two parameters in the random forest-based ML modelling: the maximum depth of the tree and the number of trees in the forest. By using the grid search optimization method provided in the scikit-learn package, we identify the optimized parameters (maximum depth of the tree as 22 and the number of trees in the forest as 31) in our ML model. Here, we set the range of grid search for maximum depth of tree as 2-30, and that for the number of trees in the forest as 1-50. To analyze the obtained ML model and exclude the influence of random splitting of training/test set to a larger extent, we also performed a cross-validation test using the k-fold strategy (number of folds set as 5 and allow shuffling), where an average cross-validation score of 0.903 can be obtained (5 cross-validation scores: 0.869, 0.910, 0.882, 0.937, 0.919). The relatively lower score than the training and test set can be attributed to the size of our SACs dataset, which will be increased in our future work. A representative python script for training the ML model was provided in the Appendix of ESI.

ESI Note 4. Feasibility of the extension of our methodology to SACs for other reactions and double-atom catalysts (DACs)

In this work, we combine the DFT calculations with machine learning for unravelling the catalytic origin of bifunctional OER/ORR for SACs. We build a three-tier computational scheme 'catalytic activity-activity descriptor-intrinsic feature importance' for investigating the structure-activity relationship for SACs, which has high adaptability for the extension to SACs for other catalytic reactions, such as hydrogen evolution reaction, nitrogen and CO_2 reduction reaction. On the one hand, scaling relationships and activity descriptors based on the electronic structure or catalyst geometry can benefit the rational design and high-throughput screening of catalysts,[13] and on the other hand, ML can link the catalytic activity with intrinsic, element-based properties, and accelerate the catalyst design.

Recently, DACs have emerged as a new frontier of heterogeneous catalysis, and due to the synergistic effect of the metal dimer site and the atomically dispersed nature, DACs may exhibit higher activity while maintaining the advantages of SACs including high atomic utilization efficiency, high selectivity, and stability.[14] The extension of SACs to DACs is natural, and computationally different mechanisms or adsorption configurations may appear, but the computational scheme is universal. We recently noticed two theoretical works on DACs combining DFT with ML: Zhu and coworkers used DFT screening to identify CuFe and NiCu DACs with ORR performance surpassing that of Pt, and identified several descriptors (electron affinity, electronegativity, and atomic radii) to describe the activity by ML;[15] Deng *et al.* also used DFT and ML to study ORR performance on a group of homo- and hetero-nuclear DACs, which is governed by geometric parameters.[16]

Table S1. DFT-calculated total energy E_{DFT}, zero point energy (ZPE), entropic contribution term TS (under the temperature of 298 K), and Gibbs free energy G for gases and adsorbates in the OER/ORR process. The entropy values for gases are taken from NIST database (https://doi.org/10.18434/T4D303). For adsorbates, only vibrational contribution is considered, and the contribution from frequency below 50 cm^{-1} is calculated by 50 cm^{-1}.[17]

Species	E_{DFT} (eV)	ZPE (eV)	TS (eV)	G (eV)
H_2 (g)	-6.767	0.269	0.404	-6.902
H_2O (g)	-14.151	0.567	0.670	-14.254
***O**	-	0.065	0.078	-
***OH**	-	0.316	0.145	-
***OOH**	-	0.420	0.197	-

Table S2. Optimized lattice parameters and magnetic moment for TM@C_2N.

TM@C_2N	*a* (Å)	*b* (Å)	TM@C_2N	*a* (Å)	*b* (Å)
Sc	8.186	8.207	**Ru**	8.254	8.324
Ti	8.145	8.247	**Rh**	8.259	8.329
V	8.193	8.262	**Pd**	8.295	8.302
Cr	8.257	8.270	**Ag**	8.284	8.285
Mn	8.278	8.287	**Cd**	8.249	8.255
Fe	8.287	8.296	**La**	8.245	8.248
Co	8.304	8.289	**Hf**	8.138	8.163
Ni	8.300	8.301	**Ta**	8.092	8.206
Cu	8.297	8.301	**W**	8.065	8.232
Zn	8.283	8.287	**Re**	8.120	8.345
Y	8.184	8.187	**Os**	8.138	8.418
Zr	8.144	8.164	**Ir**	8.143	8.424
Nb	8.094	8.196	**Pt**	8.170	8.412
Mo	8.112	8.250	**Au**	8.288	8.295

Table S3. Magnetic moment for $TM@C_2N$.

$TM@C_2N$	Magnetic moment	$TM@C_2N$	Magnetic moment
Sc	0	**Ru**	0.47
Ti	0	**Rh**	0
V	2.73	**Pd**	1.10
Cr	3.27	**Ag**	0
Mn	0	**Cd**	0
Fe	1.86	**La**	0
Co	0.07	**Hf**	1.24
Ni	1.27	**Ta**	1.19
Cu	0	**W**	0
Zn	0	**Re**	2.29
Y	0	**Os**	0.09
Zr	1.20	**Ir**	0
Nb	0	**Pt**	0
Mo	2.44	**Au**	0

Table S4. DFT-calculated adsorption Gibbs free energy for *O, *OH, and *OOH on TM@C_2N. SACs on which *OOH dissociation happens are denoted in italic.

TM@C_2N	ΔG_O (eV)	ΔG_{OH} (eV)	ΔG_{OOH} (eV)	TM@C_2N	ΔG_O (eV)	ΔG_{OH} (eV)	ΔG_{OOH} (eV)
Sc	-0.493	-1.714	1.772	**Ru**	0.685	-0.007	2.836
Ti	-2.124	-1.607	*-0.834*	**Rh**	1.877	0.558	3.480
V	-1.184	-0.864	*0.082*	**Pd**	2.765	0.828	3.844
Cr	2.861	-0.223	*3.323*	**Ag**	4.260	1.915	4.774
Mn	-0.488	-0.578	2.297	**Cd**	3.303	0.532	3.720
Fe	0.882	-0.180	2.858	**La**	0.715	-1.157	2.110
Co	1.106	0.061	2.980	***Hf***	-2.407	-2.179	*-1.362*
Ni	2.304	0.520	3.582	***Ta***	-2.155	-1.391	*-2.295*
Cu	2.664	0.721	3.790	***W***	-1.895	-2.244	*-2.318*
Zn	2.275	-0.482	2.911	**Re**		*	
Y	0.388	-1.622	1.672	**Os**	0.331	-0.231	2.652
Zr	-2.354	-1.904	*-1.172*	**Ir**	1.277	0.090	3.088
Nb	-2.152	-1.532	*-1.857*	**Pt**	1.725	0.095	3.297
Mo	-1.081	-0.689	*-0.874*	**Au**	3.012	0.989	4.072

*The values for Re@C_2N are not listed due to the failure of convergence in the DFT calculations.

Table S5. Formation energy (E_f), standard dissolution potential (U^0_{diss}), number of electrons involved in the dissolution of pure metals (N_e), and calculated dissolution potential for TM@C_2N SACs (U_{diss}, *vs.* SHE). U^0_{diss} values are extracting from references [5, 18]. Computational details are summarized in ESI Note 2.

TM@C_2N	E_f (eV)	U^0_{diss} (V)	N_e	U_{diss} (V)
Sc	-4.57	-2.08	3	-0.56
Ti	-2.48	-1.63	2	-0.39
V	-1.49	-1.13	2	-0.38
Cr	-1.19	-0.91	2	-0.31
Mn	0.44	-1.19	2	-1.41
Fe	-0.35	-0.44	2	-0.27
Co	-0.03	-0.28	2	-0.26
Ni	-0.56	-0.26	2	0.02
Cu	-0.65	0.34	2	0.66
Zn	-1.13	-0.76	2	-0.20
Y	-5.90	-2.37	2	-0.40
Zr	-3.72	-1.45	4	-0.52
Nb	-1.11	-1.10	3	-0.73
Mo	0.18	-0.20	3	-0.26
Ru	0.67	0.46	2	0.12
Rh	-0.17	0.60	2	0.69
Pd	-0.37	0.92	2	1.10
Ag	-1.26	0.80	1	2.06

Cd	-1.89	-0.40	2	0.54
La	-6.77	-2.38	3	-0.12
Hf	-3.28	1.55	4	2.37
Ta	-1.02	-0.60	3	-0.26
W	1.52	0.10	3	-0.41
Os	1.81	0.84	8	0.61
Ir	0.74	1.16	3	0.91
Pt	0.32	1.19	2	1.03
Au	-0.04	1.52	3	1.53

Table S6. All the feature values for ML modelling.

TM	Atomic number (Z)	Atomic radius (r, pm)	Outer electron number (n_e)	Electro-negativity (N)	First ionization energy (IE, kJ/mol)	Electron affinity (EA, kJ/mol)	Oxide formation enthalpy (Hox,f, eV)
Sc	21	144	3	1.36	633.1	18.1	-10.52
Ti	22	136	4	1.54	658.8	7.6	-11.45
V	23	125	5	1.63	650.9	50.6	-8.23
Cr	24	127	6	1.66	652.9	64.3	-6.39
Mn	25	139	7	1.55	717.3	-50	-5.66
Fe	26	125	8	1.83	762.5	15.7	-7.17
Co	27	126	9	1.88	760.4	63.7	-5.88
Ni	28	121	10	1.91	737.1	112	-5.09
Cu	29	138	11	1.9	745.5	118.4	-3.63
Zn	30	131	12	1.65	906.4	-58	-2.36
Y	39	162	3	1.22	600	29.6	-10.42
Zr	40	148	4	1.33	640.1	41.1	-13.09
Nb	41	137	5	1.6	652.1	86.1	-11.95
Mo	42	145	6	2.16	684.3	71.9	-7.63
Ru	44	126	8	2.2	710.2	101.3	-6.18
Rh	45	135	9	2.28	719.7	109.7	-4.93
Pd	46	131	10	2.2	804.4	53.7	-2.75

Ag	47	153	11	1.93	731	125.6	-1.01
Cd	48	148	12	1.69	867.8	-68	-1.14
La	57	169	3	1.1	538.1	48	-10.42
Hf	72	150	4	1.3	658.5	17.2	-13.60
Ta	73	138	5	1.5	761	31	-13.91
W	74	146	6	2.36	770	78.6	-9.63
Os	76	128	8	2.18	840	106.1	-6.17
Ir	77	137	9	2.2	880	151	-6.52
Pt	78	128	10	2.28	870	205.3	-3.33
Au	79	144	11	2.54	890.1	222.8	-1.08

Table S6. (continued)

Coordination	N_{sum}
N_2 (Pristine TM@C_2N)	6.08
N_1C_1	5.59
N_1S_1	5.62

Part of the data in Table S6 is extracted from the websites (Wikipedia, ptable.com) and ref.[19]. $H_{ox,f}$ values are calculated based on the DFT-calculated energies of TM crystals (based on our calculations) and the most stable bulk structures of TM oxides (from the Materials Project database calculated using DFT+U[5, 20]).

Table S7. Indicators for the ML model, including R^2 value, mean absolute error (MAE), and rooted mean square error (RMSE).

Dataset	Training set	Test set
R^2 value	0.985	0.981
MAE (eV)	0.143	0.191
RMSE (eV)	0.227	0.250

Table S8. ΔG_O and η_{sum} values for all the SACs as dataset in ML modelling.

	TM@C_2N		**N_1C_1**		**N_1S_1**	
TM	**ΔG_O (eV)**	**η_{sum} (V)**	**ΔG_O (eV)**	**η_{sum} (V)**	**ΔG_O (eV)**	**η_{sum} (V)**
Sc	-0.493	4.863	0.085	4.590	-0.43	3.830
Ti	-2.124	7.361	-2.16	7.423	-1.889	6.584
V	-1.184	5.702	-1.33	5.751	-1.125	5.903
Cr	2.861	3.307	0.28	3.129	0.107	3.398
Mn	-0.488	3.364	0.555	2.727	0.405	2.739
Fe	0.882	2.242	0.706	2.233	0.675	2.175
Co	1.106	1.879	1.549	1.950	1.435	1.712
Ni	2.304	1.264	1.841	1.577	2.299	1.487
Cu	2.664	1.222	2.802	1.230	0.85	2.894
Zn	2.275	3.239	2.537	2.751	0.454	2.970
Y	0.388	7.996	0.85	4.483	0.266	5.151

Zr	-2.354	8.309	-2.259	8.008	-2.315	8.350
Nb	-2.152	6.483	-2.44	8.971	-1.958	8.285
Mo	-1.081	2.158	-1.258	6.773	-1.296	6.060
Ru	0.685	1.046	0.634	2.198	0.985	1.905
Rh	1.877	1.109	2.205	0.713	2.077	0.889
Pd	2.765	2.199	2.04	1.547	2.605	1.261
Ag	4.26	8.461	3.795	1.767	1.329	3.228
Cd	3.303	8.606	2.183	0.972	0.698	2.580
La	0.715	3.967	1.158	7.813	0.362	8.799
Hf	-2.407	9.482	-2.315	8.564	-2.375	9.096
Ta	-2.155	2.552	-2.404	8.959	-2.338	8.854
W	-1.895	1.743	-1.667	9.067	-1.778	2.485
Os	0.331	1.535	-0.05	2.822	0.533	1.641
Ir	1.277	1.176	1.472	2.089	1.454	2.034
Pt	1.725	4.870	2.003	1.851	2.189	1.345
Au	3.012	2.354	2.582	0.750	3.173	4.573

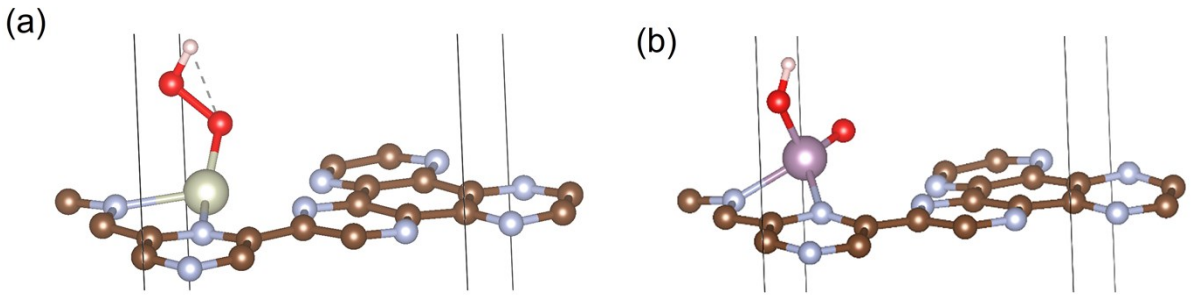

Fig. S1 Optimized structures of *OOH adsorbed on (a) Rh@C_2N and (b) Mo@C_2N. C, N, O, H, Rh, and Mo atoms are represented in brown, cyan, red, pale pink, light green, and purple, respectively. The *OOH radical dissociates on the latter (a stable *OOH adsorption geometry cannot be obtained) and keeps undissociated on the former.

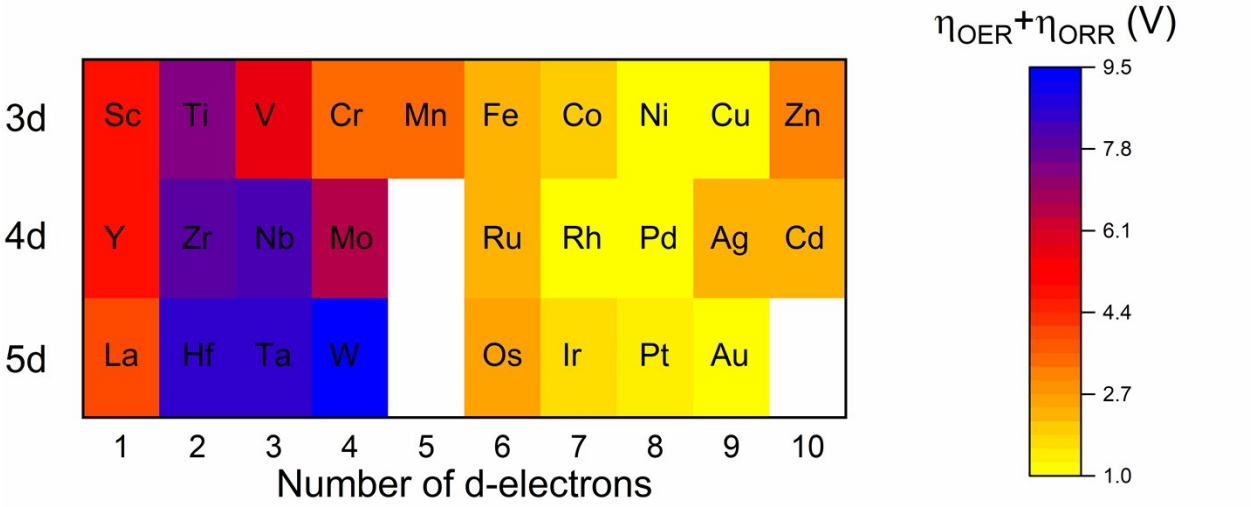

Fig. S2 Heatmap of $\eta_{sum}=\eta_{OER}+\eta_{ORR}$ on 3*d*-5*d* TM@C_2N.

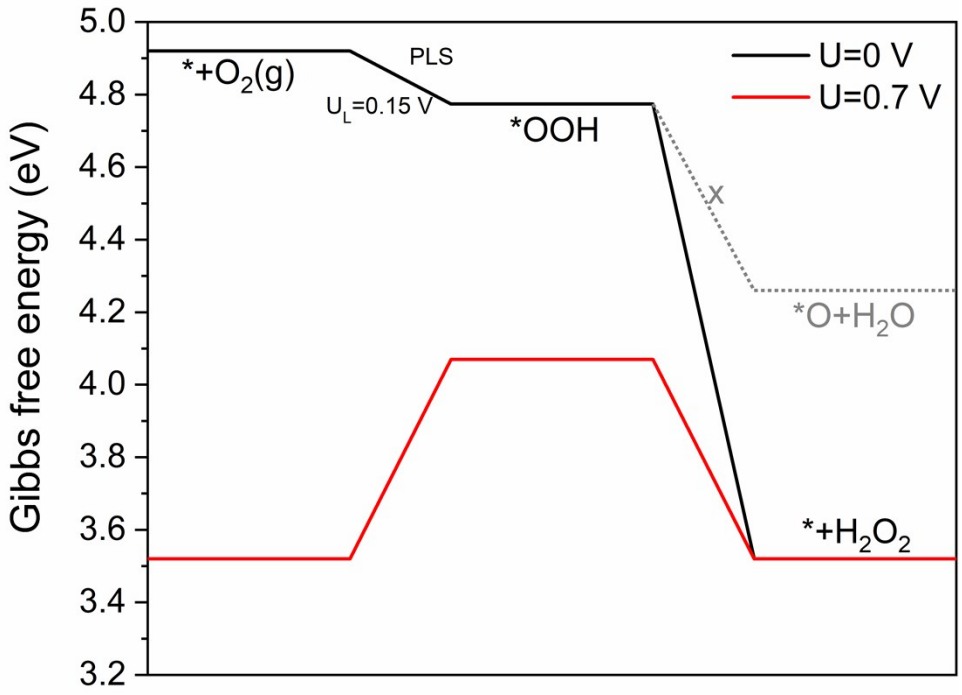

Fig. S3 Gibbs free energy diagram for $2e^-$ ORR toward H_2O_2 on $Ag@C_2N$. The potential-limiting step (PLS) and the limiting potential U_L are marked in the figures.

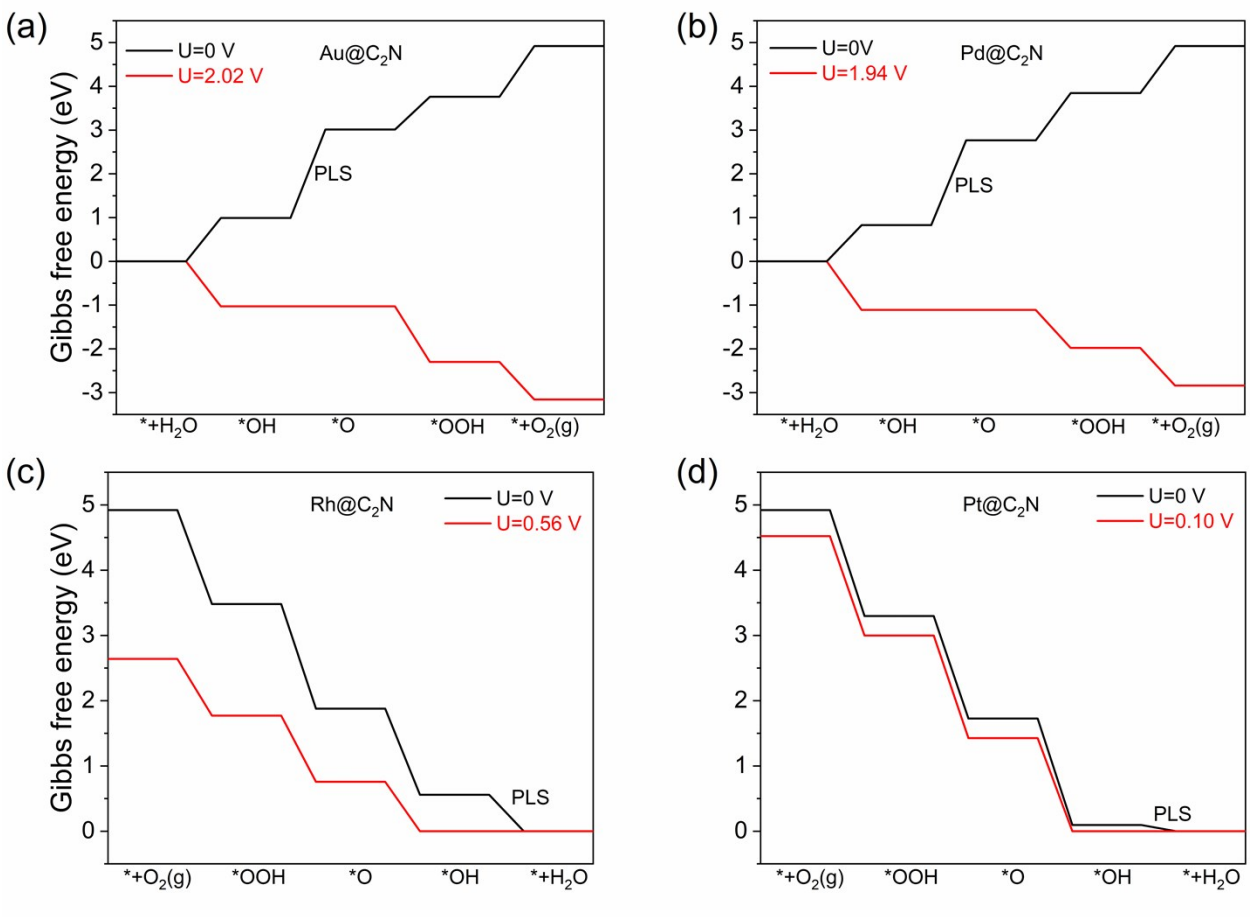

Fig. S4 Gibbs free energy diagram for OER on (a) Au and (b) Pd@C_2N and ORR on (c) Rh and (d) Pt@C_2N at zero and applied electrode potential U in acidic medium. The potential-limiting steps (PLS) are marked in the figures.

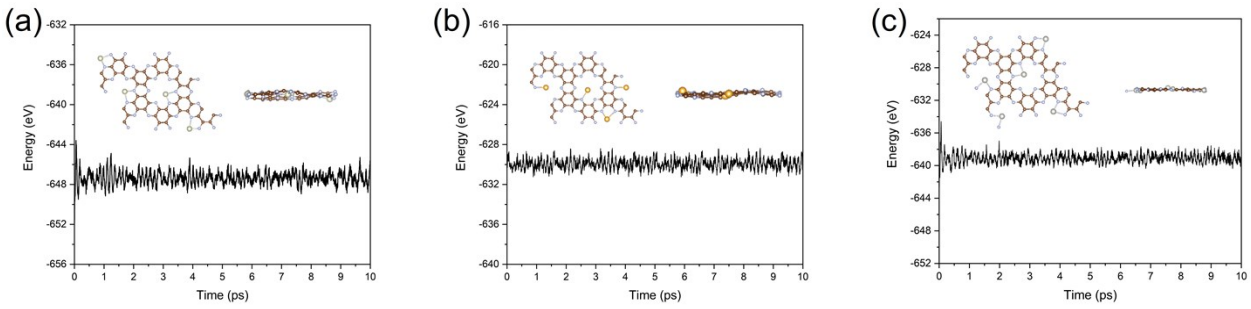

Fig. S5 Total energy as a function of time during the AIMD simulations for (a) Rh, (b) Au, and (c) Pd@C_2N at 500 K. Inset shows the side and top views of the structures after 10 ps AIMD simulations. 2×2 supercells are used in the simulations.

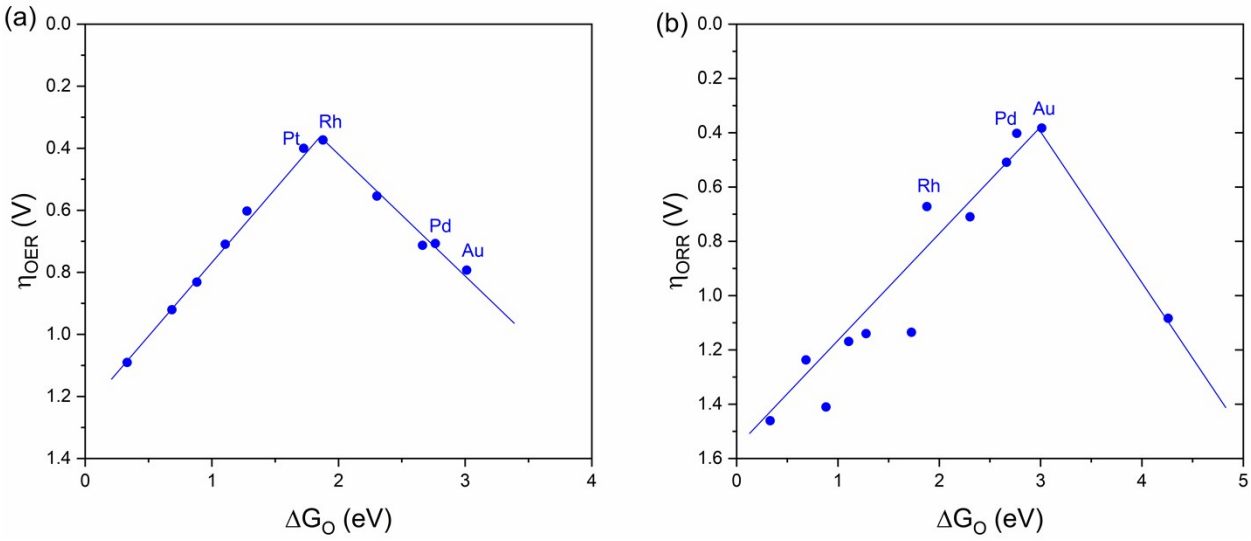

Fig. S6 Volcano-shaped plot of (a) OER and (b) ORR overpotential as a function of ΔG_O for d^7-d^9 TM@C_2N.

The reason that the highest OER and ORR activities cannot coexist in the same SAC can be attributed to the different reaction mechanisms and hence different peak positions in the volcano-shaped relationship for OER and ORR (see Fig. S6 and discussions in the main text). Indeed, none of the TM@C_2N SACs are better than both IrO_2(110) for OER (0.56 V) and Pt(111) for ORR (0.45 V). However, according to our calculations, Rh@C_2N, except for its outstanding catalytic activity for OER (η_{OER} of 0.37 V), has an ORR overpotential (0.67 V) only slightly higher than that for Pt(111), and Au/Pd@C_2N with low ORR overpotential (0.38/0.40 V) also has relatively low η_{OER} of 0.79/0.71 V. Note that there are discrepancies between DFT-calculated and experimentally measured OER/ORR overpotentials due to the simplified DFT model. After comparing with experimental reports, we found that the η_{ORR} for Rh@C_2N is close to that for Fe-single atoms supported on S, N-codoped carbon (~0.65 V under DFT-PBE level) with high experimental ORR activity (half-wave potential of 0.896 V),[21] and η_{OER} for Au/Pd@C_2N is smaller than that for all

four Mn-N_xC_y SAC models (the best one is Mn-N_2C_2 with 0.871 V under DFT-PBE level), while Mn-N_2C_2 SACs are proved to be efficient OER/ORR catalysts in the experiments.[22] Therefore, we are confident to conclude that Rh, Au, and Pd@C_2N are outstanding bifunctional OER/ORR SACs.

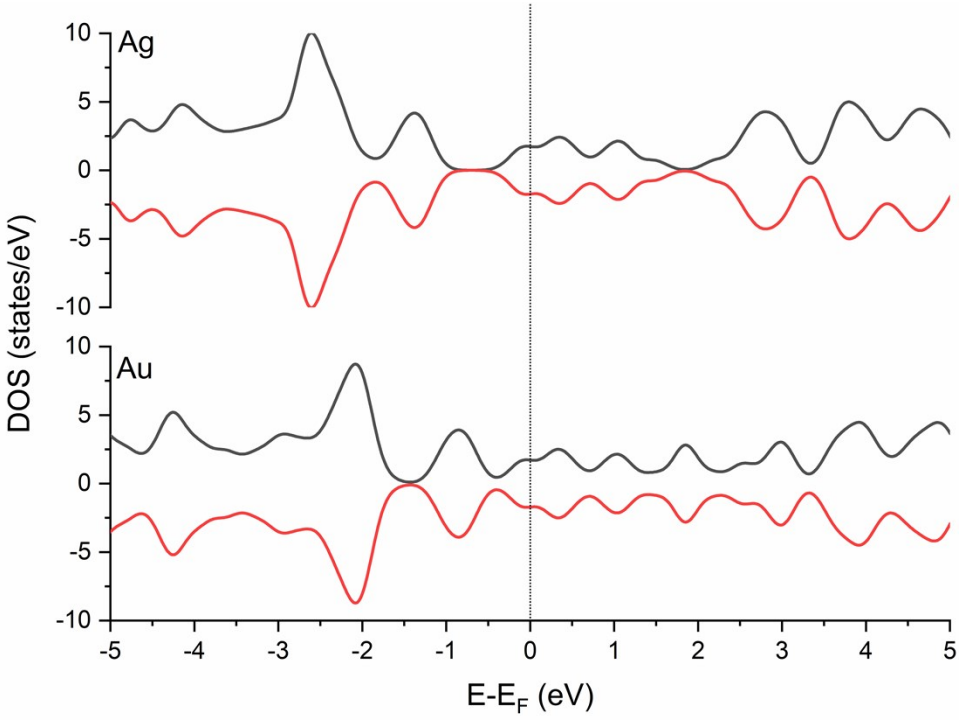

Fig. S7 Total density of states (DOSs) for Ag and Au@C_2N. Fermi level is set to zero. Contributions from spin-up and spin-down channels are denoted in black and red, respectively.

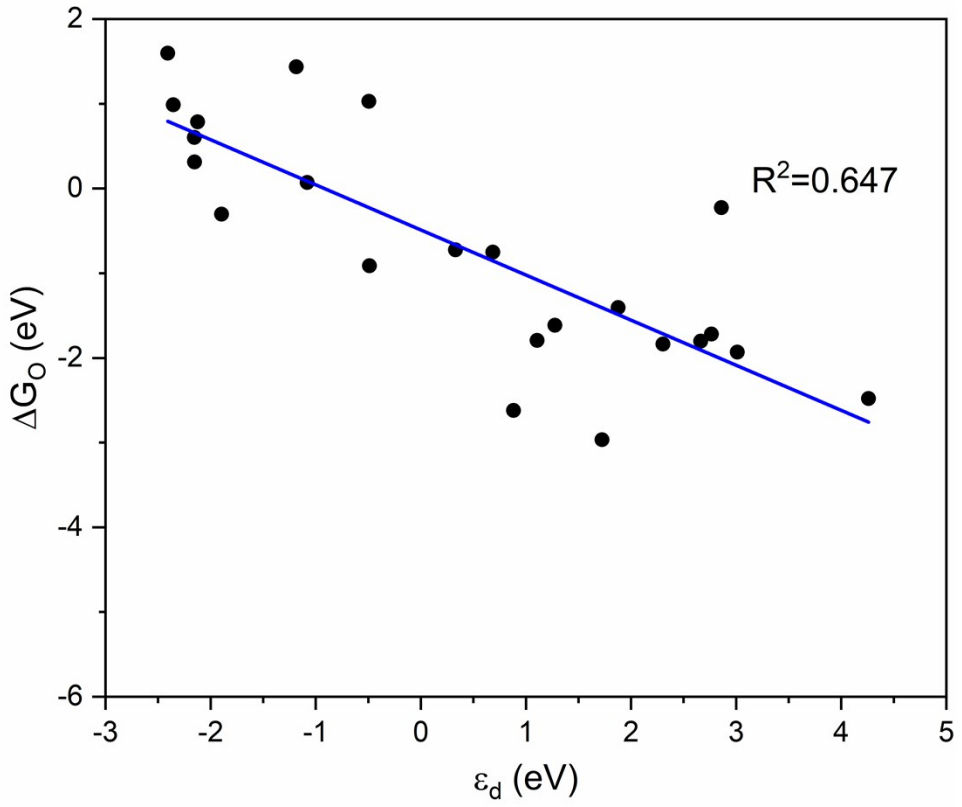

Fig. S8 Relationship between ΔG_O and *d*-band center ε_d for TM@C_2N. ε_d is defined as

$$\varepsilon_d = \frac{\int_{-\infty}^{+\infty} ED(E)dE}{\int_{-\infty}^{+\infty} D(E)dE}$$

where D(E) is TM *d*-orbital PDOS as a function of energy (as shown in Fig. 4a).

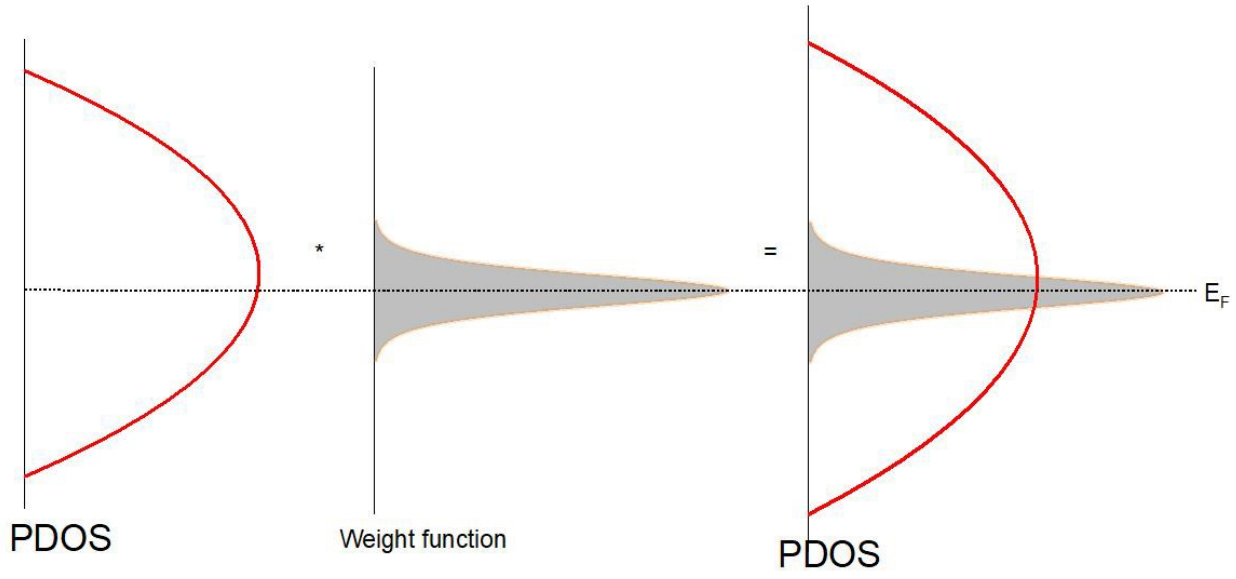

Fig. S9 Schematic illustration of the effect of weight function on the partial density of states of *d*-orbital D(E) of catalysts.

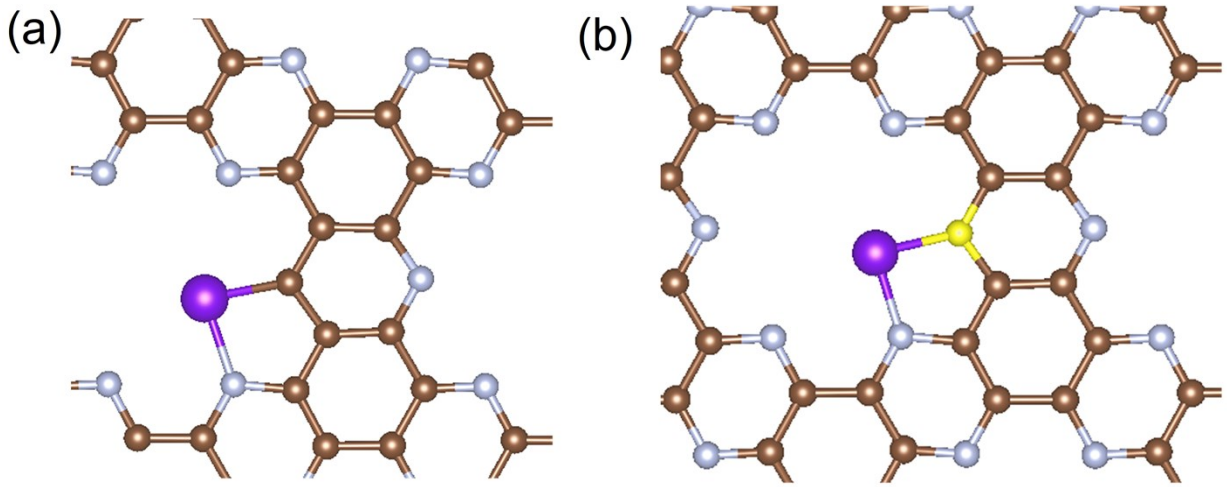

Fig. S10 Geometric structure of SACs with (a) N_1C_1 and (b) N_1S_1 coordinations. C, N, S, and TM atoms are represented in brown, cyan, yellow, and purple, respectively.

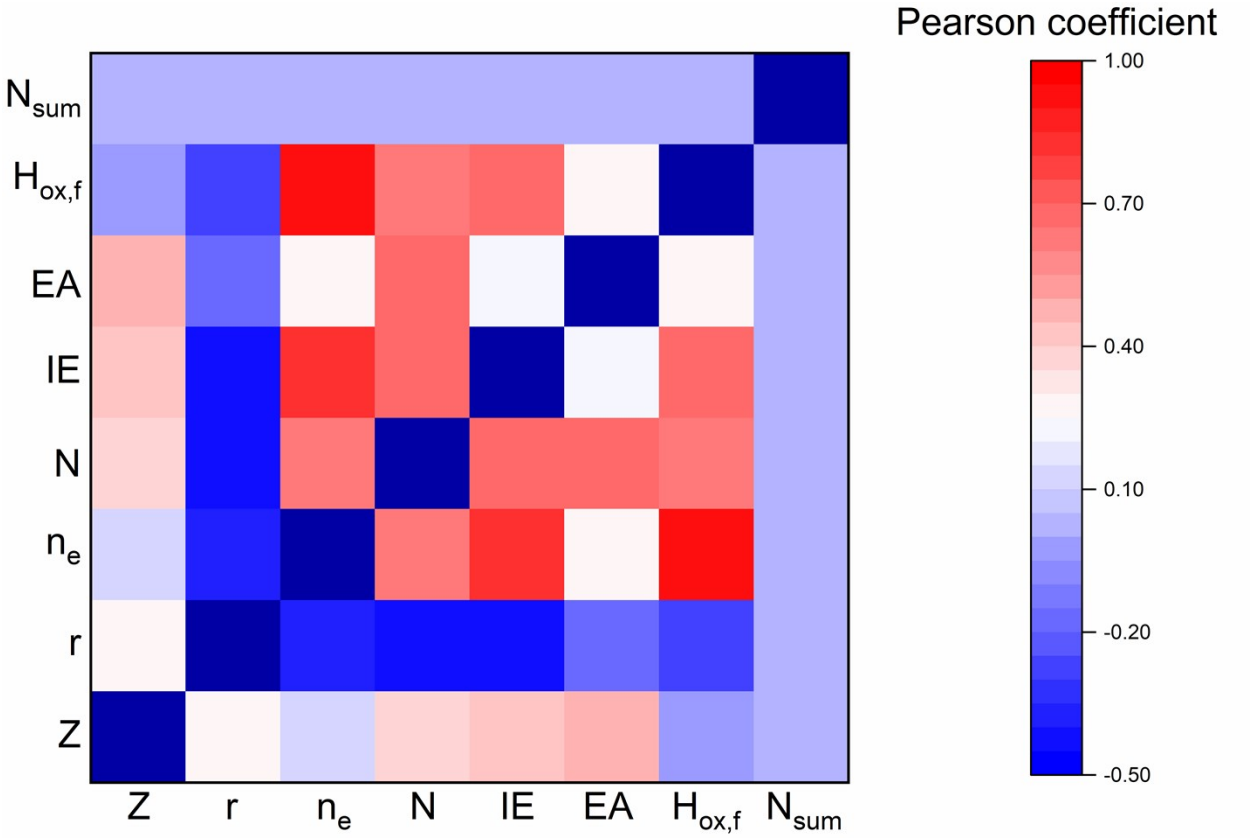

Fig. S11 Pearson correlation heatmap for features including atomic number (Z), atomic radius (r), outer electron number (n_e), electronegativity (N), first ionization energy (IE), electron affinity (EA), single-atom oxide formation enthalpy ($H_{ox,f}$), and the sum of electronegativity of coordinated atoms of TM (N_{sum}) in our ML model.

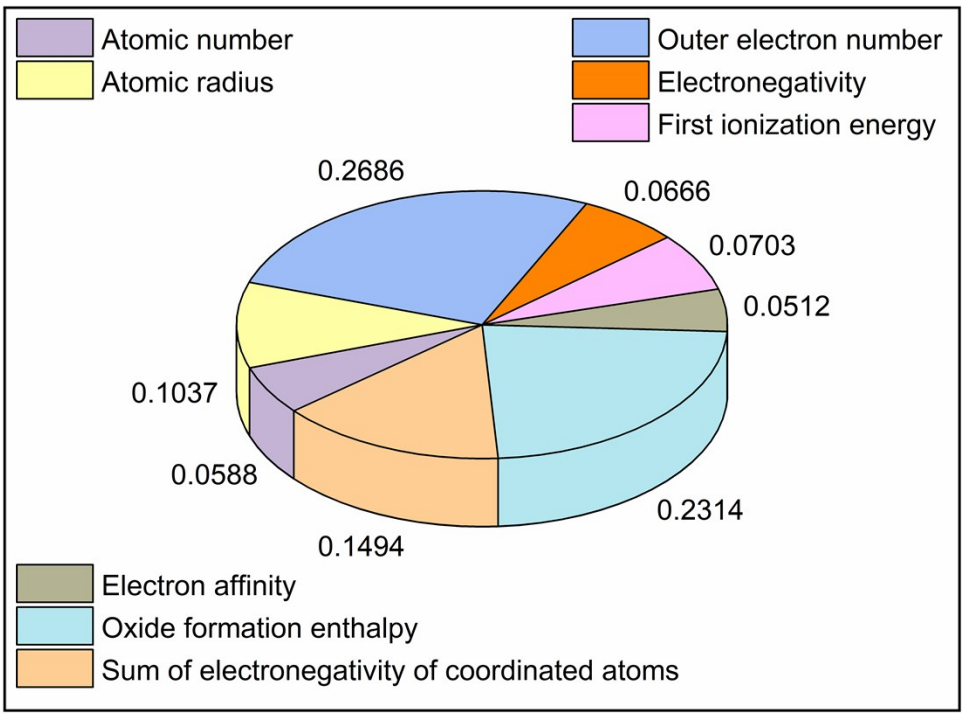

Fig. S12 Pie chart for feature importance analysis in the ML model for η_{sum}.

Appendix. A representative python script for training the ML model.

```
import numpy as np
import sklearn
from sklearn import preprocessing
#print (sklearn.__version__)
import pandas as pd

pd.set_option('display.width', 1000)
pd.set_option('display.max_columns', None)
pd.set_option('display.max_rows', None)

from sklearn.ensemble import RandomForestRegressor
from sklearn.metrics import mean_squared_error,r2_score,mean_absolute_error
import joblib
from sklearn.feature_selection import VarianceThreshold

from scipy.stats import pearsonr

from sklearn.preprocessing import StandardScaler

from sklearn.model_selection import train_test_split
from sklearn.model_selection import cross_val_score
from sklearn.model_selection import KFold
from sklearn.model_selection import GridSearchCV

dataset_url = 'Catalysis.csv'
data = pd.read_csv(dataset_url)
#print (data.head())
#print (data.shape)
#print (data.describe())

#transfer = VarianceThreshold()
#new_data = transfer.fit_transform(data)
#print('new_data:\n',new_data,new_data.shape)

data_title = data.columns.values.tolist()
r = []
for i in range(len(data_title)):
  for j in range(len(data_title)):
      #print('The Pearson correlation coefficient between %s and %s is
%f'%(data_title[i],data_title[j],pearsonr(data[data_title[i]],data[data_title[j]])[0]))
      if i==j:
         continue
      else:
         print(i,j,pearsonr(data[data_title[i]],data[data_title[j]])[0])
      r.append(pearsonr(data[data_title[i]],data[data_title[j]])[0])

#for i in range(len(r)):
#    if r[i]>0.6:
#        print(r[i])
```

```
#       print(i)
#   else:
#       continue

y = data.Quality
x = data.drop('Quality', axis=1)

x_train,x_test,y_train,y_test = train_test_split(x,y,test_size =0.2,random_state = 9)

transfer = StandardScaler()
x_train = transfer.fit_transform(x_train)
x_test = transfer.transform(x_test)

#print(y_test)
#print(x_test)

forest = RandomForestRegressor()
param_dict = {'n_estimators':[i for i in range(1,50,30)],'max_depth':[i for i in range(2,30,20)]}
forest = GridSearchCV(forest,param_grid=param_dict,cv = 3)
forest.fit(x_train, y_train)

print(y_train)
print(forest.predict(x_train))
print(y_test)
print(forest.predict(x_test))

score1 = forest.score(x_train,y_train)
print('Train:\n',score1)
score1 = forest.score(x_test,y_test)
print('Test:\n',score1)

print(forest.best_estimator_)
#joblib.dump(forest,"catalysts.pkl")

forest=RandomForestRegressor(n_estimators=31,max_depth=22)
forest.fit(x_train, y_train)

score1 = forest.score(x_train,y_train)
print('Train:\n',score1)
score1 = forest.score(x_test,y_test)
print('Test:\n',score1)
print(forest.feature_importances_)
joblib.dump(forest,"catalysts.pkl")
print('\nR2 value: Train:')
print(r2_score(y_train,forest.predict(x_train)))
print('Test:')
print(r2_score(y_test,forest.predict(x_test)))
print('RMSE: Train:')
print(mean_squared_error(y_train,forest.predict(x_train))**0.5)
print('Test:')
print(mean_squared_error(y_test,forest.predict(x_test))**0.5)
```

```
print('MAE: Train:')
print(mean_absolute_error(y_train,forest.predict(x_train)))
print('Test:')
print(mean_absolute_error(y_test,forest.predict(x_test)))

kf=KFold(n_splits=5,shuffle=True)
cv_result=cross_val_score(forest,x,y,cv=kf)
print(cv_result)
```

Supplementary References

1. I. C. Man, H. Y. Su, F. Calle-Vallejo, H. A. Hansen, J. I. Martínez, N. G. Inoglu, J. Kitchin, T. F. Jaramillo, J. K. Nørskov and J. Rossmeisl, *ChemCatChem*, 2011, **3**, 1159-1165.
2. J. K. Nørskov, J. Rossmeisl, A. Logadottir, L. Lindqvist, J. R. Kitchin, T. Bligaard and H. Jonsson, *J. Phys. Chem. B*, 2004, **108**, 17886-17892.
3. Q. Liang, G. Brocks and A. Bieberle-Hütter, *J. Phys. Energy*, 2021, **3**, 026001.
4. H. Xu, D. Cheng, D. Cao and X. C. Zeng, *Nat. Catal.*, 2018, **1**, 339-348.
5. X. Guo, S. Lin, J. Gu, S. Zhang, Z. Chen and S. Huang, *ACS Catal.*, 2019, **9**, 11042-11054.
6. X. Zhao and Y. Liu, *J. Am. Chem. Soc.*, 2021, **143**, 9423-9428.
7. V. V. Anisimov, J. Zaanen and O. K. Andersen, *Phys. Rev. B*, 1991, **44**, 943-954.
8. S. Dudarev, G. Botton, S. Savrasov, C. Humphreys and A. Sutton, *Phys. Rev. B*, 1998, **57**, 1505.
9. Y. Wang, Y. J. Tang and K. Zhou, *J. Am. Chem. Soc.*, 2019, **141**, 14115-14119.
10. X. Guo, S. Lin, J. Gu, S. Zhang, Z. Chen and S. Huang, *Adv. Funct. Mater.*, 2021, **31**, 2008056.
11. J. Greeley and J. K. Nørskov, *Electrochim. Acta* 2007, **52**, 5829-5836.
12. F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss and V. Dubourg, *J. Mach. Learn. Res.*, 2011, **12**, 2825-2830.
13. Z.-J. Zhao, S. Liu, S. Zha, D. Cheng, F. Studt, G. Henkelman and J. Gong, *Nat. Rev. Mater.*, 2019, **4**, 792-804.
14. Y. Ying, X. Luo, J. Qiao and H. Huang, *Adv. Funct. Mater.*, 2021, **31**, 2007423.
15. X. Zhu, J. Yan, M. Gu, T. Liu, Y. Dai, Y. Gu and Y. Li, *J. Phys. Chem. Lett.* , 2019, **10**, 7760-7766.
16. C. Deng, Y. Su, F. Li, W. Shen, Z. Chen and Q. Tang, *J. Mater. Chem. A*, 2020, **8**, 24563-24571.
17. V. Wang, N. Xu, J.-C. Liu, G. Tang and W.-T. Geng, *Comput. Phys. Commun.*, 2021, **267**, 108033.
18. D. R. Lide, *CRC handbook of chemistry and physics*, CRC press, 2004.
19. S. G. Bratsch and J. Lagowski, *Polyhedron*, 1986, **5**, 1763-1770.
20. A. Jain, S. P. Ong, G. Hautier, W. Chen, W. D. Richards, S. Dacek, S. Cholia, D. Gunter, D. Skinner, G. Ceder and K. A. Persson, *APL Materials*, 2013, **1**, 011002.
21. Q. Li, W. Chen, H. Xiao, Y. Gong, Z. Li, L. Zheng, X. Zheng, W. Yan, W. C. Cheong, R. Shen, N. Fu, L. Gu, Z. Zhuang, C. Chen, D. Wang, Q. Peng, J. Li and Y. Li, *Adv. Mater.* , 2018, **30**, 1800588.
22. H. Shang, W. Sun, R. Sui, J. Pei, L. Zheng, J. Dong, Z. Jiang, D. Zhou, Z. Zhuang, W. Chen, J. Zhang, D. Wang and Y. Li, *Nano Lett.* , 2020, **20**, 5443-5450.