Supporting Information

Machine learning high-throughput screening of rare earth SACs with different coordination environments for HER

Meiling Liu [1, a], Qiming Fu [1, a], Wei Zhong [a], Shaik Gouse Peera[b, *], Chao Liu [a,c, *]

[a] Faculty of Materials Metallurgy and Chemistry, Jiangxi University of Science and Technology, Ganzhou 341000, People's Republic of China

[b] Department of Environmental Science, Keimyung University, Daegu 42601, Republic of Korea

[c] Key Laboratory of Development and Application of Ionic Rare Earth Resources, Ministry of Education, Jiangxi University of Science and Technology, Ganzhou 341000, China

[1] The authors contributed equally to this work

[*] Corresponding Author' E-mail address:

E-mail: gouse@kmu.ac.kr (Shaik Gouse Peera) and liuchao198967@126.com (Chao Liu)

1 Structural Model

For double vacancies configuration, a total of 7 distinct configurations were considered. There is one configuration each for coordination N_0, N_1, N_3, and N_4. Coordination N_2 has three types: REN_2C_2-hex is characterized by two adjacent N atoms in a hexagonal ring; REN_2C_2-oppo has two N atoms arranged opposite each other; REN_2C_2-pen has two adjacent N atoms in a pentagonal ring. For the four vacancies, 16 configurations were explored. Among them, the coordination types N_0, N_1, N_5, and N_6 have only one configuration. Coordination type N_2 has four types: N_2C_4-hex indicates that there are two adjacent N atoms in a hexagonal ring, N_2C_4-qr indicates that there are two adjacent N atoms in a quadrilateral ring, N_2C_4-ip indicates that the two N atoms are in an intermediate position, and N_2C_4-oppo indicates that there are two opposite N atoms. The coordination type N_4 configurations are similar. N_3 coordination also has four types: N_3C_3-sep indicates that the three N atoms are separate, N_3C_3-hex2 means that two of the N atoms are in a hexagonal ring, N_3C_3-qr2 means that two N atoms are in a quadrilateral ring, and N_3C_3-col indicates that the three N atoms are collected together.

2 Calculation method

All spin-unrestricted calculations were performed using the density functional theory framework.[1] The generalized gradient approximation (GGA) within the Perdew-Burke-Ernzerhof (PBE) functional was employed to describe the electronic exchange and correlation effects.[2] Relativistic effects are considered using the DFT semi-core pseudopotentials (DSPP) and double numerical polarization (DNP) basis sets.[3, 4]The Grimme method was used to investigate the interactions between the catalyst surface and gas molecules.[5] The Brillouin zone sampling adopts the Monkhorst Pack method with a 5 × 5 × 1 k-point grid, and the electronic performance calculation uses a 15 × 15 × 1 dense k-point grid. The convergence thresholds for energy, maximum force, and displacement were set to 1×10^{-5} Ha, 0.002 Ha·Å^{-1}, and 0.005 Å, respectively. To accelerate convergence, a smearing value of 0.005 Ha was applied. The dielectric

constant of water was modeled using the conductor-like screening model (COSMO), with a value of 78.54 $F \cdot m^{-1}$.[6]

3 Calculation formula

The binding energy (ΔE_b) can verify the stability of the catalyst, and the calculation formula is as follows:

$$\Delta E_b = E_{RE-N-C} - E_{N-C-Gra} - E_{RE} \quad (1)$$

Where $E_{RE\text{-}N\text{-}C}$, $E_{N\text{-}C\text{-}Gra}$ and E_{RE} are the energy of optimized RE-N-C, vacancy defect graphene and an isolated rare earth atom, respectively.

The adsorption energy (ΔE_{*H}) was calculated by:

$$\Delta E_{*H} = E_{*H} - E_{*} - E_{H} \quad (2)$$

Here, E_{*H}, E_{*} and E_H are the energy of the adsorption intermediate, the original catalyst and half of H_2 energy, respectively.

The Gibbs free energy of hydrogen adsorption (ΔG_{*H}) of each electron transfer step was calculated by the following formula:[7, 8]

$$\Delta G = \Delta E_{DFT} + \Delta E_{ZPE} - T\Delta S + \Delta G_U + \Delta G_{pH} \quad (3)$$

Where ΔE_{DFT}, ΔE_{ZPE} and ΔS are the reaction heat, zero-point energy change and entropy change for each electron step; the ΔG_U and ΔG_{pH} correspond to the effects of potential and pH, respectively. $\Delta G_U = -eU$ and $\Delta G_{pH} = pH \times k_B \times T \times \ln 10$, where U, T, pH, and k_B are the applied potential, temperature, hydrogen ion exponent and the Boltzmann constant, respectively. The temperature T was set to 298.15 K.

4 Machine learning

We employed the following features: coordination number (NC), the ratio of nitrogen atoms in the active site (Nr), the distance between hydrogen and rare earth (d), the atomic radius of metal atoms (R), covalent radius (Rc), atomic mass (AM), the number of d/f orbital electrons ($N_{d/f}$), Pauling electronegativity (P), first ionization energy (IE), electron affinity (EA), and the Coulomb matrix.[9, 10] The Coulomb matrix

is defined as follows:[11]

$$M_{ij} = \begin{cases} 0.5Z_i^{2.4}, & i = j \\ \dfrac{Z_i Z_j}{|R_i - R_j|}, & i \neq j \end{cases}$$

In this context, Z represents the atomic number, while R denotes the Cartesian coordinates of the atom. The diagonal elements correspond to the atomic energies derived from the fitted charges, while the off-diagonal elements represent the Coulomb repulsion between the respective atoms. Given the symmetry of the Coulomb matrix, only the upper triangular portion is considered. However, for elements involving double vacancies (10 total) and four vacancies (21 total), the dimensionalities are inconsistent and relatively high. Therefore, the average values of the two sections are computed, designated as CM_1 (off-diagonal) and CM_2 (diagonal). This reduction simplifies the Coulomb matrix to two dimensions. All of these features are readily accessible from the periodic table, publicly available literature, and geometric optimization.

The ΔG_{*H} for 100 sets of data was first calculated. The data were then randomly split into training and testing sets in a 4:1 ratio. Three different supervised learning algorithms were selected to build ML models, including Gradient Boosting Regression (GBR), Random Forest Regression (RFR) and Bootstrap Aggregating (Bagging). For each algorithm, grid search combined with manual parameter tuning was employed to obtain the optimal hyperparameters. The performance of the models was evaluated using the coefficient of determination (R^2) and root mean square error (RMSE). The robustness of the model is evaluated more objectively by using leave-one-out cross validation (LOOCV). Finally, the SHAP method was applied to interpret the model's predictions, quantifying the contribution of each feature to the target value and exploring the dependencies between them.

Data and feature engineering determine the upper limits of machine learning. Selecting a subset of features from the intrinsic properties of hundreds of materials, which form a compact and comprehensive feature set. Our approach primarily incorporates the following strategies: (1) These features provide a thorough description

of the catalyst's electronic structure and geometric properties; (2) They are easily collected through database consultation and simple geometric optimization; (3) They are intuitively linked to physical properties, aiding in the interpretation of the model's decision-making process.

5 Fractional Coordinates and Lattice Parameters of ScN_6C_0

The atomic coordinate information of the ScN_6C_0 structure is as follows:

Lattice_Cart

12.2974996567	0.0000000000	0.0000000000
-6.1487498283	10.6499471057	0.0000000000
0.0000000000	0.0000000000	30.0000000000

Positions_Frac

C	0.065627946	0.132599177	0.501440146
C	0.134281611	0.068127774	0.500860076
C	0.269193507	0.141912502	0.500413350
C	0.337288597	0.074575161	0.500188432
C	0.472721592	0.141913115	0.500411555
C	0.533848032	0.068127795	0.500858918
C	0.666973666	0.132600094	0.501439066
C	0.733607135	0.066835151	0.501771616
C	0.866674807	0.133348090	0.501941836
C	0.933229987	0.066835620	0.501771930
C	0.065655339	0.332948257	0.501389487
C	0.129517581	0.264738608	0.501318538
C	0.261344133	0.330678703	0.500630545
C	0.669336019	0.330678753	0.500626679
C	0.735222960	0.264739094	0.501316471

C	0.867295219	0.332949184	0.501388216
C	0.933221558	0.266442131	0.501739026
C	0.066543913	0.533086308	0.500533328
C	0.134350519	0.466066762	0.500785179
C	0.269174218	0.527377588	0.500356426
C	0.858205517	0.527378646	0.500353794
C	0.931717725	0.466067081	0.500783736
C	0.063126908	0.731546005	0.500087259
C	0.133158068	0.666315184	0.500133206
C	0.268419712	0.731545379	0.500087554
C	0.337224993	0.662779467	0.500147784
C	0.472694672	0.730810124	0.500363613
C	0.669346450	0.738692408	0.500561711
C	0.858115717	0.730810433	0.500363501
C	0.925555673	0.662780040	0.500146575
C	0.066557695	0.933360441	0.500629529
C	0.133202354	0.866805574	0.500163767
C	0.268500203	0.936998352	0.500102820
C	0.333604647	0.866805280	0.500163404
C	0.466804198	0.933360528	0.500628934
C	0.533929405	0.865687372	0.500856467
C	0.667022993	0.934403724	0.501445378
C	0.735251986	0.870502696	0.501294700
C	0.867381451	0.934403784	0.501445743
C	0.931759325	0.865687648	0.500856863
C	0.330095548	0.272121107	0.500290685

N	0.542026368	0.272120303	0.500286506
N	0.330121981	0.458043458	0.500280265
N	0.727922434	0.458042836	0.500275802
N	0.541934328	0.669833253	0.500206719
N	0.727900471	0.669833397	0.500206789
Sc	0.533374592	0.466749905	0.499989409

Table S1. Catalyst and their ΔG_{*H} for ML training in groups of 100 and ΔG_{*H} of ScN_6C_0

Catalysts	ΔE_{DFT}/eV	ΔE_{ZPE}/eV	T×ΔS/eV	ΔG_{*H}
DyN_0C_4	0.48	-0.24	-0.08	0.32
EuN_0C_4	0.57	-0.23	-0.18	0.53
ScN_0C_4	1.25	-0.12	-0.35	1.48
YN_0C_4	1.29	-0.03	-0.08	1.35
CeN_1C_3	1.29	-0.03	-0.19	1.48
LaN_1C_3	1.37	-0.04	-0.23	1.55
NdN_1C_3	1.35	-0.07	-0.19	1.46
YN_1C_3	1.32	-0.05	-0.15	1.42
EuN_2C_2-hex	1.25	-0.14	-0.23	1.34
GdN_2C_2-hex	1.24	-0.06	-0.17	1.36
LuN_2C_2-hex	0.89	-0.03	-0.13	0.99
YbN_2C_2-hex	1.13	-0.04	-0.17	1.26
CeN_2C_2-oppo	1.24	-0.06	-0.16	1.34
LuN_2C_2-oppo	1.04	-0.04	-0.13	1.13
NdN_2C_2-oppo	-1.52	0.32	-0.22	-0.99
ScN_2C_2-oppo	1.08	-0.03	-0.07	1.12
EuN_2C_2-pen	1.69	-0.05	-0.09	1.73
GdN_2C_2-pen	1.38	-0.03	-0.17	1.52
YN_2C_2-pen	1.30	-0.02	-0.18	1.45
YbN_2C_2-pen	1.56	-0.05	-0.02	1.53
EuN_3C_1	1.62	-0.05	-0.19	1.76
GdN_3C_1	1.04	-0.03	-0.14	1.16
LuN_3C_1	0.63	-0.01	-0.16	0.79
YN_3C_1	0.94	-0.02	-0.17	1.09
YbN_3C_1	1.39	-0.04	-0.17	1.52

DyN_4C_0	1.16	-0.01	-0.17	1.32
EuN_4C_0	1.42	-0.03	-0.15	1.54
HoN_4C_0	0.94	-0.02	-0.16	1.08
LaN_4C_0	0.80	0.00	-0.19	0.99
ScN_4C_0	0.19	0.01	-0.14	0.35
YN_4C_0	0.50	0.02	-0.21	0.72
YbN_4C_0	1.13	-0.03	-0.15	1.25
EuN_0C_6	1.27	-0.05	-0.18	1.40
GdN_0C_6	1.09	-0.04	0.18	0.87
LaN_0C_6	0.08	-0.20	-0.23	0.11
ScN_0C_6	0.99	0.00	-0.18	1.17
YbN_0C_6	0.89	-0.04	-0.12	0.97
CeN_1C_5	1.36	0.02	-0.21	1.58
GdN_1C_5	1.27	-0.03	-0.10	1.34
NdN_1C_5	1.41	-0.04	-0.17	1.55
ScN_1C_5	0.72	-0.03	-0.14	0.83
GdN_2C_4-hex	0.98	-0.17	-0.28	1.09
HoN_2C_4-hex	1.25	-0.11	-0.13	1.27
LuN_2C_4-hex	0.76	-0.06	-0.10	0.80
YN_2C_4-hex	1.06	-0.07	-0.24	1.23
EuN_2C_4-qr	1.83	-0.03	-0.17	1.97
NdN_2C_4-qr	1.53	-0.03	-0.19	1.69
YN_2C_4-qr	1.27	0.00	-0.15	1.42
YbN_2C_4-qr	1.72	-0.04	-0.04	1.72
CeN_2C_4-ip	1.34	0.00	-0.37	1.71
GdN_2C_4-ip	1.34	-0.03	-0.14	1.45
LaN_2C_4-ip	1.55	-0.04	-0.16	1.66

YbN_2C_4-ip	1.69	-0.02	-0.16	1.83
DyN_2C_4-oppo	-1.84	0.17	-0.25	-1.42
EuN_2C_4-oppo	1.78	-0.09	-0.16	1.86
NdN_2C_4-oppo	1.61	-0.06	-0.13	1.68
YbN_2C_4-oppo	-1.84	0.21	-0.22	-1.41
CeN_3C_3-col	1.17	-0.01	-0.21	1.37
GdN_3C_3-col	1.10	-0.08	-0.20	1.22
YN_3C_3-col	0.98	-0.06	-0.25	1.17
YbN_3C_3-col	1.61	-0.02	-0.20	1.80
CeN_3C_3-qr2	1.31	-0.02	-0.17	1.46
LaN_3C_3-qr2	1.56	-0.04	-0.21	1.74
LuN_3C_3-qr2	1.08	0.02	-0.18	1.29
NdN_3C_3-qr2	1.68	-0.05	-0.05	1.68
CeN_3C_3-sep	1.53	-0.01	-0.19	1.71
EuN_3C_3-sep	1.83	-0.05	-0.04	1.82
LuN_3C_3-sep	1.24	-0.02	-0.21	1.43
NdN_3C_3-sep	1.79	-0.05	-0.09	1.82
GdN_3C_3-hex2	1.27	-0.05	-0.27	1.48
LaN_3C_3-hex2	1.32	-0.13	-0.13	1.33
LuN_3C_3-hex2	0.80	-0.05	-0.18	0.93
YN_3C_3-hex2	1.17	-0.07	-0.21	1.31
EuN_4C_2-hex	1.67	-0.10	-0.13	1.70
LaN_4C_2-hex	1.13	-0.10	-0.22	1.25
ScN_4C_2-hex	0.34	-0.05	-0.35	0.64
YbN_4C_2-hex	1.55	-0.04	-0.22	1.74
CeN_4C_2-qr	1.15	-0.01	-0.13	1.27
NdN_4C_2-qr	1.61	-0.03	-0.23	1.80

EuN_4C_2-qr	1.79	-0.01	-0.20	1.98
LaN_4C_2-qr	1.51	-0.05	-0.16	1.63
GdN_4C_2-ip	1.40	-0.04	-0.15	1.51
LuN_4C_2-ip	1.07	-0.03	-0.17	1.22
ScN_4C_2-ip	1.15	-0.01	-0.21	1.36
YN_4C_2-ip	1.37	-0.02	-0.18	1.52
EuN_4C_2-oppo	1.95	-0.07	-0.06	1.93
GdN_4C_2-oppo	1.27	0.00	-0.25	1.52
NdN_4C_2-oppo	1.69	-0.03	-0.18	1.84
YN_4C_2-oppo	1.29	0.00	-0.18	1.47
GdN_5C_1	1.20	-0.04	-0.18	1.35
LaN_5C_1	1.45	-0.05	-0.17	1.56
ScN_5C_1	0.87	0.01	-0.18	1.05
YN_5C_1	1.15	-0.01	-0.19	1.33
YbN_5C_1	-1.81	0.20	-0.22	-1.39
EuN_6C_0	1.72	-0.04	-0.16	1.84
GdN_6C_0	0.43	0.05	-0.21	0.68
LaN_6C_0	0.75	0.02	-0.20	0.97
LuN_6C_0	-0.14	0.07	-0.20	0.13
PrN_6C_0	1.15	0.00	-0.19	1.34
YbN_6C_0	1.58	-0.01	-0.22	1.79
ScN_6C_0	-0.11	0.06	-0.23	0.18

Table S2. Hyperparameters of the three algorithms

algorithms	hyperparameters
GBR	'n_estimators':100, 'max_depth':3, 'min_samples_split':3, 'learning_rate':0.1, 'loss':'squared_error'
RF	'n_estimators':200, 'max_depth':6, 'min_samples_leaf':1, 'criterion':'squared_error'
Bagging	‘n_estimators’:70, ‘max_samples’:0.9, ‘max_features’:0.8,

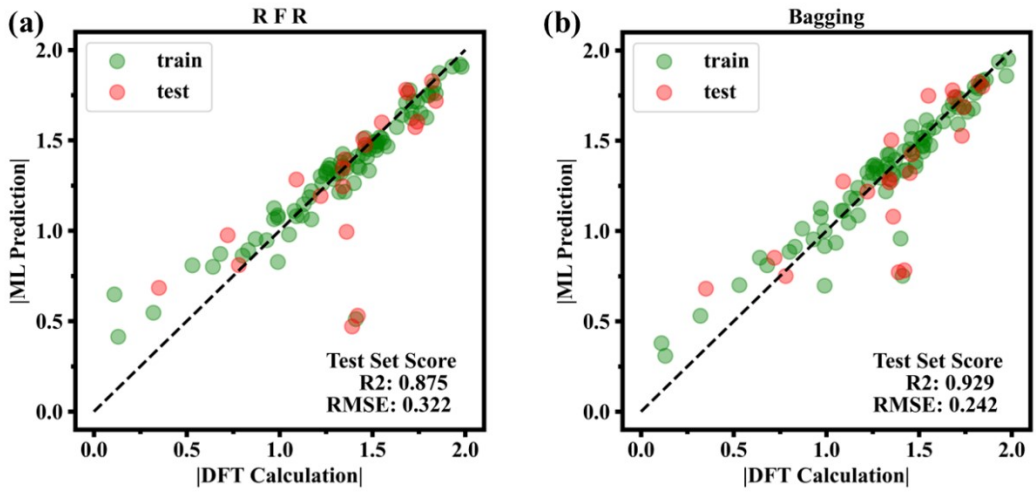

Fig. S1 Comparison of the calculated value of $|\Delta G_{*H}|$ with the predicted value of (a) the RFR and (b) Bagging.

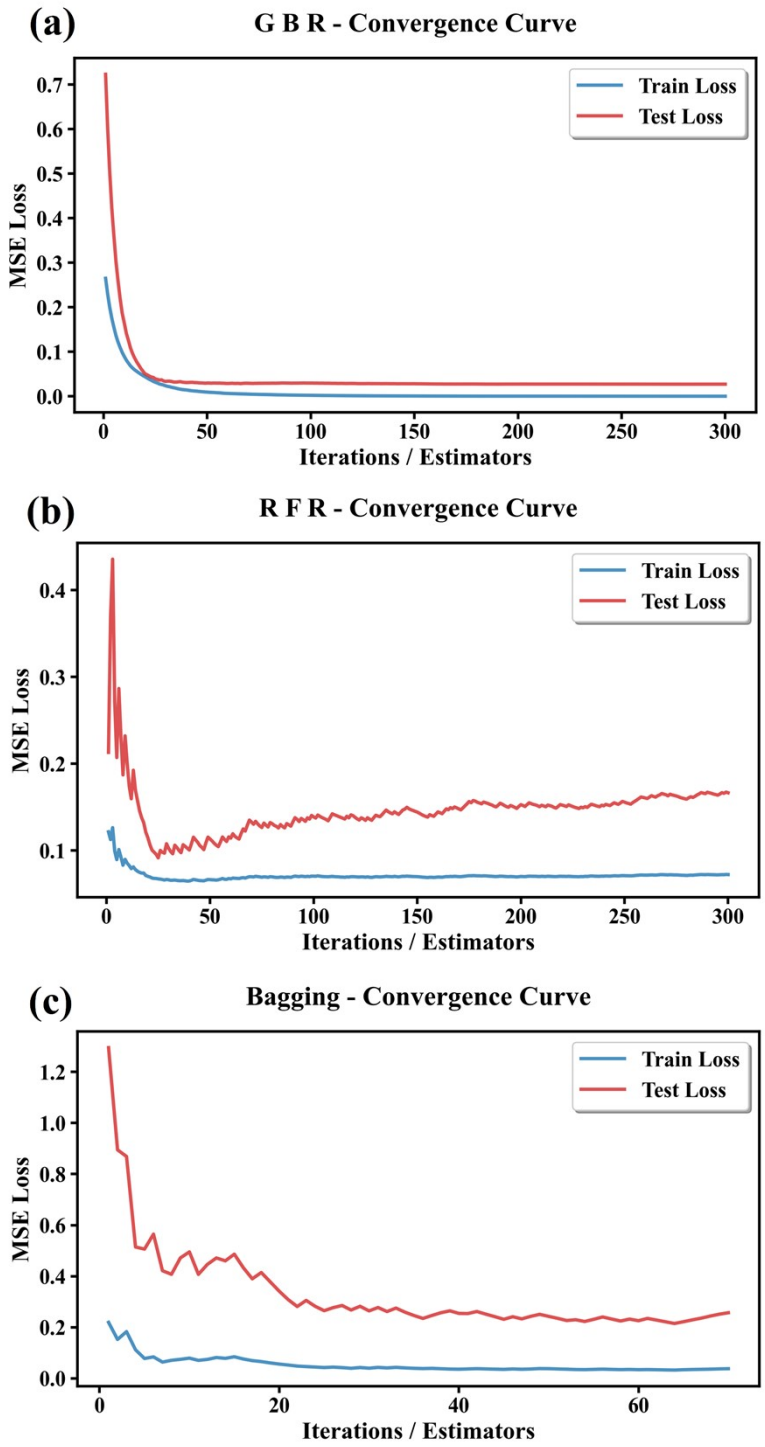

Fig. S2 The loss convergence curves on the training and testing sets of the (a) GBR, (b) RFR and (c) Bagging.

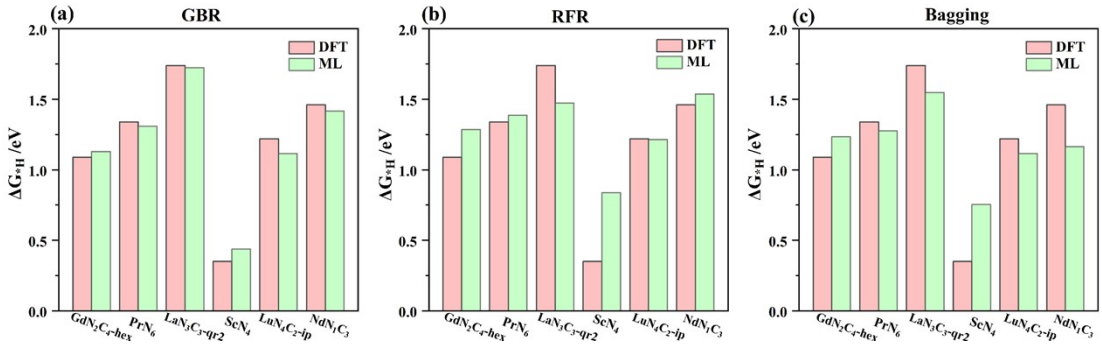

Fig. S3 Quantitative comparison of ΔG*H between DFT true values and ML predicted values for the (a) GBR, (b) RFR and (c) Bagging.

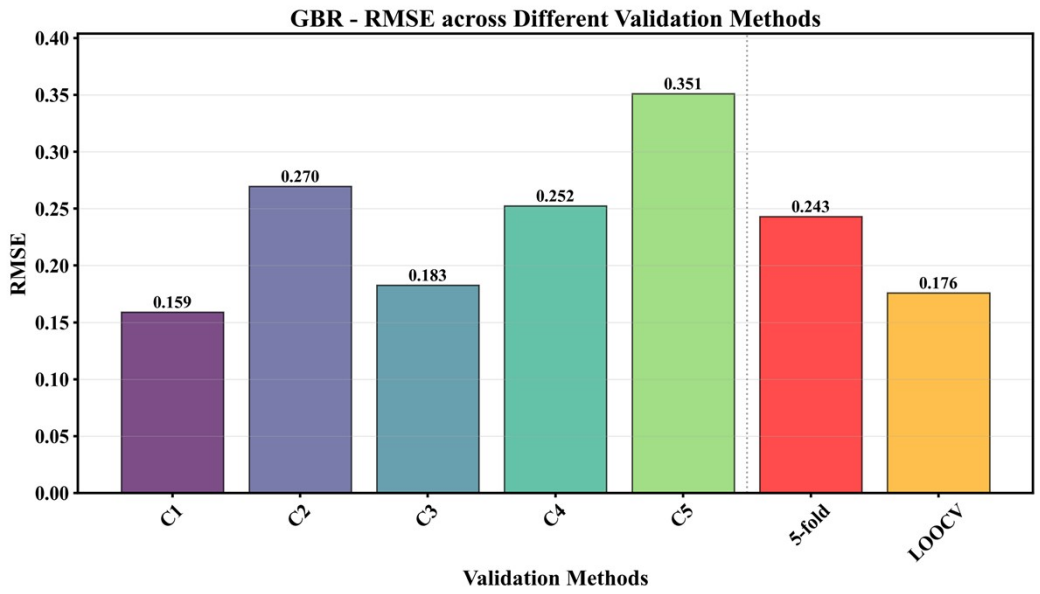

Fig. S4 The 5-fold cross-validation results of the GBR model, including the average RMSE and the RMSE of LOOCV.

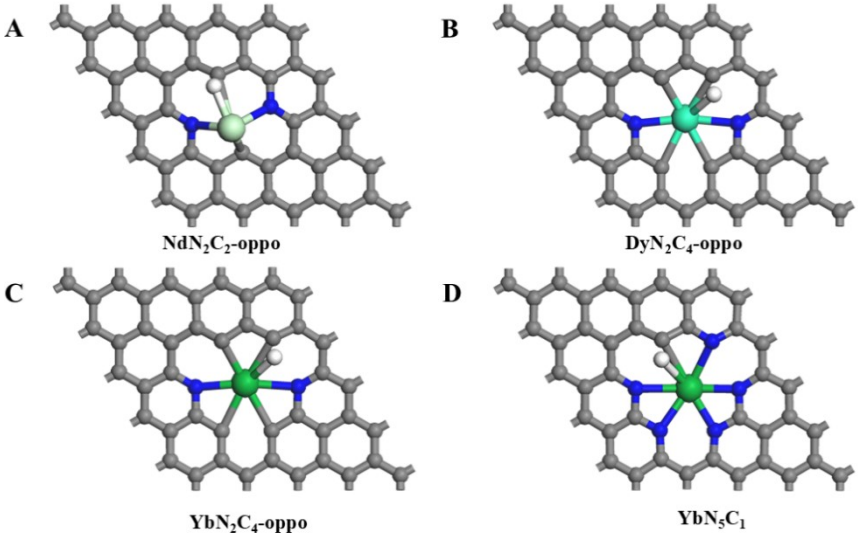

Fig. S5 H adsorbs to C atoms of (a) NdN_2C_2-oppo, (b) DyN_2C_4-oppo, (c) YbN_2C_4-oppo and (d) YbN_5C_1

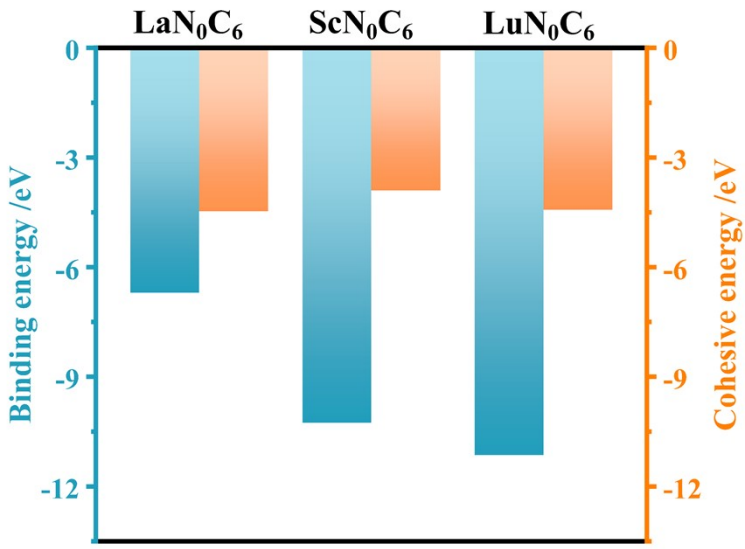

Fig. S6 Comparison of the binding energies of LaN_0C_6, ScN_6C_0 and LuN_6C_0 with the cohesion energies of metal atoms.

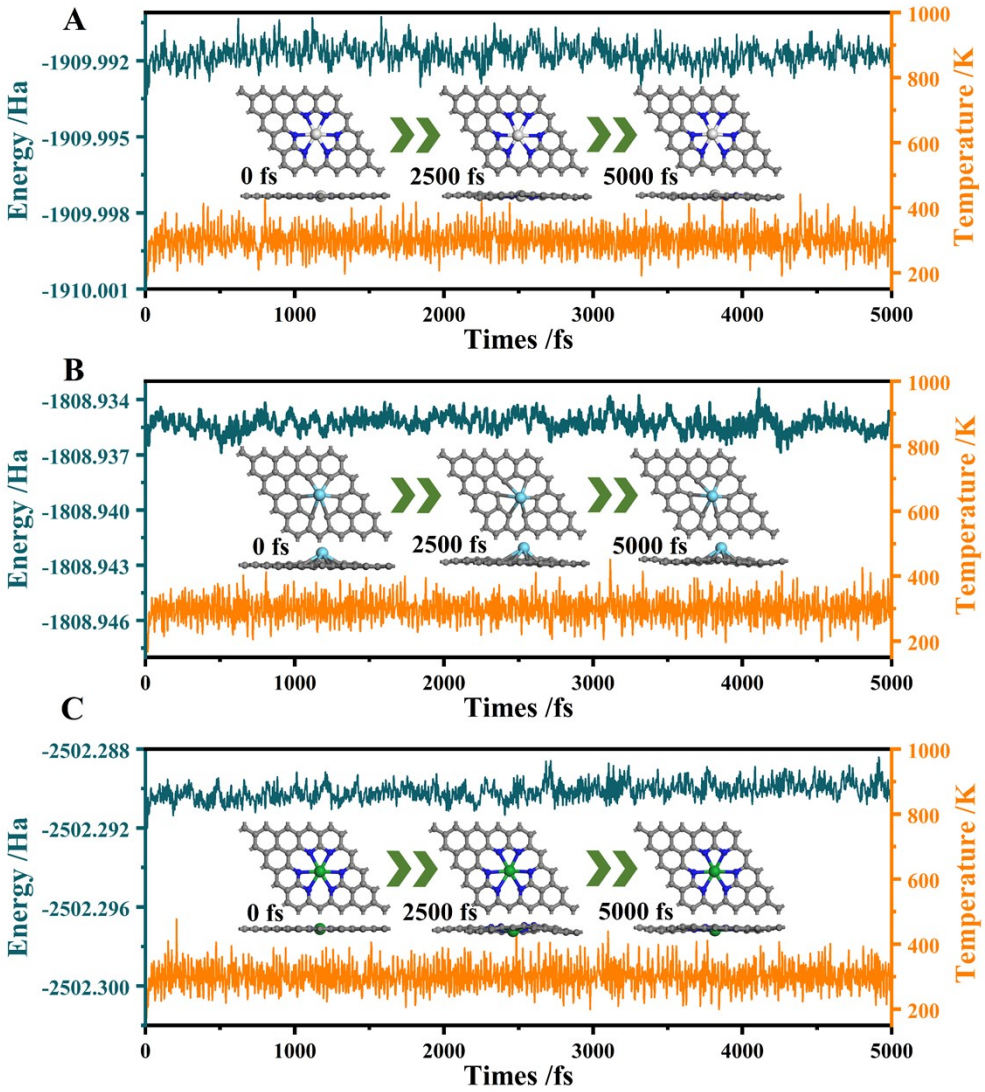

Fig. S7 The molecular dynamics plots of (a) ScN_6C_0, (b) LaN_0C_6 and (c) LuN_6C_0.

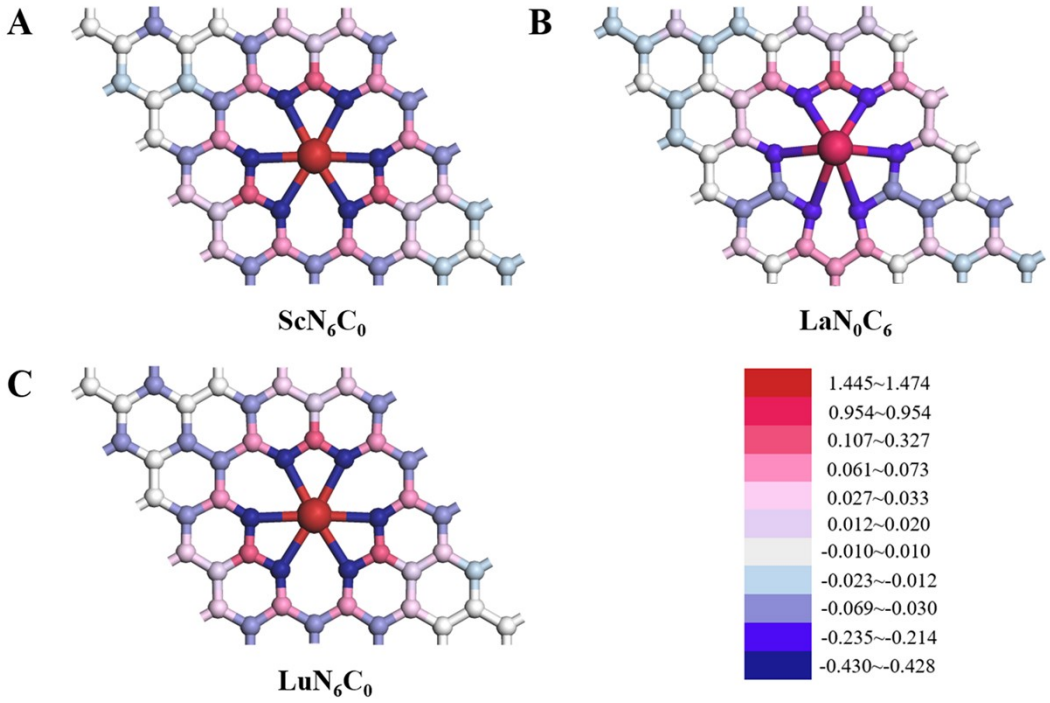

Fig. S8 Mulliken charge distribution plot of (a) ScN_6C_0, (b) LaN_0C_6 and (c) LuN_6C_0.

Reference

1. B. Delley, *J. Chem. Phys.*, 2000, **18**, 7756-7764.
2. Perdew, Burke and Ernzerhof, *Phys. Rev. Lett.*, 1996, **77**, 3865-3868.
3. A. Bergner, M. Dolg, W. Küchle, H. Stoll and H. Preu, *Mol. Phys.*, 1993, **80**, 1431-1441.
4. Delley, *Phys. Rev B*, 2002, **66**, 1551251-1551259.
5. S. Grimme, *J. Comput. Chem.*, 2006, **27**, 1787-1799.
6. B. Delley, *Mol. Simulat.*, 2006, **32**, 117-123.
7. J. K. Nørskov, J. Rossmeisl, A. Logadottir, L. Lindqvist, J. R. Kitchin, T. Bligaard and H. Jónsson, *J. Phys. Chem. B*, 2004, **108**, 17886-17892.
8. T. Zhao, K. Chen, X. Xu, X. Li, X. Zhao, Q. Cai, K. Chu and J. Zhao, *Appl. Catal. B-Environ.*, 2023, **339**, 123156.
9. C. Ning and Y. Lu, *J. Phys. Chem. Ref. Data*, 2022, **51**.
10. C. Tantardini and A. R. Oganov, *Nat. Commun.*, 2021, **12**, 2087.
11. M. Rupp, A. Tkatchenko, K.-R. Müller and O. A. von Lilienfeld, *Phys. Rev. Lett.*, 2012, **108**, 058301.