Electronic Supplementary Material (ESI) for Nanoscale.

Supporting Information

Accurate and Efficient Machine Learning Models for Predicting Hydrogen Evolution Reaction Catalysts Based on Structural and Electronic Feature Engineering in Alloys

Jingzi Zhang[a,b,1], Yuelin Wang[a,b,1], Xuyan Zhou[a,b], Chengquan Zhong[a,b], Ke Zhang[a,b], Jiakai Liu, Kailong Hu[a,b,c*], and Xi Lin[a,b,c*]

[a] *School of Materials Science and Engineering, Harbin Institute of Technology, Shenzhen 518055, Guangdong, China*

[b] *Blockchain Development and Research Institute, Harbin Institute of Technology, Shenzhen 518055, Guangdong, China*

[c] *State Key Laboratory of Advanced Welding and Joining, Harbin Institute of Technology, Harbin 150001, Heilongjiang, China*

[*]Corresponding authors.

E-mail addresses: linxi@hit.edu.cn (X. Lin);

hukailong@hit.edu.cn (K. Hu);

[1]These authors contributed equally to this work.

Method

Support Vector Regression

The SVR model represents instances as points in space, with the examples of different categories separated by a large gap. The SVR training algorithm creates a model that assigns new instances to one of two categories, making it a non-probabilistic binary linear classifier, based on a series of training examples that are individually labeled as belonging to one of two categories. New instances are then mapped into the same area and assigned to one of the categories based on which side of the gap they fall on. SVR can perform both linear and non-linear regression by utilizing the kernel trick, which involves implicitly transforming inputs into high-dimensional feature spaces.[1]

Gradient boosting decision tree

GBDT model is a powerful tool for optimizing arbitrary differentiable loss functions. It works by constructing an additive model in a forward stage-wise fashion, where in each stage a regression tree is fit on the negative gradient of the given loss function. The boosting tree model was optimized using an additive model and a forward stagewise algorithm. During training, the negative gradient of the loss function was used to fit the approximate value of the loss in each iteration, resulting in a continuous reduction of the error term generated in the training process [2,3].

K-nearest neighbor

KNN model has been demonstrated to be effective in cases where the data labels are continuous rather than discrete variables. This method utilizes an unsupervised learner to identify the nearest neighbors of a query point, and then assigns a label to the query point based on the mean of the labels of its nearest neighbors.[4]

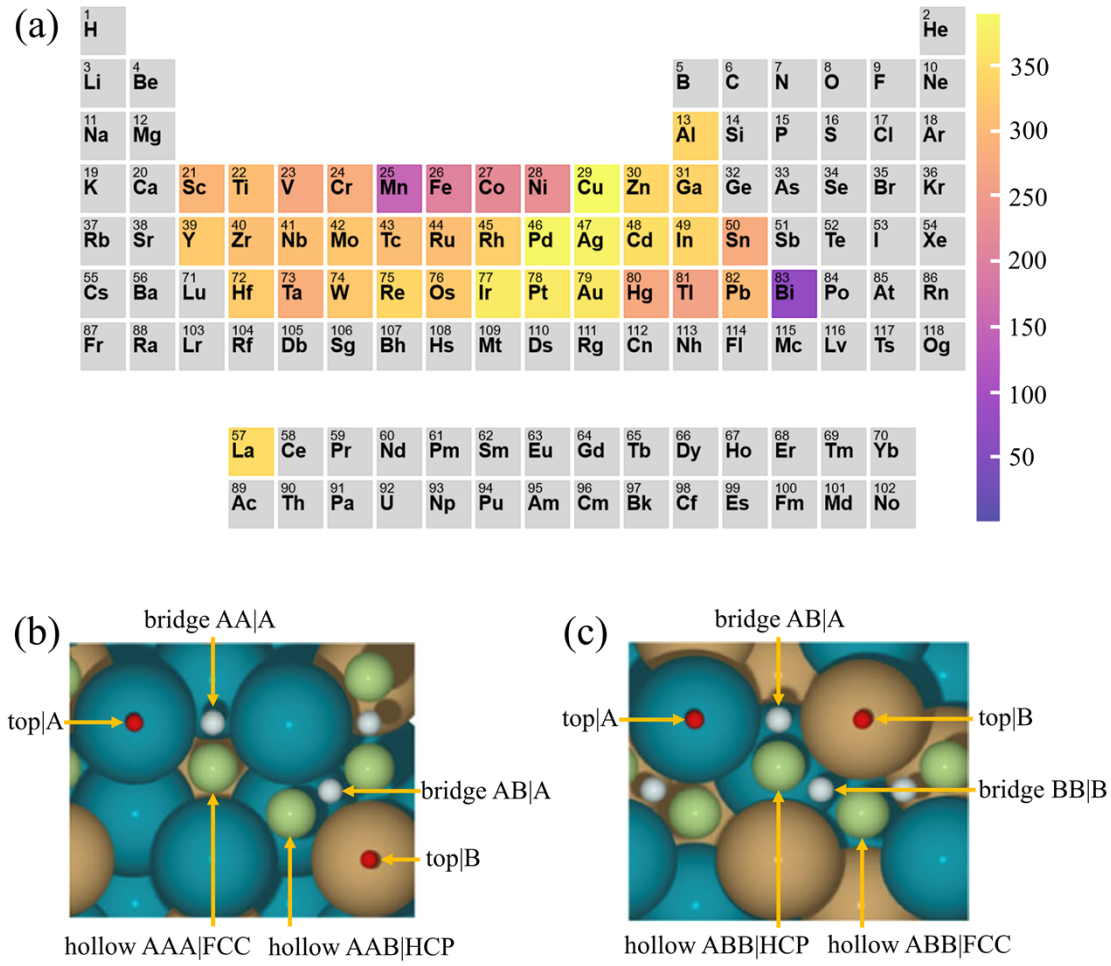

Fig. S1. (a) The periodic table outlining H adsorbate elements and the 37 metals included in the dataset. Structure models and enumerated adsorption sites: (b) A_3B alloy, (c) AB alloy. The A and B atoms are represented dark green and brown sphere. The top, bridge, and hollow sites are shown in red, white, and green, respectively.

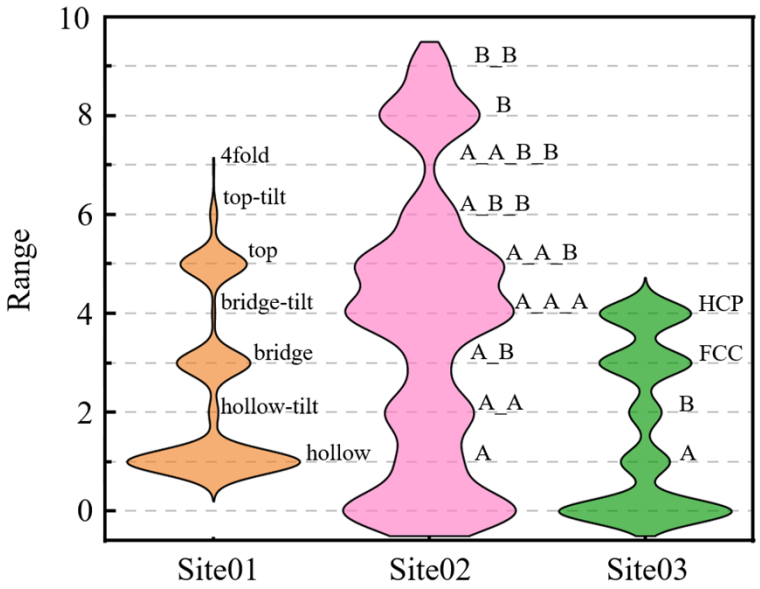

Fig. S2. The violins distribution represents three adsorption site features.

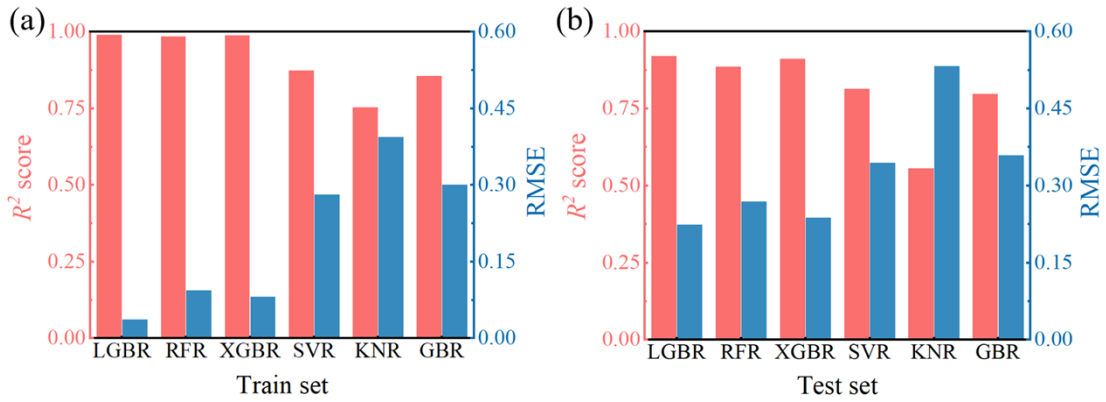

Fig. S3. Comparison of the R^2 score and the RMSE for six ML models on (a) the train set and (b) the test set.

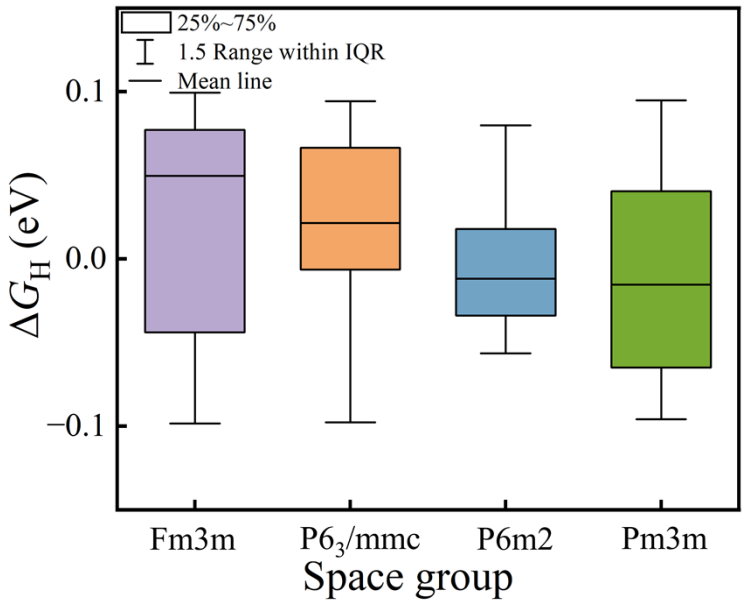

Fig. S4. Box plots of ΔG_{H*} distributions for different ranges of the space group.

Table S1 Detailed parameters of eleven ML models.

ML model name	Parameters
SVR	Degree: 3 C: 1e3 Gamma: scale
KNN	N_neighbors: 10 Weights: uniform Leaf_size: 45
LGB	N_estimators: 1000 Num_leaves: 15 Max_depth: -1 Subsample: 0.6 Random_state: 2022
RF	N_estimators: 500 Max_depth: 8 Random_state: 2022
GBDT	Max_depth: 7 Min_samples_leaf: 8 N_estimators: 600 Random_state: 2022
XGB	Colsample_bytree: 0.9 Max_depth: 5 N_estimators: 700 Objective: 'reg: squarederror'

Random_state: 2022

Table S2 Dataset for HER alloy materials.

Number	Formula	Site information	ΔG_{H^*} (eV)
1	Pt_3Ti	bridge AA|B	-0.052
12	Sc_3Cu	top B	0.095
…	…	…	…
8856	MnRu	hollow AAB|FCC	0.026

Table S3 Calculation methods of element features in compounds. The "i" represents the element number of alloys. The "t" represents the property of element. The "p" represents the weighted score. In the calculation of the weight entropy of mixing $\omega = t_i/(t_1 + ... + t_i)$.

Features description	Computational formula
The average	$= \mu = (t_1 + ... + t_i)/i$
Weighted mean	$= \nu = (p_1 \times t_1) + ... + (p_i \times t_i)$
Geometric mean	$= (t_1 \times ... \times t_i)^{1/i}$
Weighted geometric mean	$= (t_1)^{p1} \times ... \times (t_i)^{pi}$
The entropy of mixing	$= -\omega_1 In(\omega_1) - ... - \omega_i In(\omega_i)$
Weighted entropy of mixing	$= -\frac{p_1\omega_1}{p_1\omega_1 + ... + p_i\omega_i} In\left(\frac{p_1\omega_1}{p_1\omega_1 + ... + p_i\omega_i}\right) ... - \frac{p_i\omega_i}{p_1\omega_1 + ... +}$
Extreme value range	$= t_1 - t_2\ (t_1 > t_2)$
Weighted range	$= p_1t_1 - p_2t_2$
The standard deviation	$= [(1/2)((t_1 - \mu)^2 + (t_2 - \mu)^2)]^{1/2}$
Weighted standard deviation	$= [p_1(t_1 - \nu)^2 + p_2(t_2 - \nu)^2]^{1/2}$
The maximum	$= Maximum\ (t_1,t_2)$
The minimum value	$= Minimum\ (t_1,t_2)$
The mode	$= Mode\ (count(t_1,t_2))$
Element species number	$= Unique\ ([t_1,t_2])$

Table S4 The explication and LabelEncoder method for adsorption site features in the

dataset.

Num	Site name	Site feature description	LabelEncoder S_1	LabelEncoder S_2	LabelEncoder S_3
1	top|A	The top site of atom A	3	1	0
2	top|B	The top site of atom B	3	8	0
3	bridge AA|A	The bridge site between atom A and atom A, and near the atom A	2	2	1
4	bridge AB|B	The bridge site between atom A and atom A, and near the atom B	2	3	2
5	bridge AB|A	The bridge site between atom A and atom B, and near the atom A	2	3	1
6	bridge BB|B	The bridge site between atom B and atom B, and near the atom B	2	9	2
7	hollow AAB|FCC	The hollow site between two atom A and one atom B, and there are no atoms directly below the hollow site	1	5	3
8	hollow AAB|HCP	The hollow site between two atom A and one atom B, and there are atoms directly below the hollow site	1	5	4
9	hollow AAA|FCC	The hollow site between three atoms A, and there is atom directly below it	1	4	3
10	hollow AAA|HCP	The hollow site between three atoms A, and there is atom directly below it	1	4	4
11	4fold	The hollow between four atoms	4	0	0

Table S5 The S_1 feature coded by LabelEncoder method.

S_1 site types	LabelEncoder number
hollow	1
bridge	2
top	3
4fold	4
hollow-tilt	5
top-tilt	6
bridge-tilt	7

Table S6 The S_2 feature coded by LabelEncoder method.

S_2 site types	LabelEncoder number
A	1
AA	2
AB	3
AAA	4
AAB	5
ABB	6
AABB	7
B	8
BB	9

Table S7 The S_3 feature coded by LabelEncoder method.

S_3 site types	LabelEncoder number
A	1
B	2
FCC	3
HCP	4

Table S8 The explication of 20 features in SHAP value plot.

Num	Feature name	Feature description
1	M_c	Mean column: mean of group number of elements in the composition
2	S_2	Site02: the position of atoms around the adsorption site
3	M_p	Mean melting point: mean of melting point of elements in the composition
4	S_1	Site01: the position information of the adsorption site
5	M_d	Mean NdUnfilled: mean of number of unfilled *d*-orbitals among elements in the composition
6	M_u	Mean Nunfilled: mean of number of unfilled valence orbitals among elements in the composition
7	S_3	Site03: the atoms position of adsorbed substance H around the adsorption site
8	D_f	Std_dev heat of formation: standard deviation of heat formation among elements in the composition
9	M_r	Mean covalent radius: mean of covalent radius among elements in the composition
10	M_v	Mean ndvalence: mean of number of valence *d*-orbitals among elements in the composition
11	M_e	Mean electronegativity: mean of electronegativity among elements in the composition
12	E_f	Entropy heat of formation: weighted entropy of heat formation among elements in the composition

13	D_c	Std_dev lattice constant: standard deviation of lattice constant among elements in the composition
14	E_r	Entropy metallic radius: weighted entropy of metallic radius among elements in the composition
15	M_t	Mean thermal conductivity: mean of thermal conductivity among elements in the composition
16	M_n	Mean Mendeleev number: mean of Mendeleev number among elements in the composition
17	E_a	Entropy atomic volume: weighted entropy of atomic volume among elements in the composition
18	E_h	Entropy fusion heat: weighted entropy of fusion heat among elements in the composition
19	M_a	Mean atomic weight: mean of atomic weight among elements in the composition
20	A_e	Avg_dev electronegativity: average deviation of electronegativity among elements in the composition

Table S9 Comparison of experimentally measured η^{HER} with the values predicted by LGB model with three HER alloy electrocatalysts in the dataset.

Alloy	**DFT η^{HER} (V)**	**ML predicted η^{HER} (V)**	**Experimentally measured η^{HER} (V)**
CoNi	0.075	0.083	0.098 [5]
PtRu	0.146	0.131	0.113 [6]
RuCo	0.017	0.016	0.010 [7]
FeNi	0.124	0.126	0.126 [8]

Table S10 The ΔG_{H^*} values of fourteen alloys from MP database using ML model prediction and the DFT calculation.

Num	**Mp-id**	**Formula**	**ML-ΔG_{H^*} (eV)**	**DFT-ΔG_{H^*} (eV)**	**ML-η_{HER} (V)**	**DFT-η_{HER} (V)**	**Site**
1	mp-437	MgAu	-0.095	0.081	0.095	0.081	bridge
2	mp-721	TbCd	-0.021	0.013	0.021	0.013	bridge
3	mp-1172	MgRh	-0.105	0.020	0.105	0.020	bridge
4	mp-1857	YbCd	-0.043	0.016	0.043	0.016	bridge
5	mp-2165	SmZn	-0.038	0.020	0.038	0.020	bridge
6	mp-2525	PrAg	-0.041	0.050	0.041	0.050	top
7	mp-2724	TbSb	0.042	0.051	0.042	0.051	bridge
8	mp-7576	CrSi	-0.037	0.050	0.037	0.050	bridge
9	mp-11256	ScAu	0.038	0.036	0.038	0.036	top
10	mp-12793	NdAl	-0.077	-0.024	0.077	0.024	bridge

11	mp-20582	LaIn	-0.047	0.019	0.047	0.019	bridge
12	mp-574283	GdTe	-0.040	0.097	0.040	0.097	bridge
13	mp-998985	TiAu	-0.048	0.012	0.048	0.012	top
14	mp-1079910	SiTc	0.033	0.047	0.033	0.047	top

References

[1] H. Drucker, C.J. Burges, L. Kaufman, A. Smola, V. Vapnik, Support vector regression machines, Adv. Neural. Inf. Process. Syst. 9 (1996)

[2] H. Rao, X. Shi, A.K. Rodrigue, J. Feng, Y. Xia, M. Elhoseny, X. Yuan, L. Gu, Feature selection based on artificial bee colony and gradient boosting decision tree, Appl. Soft Comput. 74 (2019) 634−642.

[3] J. Feng, Y. Yu, Z.-H. Zhou, Multi-layered gradient boosting decision trees, Adv. Neural. Inf. Process. Syst. 31 (2018)

[4] J.M. Keller, M.R. Gray, J.A. Givens, A fuzzy k-nearest neighbor algorithm, IEEE Trans. Syst. Man Cybern. (1985) 580−585.

[5] Tan X, Geng S, Ji Y, et al. Closest packing polymorphism interfaced metastable transition metal for efficient hydrogen evolution. Advanced Materials, 2020, 32(40): 2002857.

[6] Zhang D, Wang Z, Wu X, et al. Noble metal (Pt, Rh, Pd, Ir) doped Ru/CNT ultra‐small alloy for acidic hydrogen evolution at high current density. Small, 2022, 18(3): 2104559.

[7] Cai C, Liu K, Zhu Y, et al. Optimizing hydrogen binding on Ru sites with RuCo alloy nanosheets for efficient alkaline hydrogen evolution. Angewandte Chemie International Edition, 2022, 61(4): e202113664.

[8] Ren J T, Chen L, Wang Y S, et al. FeNi nanoalloys encapsulated in N-doped CNTs tangled with N-doped carbon nanosheets as efficient multifunctional catalysts for overall water splitting and rechargeable Zn–Air batteries. ACS Sustainable Chemistry & Engineering, 2019, 8(1): 223-237.