



A machine learning model with minimize feature parameters for multi-type hydrogen evolution catalyst prediction



Chao Wang, Bing Wang, Changhao Wang, Aojian Li, Zhipeng Chang & Ruzhi Wang

The vast chemical compositional space presents challenges in catalyst development using traditional methods. Machine learning (ML) offers new opportunities, but current ML models are typically limited to screening a single catalyst type. In this work, we developed an efficient ML model to predict hydrogen evolution reaction (HER) activity across diverse catalysts. By minimizing features, we introduced a key energy-related feature $\varphi = \text{Nd}0^2/\psi_0$, which correlates with HER free energy. Using just ten features, the Extremely Randomized Trees model achieved $R^2 = 0.922$. We predicted 132 new catalysts from the Material Project database, among which several exhibited promising HER performance. The time consumed by the ML model for predictions is one 200,000th of that required by traditional density functional theory (DFT) methods. The model provides an efficient approach for discovering high-performance HER catalysts using a small number of key features and offers insights for the development of other catalysts.

With increasing concern over environmental pollution and the depletion of fossil fuels, the search for a clean and sustainable energy source has become urgent^{1–3}. Hydrogen (H₂) is considered one of the most promising alternative energy sources due to its high energy density and zero carbon emissions⁴. Currently, hydrogen production via water electrolysis powered by renewable energy is a highly promising technology^{5,6}. Hydrogen production from water electrolysis is controlled by the hydrogen evolution reaction (HER). However, the electrochemical reactions exhibit slow kinetics, resulting in high overpotentials for water electrolysis^{7–9}. Therefore, it is essential to develop efficient catalysts to enhance electrochemical reactions and reduce overpotentials. Noble metals (such as Pt, Ir, and Ru) and their derivatives exhibit excellent conductivity for water electrolysis, and the adsorption free energy of hydrogen atoms on noble metal surfaces is close to zero, hence noble metals are regarded as the most effective catalysts for the HER^{10,11}. However, noble metal materials have drawbacks such as high cost and limited availability, which restrict their large-scale commercial application. Therefore, the design and development of low-cost, high-efficiency electrocatalyst materials are crucial from the perspective of production cost and efficiency^{12–14}.

Various types of hydrogen evolution catalysts (HECs) have been developed, showing certain catalytic activities, such as alloys, carbides, nitrides, oxides, phosphides, sulfides, and perovskites^{15–18}. However, developing catalysts based on traditional experimental methods faces several issues, including long development cycles and significant randomness.

Additionally, high-throughput computational methods using density functional theory (DFT) to develop efficient catalysts require substantial computational resources¹⁹. Therefore, developing excellent catalysts from a vast compositional space using empirical experiments and DFT calculations remains a significant challenge^{20,21}.

Machine learning (ML) is a powerful statistical method that constructs models based on input data and provides target values through computational algorithms. ML can be used to analyze the complex relationships between input features and target performance. Additionally, ML can assess the importance of each input feature and predict the catalytic activities of numerous unknown catalysts. The robust capabilities of ML have been applied to the rapid screening of excellent catalysts^{22–27}. Powerful ML algorithms can help uncover the relationships between the physicochemical properties of catalysts and their HER activity, thereby accelerating the discovery of efficient HER electrocatalysts^{28,29}. For instance, Chandra Veer Singh et al. developed a neural network model for designing high-entropy alloy (HEA) catalysts by decoupling ligand and coordination effects³⁰, achieving a test set prediction accuracy of MAE = 0.09 eV and RMSE = 0.12 eV. Lin et al. developed an ML model using 147 features to rapidly predict the activity of binary alloy HEAs³¹, with a test set prediction accuracy of $R^2 = 0.921$ and RMSE = 0.224 eV. S. Kim et al. used 20 features to build a CatBoost regression model for transition metal single-atom-based superb hydrogen evolution electrocatalysts³², with a test set prediction accuracy of $R^2 = 0.88$ and RMSE = 0.18. Mu et al. used 13 features to develop

a random forest regression model for double-atom catalysts with H₂ evolution activity supported on graphene³³, achieving a test set prediction accuracy of $R^2 = 0.871$ and MSE = 0.150. However, current ML methods for exploring efficient HECs are only applicable to the design of a single type of HECs, and they suffer from issues such as the use of numerous features and low model accuracy. This limitation stems from the significant variations in features required for multi-types of HECs. Therefore, it is urgently necessary to develop ML models that use fewer features, possess higher accuracy, and can predict the activity of various HECs.

In this work, we developed a high-precision ML model to design highly active HECs. We obtained atomic structure features and hydrogen adsorption free energy ΔG_H data for 10,855 HECs from Catalysis-hub for training and prediction³⁴. The dataset includes various types of HECs, such as pure metals, transition metal intermetallic compounds, light metal intermetallic compounds, non-metallic compounds, and perovskite. Using only 23 features based on atomic structure and electronic information of the catalyst active sites, without the need for additional DFT calculations, we established six ML models: Random Forest Regression (RFR), Gradient Boosting Regression (GBR), Extreme Gradient Boosting Regression (XGBR), Decision Tree Regression (DTR), Light Gradient Boosting Machine Regression (LGBMR), and Extremely Randomized Trees Regression (ETR). The ETR model achieved an R-Square (R^2) score of 0.921 for predicting ΔG_H , outperforming the other ML models. Through feature importance analysis and feature engineering, we reselected and identified more relevant features, reducing the number of features from 23 to 10 and improving the R^2 score to 0.922. Furthermore, a comparison between two deep learning (DL) models, the Crystal Graph Convolutional Neural Network (CGCNN) and the Orbital Graph Convolutional Neural Network (OGCNN), demonstrated that the ETR model outperforms these DL models in accuracy, indicating the crucial role of feature selection in achieving high predictive performance. Finally, we predicted the performance of 132 different HECs and further validated the ML model's prediction accuracy using DFT methods. This work provides

interpretable insights for accelerating the compositional design of high-performance HECs.

Results

The schematic diagram for implementing the workflow is shown in Fig. 1. The process includes data collection, feature extraction, ML model training and testing, feature engineering and ML optimization, catalyst prediction and screening. Data is the foundation of ML. In order to generate predictions using ML models, credible and sufficient data sets are essential.

Data collection

We obtained 11,068 HER free energies and corresponding adsorption structures from the Catalysis-hub database³⁵. The data in this database are sourced from published literature, peer-reviewed, and validated to ensure data accuracy. The dataset includes various types of HECs, such as pure metals, transition metal intermetallic compounds, light metal intermetallic compounds, non-metallic compounds, and perovskites, all data in the dataset are derived from DFT calculations. As shown in Fig. S2, transition metal intermetallic compounds, light metal intermetallic compounds, and non-metallic compounds together account for over 90% of the dataset, aligning with the current research focus in catalyst development. In the dataset, hydrogen adsorption sites include top sites, bridge sites, and vacancy sites, with vacancy-site adsorption being the most prevalent. Here, we present the top and side views of three representative structures from the dataset in Fig. 2: TaIr₃, LaIr₃, and PtCo₃, which adsorb hydrogen atoms at the top site, bridge site, and hollow site, respectively. Generally, when a hydrogen atom adsorbs on a surface, the distance between the hydrogen atom and the surface atoms typically falls within the range of 1.5 Å to 2.5 Å. This range indicates that there is sufficient interaction between the hydrogen atom and surface atoms to facilitate adsorption. In our dataset, the distances for hydrogen adsorption on the surface also lie within this range. As shown in Fig. 2a, the bond length for hydrogen adsorption at the top site of TaIr₃ is 1.892 Å; in Fig. 2b, the bond lengths for hydrogen adsorption at the bridge

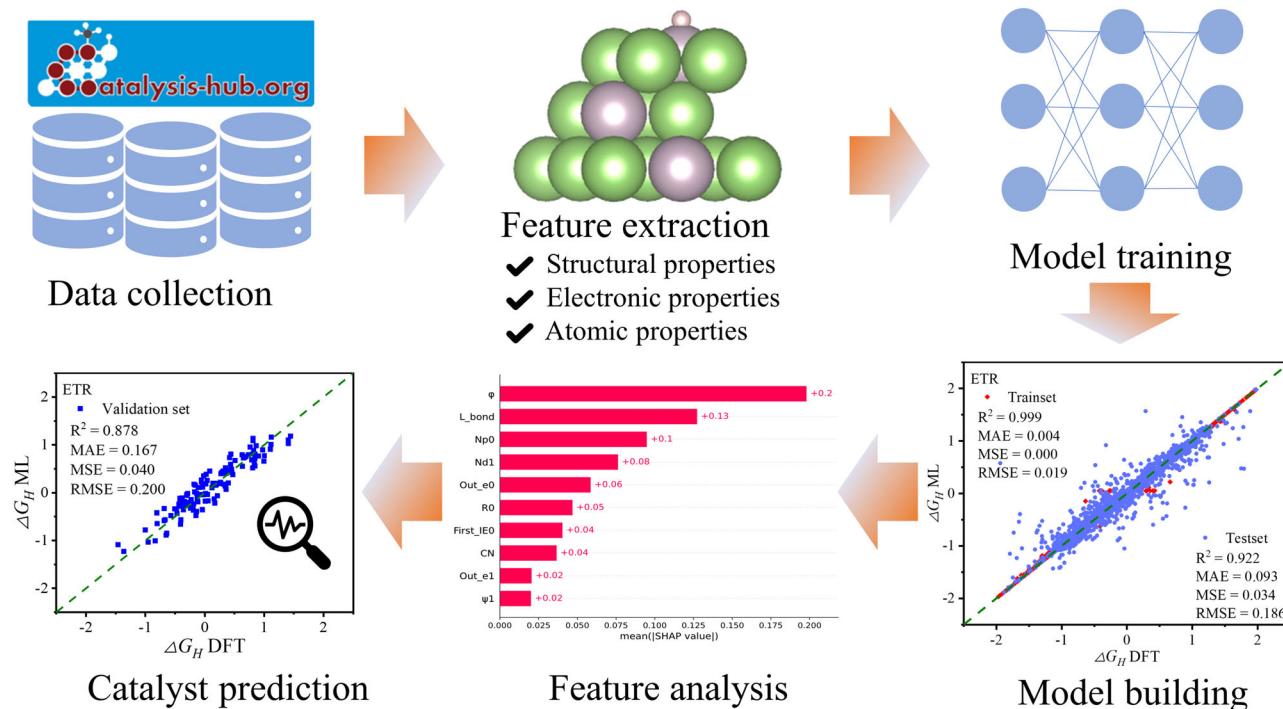
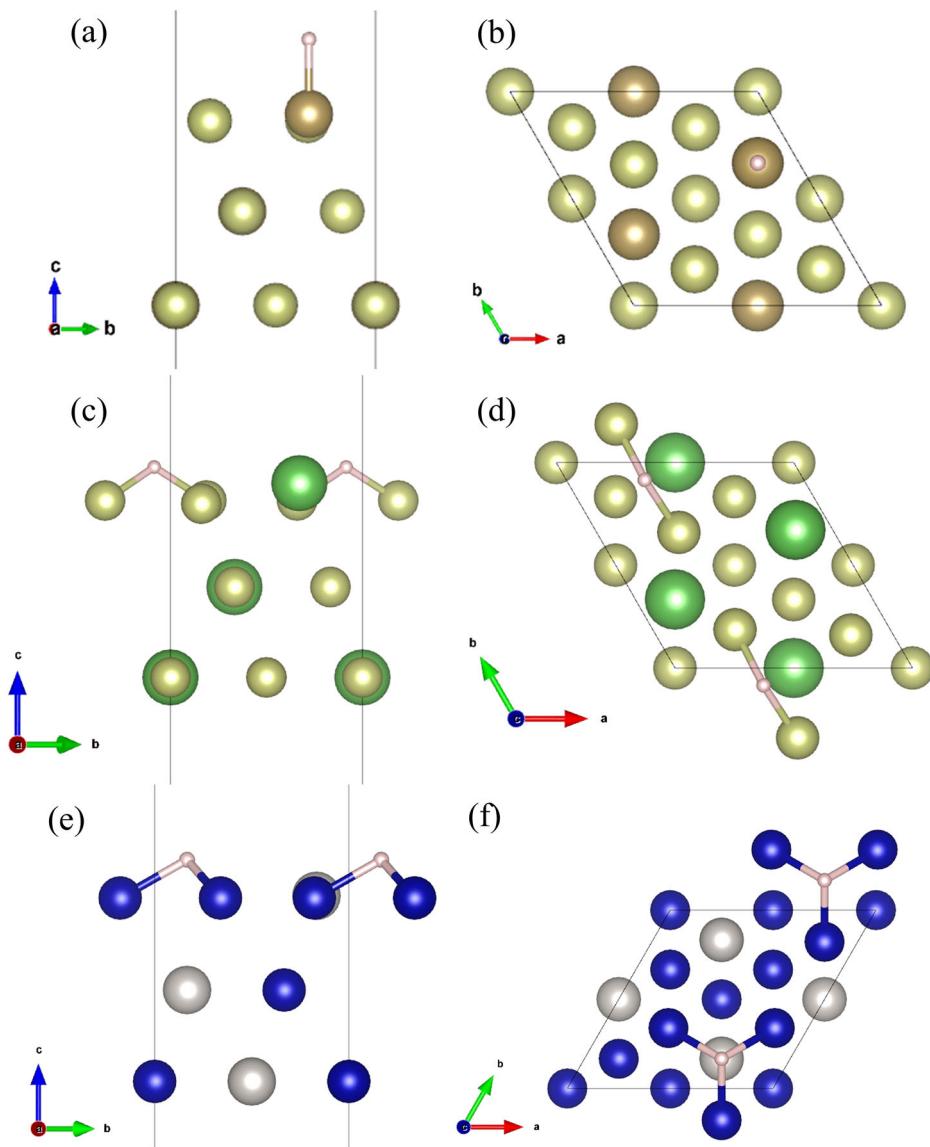


Fig. 1 | Workflow of the ML-accelerated HER catalytic activity prediction. The process includes data collection, feature extraction, model training, model building, feature analysis, and model prediction. Data collection: collect the atomic structures and ΔG_H for HER. Feature extraction: extract structural features, electronic features, and atomic features from the atomic structures of the HER. Model training: improve

ML models accuracy through hyperparameter tuning. Model building: use the ML models, fitted on the training set, to make predictions on the test set. Feature analysis: analyze feature importance and correlations, and use feature engineering to reduce the feature set while introducing key features to enhance model accuracy. Model prediction: use the ML model to predict potential HECs.

Fig. 2 | Structural schematics illustrating hydrogen adsorption at three different atomic sites. It shows the front views of hydrogen adsorption on TaIr_3 , LaIr_3 , and PtCo_3 in (a, c, e) respectively, and the top views of hydrogen adsorption on TaIr_3 , LaIr_3 , and PtCo_3 in (b, d, f) respectively.



site of LaIr_3 are 1.795 Å and 1.796 Å; and in Fig. 2c, the bond lengths for hydrogen adsorption at the vacancy site of PtCo_3 are all 1.758 Å. Additionally, some catalysts exhibit relatively larger adsorption distances for hydrogen atoms, such as Y_3Sc , where the bond lengths for hydrogen adsorption at the vacancy site are 2.262 Å, 2.268 Å, and 2.269 Å. To accurately describe the number of surface atoms involved in hydrogen adsorption, we set a cutoff distance of 2.4 Å between the surface atoms and hydrogen atoms in our feature extraction script, considering all atoms within this range as active center atoms.

The distribution of the free energies of the HECs in the dataset is shown in Fig. 3a, with a range of $[-12.4, 22.1]$ eV, the inset represents a magnification of certain regions.

According to Nørskov's work⁵, the HER catalytic activity is optimal when the absolute value of ΔG_H is zero. Notably, 95.5% of the data falls within the range of $[-2, 2]$ eV. We narrowed the hydrogen adsorption free energy range to $[-2, 2]$ eV and removed unreasonable hydrogen adsorption structures. The adjusted distribution of free energy in the hydrogen adsorption dataset is shown in Fig. 3b. The total number of adjusted data points is 10,855, involving 42 elements, as shown in Fig. S3, which includes most transition metals. Additionally, the surface coverage of hydrogen atoms in this dataset is $\leq 25\%$. According to J.K. Nørskov's calculations on the HER free energy of pure metals³⁶, the HER free energies calculated at low

coverage align well with experimental results. Therefore, the HER free energies calculated using this dataset are expected to accurately reflect the catalytic activity of the real catalyst.

Feature extraction and ML model building

Feature extraction is an indispensable component for ML, and designing appropriate and comprehensive features is the most crucial stage in constructing ML models. Therefore, it is crucial to establish catalytic reaction descriptors primarily based on adsorption structures and electronic properties. The surface structure of catalysts plays a significant role in studying the catalytic effects of HER. The feature extraction scripts used in this study utilize the Python module of the Atomic Simulation Environment (ASE) to automatically identify adsorbed hydrogen atoms and material surface structures, extracting the relevant features. Based on previous research, we designed electronic and elemental feature attributes for the active sites of catalysts and their nearest neighbors^{37–43}.

The features we collected include properties of the active site atom where hydrogen adsorption occurs and the surrounding atoms near the active site. The specific features are as follows: d-band electron count, p-band electron count, s-band electron count, valence electron count, electronegativity, first ionization energy, atomic radius, adsorption bond length between the active site and the hydrogen atom, and the geometric

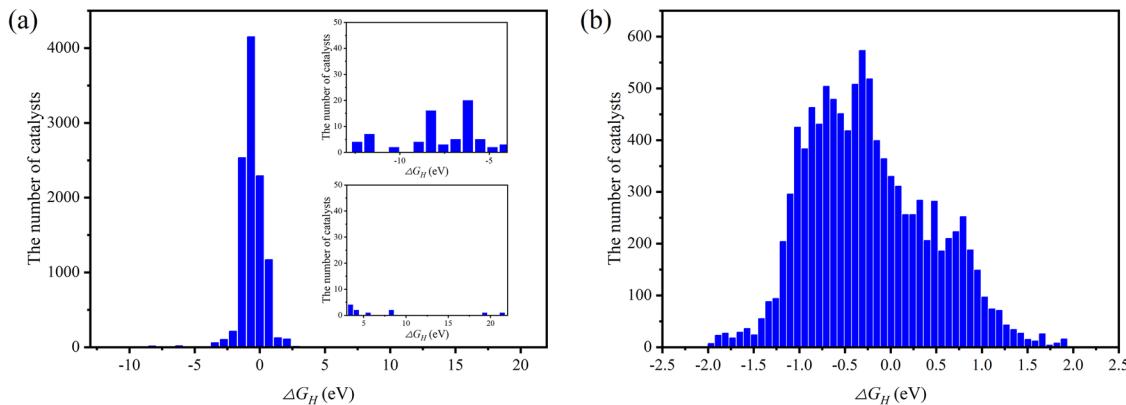


Fig. 3 | Distribution diagram of hydrogen evolution free energy. Distribution of hydrogen evolution free energies **a** within the entire dataset, **b** within the range of [-2, 2] eV. The inset in (a) represents a magnification of certain regions.

mean of the coordination number of atoms closest to the adsorption center. All selected features have clear physical and chemical significance.

Previous studies have shown that the d-band and p-band electron counts directly influence the position of the d-band and p-band centers of catalysts^{44–46}. Generally, a higher d-electron count correlates with more positive adsorption energy, while a higher p-electron count correlates with more negative adsorption energy. The inclusion of s-electron count is primarily to differentiate the HER (hydrogen evolution reaction) performance of Cu, Ag, Au, and Pt from other elements, as these metals have only one electron in their outermost shell. To differentiate the catalytic activity of main group elements and transition metals, the characteristics associated with the s- and p-orbitals play a critical role. Unlike transition metals, the chemical properties of main group elements are primarily determined by the electronic structure of their s- and p-orbitals, with minimal contribution from d-orbitals. The introduced features of s, p-electron count effectively distinguish the impact of main group elements and transition metals on surface catalysis, facilitating the capture of the chemical behavior and catalytic performance of main group elements. The valence electron count overlaps somewhat with the selection of s, p, d electron counts and was initially included in the model to explore feature suitability. Electronegativity and first ionization energy represent an atom's ability to gain and lose electrons, respectively, and electron transfer plays a crucial role in catalytic reactions. The atomic radius influences the arrangement of active atoms on the catalyst surface, electronic distribution, and surface density of states. A smaller atomic radius typically results in shorter interatomic distances on the surface, leading to higher surface electron densities, which can strengthen the bonding between adsorbed hydrogen and the catalyst. This scenario makes the ΔG_H more negative, hindering hydrogen desorption and thereby impeding H₂ production. Conversely, a larger atomic radius often increases the interatomic distances on the surface, reducing the surface electron density and weakening hydrogen adsorption, making ΔG_H more positive. However, if adsorption is too weak, hydrogen cannot stably adsorb on the catalyst surface, which is also unfavorable for H₂ production.

Adsorption bond length refers to the distance between the hydrogen atom and the active site on the catalyst surface. This directly affects the interaction strength between the hydrogen atom and the catalyst surface. Shorter adsorption bond lengths typically correspond to stronger adsorption due to stronger interactions between the hydrogen atom and the catalyst surface atoms. In this case, the hydrogen atom binds more tightly to the surface, resulting in a more negative ΔG_H . Longer adsorption bond lengths, on the other hand, correspond to weaker adsorption due to weaker interactions, making ΔG_H more positive. Guided by prior knowledge, we processed the features using the geometric mean method. This approach averages the involved features and ensures that the processed features of all catalysts remain on the same order of magnitude, facilitating the exploration of factors influencing hydrogen evolution performance. The formula for calculating the geometric mean is shown in Eq. 1. The significance of each

processed feature is detailed in Table S1.

$$X = \left(\prod_{i=1}^n x_i \right)^{\frac{1}{n}} \quad (1)$$

After obtaining the features of HER catalysts, due to the robustness and stability of tree models, we employed six tree-based ML algorithms in this study, namely RFR, GBR, XGBR, DTR, LGBMR and ETR, the detailed parameters of the six ML models are provided in Table S2. We assessed the fitting performance of these six models using metrics such as R-squared (R^2), Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE), which represent the accuracy score, average absolute error, mean squared error, and average error of the models, respectively. The specific prediction results are presented in Fig. 4.

The DTR algorithm achieved an R^2 value of 0.855 on the test set, with MAE, MSE, and RMSE values of 0.131, 0.064, and 0.253, respectively. The GBR algorithm obtained an R^2 value of 0.881 on the test set, with corresponding MAE, MSE, and RMSE values of 0.112, 0.055, and 0.235. RFR achieved an R^2 value of 0.912 on the test set, with MAE, MSE, and RMSE values of 0.110, 0.040, and 0.200, respectively. XGBR yielded an R^2 value of 0.912 on the test set, with MAE, MSE, and RMSE values of 0.107, 0.040, and 0.200. LGBMR resulted in an R^2 value of 0.913 on the test set, with MAE, MSE, and RMSE values of 0.117, 0.039, and 0.198. ETR achieved the highest R^2 value of 0.921 on the test set, with corresponding MAE, MSE, and RMSE values of 0.104, 0.036, and 0.189. Notably, all six tree-based models achieved an R^2 value greater than 0.85 on the test set, indicating their ability to effectively describe the relationship between the selected features and ΔG_H . This suggests that the chosen features can effectively establish the relationship between catalyst features and ΔG_H . Among the six tree model, the ETR algorithm model achieved the highest R^2 value on the test set. Therefore, we selected the ETR model for subsequent predictions and feature importance analysis. In addition, we also compared two non-tree-based algorithms MLP model and SVM model, the predictive accuracy of these two models is shown in Fig. S4. For the MLP model, the fitting accuracy on the training set is $R^2 = 0.847$, with MAE, MSE, and RMSE values of 0.156, 0.070, and 0.264, respectively. On the test set, the predictive accuracy is $R^2 = 0.815$, with MAE, MSE, and RMSE values of 0.178, 0.084, and 0.290, respectively. For the SVM model, the fitting accuracy on the training set is $R^2 = 0.829$, with MAE, MSE, and RMSE values of 0.148, 0.078, and 0.279, respectively. On the test set, the predictive accuracy is $R^2 = 0.809$, with MAE, MSE, and RMSE values of 0.164, 0.086, and 0.294, respectively. The detailed parameters of the MLP model and SVM models are provided in Table S2. The results indicate that the MLP and SVM models exhibit relatively poor fitting performance on both the training and test sets, primarily due to their inability to effectively capture complex nonlinear relationships. In contrast, ensemble tree models, which combine multiple decision trees or optimize

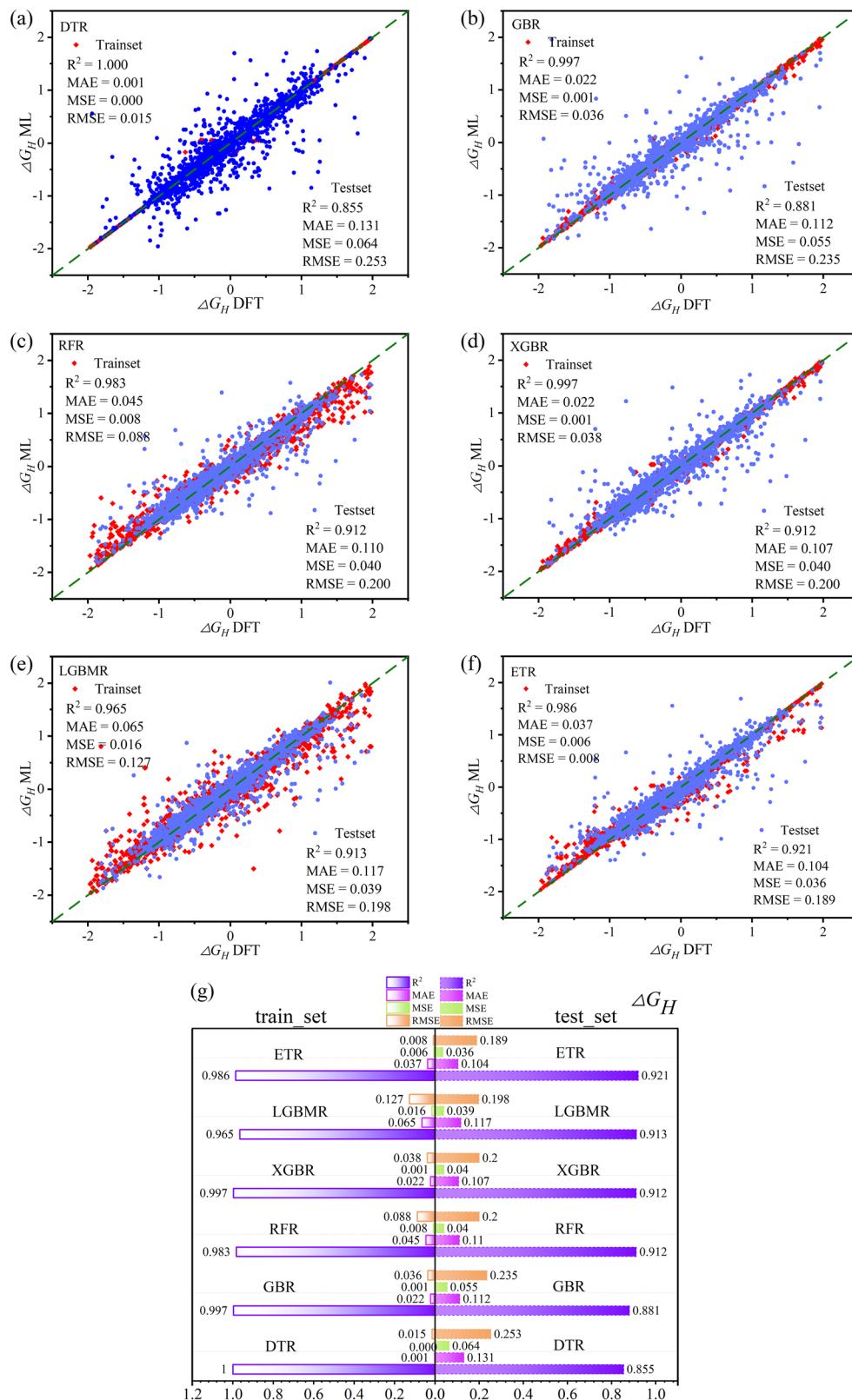


Fig. 4 | Prediction performance of six tree-based ML models on the training and test sets. a–f R^2 score, MAE, MSE, and RMSE for each model: DTR, GBR, RFR, XGBR, LGBMR, and ETR on both the training and test sets, **g** Overall comparison of prediction accuracy among the six ML models.

based on residuals, excel at capturing such intricate nonlinear patterns. Additionally, these models demonstrate strong robustness to feature noise, missing values, and high-dimensional data.

In order to assess the correlation between the selected features and ΔG_H , we utilized Pearson correlation coefficient method (PCCM) to represent their

relationship. Figure S5 presents a heatmap depicting the relationship between features and between features and ΔG_H . Pearson correlation coefficient close to 1 indicates a high correlation between two variables. The examination of Pearson correlation coefficients between the selected features and ΔG_H revealed that the three features with the highest Pearson correlation

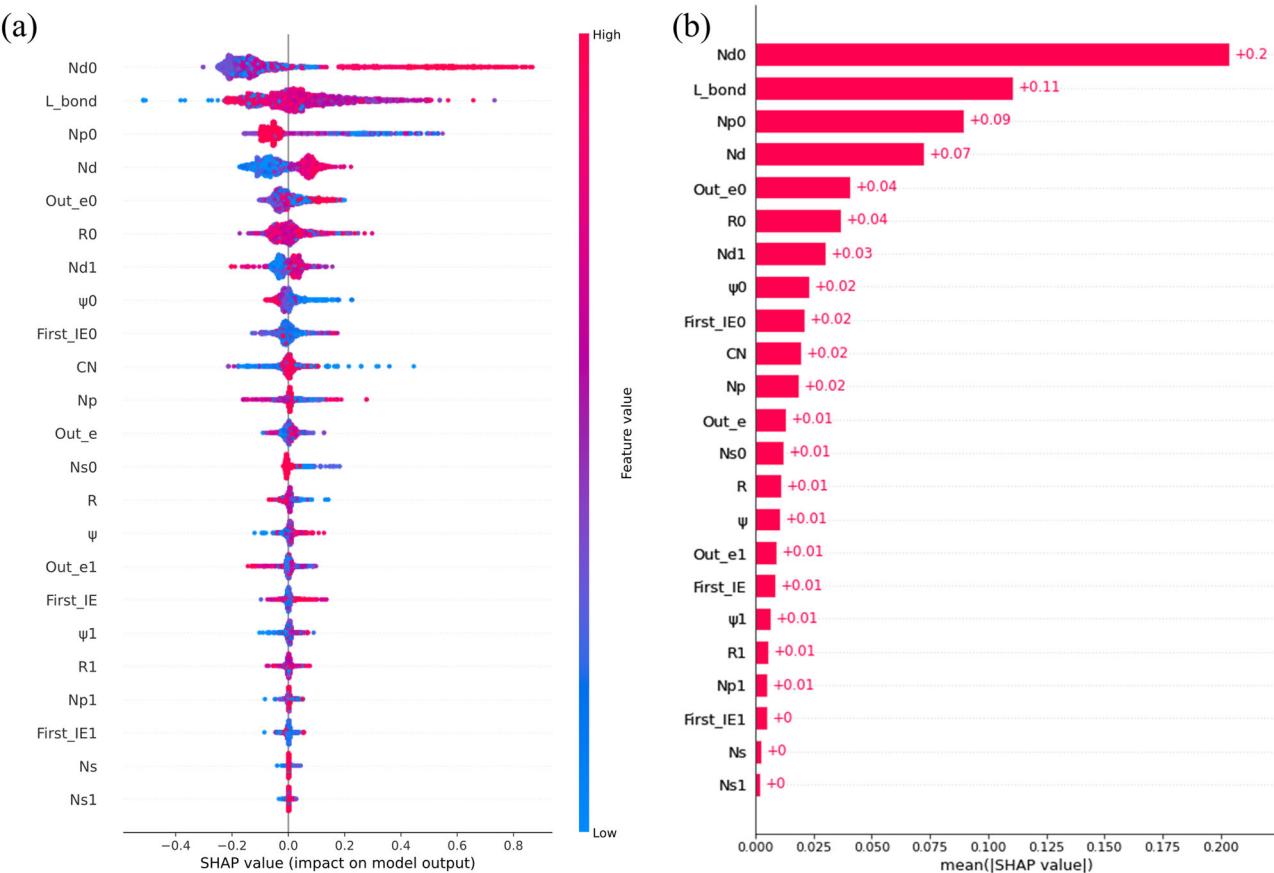


Fig. 5 | SHAP value distribution analysis for evaluating the importance of all features. **a** Global interpretation (average feature importance) and local interpretation (SHAP value distribution) of the ETR model for ΔG_H . **b** The SHAP values of each input feature importance on ETR model for ΔG_H .

coefficients, Nd0, Nd1, and Nd, all exceeded or equaled 0.4, indicating a significant association between the d-electron feature and ΔG_H . This finding is consistent with Nørskov's d-band center theory^{47,48}. As the number of d-electrons increases, the d-band center lowers, resulting in more d-electrons occupying the antibonding orbitals. This leads to increased instability of the catalyst, further weakening the adsorption of H atoms and making ΔG_H more positive. To more intuitively determine the impact of each feature on ΔG_H , we employed the SHAP method to analyze the importance of each feature. The SHAP values provided new insights into the ranking of feature importance. Given that the ETR model performed best on the test set ($R^2 = 0.921$), we conducted a SHAP evaluation on the ETR model. As depicted in Fig. 5a, b, the importance of all features is showcased, with the degree of influence determined by the mean absolute SHAP values across all data points in the dataset. The features are ranked based on their impact on the model output, with feature importance decreasing from top to bottom. In Fig. 5a, if the SHAP value of a feature increases with an increase in the feature value, it indicates a positive correlation with ΔG_H . For instance, the feature Nd0 (The geometric mean of the d electron count of the adsorption center atom) exhibits a positive correlation with ΔG_H , consistent with our aforementioned discussion: as the number of d-electron increases in the active center atom, more d-electrons occupy antibonding orbitals, resulting in weaker adsorption of H atoms and a more positive value of ΔG_H . One noteworthy feature is the feature ψ_0 (The geometric mean of the electronegativity of the adsorption center atom), which exhibits a negative correlation with ΔG_H . A larger value of feature ψ_0 indicates a stronger electron-attracting capability of the active center^{49–51}. When adsorbing hydrogen atoms, the electrons of the hydrogen atom transfer to the catalyst surface. A higher value of feature ψ_0 implies a greater electron transfer from the hydrogen atom to the catalyst surface, resulting in a stronger adsorption of hydrogen atoms by the catalyst and consequently a more negative ΔG_H .

Figure 5b shows that the mean absolute SHAP values of all features were employed as inputs for the voting regressor on database, with Nd0 having the most significant impact on the prediction of ΔG_H values, followed by L_bond (The bond length between the hydrogen atom and the adsorption center). The mean absolute SHAP values of these two features exceed 0.1 eV, indicating that, on average, these features contribute to a prediction variation greater than 0.1 eV.

To clarify the impact of the features Nd0 and ψ_0 on the prediction of ΔG_H , we further analyzed the SHAP values of these features. Figure S6a, b illustrates the impact of features Nd0 and ψ_0 on SHAP values, respectively. It is visually apparent that as the value of feature Nd0 increases, the SHAP values exhibit an upward trend (positively correlated with ΔG_H). Conversely, for feature ψ_0 , an increase in its value leads to a downward trend in SHAP values (negatively correlated with ΔG_H). The SHAP feature analysis results validate the effectiveness of the electronic and structural features constructed in this study. However, given the current use of a large number of features, there is a need for optimization, and more relevant new features await discovery. To further optimize the ML model, we conducted feature engineering by reducing and combining the existing features.

Feature engineering and optimization of ML models

Feature engineering is indispensable for ML, and designing appropriate and comprehensive features is crucial for constructing ML models. Thus, establishing efficient catalytic features is essential for enhancing model accuracy and exploring factors influencing reactions. Feature engineering transforms individual electronic and elemental properties into composite features and develops suitable features for catalytic reactions. Analysis of feature importance using the SHAP method reveals positive correlations between the geometric mean Nd0 of the number of d electrons in active centers and the geometric mean ψ_0 of the electronegativity of active centers

with ΔG_H , while negative correlations are observed. Based on the above discussions, we introduce a new energy-related feature φ , defined as follows:

$$\varphi = \frac{Nd0^2}{\psi 0} \quad (2)$$

Analysis of this formula reveals that $Nd0^2$ and $1/\psi 0$ are positively correlated with ΔG_H , indicating that the newly introduced feature φ is also positively correlated with ΔG_H .

Based on the d-band model and Muffin-Tin orbital theory^{52,53}, we established a relationship between the adsorption energy and the square of the d-electron count. According to these theories, the adsorption energy (E_{ad}) on a metal surface is closely related to the coupling strength of the metal's d-orbitals with the adsorbate. This coupling strength is represented by the coupling Hamiltonian element V_{ad} , which is influenced by the spatial extent of the metal d-orbitals (r_d) and the distance between the adsorbate and the metal surface (L): $E_{ad} \propto (V_{ad})^2 \propto \frac{(r_d)^3}{L^2}$. Here, r_d is the spatial extent of the metal's d-orbitals, which is directly linked to the number of d-electrons (Nd0) at the active center atom. Specifically, Nd0 affects both the position of the metal's d-band center and the spatial extent of the d-orbitals. Thus, r_d can be indirectly reflected by the electronic structure of the metal.

From the above relationship, we observe that the adsorption energy E_{ad} is proportional to the spatial extent r_d , which is influenced by the d-electron count. Specifically, as the d-electron count increases, the spatial extent r_d of the d-orbitals grows, enhancing the coupling strength between the metal and the adsorbate. Consequently, the adsorption energy exhibits a quadratic relationship with the d-electron count, $E_{ad} \propto (V_{ad})^2$. This quadratic relationship reflects the synergistic effects among the metal's d-orbital electrons and how the overlap of the electron clouds impacts the strength of electronic interactions between the adsorbate and the metal surface. As the d-electron count increases, the metal's d-orbital electron cloud expands, intensifying the overlap of electron clouds. This overlap is not merely additive but results in a nonlinear increase in adsorption strength due to electron-electron interactions, such as electron repulsion and orbital overlap effects, which are captured by the quadratic relationship. Additionally, the adsorption distance L in our model is indirectly estimated using the metal's electronegativity, further linking the electronic structure of the metal to the adsorption energy. This approach also incorporates the relationship between the metal's electronegativity and the d-electron count into the adsorption energy model, the physical and chemical significance of the newly introduced feature φ has been further clarified.

Furthermore, PCCM analysis found a high correlation between features of the nearest active site and the overall features of active sites. However, an excessive number of features during ML training can diminish training efficiency and impact prediction accuracy. To address this issue, appropriate measures must be taken to eliminate redundant information in the dataset. It is noteworthy that, due to the "curse of dimensionality," a relatively large feature space does not necessarily result in more accurate predictions. In high-dimensional spaces, data become sparse, leading to model overfitting and increased computation time. Therefore, we implemented a rigorous feature selection process to eliminate features with low importance or high correlation.

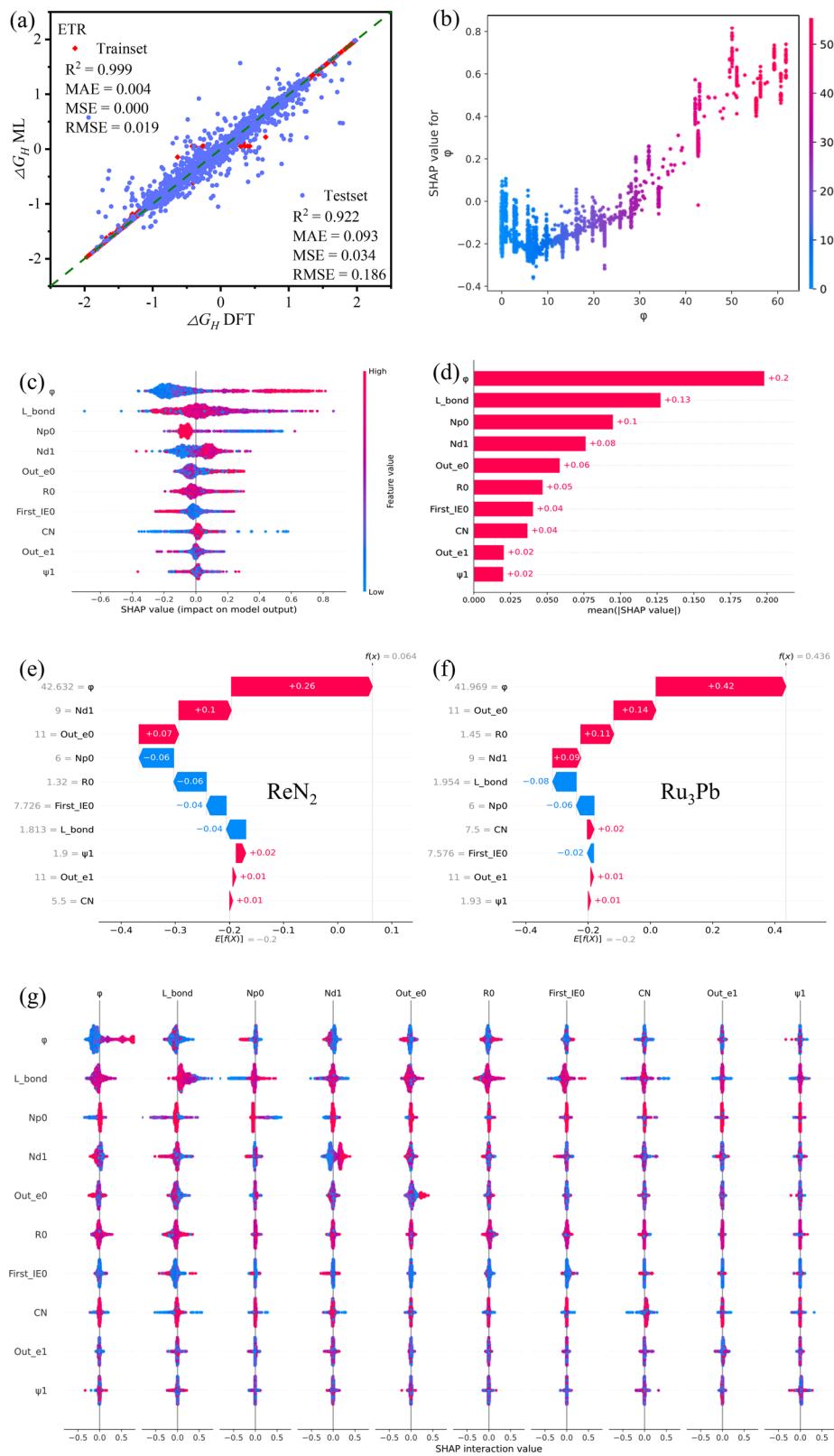
The features adjustments are as follows: 1) Features with a Pearson correlation coefficient greater than 85% were removed if their SHAP values were low. 2) A new feature, φ , was introduced, and the two correlated features, Nd0 and $\psi 0$, were removed. 3) Features with low SHAP values were eliminated. After feature engineering, only 10 features remained in Fig. 6c. Subsequently, an ETR model was constructed using these ten features, as illustrated in Fig. 6a. The ML model built solely using these 10 features achieved an R^2 value of 0.922, with MAE, MSE, and RMSE values of 0.039, 0.034, and 0.186 on the test set, respectively. This model outperformed the ML model using 23 features, demonstrating the effectiveness of our feature engineering and the newly constructed feature φ has a higher correlation

with ΔG_H and plays a crucial role in predicting the activity of various HECs, the detailed parameters of the ETR model with ten features are provided in Table S3. Figure 6b presents the SHAP values corresponding to the newly introduced feature φ , showing that the SHAP values increase as the feature φ increases. This indicates a clear positive correlation between feature φ and ΔG_H . As the value of feature φ increases, the ΔG_H becomes larger, implying a weaker adsorption capacity of the catalyst for H atoms. The SHAP values for importance analysis of 10 features are shown in Fig. 6c, d. The figures also illustrate a strong positive correlation between the energy-related feature φ and ΔG_H . Moreover, φ has the highest mean absolute SHAP value, indicating that it has the most significant impact on ΔG_H . In addition to dimensionality reduction based on physical and chemical insights and SHAP value analysis, the effects of alternative dimensionality reduction methods on the accuracy of the ETR model were investigated. L1 regularization mitigates these challenges by zeroing out the weights of certain features, thereby reducing the number of features and alleviating the risk of overfitting in high-dimensional spaces. Additionally, dimensionality reduction techniques, such as Principal Component Analysis (PCA) and t-Distributed Stochastic Neighbor Embedding (t-SNE), can effectively lower model complexity and improve performance. Using L1 regularization, the feature set was reduced to 9 dimensions, retaining the features Nd0, Nd1, Nd, Np0, Np, Out_e0, Out_e1, $\psi 0$, and First_IE, with the ETR model achieving a test set prediction accuracy of $R^2 = 0.786$. PCA reduced the feature set to ten dimensions, resulting in a test set prediction accuracy of $R^2 = 0.890$. For t-SNE, the feature set was reduced to 2 and 3 dimensions, yielding test set prediction accuracies of $R^2 = 0.873$ and $R^2 = 0.876$, respectively. Detailed model performance metrics are provided in Fig. S7. These comparisons demonstrate that dimensionality reduction to 10 dimensions, guided by physical and chemical insights and SHAP value analysis, is a reasonable approach that further enhances the accuracy of the ETR model.

To demonstrate the predictive performance of the features on the ΔG_H of catalysts, we selected two catalysts, ReN_2 and Ru_3Pb , from the test set and predicted their ΔG_H using ETR model. The predicted results are shown in Fig. 6e, f. The DFT-calculated (ML-predicted) ΔG_H values for ReN_2 and Ru_3Pb are 0.063 eV (0.064 eV) and 0.430 eV (0.436 eV), respectively. Among the features, φ made the largest contribution to the predicted ΔG_H values, highlighting the importance of φ in predicting ΔG_H . Figure 6g displays the interaction plot for all features, where φ and Nd1 show a strong positive correlation with ΔG_H , and no significant correlations are observed between the features themselves.

To further demonstrate the predictive accuracy of the ML model, we compared it with two DL graph neural network models. Here, we selected the deep learning models CGCNN and OGCNN for comparison because both have demonstrated outstanding performance in predicting crystal properties, particularly in the fields of solid-state materials and materials science^{54,55}. For example, in the original literature on CGCNN and OGCNN, the CGCNN model achieves excellent prediction accuracy, with errors of 0.039 eV and 0.072 eV for material formation energy and absolute total energy from the Materials Project database, respectively. Additionally, the prediction errors for the band gap, Fermi energy, and Poisson ratio are 0.388 eV, 0.363 eV, and 0.030, respectively. The OGCNN model also demonstrates outstanding predictive performance, with R^2 values of 0.996, 0.91, and 0.91 for material formation energy, Fermi energy, and band gap, respectively. Furthermore, recent works have also demonstrated their superior performance. For instance, Jin-Soo Kim and colleagues utilized the CGCNN model to predict the histogram of decomposition enthalpy and energy bands for inorganic perovskites⁵⁶, achieving R^2 values of 1 and 0.986, with mean absolute errors of 0.449 meV atom⁻¹ and 0.037 eV, respectively. These models are considered benchmark DL models in the field of crystal structure prediction in recent years, excelling in both accuracy and generalization capability. Both DL models were fully trained. Figure 7a, c shows the loss function curves on training set and validation set for CGCNN and OGCNN, respectively. As can be seen from these figures, after 300 epochs, the loss of both models converged, indicating that both models had

Fig. 6 | Model performance and feature analysis after feature engineering. **a** Accuracy of the ETR model on the training and testing sets with ten features. **b** The SHAP value distribution of feature φ . **c** Global interpretation (average feature importance) and local interpretation (SHAP value distribution) and **d** SHAP values of each input feature importance on the ETR model with ten features ΔG_H . The contribution of each feature to ΔG_H predictions for **e** ReN_2 and **f** Ru_3Pb . **g** The interaction plot for all features based on importance.



adequately learned the dataset. Figure 7b, d displays the learning curves of CGCNN and OGCNN on the training, test, and validation sets. The R^2 values of the CGCNN and OGCNN models for the test set predictions of ΔG_H are 0.913 and 0.921, respectively, both lower than the R^2 of 0.922 for the ETR model, further confirming the high accuracy of the proposed model and the importance of the selected features.

Additionally, we independently assessed the impact of the two most significant features, φ and L_bond , on the accuracy of the ETR model, we established a ML model using only one feature φ to predict the test set. As shown in Fig. S8a, the model's accuracy with just one feature φ was $R^2 = 0.51$. The prediction of ΔG_H for the test set data using only feature φ exhibited a clear linear relationship, indicating the potential application value of feature

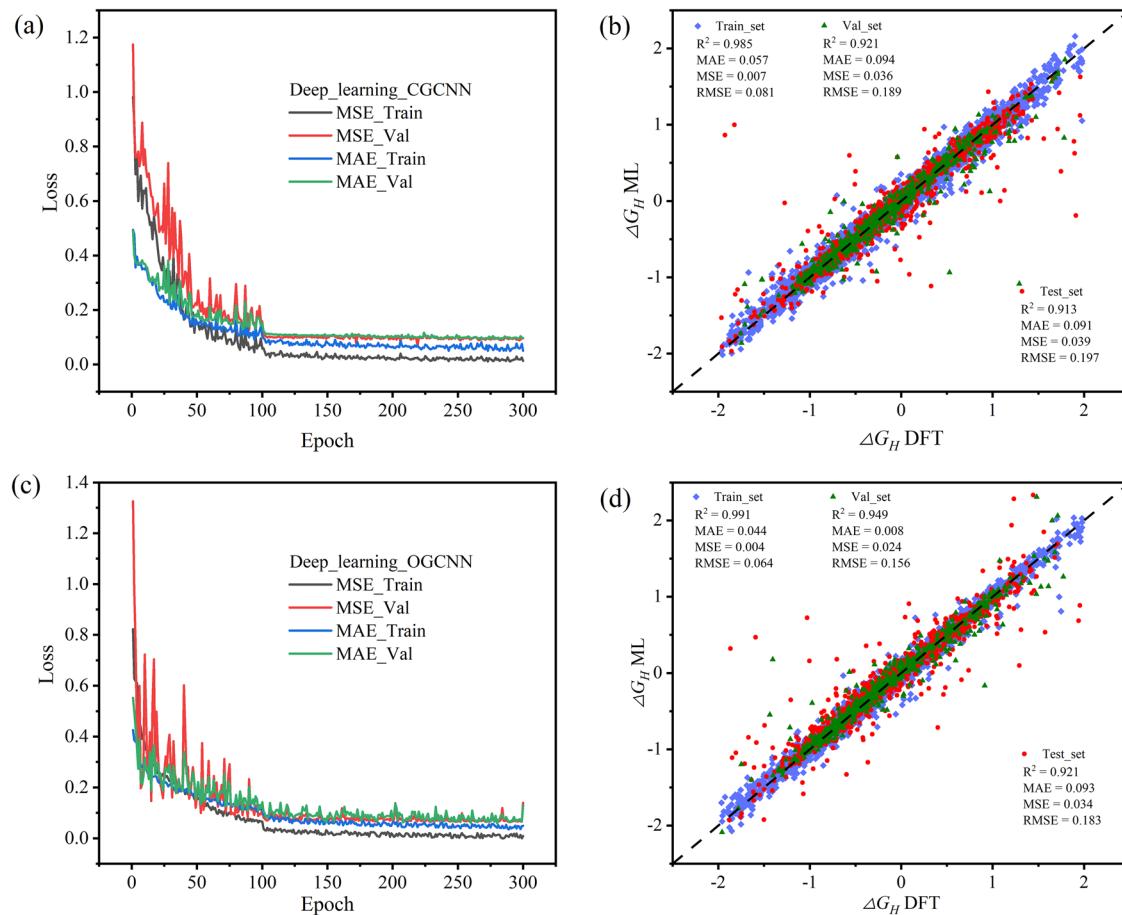


Fig. 7 | Prediction accuracy of two DL models based on graph convolutional networks. Loss function curves on training set and validation set for **a** CGCNN and **c** OGCNN, learning curves of **b** CGCNN and **d** OGCNN on the training, test, and validation sets.

φ in predicting ΔG_H . To further explore the key factors influencing the performance of HECs, we established a ML model using only two features, φ and L_{bond} , to predict the test set, as shown in Fig. S8b. The predictive accuracy achieved an R^2 value of 0.741. Using only these two features, the hydrogen evolution performance of the catalyst was accurately predicted, indicating that the feature φ and L_{bond} are crucial in determining the hydrogen evolution performance of the catalyst.

High-Activity Catalyst Screening

The purpose of this ML model is to predict the catalytic activity of potential catalysts for HER. The ETR model, developed using the 10 features we devised, was utilized to predict the ΔG_H of various HECs obtained from the Material Project database⁵⁷. Subsequently, we validated the predictions of the ML model using DFT calculations. To ensure consistency between the DFT computational settings and those of the ML dataset, we decided not to account for the effects of explicit or implicit water in this study. The catalysts encompassed various types, including transition metal intermetallic compounds, light metal intermetallic compounds, non-metallic compounds, and perovskites, a total of 132 adsorption site Gibbs free energy calculations were performed. It is noteworthy that the validation set we used does not overlap with the training set utilized to construct the ML model. Predictions made on this validation set serve to validate the accurate extrapolation capability of our ML model. We employed the ML model to predict the ΔG_H of these HECs, followed by detailed calculations using DFT. To validate the accuracy of our ML model, we compared the results of DFT calculations with the predictions of the ML model. Figure 8a illustrates the comparative analysis between DFT calculations and ML predictions of ΔG_H for the selected catalysts, while Table S4 enumerates the results of DFT

calculations and ML model predictions for each catalyst. The R^2 value of the ML model predictions on the validation set is 0.878, with MAE, MSE, and RMSE values of 0.167, 0.040, and 0.200, respectively. These results demonstrate a high degree of consistency between our ML predictions and the DFT calculations of ΔG_H , affirming the accuracy of our model. Based on the data predicted by the ML model, several high-activity HECs have been identified. To provide a more intuitive evaluation of the HER performance of these catalysts, we calculated the theoretical overpotential of the catalysts. The relationship between the theoretical overpotential and $\eta = |\Delta G_H|/e$. Figure 8b-d illustrate three excellent HECs, representing AB-type, AB₃-type, and perovskite-type structures: ReIr (id: mp-1219533) with vacancy H adsorption, Re₃W (id: mp-974416) with atop H adsorption, and BaNdO₃ (id: mp-54307) with atop H adsorption, with the numbers in parentheses corresponding to their IDs in the MP database. The DFT-calculated ΔG_H values for ReIr, Re₃W, and BaNdO₃ are 0.009 eV, 0.014 eV, and -0.416 eV, respectively, while the ML model predicts ΔG_H as -0.174 eV, -0.017 eV, and -0.117 eV. The differences in ΔG_H between the DFT calculations and ML predictions for these three catalysts are 0.185 eV, 0.031 eV, and 0.299 eV, respectively. The MAE for the three catalysts is 0.171 eV, which is very close to the MAE of 0.167 in Fig. 8(a), reflecting the universality and accuracy of the ML model's predictions, and further indicating that ReIr and Re₃W are promising HECs. Moreover, efficiency comparisons between DFT calculations and ETR model predictions reveal a 200,000-fold enhancement in efficiency using ML model for predicting ΔG_H of 132 adsorption site in the validation set in Figure S9, facilitating rapid screening of highly active HECs and substantially reducing time and computational costs, addressing a long-standing challenge in identifying superior catalysts from a large pool of candidates using traditional DFT and

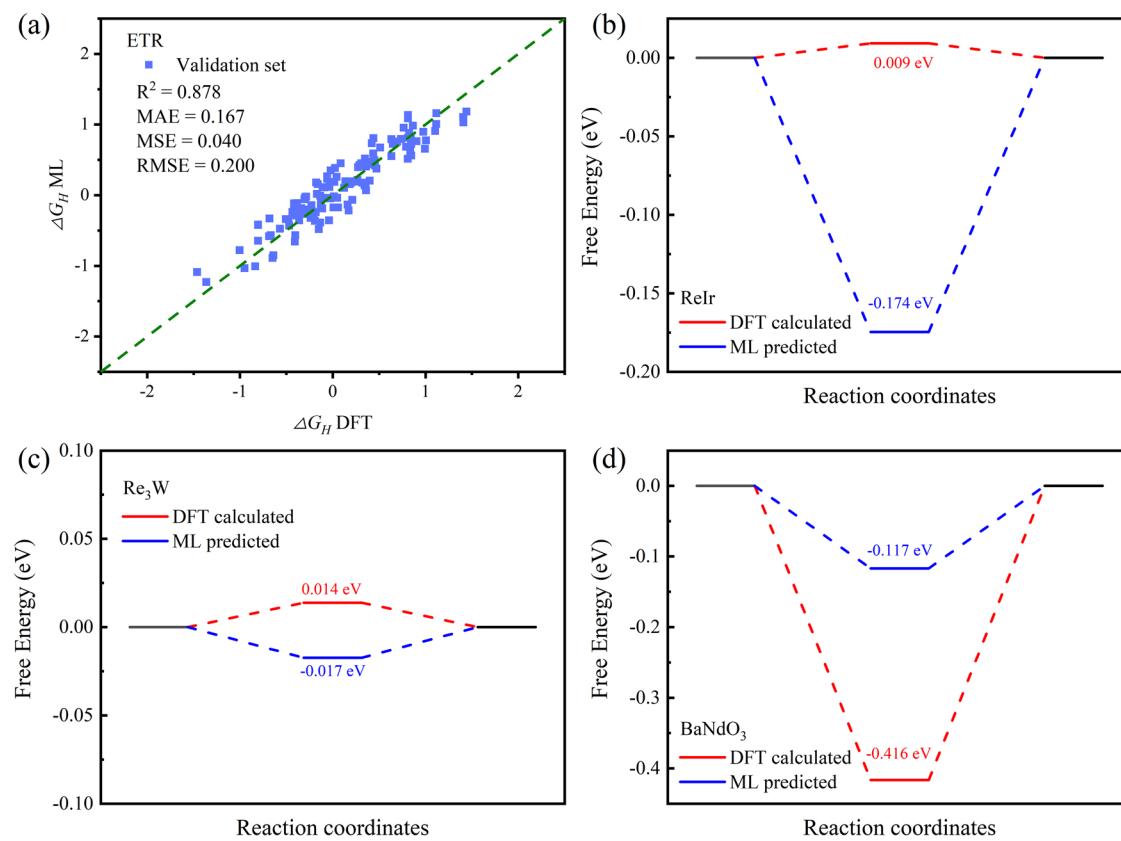


Fig. 8 | ML model prediction of potential HERCs. **a** Comparison plot of ETR model predictions and DFT calculations for the hydrogen evolution free energy ΔG_H of 132 catalysts. Comparison of the DFT-calculated and ML-predicted reaction pathway and ΔG_H values for the HER of **b** ReIr, **c** Re_3W , and **d** $BaNdO_3$, respectively.

experimental approaches, thus aiding in the accelerated development and practical deployment of catalysts.

Discussion

In this work, a dataset comprising 10,855 HER catalysis data are collected from the Catalysis-Hub database for training and testing six ML models. Feature importance analysis and feature engineering techniques were employed to minimize the number of features and introduce a new composite feature, $\varphi = Nd^2/\psi_0$, which exhibited a strong positive correlation with the HER ΔG_H and had a clear physical interpretation regarding the HER activity. A precise and efficient ML model was established using only 10 features based on active sites, without requiring additional DFT calculations, to predict various types of HER catalysts. Through ten-fold cross-validation, the ETR model achieved an R^2 score of 0.922 on the test set, with MAE, MSE, and RMSE values of 0.039, 0.034, and 0.186, respectively. Additionally, we compared two deep learning models, CGCNN and OGCNN, and found that the prediction accuracy of our ML model surpassed both. To further validate the effectiveness of the ML model in predicting various types of HER catalysts, the ΔG_H values of 132 catalysts were predicted and compared with DFT-calculated results. Promising HERCs were identified based on the predictions of the ML model. Compared to costly DFT calculations, the ML model achieved a 200,000-fold increase in time efficiency in predicting HER catalysts. The ML model developed in this work can predict the adsorption free energy for various types of catalysts, as well as the adsorption free energy for different adsorption sites, thereby aiding in the screening of potential HER catalysts. By using the ML model to predict the HER free energy at each adsorption site and evaluating the adsorption strength of these sites, it is possible to identify the optimal adsorption site for the catalyst and assess its catalytic activity. This work presents practical solutions for discovering high-performance HER catalysts within the vast space of catalysts and provides insights for the design of other electrocatalysts.

Methods

Density functional theory calculations

All calculations were performed based on density functional theory (DFT) by means of the Vienna ab initio simulation package (VASP) and DS-PAW^{58–60}. The projector-augmented wave method is used to describe the ion–electron interaction with a cutoff energy of 500 eV⁶⁰, which is tested to be precise enough with high efficiency. Figure S1 shows the variation in total energy of AgPd surface as a function of cutoff energy ranging from 50 eV to 500 eV, with a step size of 50 eV. The results indicate that a cutoff energy of 500 eV provides sufficient accuracy. The exchange–correlation interaction is determined by the Perdew–Burke–Ernzerhof functional on the framework of the generalized gradient approximation⁶¹. For all the calculations, the vacuum space in the z-direction was set as 15 Å to avoid potential interaction between periodic surfaces. The bottom two atomic layers were fixed, while the remaining atoms were fully relaxed to allow structural optimization. This approach ensures an accurate representation of the surface structure during catalytic reactions and maintains the simulated system in a physically realistic state. In this study, all our hydrogen adsorption models are designed under low-coverage conditions, with hydrogen atom coverages below 25%. According to the Langmuir adsorption theory and the findings of Nørskov et al.^{36,62}, the interactions between adsorbed atoms are negligible at low coverage, and the coverage effect can generally be disregarded. The energy convergence criterion is 10^{-5} eV, while the atomic force is set as 0.01 eV/Å during the relaxing process. A $3 \times 3 \times 1$ gamma-centered k-point grid is used. The effect of van der Waals (vdW) interactions was considered using the DFT-D3 method⁶³.

The catalyst's activity trend towards the HER can be elucidated by the Gibbs free energy associated with hydrogen adsorption on its surface. Following the Sabatier principle, optimal catalytic activity is achieved when the Gibbs free energy of hydrogen adsorption approaches the thermoneutral value close to zero⁶⁴. The Gibbs free energy of hydrogen adsorption, denoted

as ΔG_H in the following, can be calculated as:

$$\Delta G_H = \Delta E_H + \Delta E_{ZPE} - T \Delta S_H \quad (3)$$

where ΔE_H is the hydrogen adsorption energy, defined as:

$$\Delta E_H = \Delta E_{\text{slab}-H} - \Delta E_{\text{slab}} - \frac{1}{2} E_{H^2} \quad (4)$$

where $\Delta E_{\text{slab}-H}$ is the total energy of a hydrogen-adsorbed slab structure, ΔE_{slab} is the total energy of slab structure without a hydrogen atom, and E_{H^2} denotes the energy of an isolated hydrogen gas molecule. ΔE_{ZPE} and ΔS_H are the changes in the zero-point energies (ZPE) and entropy of hydrogen in the adsorbed state, respectively, obtained from the vibrational frequency calculations. For the calculation of vibrational frequencies of H atoms, all other atoms were fixed except for the hydrogen atoms, and the vibrational frequencies of hydrogen were computed. After obtaining the frequency calculation results, the free energy corrections of the system were determined using the VASP Toolkit⁶⁵. Specifically, the frequencies obtained from the calculation were used to compute the zero-point energy correction, which was then combined with thermodynamic formulas to estimate the temperature-dependent free energy corrections. T represents the temperature and was set as 298.15 K. In Nørskov's work, When the absolute value of ΔG_H is zero, the HER catalytic activity is the best^{64,66}.

ML models

Efficient and suitable ML models are essential to explore the relationship between structure and catalytic activity, and in order to ensure the stability, reliability, and nonlinear processing capability of the models. The robustness and interpretability of tree models are widely employed in ML models. The ML process is performed using six tree-based algorithms, namely RFR, GBR, XGBR, DTR, LGBMR, and ETR. In addition, we also compared two non-tree-based algorithms: the Multilayer Perceptron (MLP, a neural network model) model and the Support Vector Machines (SVM) model. These algorithms can be seamlessly integrated with the open-source Scikit-learn library. The input data from DFT calculations are randomly divided into a training set and a testing set, with a ratio of 4:1. To demonstrate the stability and accuracy of the ML model, a 10-fold cross-validation (CV) was applied to all algorithms to find the optimal combination of hyperparameters for all model. Grid search is an exhaustive search method that sets a group of candidate values for each hyperparameter, then generates the Cartesian product of these candidate values to form a grid of hyperparameter combinations. Subsequently, grid search trains and evaluates the model for each combination of hyperparameters to find the combination with the best performance, thus improving the prediction accuracy and robustness of the model. The optimal hyperparameters for specific algorithm models are detailed in the supplementary Information. In addition, two graph neural network-based DL models, CGCNN and OGCNN, were employed. Both models were trained for a maximum of 300 epochs with a batch size of 512, and the Adam optimizer was used for parameter optimization. The dataset was split into training, testing, and validation sets, comprising 80%, 15%, and 5% of the data, respectively. The accuracy evaluation metrics for the ML and DL models include four measures, R^2 -Square (R^2), Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE). The calculation formula of the four evaluation indicators is as follows:

$$R^2 = 1 - \frac{\sum_i^n (y_i - f(x_i))^2}{\sum_i^n (y_i - \bar{y})^2} \quad (5)$$

$$\text{MAE} = \frac{1}{n} \sum_i^n |f(x_i) - y_i| \quad (6)$$

$$\text{MSE} = \frac{1}{n} \sum_i^n (f(x_i) - y_i)^2 \quad (7)$$

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_i^n (f(x_i) - y_i)^2} \quad (8)$$

where $f(x_i)$ is the predicted value of the model; y_i is the true value; \bar{y} is the mean value; MAE, MSE and RMSE can be regarded as the prediction error and R^2 can be approximately regarded as the accuracy of regression fitting.

In order to evaluate the dependence between our ML features, the PCCM was applied to evaluate the relevance between two features⁶⁷, which can be expressed as:

$$p = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2} \sqrt{\sum_i (y_i - \bar{y})^2}} \quad (9)$$

where x and y are two features and \bar{x} and \bar{y} are the corresponding mean values. p ranges from -1 to 1. When p approaches 1, there is a linear relationship between the two features.

Subsequently, Shapley Additive Explanations (SHAP) analysis was conducted on the model⁶⁸. SHAP stands out as one of the widely accepted methods for elucidating ML models. In this approach, each feature is assigned an importance scale, where a higher absolute SHAP value denotes a more substantial contribution to the outcomes of ML models. Moreover, a positive or negative SHAP value signifies that the feature exerts a positive or negative effect on the prediction.

Data availability

The data that support the findings of this study are available from https://github.com/wangchaobjut/Multi_Type_HERs.git.

Code availability

The data that support the findings of this study are available from https://github.com/wangchaobjut/Multi_Type_HERs.git.

Received: 9 September 2024; Accepted: 12 March 2025;

Published online: 24 April 2025

References

1. Chu, S. & Majumdar, A. Opportunities and challenges for a sustainable energy future. *Nature* **488**, 294–303 (2012).
2. Tachibana, Y., Vayssières, L. & Durrant, J. R. Artificial photosynthesis for solar water-splitting. *Nat. Photonics* **6**, 511–518 (2012).
3. Lewis, N. S. Research opportunities to advance solar energy utilization. *Science* **351**, aad1920 (2016).
4. Zhu, W., Bian, Z. & Lu, Y. Environmental control system for pig farm based on mobile coordinator routing algorithm. Precision Livestock Farming 2019—Papers Presented at the 9th European Conference on Precision Livestock Farming, ECPLF 2019, 851–857 (2019).
5. Nørskov, J. K. & Christensen, C. H. Toward efficient hydrogen production at surfaces. *Science* **312**, 1322–1323 (2006).
6. Patra, B. C. et al. A metal-free covalent organic polymer for electrocatalytic hydrogen evolution. *ACS Catal.* **7**, 6120–6127 (2017).
7. Liu, F. et al. Rational design of better hydrogen evolution electrocatalysts for water splitting: a review. *Adv. Sci.* **9**, 2200307 (2022).
8. Zhang, J. et al. Electrocatalysts design guided by active intermediates of hydrogen evolution reaction. *Adv. Energy Mater.* **13**, 2302436 (2023).
9. Strmcnik, D., Lopes, P. P., Genorio, B., Stamenkovic, V. R. & Markovic, N. M. Design principles for hydrogen evolution reaction catalyst materials. *Nano Energy* **29**, 29–36 (2016).

10. Lu, J., Yin, S. & Shen, P. K. Carbon-encapsulated electrocatalysts for the hydrogen evolution reaction. *Electrochem. Energy Rev.* **2**, 105–127 (2019).
11. Zhong, W. et al. Ultralow-temperature assisted synthesis of single platinum atoms anchored on carbon nanotubes for efficiently electrocatalytic acidic hydrogen evolution. *J. Energy Chem.* **51**, 280–284 (2020).
12. Feidenhans'l, A. A. et al. Precious metal free hydrogen evolution catalyst design and application. *Chem. Rev.* **124**, 5617–5667 (2024).
13. Liu, J. et al. Rationally designing efficient electrocatalysts for direct seawater splitting: challenges, achievements, and promises. *Angew. Chem. Int. Ed.* **61**, e202210753 (2022).
14. Mahmood, N. et al. Electrocatalysts for hydrogen evolution in alkaline electrolytes: mechanisms, challenges, and prospective solutions. *Adv. Sci.* **5**, 1700464 (2018).
15. Tran, K. & Ulissi, Z. W. Active learning across intermetallics to guide discovery of electrocatalysts for CO_2 reduction and H_2 evolution. *Nat. Catal.* **1**, 696–703 (2018).
16. Gu, Y. et al. Single atom-modified hybrid transition metal carbides as efficient hydrogen evolution reaction catalysts. *Adv. Funct. Mater.* **31**, 2104285 (2021).
17. Wang, Q. et al. Recent progress in high-entropy alloy electrocatalysts for hydrogen evolution reaction. *Adv. Mater. Interfaces* **11**, 2301020 (2024).
18. Mei, J., Deng, Y., Cheng, X., Wang, X. & Wu, Q. Recent advances in iron-based sulfides electrocatalysts for oxygen and hydrogen evolution reaction. *Chin. Chem. Lett.* **35**, 108900 (2024).
19. Brunin, G., Ricci, F., Ha, V.-A., Rignanese, G.-M. & Hautier, G. Transparent conducting materials discovery using high-throughput computing. *npj Comput. Mater.* **5**, 63 (2019).
20. Yang, C. et al. Overcoming immiscibility toward bimetallic catalyst library. *Sci. Adv.* **6**, eaaz6844 (2020).
21. Freeze, J. G., Kelly, H. R. & Batista, V. S. Search for catalysts by inverse design: artificial intelligence, mountain climbers, and alchemists. *Chem. Rev.* **119**, 6595–6612 (2019).
22. Wang, M. & Zhu, H. Machine learning for transition-metal-based hydrogen generation electrocatalysts. *ACS Catal.* **11**, 3930–3937 (2021).
23. Wang, C. et al. Efficient machine learning model focusing on active sites for the discovery of bifunctional oxygen electrocatalysts in binary alloys. *ACS Appl. Mater. Interfaces* **16**, 16050–16061 (2024).
24. Mai, H., Le, T. C., Chen, D., Winkler, D. A. & Caruso, R. A. Machine learning for electrocatalyst and photocatalyst design and discovery. *Chem. Rev.* **122**, 13478–13515 (2022).
25. Li, J. et al. Machine learning-assisted low-dimensional electrocatalysts design for hydrogen evolution reaction. *Nano-Micro Lett.* **15**, 227 (2023).
26. Sun, X. et al. Machine-learning-accelerated screening of hydrogen evolution catalysts in MBenes materials. *Appl. Surf. Sci.* **526**, 146522 (2020).
27. Chen, L. et al. A universal machine learning framework for electrocatalyst innovation: a case study of discovering alloys for hydrogen evolution reaction. *Adv. Funct. Mater.* **32**, 2208418 (2022).
28. Ren, C. et al. A universal descriptor for complicated interfacial effects on electrochemical reduction reactions. *J. Am. Chem. Soc.* **144**, 12874–12883 (2022).
29. Gao, W. et al. Determining the adsorption energies of small molecules with the intrinsic properties of adsorbates and substrates. *Nat. Commun.* **11**, 1196 (2020).
30. Lu, Z., Chen, Z. W. & Singh, C. V. Neural network-assisted development of high-entropy alloy catalysts: decoupling ligand and coordination effects. *Matter* **3**, 1318–1333 (2020).
31. Zhang, J. et al. Accurate and efficient machine learning models for predicting hydrogen evolution reaction catalysts based on structural and electronic feature engineering in alloys. *Nanoscale* **15**, 11072–11082 (2023).
32. Umer, M. et al. Machine learning assisted high-throughput screening of transition metal single-atom-based superb hydrogen evolution electrocatalysts. *J. Mater. Chem. A* **10**, 6679–6689 (2022).
33. Wei, C. et al. Data-driven design of double-atom catalysts with high H_2 evolution activity/ CO_2 reduction selectivity based on simple features. *J. Mater. Chem. A* **11**, 18168–18178 (2023).
34. Winther, K. T. et al. Catalysis-Hub.org, an open electronic structure database for surface reactions. *Sci. Data* **6**, 75 (2019).
35. Torres-Beltrán, M. et al. Author Correction: a compendium of geochemical information from the Saanich Inlet water column. *Sci. Data* **6**, 1 (2019).
36. Nørskov, J. K. et al. Trends in the exchange current for hydrogen evolution. *J. Electrochem. Soc.* **152**, J23 (2005).
37. Li, H. et al. Computational design of (100) alloy surfaces for the hydrogen evolution reaction. *J. Mater. Chem. A* **8**, 17987–17997 (2020).
38. Fung, V., Hu, G., Wu, Z. & Jiang, D.-E. Descriptors for hydrogen evolution on single atom catalysts in nitrogen-doped graphene. *J. Phys. Chem. C* **124**, 19571–19578 (2020).
39. Wexler, R. B., Martinez, J. M. P. & Rappe, A. M. Chemical pressure-driven enhancement of the hydrogen evolving activity of Ni2P from nonmetal surface doping interpreted via machine learning. *J. Am. Chem. Soc.* **140**, 4678–4683 (2018).
40. Lin, S., Xu, H., Wang, Y., Zeng, X. C. & Chen, Z. Directly predicting limiting potentials from easily obtainable physical properties of graphene-supported single-atom electrocatalysts by machine learning. *J. Mater. Chem. A* **8**, 5663–5670 (2020).
41. Ge, L. et al. Predicted optimal bifunctional electrocatalysts for the hydrogen evolution reaction and the oxygen evolution reaction using chalcogenide heterostructures based on machine learning analysis of *in silico* quantum mechanics based high throughput screening. *J. Phys. Chem. Lett.* **11**, 869–876 (2020).
42. Zheng, J. et al. High-throughput screening of hydrogen evolution reaction catalysts in MXene materials. *J. Phys. Chem. C* **124**, 13695–13705 (2020).
43. Parker, A. J., Opletal, G. & Barnard, A. S. Classification of platinum nanoparticle catalysts using machine learning. *J. Appl. Phys.* **128**, 014301 (2020).
44. Norsko, J. K. Chemisorption on metal surfaces. *Rep. Prog. Phys.* **53**, 1253 (1990).
45. Nørskov, J. K. Electronic factors in catalysis. *Prog. Surf. Sci.* **38**, 103–144 (1991).
46. Oya, N., Ikezaki, T. & Yoshie, N. A crystalline supramolecular polymer with self-healing capability at room temperature. *Polym. J.* **45**, 955–961 (2013).
47. Nørskov, J. K. Theory of chemisorption and heterogeneous catalysis. *Physica B+C* **127**, 193–202 (1984).
48. Ruban, A., Hammer, B., Stoltze, P., Skriver, H. L. & Nørskov, J. K. Surface electronic structure and reactivity of transition and noble metals. *J. Mol. Catal. A: Chem.* **115**, 421–429 (1997).
49. Singh, H., Prendergast, D. & Nath, M. Modulation of electrocatalytic activity by tuning anion electronegativity: case study with copper chalcogenides. *J. Phys. Energy* **5**, 045016 (2023).
50. Wang, M. et al. Boosting electrocatalysis activities of 2D ultrathin BiOX/rGO ($\text{X}=\text{F}, \text{Cl}, \text{Br}, \text{I}$) nanosheets as sulfur hosts: insight into the electronegativity effect of halogenated elements on the electrochemical performances of lithium-sulfur batteries. *Inorganic Chem. Front.* **11**, 4277–4287 (2024).
51. Pan, S. et al. Rational modulating electronegativity of substituents in amorphous metal-organic frameworks for water oxidation catalysis. *Int. J. Hydrog. Energy* **45**, 9723–9732 (2020).

52. Calle-Vallejo, F. et al. Number of outer electrons as descriptor for adsorption processes on transition metals and their oxides. *Chem. Sci.* **4**, 1245–1249 (2013).
53. Xin, H., Holewinski, A. & Linic, S. Predictive structure–reactivity models for rapid screening of Pt-based multimetallic electrocatalysts for the oxygen reduction reaction. *ACS Catal.* **2**, 12–16 (2012).
54. Xie, T. & Grossman, J. C. Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties. *Phys. Rev. Lett.* **120**, 145301 (2018).
55. Karamad, M. et al. Orbital graph convolutional neural network for material property prediction. *Phys. Rev. Mater.* **4**, 093801 (2020).
56. Kim, J.-S., Noh, J. & Im, J. Machine learning-enabled chemical space exploration of all-inorganic perovskites for photovoltaics. *npj Comput. Mater.* **10**, 97 (2024).
57. Ward, L., Agrawal, A., Choudhary, A. & Wolverton, C. A general-purpose machine learning framework for predicting properties of inorganic materials. *npj Comput. Mater.* **2**, 16028 (2016).
58. Kresse, G. & Furthmüller, J. Efficient iterative schemes for ab initio total-energy calculations using a plane-wave basis set. *Phys. Rev. B* **54**, 11169–11186 (1996).
59. Kresse, G. & Joubert, D. From ultrasoft pseudopotentials to the projector augmented-wave method. *Phys. Rev. B* **59**, 1758–1775 (1999).
60. Blöchl, P. E. Projector augmented-wave method. *Phys. Rev. B* **50**, 17953–17979 (1994).
61. Perdew, J. P., Burke, K. & Ernzerhof, M. Generalized gradient approximation made simple. *Phys. Rev. Lett.* **77**, 3865–3868 (1996).
62. Langmuir, I. The constitution and fundamental properties of solids and liquids. Part I. Solids. *J. Am. Chem. Soc.* **38**, 2221–2295 (1916).
63. Grimme, S. Semiempirical GGA-type density functional constructed with a long-range dispersion correction. *J. Comput. Chem.* **27**, 1787–1799 (2006).
64. Nørskov, J. K., Bligaard, T., Rossmeisl, J. & Christensen, C. H. Towards the computational design of solid catalysts. *Nat. Chem.* **1**, 37–46 (2009).
65. Wang, V., Xu, N., Liu, J.-C., Tang, G. & Geng, W.-T. VASPKIT: A user-friendly interface facilitating high-throughput computing and analysis using VASP code. *Comput. Phys. Commun.* **267**, 108033 (2021).
66. Valdés, Á., Qu, Z. W., Kroes, G. J., Rossmeisl, J. & Nørskov, J. K. Oxidation and photo-oxidation of water on TiO₂ surface. *J. Phys. Chem. C* **112**, 9872–9879 (2008).
67. Armstrong, R. A. Should Pearson's correlation coefficient be avoided? *Ophthalmic Physiol. Opt.* **39**, 316–327 (2019).
68. Vega García, M. & Aznarte, J. L. Shapley additive explanations for NO₂ forecasting. *Ecol. Inform.* **56**, 101039 (2020).

Acknowledgements

This work was financially supported by the National Key R&D Program of China (Grant No.2021YFB3500403) and the Youth Fund of the National Natural Science Foundation of China (Grant no. 52305443). We gratefully acknowledge HZWTECH for providing computational facilities.

Author contributions

R.Z.W. proposed the workflow and scientific insight. C.W. designed the code, performed the calculations, analyzed the data and wrote the original paper. B.W. and C.H.W. reviewed and edited the paper. Z.P.C. and A.J.L. collected the validation dataset. All authors contributed to the discussions.

Competing Interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41524-025-01607-4>.

Correspondence and requests for materials should be addressed to Ruzhi Wang.

Reprints and permissions information is available at <http://www.nature.com/reprints>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2025