THE JOURNAL OF PHYSICAL CHEMISTRY C
A JOURNAL OF THE AMERICAN CHEMICAL SOCIETY

pubs.acs.org/JPCC

Article

A Surrogate Machine Learning Model for the Design of Single-Atom Catalyst on Carbon and Porphyrin Supports towards Electrochemistry

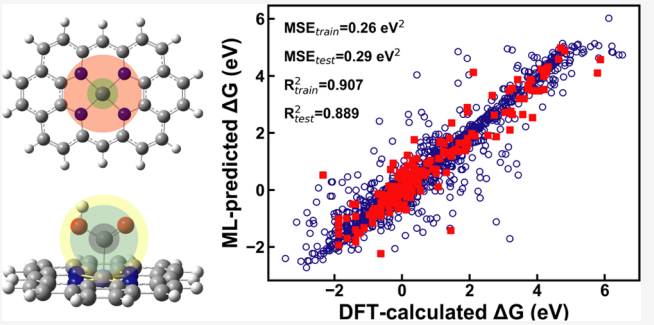
Mohsen Tamtaji, Shuguang Chen, Ziyang Hu, William A. Goddard III,* and GuanHua Chen*

Cite This: *J. Phys. Chem. C* 2023, 127, 9992−10000

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: We apply the machine learning (ML) tool to calculate the Gibbs free energy (ΔG) of reaction intermediates rapidly and accurately as a guide for designing porphyrin- and graphene-supported single-atom catalysts (SACs) toward electrochemical reactions. Based on the 2105 DFT calculation data from the literature, we trained a support vector machine (SVR) algorithm. The hyperparameters were optimized using Bayesian optimization along with 10-fold cross-validation to avoid overfitting. Based on the Shapley Additive exPlanation (SHAP) and permutation methods, the feature importance analysis suggests that the most important parameters are the number of pyridinic nitrogen (Npy), the number of d electrons (θ_d), and the number of valence electrons of reaction intermediates. Inspired by this feature importance analysis and the Pearson correlation coefficient, we found a linear dependent, simple, and general descriptor (φ) to describe ΔG of reaction intermediates (*e.g.*, $\Delta G_{OH^*} = 0.020\varphi - 2.190$). Using the trained SVR algorithm, ΔG_{OH^*}, ΔG_{O^*}, ΔG_{OOH^*}, ΔG_{OO^*}, ΔG_{H^*}, ΔG_{COOH^*}, ΔG_{CO^*}, and ΔG_{N2^*} intermediates are predicted for the oxygen reduction reaction (ORR), the oxygen evolution reaction (OER), the hydrogen evolution reaction (HER), and the CO_2 reduction reaction (CO_2RR). The SVR model predicts an ORR overpotential of 0.51 V and an HER overpotential of 0.22 V for FeN4-SAC. Moreover, we used the SVR algorithm for high-throughput screening of SACs, suggesting new SACs with low ORR overpotentials. This strategy provides a data-driven catalyst design method that significantly reduces the costs of DFT calculations while providing the means for designing SACs for electrocatalysis and beyond.

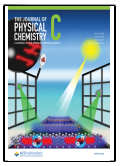


INTRODUCTION

Single-atom catalysts (SACs) are extensively applied for various electrochemical reactions to produce value-added chemicals[1−3] due to their high atom utilization efficiency along with their unique properties.[4−6] According to their widespread applications, the rational design of SACs has received a lot of interest in improving the feasibility and efficiency of optimizing the desired products.[7,8] Density functional theory (DFT) calculations are mostly applied for the rational design of SACs with a focus on high activities and selectivities.[9] However, DFT calculations are computationally expensive and time-consuming[10,11] due to the fact that the complexity of structure−activity relationships requires a huge number of nontrivial DFT calculations in a vast dimensional space, such as environmental coordination, SAC type, and reactants.[12] In addition, designing advanced SACs requires fundamental understanding and deep analysis of the DFT-calculated data through data analysis. To address these issues, machine learning (ML), as a data-intensive tool, provides researchers the ability to accelerate time-consuming DFT calculations to predict the catalytic activity in a large parameter space of SACs.[13,14] For example, the DFT-predicted data along with input features are previously applied to train ML algorithms.[15] Trained ML algorithms can then be used for predicting the optimal SAC with high activity and also for performing feature importance analysis and introducing new descriptors.[16,17] Subsequently, optimized SACs can be used for the desired electrochemical reaction for metal−air batteries and for producing valuable chemicals and fuels. Although ML has recently been used to predict the properties of SACs,[18−21] it is still in an early stage.[22] Here, ML aims to guide and reduce the number of DFT calculations in the search for the ideal catalysts.[23,24] However, major issues for applying ML in SAC design are the lack of a universal ML algorithm, a consistent database, and appropriate descriptors and input features.[25]

Herein, we propose the use of support vector machine (SVR) method as a supervised ML algorithm for predicting

Received: February 3, 2023
Revised: April 4, 2023
Published: May 19, 2023



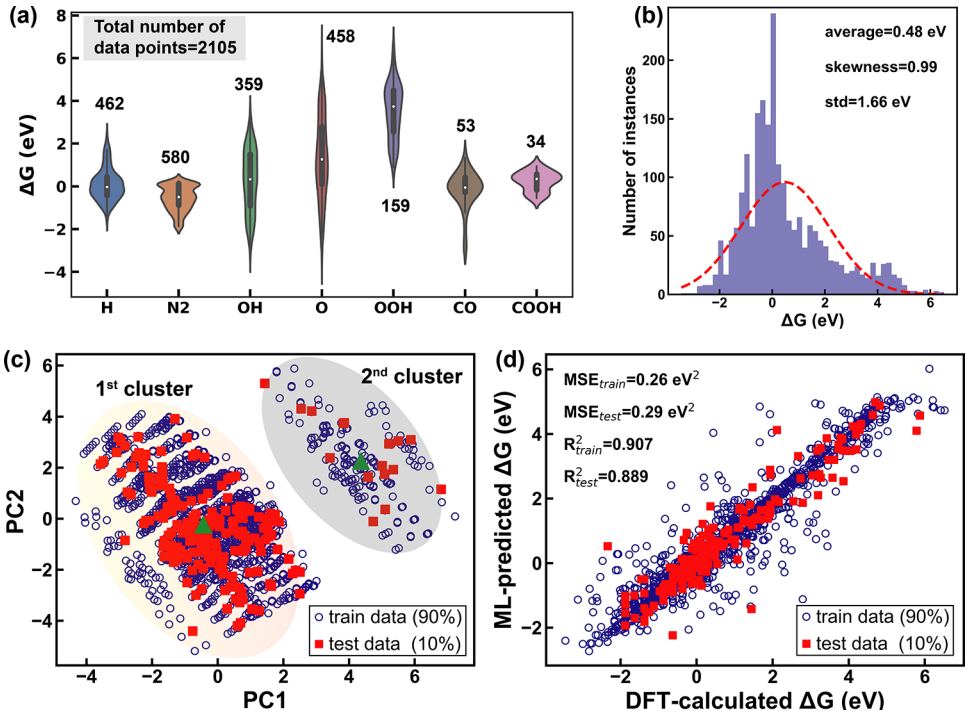


Figure 1. Input data analysis for machine learning (ML). (a) Violin plot of input data distribution for H*, N_2*, OH*, O*, OOH*, CO*, and COOH* reaction intermediates, indicating a total of 2105 input data. (b) The histogram of variation of Gibbs free energies (ΔG) indicates a standard deviation (std) and an average of 1.66 eV and 0.48 eV. (c) Principal component analysis (PCA) and clustering for the training data (90%, blue unfilled circles) and test data (10%, red solid square) projected onto the PC1−PC2 plane. PC1 and PC2 stand for the first and second principal components, respectively. The k-means clustering method suggests two main clusters in the PC1−PC2 plane, with the centers shown in the green solid up triangle. (d) The parity plot of ML-predicted versus DFT-calculated Gibbs free energy of reaction intermediates such as H*, OH*, O*, OOH*, CO*, and COOH* for single-atom catalysts (SACs). The support vector regression (SVR) algorithm shows satisfactory MSE and R^2 values for both training and test data without any signs of underfitting.

Gibbs free energies of different reaction intermediates on graphene- and porphyrin-supported SACs. We use the trained ML algorithm to predict free energy, feature−feature correlation coefficients, and Pearson coefficients, introduce a new descriptor, perform feature importance analysis, and perform high-throughput screening. Moreover, for the first time, the properties of reaction intermediates are successfully applied as the input features. Our findings on the input features and their impact on the model output are evaluated using Shapley Additive exPlanation (SHAP) and permutation methods, indicating that the most important and informative parameters are the number of pyridinic nitrogen (Npy), the number of d electrons of the metal center (θ_d), and the number of valence electrons of the reaction intermediate. Inspired by this result, we introduced a new descriptor to describe the adsorption energy of various intermediates. Subsequently, we applied the SVR model to several examples including the oxygen reduction reaction (ORR), the oxygen evolution reaction (OER), the CO_2 reduction reaction (CO_2RR), and the hydrogen evolution reaction (HER) to show the potential application of our proposed ML model for the rapid design and discovery of SACs toward electrochemical reactions. Ultimately, the SVR algorithm is used for the high-throughput screening of SACs for ORR overpotential, suggesting some new catalysts with high performance.

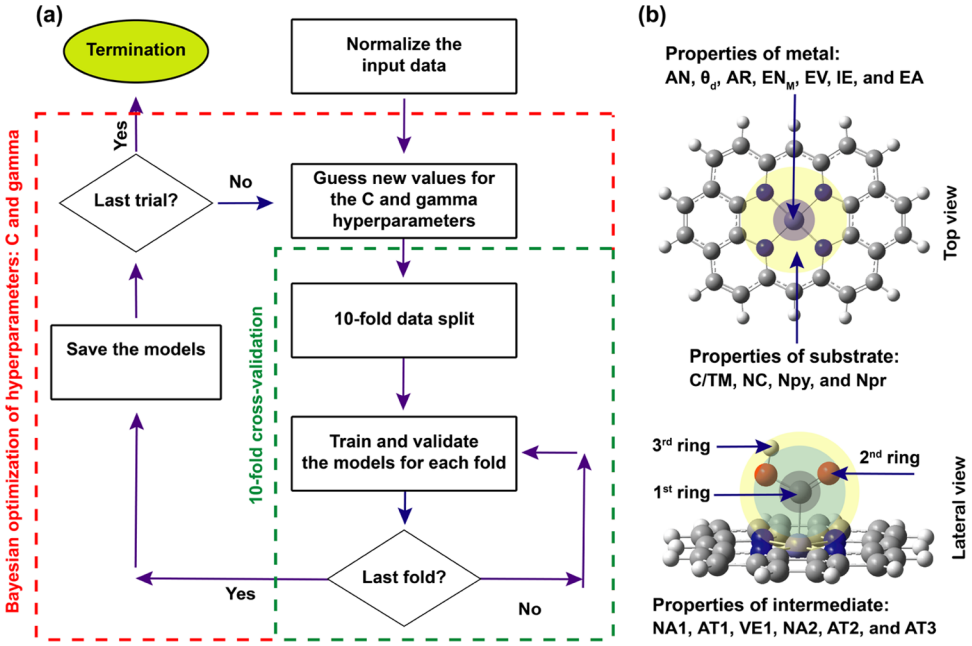
METHODS

Input Data Collection. The data used for the training of our ML model is composed of DFT-calculated data collected from the literature. To consider the data consistency for ML applications, all of the DFT data are calculated using the Vienna ab initio simulation package (VASP) with the Perdew−Burke−Ernzerhof (PBE) functional. In addition, the duplicated and outlier data, which are significantly different from other data points in the data set, are deleted. The data contains 2105 data points for the Gibbs free energy (ΔG) of reaction intermediates such as OH*, O*, OOH*, O_2*, H*, CO_2*, COOH*, CO*, and N_2* (please see the excel file in the Supporting Information for more details). The data is based on the 3d, 4d, and 5d transition metals on graphene and porphyrin supports. The data is sorted in an Excel file in such a way that the input features, the SAC structure, and the Gibbs free energies are provided for the training of our ML algorithm (see the Supporting Information). Figure 1a shows the violin plot of Gibbs free energy distribution for H*, N_2*, OH*, O*, OOH*, CO*, and COOH* reaction intermediates. The violin plot also displays the number of data points for each reaction intermediates, leading to a total of 2105 input data. This indicates a fair balanced data distribution among the reaction intermediates, suggesting the quality and consistency of the input data. The Gibbs free energy is distributed from −4 to +7 eV. More specifically, the Gibbs free energy of the OOH* intermediate is between 0 and +7 eV with an average of around +4 eV and a median of around +5.5 eV. Figure 1b displays the histogram for the variation of Gibbs free energies, indicating the average, standard deviation (std), and skewness of 0.48 eV, 1.66 eV, and 0.99, respectively.

Input Feature Selection. In order to have high-quality data analysis and an interpretable ML algorithm, we need

https://doi.org/10.1021/acs.jpcc.3c00765
J. Phys. Chem. C 2023, 127, 9992−10000

Scheme 1. Construction of ML for the Design of Single-Atom Catalyst[a]



[a](a) Flowchart for the automated hyperparameter tuning of the support vector regression (SVR) model using Bayesian optimization along with a 10-fold cross-validation. (b) Top view and lateral view of the structure of single-atom catalysts (SACs) along with the input features including the properties of transition metal, substrate, and intermediates.

reasonable input feature selection.[26] Recent works have utilized several input features including the enthalpy of vaporization, d-band center, Bader charge, electron affinity, ionization energy, covalent radius, the number of electrons in d orbitals, formation energy, and oxide formation enthalpy to describe the reactivity of SACs.[19,20,26−32] Unfortunately, one of the major limitations of applying ML in the design of SACs is the lack of suitable input features. A suitable input feature requires simultaneously high simplicity, reasonable feature importance value, and physical interpretation.[26] For example, the black-box nature of ML algorithms sometimes makes a physical interpretation of input features, including the d-band center and enthalpy of vaporization, impossible.[33] For example, the d-band center is normally used as an input feature with a high feature importance value to describe the activity of SACs. Nevertheless, the d levels of atomically dispersed SACs on graphene and porphyrin substrates might not form a band, making the evaluation of the position of the d-band center not possible. In addition, the simplicity of input features requires using the properties of metal atoms and substrates, which are easily available without requiring expensive DFT calculations. In contrast to the density of states and Bader charge, input features including the ionization energy, number of electrons in the d orbital, atomic number, and coordination number of metal atoms satisfy the simplicity requirement. Scheme 1b shows the top view and lateral view of the structure of a typical SAC along with the list of input features including the intrinsic properties of the metal atom (M) along with the properties of substrate and reaction intermediates. In fact, to the best of our knowledge, for the first time in this work, the use of properties of reaction intermediates is successfully shown as the input feature, which enables us to generalize the ML model for several electrochemical reactions. Therefore, the input features composed of atomic number (AN), number of d electrons (θ_d) atomic radius (AR), electronegativity (EN_M), enthalpy of vaporization (EV), first ionization energy (IE), electron affinity (EA), number of C atoms of substrate per one transition metal (C/TM), number of C atoms bonded to transition metal (NC), number of pyrrolic nitrogen bonded to transition metal (Npy), number of pyridinic nitrogen bonded to transition metal (Npr), number of intermediate atoms in the 1st ring (NA1), sum of atomic number of intermediate atoms in the 1st ring (AT1), valence electron of intermediate atoms in the 1st ring (VE1), number of intermediate atoms in the 2nd ring (NA2), sum of atomic number of intermediate atoms in the 2nd ring (AT2), and sum of atomic number of intermediate atoms in the 3rd ring (AT3).

Machine Learning (ML) Implementation. We use SVR as a supervised machine learning (ML) algorithm[34] to calculate the Gibbs free energy of different intermediates using data from the DFT calculations. We use Scikit-learn, NumPy, Matplotlib, SHAP, Pickle, and SciPy libraries in Python 3.6 to read and process the data, train and save the ML algorithm, and perform feature importance analysis. The data is normalized to ensure that the range of values is consistent across all columns in the data set to keep the consistency of input data. This step is important if the ML algorithm being used is sensitive to the range of values. We applied the normalized input data along with the input features for the construction of the ML model and then we used the mean-squared error (MSE) and R^2 value to evaluate the performance of the SVR model as follows[34]

$$\mathrm{MSE} = \frac{1}{n}\sum_{n}^{i=1}(\Delta G_{\mathrm{DFT},i} - \Delta G_{\mathrm{ML},i})^2 \quad (1)$$

where $\Delta G_{\mathrm{DFT},i}$ and $\Delta G_{\mathrm{ML},i}$ are the DFT- and ML-predicted Gibbs free energies, respectively, for intermediate i, and n is the number of instances in the training and test data sets.

https://doi.org/10.1021/acs.jpcc.3c00765
J. Phys. Chem. C 2023, 127, 9992−10000

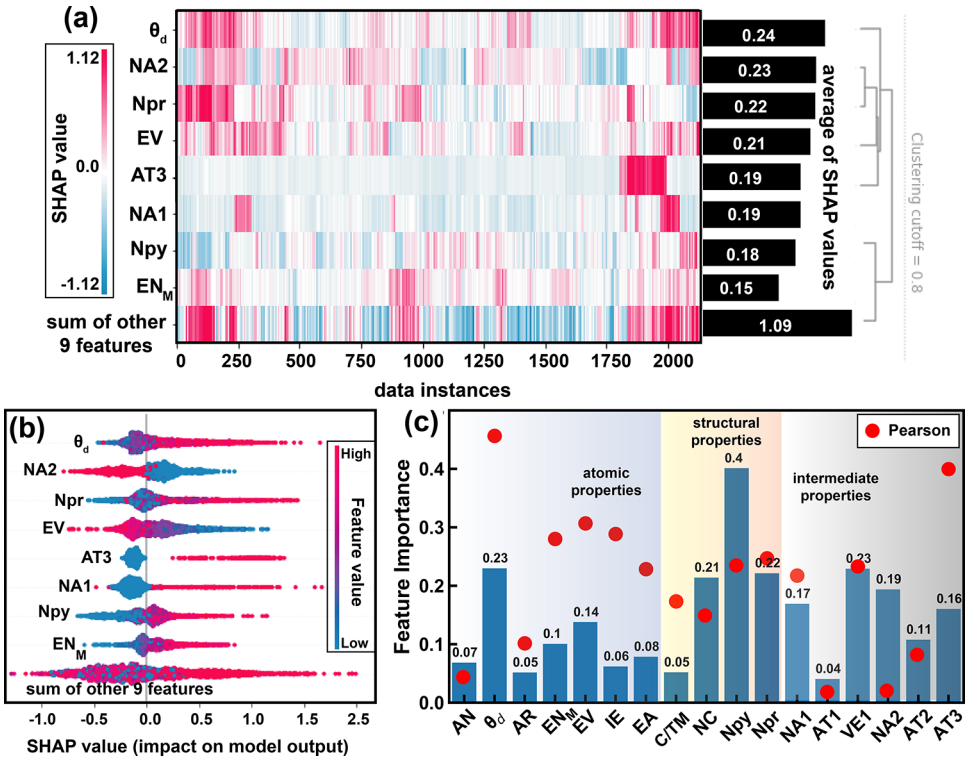


Figure 2. Feature importance analysis. (a) The heatmap of SHAP values of input feature for the whole 2105 data instances in the order of feature importance. The black horizontal bar plot displays the average of SHAP values for each input feature across the data instances, indicating the number of d electrons (θ_d) as the most important parameter. (b) The influence of each input feature on the model output for the whole 2105 data instances using SHAP value in the order of feature importance, colored based on the features' values. (c) Feature importance analysis on the Gibbs free energy (ΔG) of reaction intermediates based on the permutation method and the corresponding Pearson correlation coefficients (solid red circle). This indicates the number of pyridinic nitrogens (Npy) and the number of d electrons (θ_d) as the most important parameters.

Feature Importance Analysis. Based on the trained SVR model, several methods such as SHAP,[35,36] permutation,[37] and Pearson correlation coefficient[38] have been applied to screen the impact of each input feature on the model output. The SHAP technique is a good feature attribution algorithm with the ability to screen the impact of each data point on the model output.[39] This method assigns equal weights to feature coalitions of all sizes.[35] Permutation method is another way to perform feature importance analysis. As a model inspection technique, permutation can be applied to any fitted estimator, especially the nonlinear ones, once the data is tabular.[37] The permutation feature importance breaks the relationship between a single feature and the model output by randomly shuffling the feature value and defining how much the feature affects the model output. This method has the advantage of being model agnostic, which can be evaluated several times with various permutations of the features. Moreover, we use the Pearson correlation coefficient to quantify the linear dependencies between pairs of variables such as input features and model output. The Pearson correlation defines the direction and strength of the linear relationship between two features by giving a number between −1 and 1.[38]

RESULTS AND DISCUSSION

We apply SVR model to establish the structure−activity relationship for the prediction of Gibbs free energy. Scheme 1a shows the flowchart for the construction of our SVR model for SACs. Scheme 1b displays the top view and lateral view of the structure of a typical SAC. According to Scheme 1a, first, the input data was randomly partitioned into the train set (90%, 1894 data points) and test set (10%, 211 data points). The input features, as shown in Scheme 1b, were then normalized by the MinMaxScaler function of Sklearn, and the initial guess for the hyperparameters of the SVR algorithm (C and gamma) was given to the model. After that, the optimized value for the hyperparameters was obtained using Bayesian optimization by minimizing the MSE of the test set as the activation function. To avoid overfitting, a 10-fold cross-validation loop was applied inside the Bayesian optimization loop, in which the training data was randomly split into 10 groups. Subsequently, the ML model was trained and validated for each group and the optimized ML model was saved for further predictions and feature importance analysis.

Figure 1a displays the violin plot of input data distribution, and Figure 1b displays the histogram for variations of Gibbs free energies. A standard deviation of 1.66 eV indicates that the data is spread out and not clustered around the mean (0.48 eV). The positive value of skewness (0.99) in Figure 1b suggests the deviation of Gibbs free energy values from the average toward the positive values and the curve is longer toward the right trail.[38] As shown in Figure S1, we found the natural log expression of ln(ΔG+4.36) leading to a minimized skewness of 0.00 with a rather symmetrical distribution. It is worth mentioning that we used the bare value of ΔG for the training of our ML model due to the higher accuracy. To shed more light on the input data visualization and find out the distribution of test data among the training data, we perform principal component analysis (PCA). PCA, which is a statistical method, provides the reduction in the dimensionality of the system by linearly transforming the data set into a new

https://doi.org/10.1021/acs.jpcc.3c00765
J. Phys. Chem. C 2023, 127, 9992−10000

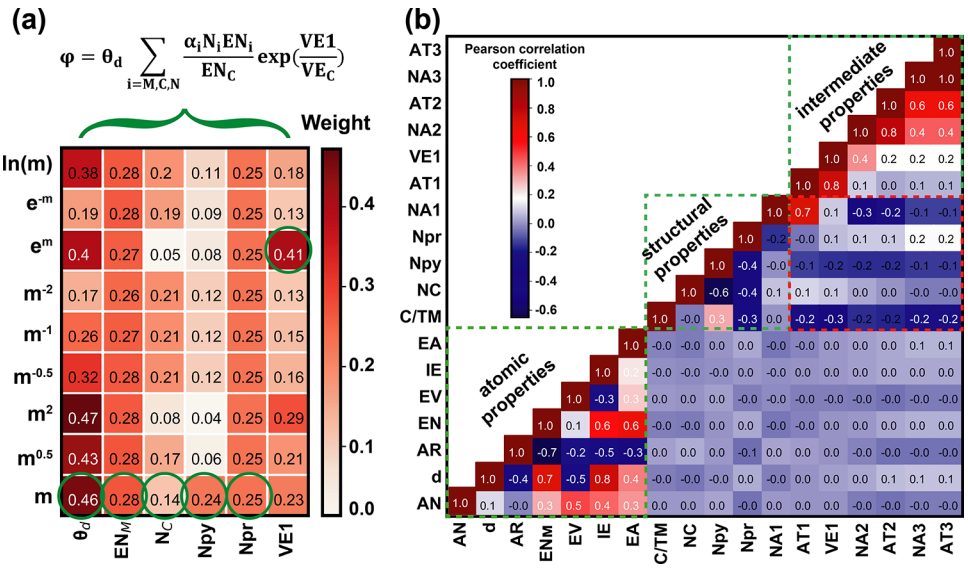


Figure 3. Correlation heatmap. (a) The heatmap for Pearson coefficient weights in the polynomial, exponential, and logarithmic forms for the most important parameters. (b) Feature−feature correlation map of the input features. This indicates that there is a linear relationship inside each group, namely, atomic, structural, and intermediate properties. However, there is no linear relationship between atomic properties with structural and intermediate properties.

coordinate system while preserving most of the variation in the data and giving more insight into the data analysis and visualization. Here, we applied PCA to reduce the dimensionality of the data set from 17 dimensions down to 2 dimensions. The first and second principal components, PC1 and PC2, respectively, were chosen for the projection, retaining 38% of the variation in the data. Figure 1c shows the PCA analysis for the training and test data projected onto the PC1−PC2 plane, suggesting two main regions for the input data with the centers shown in green triangles. The training and test data are shown to be randomly distributed inside both regions.

After data analysis and visualization, all 2105 data points were used to make the final SVR model. Figure 1d shows the parity plot of ML- versus DFT-predicted Gibbs free energy of reaction intermediates for SACs. The SVR algorithm shows satisfactory MSE values of 0.26 and 0.29 eV^2 and R^2 values of 0.907 and 0.889, respectively, for the training and test data, with no signs of underfitting and overfitting. According to Figure 1c, as the test data is randomly distributed along the whole data, the MSE_{test} measures the interpolative prediction ability of the trained SVR model.

In addition to establishing a deep structure−activity relationship for the prediction of Gibbs free energies, more information can be extracted from the trained SVR model such as feature importance analysis, introduction of new descriptors,[34,40] and high-throughput screening,[41] which can be of great help for the rational design of SACs.[42]

Although the feature importance analysis has previously been reported for a specific reaction with a small data set,[26] there is a strong need for such analysis to be conducted across a wide range of electrochemical reactions with a large data set. To perform feature importance analysis, one can utilize the SHAP method, which calculates the contribution of each feature to the predicted outcome for a given data point in a data set. Figure 2a shows the heatmap of SHAP values of each input feature for the whole 2105 data instances in the order of feature importance. Besides, the heatmap indicates the normal distribution of SHAP values along the data sets for each input feature, suggesting appropriate distributions of data for the training and test data. The black horizontal bar plot displays the average of SHAP values for each input feature across the data instances, indicating the number of d electrons (θ_d) as the most important parameter, in agreement with the previous reports.[13,18,43−45] Indeed, by increasing the d electrons of the active metal site, the charge transfer and adsorption strength of reaction intermediates can be directly affected. Figure 2b displays the influence of each input feature on the model output for the whole 2105 data points using the SHAP value in the order of feature importance, colored based on the features' values. This indicates that the higher values of θ_d, Npr, AT3, Npy, and EN_M possess higher SHAP values with a higher impact on the model output. In contrast, the lower values of NA2 and EV possess higher SHAP values with a higher impact on the model output. This can be interpreted based on the fact that higher θ_d, Npr, Npy, and EN_M provide more electrons to the metal center, which can have more impact on the adsorption of reaction intermediates.

In addition to the SHAP method, there exist alternative methods for assessing feature importance. Figure 2c illustrates the permutation-based feature importance analysis, indicating that all of the input features possess a reasonable value for feature importance. More specifically, the number of pyridinic nitrogen (Npy), the number of d electrons (θ_d), and the number of valence electrons of reaction intermediates (VE1) are the most important, informative, and interpretable parameters, in agreement with the SHAP feature importance analysis. The obtained feature importance can be also physically interpreted. For example, as the number of nitrogen increases, the number of charges and electronic properties of active metal sites highly change, which can hugely affect the adsorption energy of reaction intermediates. Moreover, by changing the valence electrons of reaction intermediates, the hybridization of p orbitals in the reaction intermediate with the d orbital in the active metal site is strongly affected.

Moreover, in order to find whether the relationship between the Gibbs free energies and the input features is linear or not,

https://doi.org/10.1021/acs.jpcc.3c00765
J. Phys. Chem. C 2023, 127, 9992−10000

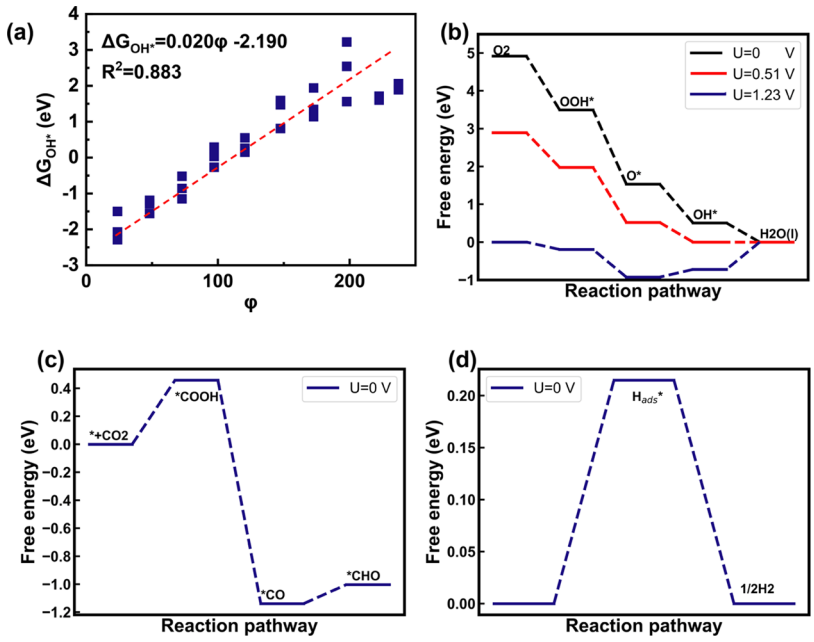


Figure 4. Prediction performance. (a) Gibbs free energy of OH* intermediate (ΔG_{OH^*}) versus the new descriptor (φ), suggesting that the new descriptor satisfactorily defines the linear relationship with the ΔG of reaction intermediates for SACs. Free-energy diagram predicted for (b) ORR, (c) CO_2RR, and (d) HER for FeN4-SAC.

the Pearson correlation coefficients were calculated. Based on this, θ_d has the most coefficient weight with higher linear dependency on the model output (Figure 2c). Therefore, θ_d is the only feature with both high feature importance and high linear dependency. However, the results indicate that the relationship between the Gibbs free energies and the other input features is not necessarily linear, in spite of their high feature importance.[43] It indicates that most portion of their feature importance is not linear. Therefore, we established a hypothesis space by using linear, exponential, natural logarithm, and polynomial forms of the primary descriptors (θ_d, EN_M, NC, Npy, Npr, and VE1) to find the linear relationships with the secondary descriptors. The mathematical form of the model is composed of 54 secondary descriptors ($x_{i=1,\cdots,54}$) as follows (see Table S1 of the Supporting Information for more details)[34]

$$\Delta G = \sum_{i=1}^{m} \beta_i x_i \tag{2}$$

The magnitude of the Pearson correlation coefficient weights (β) is displayed in the heatmap demonstrated in Figure 3a. As such, the linear term of θ_d, EN_M, NC, Npy, and Npr and the exponential term of VE1 are the most weighted and informative secondary descriptors. Taking into account the above results, a general descriptor (φ) is proposed as follows

$$\varphi = \theta_d \sum_{i=M,C,N} \frac{\alpha_i N_i EN_i}{EN_C} \exp\left(\frac{VE1}{VE_C}\right) \tag{3}$$

where α is a constant that depends on the environments of the metal center (α = 1 for pyridinic and α = 1.25 for pyrrolic nitrogen). N_i is the number of atoms bonded to the metal center, EN_i is the electronegativity of atom i, i stands for the C, N, and metal (M) atoms, EN_C is the electronegativity of carbon (=2.55), VE_C is the valence electron of carbon (=4), VE1 is the valence electron of intermediate atoms in the 1st ring, and θ_d is the number of d electrons of the metal center. The descriptor considers simultaneously the intrinsic properties of the metal atom along with the structural and intermediate properties, which can be simplified as

$$\varphi = k\theta_d e^{0.25\ VE1} \tag{4}$$

where k is the electronegativity coefficient (=$0.5EN_M$+1.5Npy +1.2Npr+N_C). The simplified descriptor bears physical meaning. For instance, with an increase in the number of nitrogens, electrons in the d orbital, and valence electrons in the intermediate's first ring, the likelihood of electron sharing between the metal and intermediate enhances. This gives rise to an increase in the descriptor value that could lead to a direct impact on the adsorption energy of the reaction intermediate. Figure 3b shows the feature−feature correlation map of the input features. This indicates that atomic properties are likely related to each other. Similarly, structural and intermediate properties are likely related, while there is no linear relationship between atomic properties with structural and intermediate properties.

Figures 4a and S4 display the ORR overpotential (η^{ORR} eV) along with Gibbs free energy of OH* OOH*, O*, and H* intermediates (ΔG_{OH^*}, ΔG_{OOH^*}, ΔG_{O^*}, and ΔG_{H^*}) versus the new descriptor for the MN4-SAC structure, in which M stands for all of the 3d, 4d, and 5d transition metals (see Figure S3 of the Supporting Information for more details on the MN4-SAC structure). It suggests that the new descriptor defines satisfactory linear relationships, for instance, $\Delta G_{OH^*} = 0.020\varphi - 2.190$ with an R^2 value of 0.883. This linear relationship indicates that an increase in φ corresponds to a change in adsorption energy and can be explained by the electronic structure of the active metal center bonded with the OH* intermediate.[7] Moreover, a large value for φ suggests that the metal center possesses more valence electrons in its d orbital after charge redistribution.[7] In other words, SACs with higher work function possess more valence electrons in their d

https://doi.org/10.1021/acs.jpcc.3c00765
J. Phys. Chem. C 2023, 127, 9992−10000

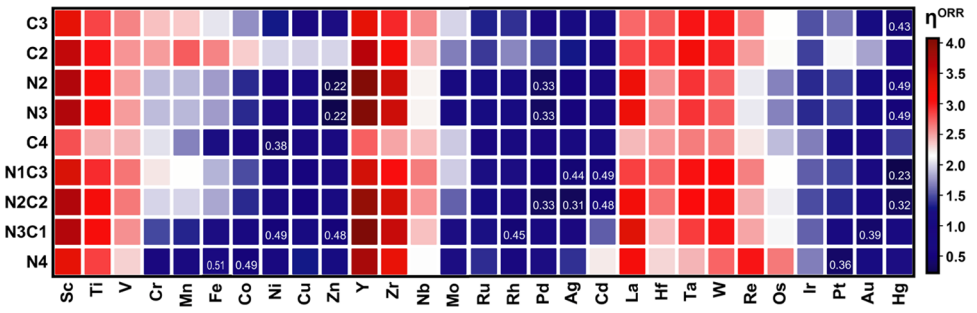


Figure 5. High-throughput screening. High-throughput screening for the overpotential of ORR, indicating low overpotential of, for example, 0.33 V for PdN2C2-SAC (see Figure S3 of the Supporting Information for more details on N4, N3C1, N2C2, N1C3, C4, N3, N2, C2, and C3 structures).

orbital, leading to a higher value for φ and higher adsorption energy for OH*.[7]

The prediction performance of the trained SVR model is evaluated for FeN4-SAC, a well-known SAC, through Figure 4b–d which shows the free-energy diagram predicted for ORR, CO_2RR, and HER. Figure 4b displays an overpotential of 0.51 V for ORR, which agrees with the theoretically obtained ORR overpotential of 0.59 V from the first-principles DFT calculations.[46] As for CO_2RR, Figure 4c shows rather a high adsorption energy of FeN4-SAC for the CO intermediate (−1.15 eV), which makes it more efficient at producing CH_4 and multicarbon products instead of just CO. This is because the CO intermediate can remain stable on the catalyst, allowing for proton transfers to occur and create more complex products.[47] This is a promising aspect of the catalyst for CO_2RR, as multicarbon products have the potential to be more useful and valuable than simpler products like CO or CH_4. In addition, based on our results, introducing one N vacancy in FeN3-SAC (see Figure S3 of the Supporting Information for more details on the MN3-SAC structure) lowers the Gibbs free energy of COOH* and CO* intermediates, which agrees with recent findings.[48] It is worth mentioning that the data for vacancy defects is not in the training data set, suggesting that the proposed ML algorithm is applicable for extrapolative predictions as well.

As for HER, an overpotential of 0.22 V is predicted for FeN4-SAC (Figure 4d), which agrees with the HER overpotential of 0.25 V obtained from DFT calculations.[10] This suggests a good performance for FeN4-SAC in HER, which is consistent with its good experimental HER performance.[10,46]

The proposed SVR model also indicates the potential to perform high-throughput screening for new SACs.[45,49−52] Figure 5 shows the high-throughput screening for the overpotential of ORR (η^{ORR}), indicating some catalysts with low overpotentials. More specifically, FeN4-SAC (0.51 V), CoN4-SAC (0.49 V), NiN3N1-SAC (0.49 V), NiC4-SAC (0.38 V), ZnN3C1-SAC (0.48 V), ZnN3-SAC (0.22 V), ZnN2-SAC (0.22 V), RhN3C1-SAC (0.45 V), PdN2C2-SAC (0.33 V), PdN3-SAC (0.33 V), PdN2-SAC (0.33 V), AgN2C2-SAC (0.31 V), AgN1C3-SAC (0.44 V), CdN2C2-SAC (0.48 V), CdN1C3-SAC (0.49 V), PtN4-SAC (0.36 V), AuN3C1-SAC (0.39 V), HgN2C2-SAC (0.32 V), HgN1C3-SAC (0.23 V), HgN3-SAC (0.49 V), HgN2-SAC (0.49 V), and HgC3-SAC (0.43 V) lead to low overpotentials, which can be applicable for metal−air battery and fuel cell applications. Besides, the lower overpotential obtained for CoN4-SAC in comparison with that of FeN4-SAC indicates its better ORR performance, which is interestingly in agreement with experimental results reported in the literature.[46] As mentioned, FeN4-SAC led to a low overpotential of 0.51 V, while the FeC2-SAC structure led to a high overpotential of 2.7 V, suggesting that the coordination environment has a significant effect on the activity of these SACs for ORR.[12,53] Moreover, the study found that certain metals (Sc, Ti, V, Y, Zr, Nb, La, Hf, Ta, W, Re, and Os) led to a high ORR overpotential in the examined catalyst structures with a poor ORR activity due to their strong binding of the OH* intermediate. Recently, it has been shown that YN4-SAC can be activated toward ORR by an axial chlorine ligand due to the weakening of the binding energy of the OH* intermediate.[54]

CONCLUSIONS

In our study, we utilized the support vector machine (SVR) algorithm to design single-atom catalysts (SACs) supported by both porphyrin and graphene by calculating the Gibbs free energy (ΔG) of reaction intermediates toward electrochemical reactions. Bayesian optimization, coupled with 10-fold cross-validation, was used to optimize the hyperparameters of our SVR model trained with 2105 DFT-calculated data points from the relevant literature. Through the use of SHAP and permutation methods, we determined the number of pyridinic nitrogens (Npy), the number of d electrons (θ_d), and the number of valence electrons of the reaction intermediate as the most significant factors via feature importance analysis. Inspired by the feature importance analysis and the Pearson correlation coefficient, we represented ΔG of reaction intermediates using a general and linear dependent descriptor (φ). Our trained SVR algorithm enabled us to calculate ΔG of OH*, O*, OOH*, H*, COOH*, CO*, and N_2* intermediates for the purpose of ORR, OER, HER, and CO_2RR. More specifically, we applied the ML algorithm for high-throughput screening of SACs for ORR overpotential, proposing several new candidates such as ZnN3-SAC, ZnN2-SAC, and HgN1C3-SAC with the ORR overpotential of below 0.30 V. This strategy proposes a data-driven method for catalyst design that can remarkably lower the DFT calculation time while providing the means for designing SACs for electrochemical reactions.

ML Code Availability. The ML algorithm is available free of charge at https://github.com/MohsenTamtaji/PySACs.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jpcc.3c00765.

Input data histogram, the structure of single-atom catalysts, secondary descriptors (PDF)

https://doi.org/10.1021/acs.jpcc.3c00765
J. Phys. Chem. C 2023, 127, 9992−10000

Collected DFT data from the literature (XLSX)

AUTHOR INFORMATION

Corresponding Authors

William A. Goddard III − *Materials and Process Simulation Center (MSC), MC 139-74, California Institute of Technology, Pasadena, California 91125, United States;* orcid.org/0000-0003-0097-5716; Email: wag@caltech.edu

GuanHua Chen − *Department of Chemistry, The University of Hong Kong, Hong Kong SAR 999077, China; Hong Kong Quantum AI Lab Limited, Hong Kong SAR 999077, China;* Email: ghc@everest.hku.hk

Authors

Mohsen Tamtaji − *Department of Chemistry, The University of Hong Kong, Hong Kong SAR 999077, China; Hong Kong Quantum AI Lab Limited, Hong Kong SAR 999077, China;* orcid.org/0000-0001-9118-5474

Shuguang Chen − *Department of Chemistry, The University of Hong Kong, Hong Kong SAR 999077, China; Hong Kong Quantum AI Lab Limited, Hong Kong SAR 999077, China*

Ziyang Hu − *Department of Chemistry, The University of Hong Kong, Hong Kong SAR 999077, China; Hong Kong Quantum AI Lab Limited, Hong Kong SAR 999077, China;* orcid.org/0000-0002-7693-5457

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jpcc.3c00765

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

G.H.C. acknowledges financial support from the General Research Fund (Grant No. 17309620) and Hong Kong Quantum AI Lab, AIR@InnoHK of the Hong Kong Government. W.A.G. thanks the U.S. National Science Foundation (CBET-2005250) for support.

REFERENCES

(1) Huang, Y.; Rehman, F.; Tamtaji, M.; Li, X.; Huang, Y.; Zhang, T.; Luo, Z. Mechanistic Understanding and Design of Non-Noble Metal-Based Single-Atom Catalysts Supported on Two-Dimensional Materials for CO2 Electroreduction. *J. Mater. Chem. A* **2022**, *10*, 5813−5834.

(2) Abbas, S. A.; Song, J. T.; Tan, Y. C.; Nam, K. M.; Oh, J.; Jung, K. D. Synthesis of a Nickel Single-Atom Catalyst Based on Ni-N4-XCxActive Sites for Highly Efficient CO2Reduction Utilizing a Gas Diffusion Electrode. *ACS Appl. Energy Mater.* **2020**, *3*, 8739−8745.

(3) Yang, S.; Liu, X.; Niu, F.; Wang, L.; Su, K.; Liu, W.; Dong, H.; Yue, H.; Yin, Y. 2D Single-Atom Fe-N-C Catalyst Derived from a Layered Complex as an Oxygen Reduction Catalyst for PEMFCs. *ACS Appl. Energy Mater.* **2022**, *5*, 8791−8799.

(4) Chen, C.; Zhang, Z.; Li, G.; Li, L.; Lin, Z. Recent Advances on Nanomaterials for Electrocatalytic CO2Conversion. *Energy Fuels* **2021**, *35*, 7485−7510.

(5) Sun, K.; Xu, W.; Lin, X.; Tian, S.; Lin, W. F.; Zhou, D.; Sun, X. Electrochemical Oxygen Reduction to Hydrogen Peroxide via a Two-Electron Transfer Pathway on Carbon-Based Single-Atom Catalysts. *Adv. Mater. Interfaces* **2021**, *8*, No. 2001360.

(6) Wang, A.; Li, J.; Zhang, T. Heterogeneous Single-Atom Catalysis. *Nat. Rev. Chem.* **2018**, *2*, 65−81.

(7) Xu, H.; Cheng, D.; Cao, D.; Zeng, X. C. A Universal Principle for a Rational Design of Single-Atom Electrocatalysts. *Nat. Catal.* **2018**, *1*, 339−348.

(8) Kraushofer, F.; Parkinson, G. S. Single-Atom Catalysis: Insights from Model Systems. *Chem. Rev.* **2022**, *122*, 14911−14939.

(9) Tao, X.; Nan, B.; Li, Y.; Du, M.; Guo, L. L.; Tian, C.; Jiang, L.; Shen, L.; Sun, N.; Li, L. N. Highly Active Isolated Single-Atom Pd Catalyst Supported on Layered MgO for Semihydrogenation of Acetylene. *ACS Appl. Energy Mater.* **2022**, *5*, 10385−10390.

(10) Hossain, M. D.; Liu, Z.; Zhuang, M.; Yan, X.; Xu, G. L.; Gadre, C. A.; Tyagi, A.; Abidi, I. H.; Sun, C. J.; Wong, H.; et al. Rational Design of Graphene-Supported Single Atom Catalysts for Hydrogen Evolution Reaction. *Adv. Energy Mater.* **2019**, *9*, No. 1803689.

(11) Li, L.; Chang, X.; Lin, X.; Zhao, Z. J.; Gong, J. Theoretical Insights into Single-Atom Catalysts. *Chem. Soc. Rev.* **2020**, *49*, 8156−8178.

(12) Liu, D.; He, Q.; Ding, S.; Song, L. Structural Regulation and Support Coupling Effect of Single-atom Catalysts for Heterogeneous Catalysis. *Adv. Energy Mater.* **2020**, *10*, No. 2001482.

(13) Wu, L.; Guo, T.; Li, T. Machine Learning-Accelerated Prediction of Overpotential of Oxygen Evolution Reaction of Single-Atom Catalysts. *iScience* **2021**, *24*, No. 102398.

(14) Lian, Z.; Yang, M.; Jan, F.; Li, B. Machine Learning Derived Blueprint for Rational Design of the Effective Single-Atom Cathode Catalyst of the Lithium-Sulfur Battery. *J. Phys. Chem. Lett.* **2021**, *12*, 7053−7059.

(15) Steinmann, S. N.; Wang, Q.; Seh, Z. W. How Machine Learning Can Accelerate Electrocatalysis Discovery and Optimization. *Mater. Horiz.* **2023**, *10*, 393−406.

(16) Suvarna, M.; Preikschas, P.; Pérez-Ramírez, J. Identifying Descriptors for Promoted Rhodium-Based Catalysts for Higher Alcohol Synthesis. *ACS Catal.* **2022**, *12*, 15373−15385.

(17) Andersen, M.; Reuter, K. Adsorption Enthalpies for Catalysis Modeling through Machine-Learned Descriptors. *Acc. Chem. Res.* **2021**, *54*, 2741−2749.

(18) 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* **2020**, *8*, 5663−5670.

(19) Niu, H.; Wan, X.; Wang, X.; Shao, C.; Robertson, J.; Zhang, Z.; Guo, Y. Single-Atom Rhodium on Defective g-C3N4: A Promising Bifunctional Oxygen Electrocatalyst. *ACS Sustainable Chem. Eng.* **2021**, *9*, 3590−3599.

(20) Guo, X.; Lin, S.; Gu, J.; Zhang, S.; Chen, Z.; Huang, S. Simultaneously Achieving High Activity and Selectivity toward Two-Electron O2 Electroreduction: The Power of Single-Atom Catalysts. *ACS Catal.* **2019**, *9*, 11042−11054.

(21) Deng, C.; Su, Y.; Li, F.; Shen, W.; Chen, Z.; Tang, Q. Understanding Activity Origin for the Oxygen Reduction Reaction on Bi-Atom Catalysts by DFT Studies and Machine-Learning. *J. Mater. Chem. A* **2020**, *8*, 24563−24571.

(22) Toyao, T.; Maeno, Z.; Takakusagi, S.; Kamachi, T.; Takigawa, I.; Shimizu, K. I. Machine Learning for Catalysis Informatics: Recent Applications and Prospects. *ACS Catal.* **2020**, *10*, 2260−2297.

(23) Chen, C.; Zuo, Y.; Ye, W.; Li, X.; Deng, Z.; Ong, S. P. A Critical Review of Machine Learning of Energy Materials. *Adv. Energy Mater.* **2020**, *10*, No. 1903242.

(24) Di Liberto, G.; Di; Tosoni, S.; Cipriano, L. A.; Pacchioni, G. A Few Questions about Single-Atom Catalysts: When Modeling Helps. *Acc. Mater. Res.* **2022**, *3*, 986−995.

(25) Chanussot, L.; Das, A.; Goyal, S.; Lavril, T.; Shuaibi, M.; Riviere, M.; Tran, K.; Heras-Domingo, J.; Ho, C.; Hu, W.; et al. Open Catalyst 2020 (OC20) Dataset and Community Challenges. *ACS Catal.* **2021**, *11*, 6059−6072.

(26) Tamtaji, M.; Gao, H.; Hossain, M. D.; Galligan, P. R.; Wong, H.; Liu, Z.; Liu, H.; Cai, Y.; Goddard, W. A.; Luo, Z. Machine Learning for Design Principles for Single Atom Catalysts towards Electrochemical Reactions. *J. Mater. Chem. A* **2022**, *10*, 15309−15331.

(27) Lu, Z.; Yadav, S.; Singh, C. V. Predicting Aggregation Energy for Single Atom Bimetallic Catalysts on Clean and O* Adsorbed Surfaces through Machine Learning Models. *Catal. Sci. Technol.* **2020**, *10*, 86−98.
(28) Sun, M.; Wu, T.; Xue, Y.; Dougherty, A. W.; Huang, B.; Li, Y.; Yan, C. H. Mapping of Atomic Catalyst on Graphdiyne. *Nano Energy* **2019**, *62*, 754−763.
(29) Rao, K. K.; Do, Q. K.; Pham, K.; Maiti, D.; Grabow, L. C. Extendable Machine Learning Model for the Stability of Single Atom Alloys. *Top. Catal.* **2020**, *63*, 728−741.
(30) Ha, M.; Kim, D. Y.; Umer, M.; Gladkikh, V.; Myung, C. W.; Kim, K. S. Tuning Metal Single Atoms Embedded in NxCy Moieties toward High-Performance Electrocatalysis. *Energy Environ. Sci.* **2021**, *14*, 3455−3468.
(31) Sun, X.; Zheng, J.; Gao, Y.; Qiu, C.; Yan, Y.; Yao, Z.; Deng, S.; Wang, J. Machine-Learning-Accelerated Screening of Hydrogen Evolution Catalysts in MBenes Materials. *Appl. Surf. Sci.* **2020**, *526*, No. 146522.
(32) Ying, Y.; Fan, K.; Luo, X.; Qiao, J.; Huang, H. Unravelling the Origin of Bifunctional OER/ORR Activity for Single-Atom Catalysts Supported on C 2 N by DFT and Machine Learning. *J. Mater. Chem. A* **2021**, *9*, 16860−16867.
(33) Di Liberto, G.; Di; Cipriano, L. A.; Pacchioni, G. Universal Principles for the Rational Design of Single Atom Electrocatalysts? Handle with Care. *ACS Catal.* **2022**, *12*, 5846−5856.
(34) Tamtaji, M.; Guo, X.; Tyagi, A.; Galligan, P. R.; Liu, Z.; Roxas, A.; Liu, H.; Cai, Y.; Wong, H.; Zeng, L.; et al. Machine Learning-Aided Design of Gold Core-Shell Nanocatalysts toward Enhanced and Selective Photooxygenation. *ACS Appl. Mater. Interfaces* **2022**, *14*, 46471−46480.
(35) Anker, A. S.; Kjær, E. T. S.; Kjaer, E. T. S.; Juelsholt, M.; Christiansen, T. L.; Skjærvø, S. L.; Linn Skjaervø, S.; Jørgensen, M. R. V.; Ry, M.; Kantor, I.; Jørgensen, V.; Sørensen, D. R.; Kantor, I.; Billinge, S. J. L.; Sørensen, D. R.; Selvan, R.; Billinge, S. J. L. Extracting Structural Motifs from Pair Distribution Function Data of Nanostructures Using Explainable Machine Learning. *npj Comput. Mater.* **2022**, *8*, 3−5.
(36) Shin, D.; Choi, G.; Hong, C.; Han, J. W. Surface Segregation Machine-Learned with Inexpensive Numerical Fingerprint for the Design of Alloy Catalysts. *Mol. Catal.* **2023**, *541*, No. 113096.
(37) Mi, X.; Zou, B.; Zou, F.; Hu, J. Permutation-Based Identification of Important Biomarkers for Complex Diseases via Machine Learning Models. *Nat. Commun.* **2021**, *12*, No. 3008.
(38) Panapitiya, G.; Avendano-Franco, G.; Ren, P.; Wen, X.; Li, Y.; Lewis, J. P. Machine-Learning Prediction of CO Adsorption in Thiolated, Ag-Alloyed Au Nanoclusters. *J. Am. Chem. Soc.* **2018**, *140*, 17508−17514.
(39) Zong, X.; Vlachos, D. G. Exploring Structure-Sensitive Relations for Small Species Adsorption Using Machine Learning. *J. Chem. Inf. Model.* **2022**, *62*, 4361−4368.
(40) Jonayat, A. S. M.; Van Duin, A. C. T.; Janik, M. J. Discovery of Descriptors for Stable Monolayer Oxide Coatings through Machine Learning. *ACS Appl. Energy Mater.* **2018**, *1*, 6217−6226.
(41) Ioannidis, E. I.; Gani, T. Z. H.; Kulik, H. J. MolSimplify: A Toolkit for Automating Discovery in Inorganic Chemistry. *J. Comput. Chem.* **2016**, *37*, 2106−2117.
(42) Fu, Z.; Wu, M.; Li, Q.; Ling, C.; Wang, J. A Simple Descriptor for Nitrogen Reduction Reaction over Single Atom Catalysts. *Mater. Horiz.* **2023**, *10*, 852−858.
(43) 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* **2020**, *124*, 19571−19578.
(44) Wan, X.; Yu, W.; Niu, H.; Wang, X.; Zhang, Z.; Guo, Y. Revealing the Oxygen Reduction/Evolution Reaction Activity Origin of Carbon-Nitride-Related Single-Atom Catalysts: Quantum Chemistry in Artificial Intelligence. *Chem. Eng. J.* **2022**, *440*, No. 135946.
(45) Zheng, G.; Li, Y.; Qian, X.; Yao, G.; Tian, Z.; Zhang, X.; Chen, L. High-Throughput Screening of a Single-Atom Alloy for Electroreduction of Dinitrogen to Ammonia. *ACS Appl. Mater. Interfaces* **2021**, *13*, 16336−16344.
(46) Khan, K.; Liu, T.; Arif, M.; Yan, X.; Hossain, M. D.; Rehman, F.; Zhou, S.; Yang, J.; Sun, C.; Bae, S. H.; et al. Laser-Irradiated Holey Graphene-Supported Single-Atom Catalyst towards Hydrogen Evolution and Oxygen Reduction. *Adv. Energy Mater.* **2021**, *11*, No. 2101619.
(47) Li, H.; Liu, T.; Wei, P.; Lin, L.; Gao, D.; Wang, G.; Bao, X. High-Rate CO_2 Electroreduction to C_{2+} Products over a Copper-Copper Iodide Catalyst Angewandte. *Angew. Chem., Int. Ed.* **2021**, *133*, 14450−14454.
(48) An, B.; Zhou, J.; Zhu, Z.; Li, Y.; Wang, L.; Zhang, J. Uncovering the Coordination Effect on the Ni Single-Atom Catalysts for CO2 Reduction Including Vacancy Defect and Non-Vacancy Defect Structures. *Fuel* **2022**, *310*, No. 122472.
(49) Zafari, M.; Kumar, D.; Umer, M.; Kim, K. S. Machine Learning-Based High Throughput Screening for Nitrogen Fixation on Boron-Doped Single Atom Catalysts. *J. Mater. Chem. A* **2020**, *8*, 5209−5216.
(50) Palizhati, A.; Zhong, W.; Tran, K.; Back, S.; Ulissi, Z. W. Toward Predicting Intermetallics Surface Properties with High-Throughput DFT and Convolutional Neural Networks. *J. Chem. Inf. Model.* **2019**, *59*, 4742−4749.
(51) Sun, H.; Li, Y.; Gao, L.; Chang, M.; Jin, X.; Li, B.; Xu, Q.; Liu, W.; Zhou, M.; Sun, X. High Throughput Screening of Single Atomic Catalysts with Optimized Local Structures for the Electrochemical Oxygen Reduction by Machine Learning. *J. Energy Chem.* **2023**, *81*, 349−357.
(52) Boonpalit, K.; Wongnongwa, Y.; Prommin, C.; Nutanong, S.; Namuangruk, S. Data-Driven Discovery of Graphene-Based Dual-Atom Catalysts for Hydrogen Evolution Reaction with Graph Neural Network and DFT Calculations. *ACS Appl. Mater. Interfaces* **2023**, *15*, 12936−12945.
(53) Sathishkumar, N.; Chen, H. T. Regulating the Coordination Environment of Single-Atom Catalysts Anchored on Thiophene Linked Porphyrin for an Efficient Nitrogen Reduction Reaction. *ACS Appl. Mater. Interfaces* **2023**, *15*, 15545−15560.
(54) Ji, B.; Gou, J.; Zheng, Y.; Pu, X.; Wang, Y.; Kidkhunthod, P.; Tang, Y. Coordination Chemistry of Large-size Yttrium Single-atom Catalysts for Oxygen Reduction Reaction. *Adv. Mater.* **2023**, No. 2300381.

https://doi.org/10.1021/acs.jpcc.3c00765
J. Phys. Chem. C 2023, 127, 9992−10000