ACS OMEGA

Open Access

http://pubs.acs.org/journal/acsodf

Article

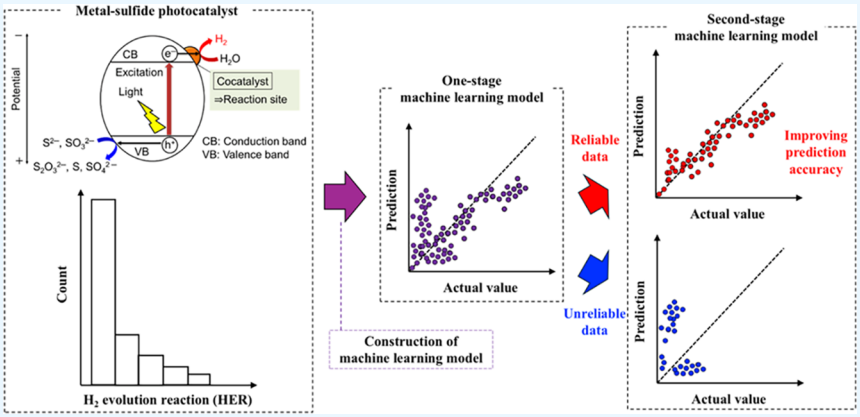
Data-Driven Approach Considering Imbalance in Data Sets and Experimental Conditions for Exploration of Photocatalysts

Wataru Takahara, Ryuto Baba, Yosuke Harashima,* Tomoaki Takayama, Shogo Takasuka, Yuichi Yamaguchi, Akihiko Kudo,* and Mikiya Fujii*

Cite This: *ACS Omega* 2025, 10, 14626−14639

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information



ABSTRACT: In the field of data-driven material development, an imbalance in data sets where data points are concentrated in certain regions often causes difficulties in building regression models when machine learning methods are applied. One example of inorganic functional materials facing such difficulties is photocatalysts. Therefore, advanced data-driven approaches are expected to help efficiently develop novel photocatalytic materials even if an imbalance exists in data sets. We propose a two-stage machine learning model aimed at handling imbalanced data sets without data thinning. In this study, we used two types of data sets that exhibit the imbalance: the Materials Project data set (openly shared due to its public domain data) and the in-house metal-sulfide photocatalyst data set (not openly shared due to the confidentiality of experimental data). This two-stage machine learning model consists of the following two parts: the first regression model, which predicts the target quantitatively, and the second classification model, which determines the reliability of the values predicted by the first regression model. We also propose a search scheme for variables related to the experimental conditions based on the proposed two-stage machine learning model. This scheme is designed for photocatalyst exploration, taking experimental conditions into account as the optimal set of variables for these conditions is unknown. The proposed two-stage machine learning model improves the prediction accuracy of the target compared with that of the one-stage model.

1. INTRODUCTION

Data-driven material development has attracted significant attention in recent years. This scientific field is often referred to as materials informatics (MI) and has been applied to various material systems.[1−5] Machine learning methods are applied at the MI to predict the material properties. Libraries in Python, a programming language commonly used in machine learning and available for MI, are also being developed.[6−13] In addition, open databases that can be used for machine learning exist.[14−25] Machine learning methods have the potential to predict target material properties, even in complex material systems, for which theoretical equations and computational simulations are sometimes difficult to apply. The following are examples of the application of machine learning methods to complex material systems. We (Takahara and coauthors) studied thermosetting resin composite materials.[26] Okuyama and coauthors investigated alternative composite materials.[27] However, in the field of materials development, an imbalance in data sets where data points are concentrated in certain regions presents difficulties when building prediction models using machine learning. The imbalance in data sets often reduces the reliability of the

Received: July 30, 2024
Revised: March 10, 2025
Accepted: March 14, 2025
Published: April 10, 2025

https://doi.org/10.1021/acsomega.4c06997
ACS Omega 2025, 10, 14626−14639

model and results in outputs that include both reliable and unreliable predictions. Training a machine learning model on imbalanced data sets often causes the model to overfit specific patterns in the data and reduces its ability to generalize to unseen data.

One example of inorganic functional materials facing such difficulty is photocatalysts. Recently, photocatalytic water splitting has attracted attention as a promising technology for producing low-cost green hydrogen.[28−30] The development of visible light-responsive photocatalysts is important for the efficient utilization of solar light. Metal sulfide is an attractive photocatalyst group because many of them show a high efficiency for the photocatalytic H_2 evolution reaction (HER) using visible light in the solar spectrum. A metal-sulfide photocatalyst plays an important role as an H_2-evolving photocatalyst when a Z-scheme system is constructed for water splitting under visible light irradiation. Figure 1 shows

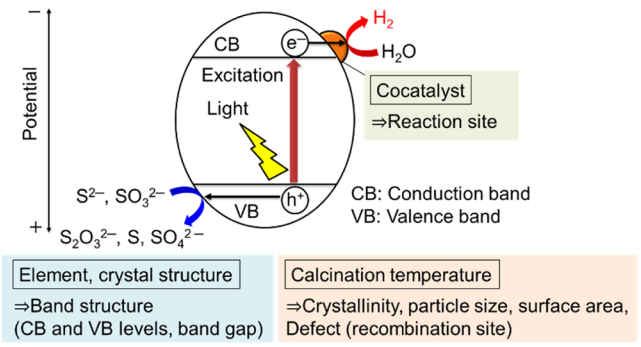


Figure 1. Sacrificial H_2 evolution over a metal-sulfide photocatalyst and factors affecting the photocatalytic performance.

sacrificial H_2 evolution using a semiconductor particulate metal-sulfide photocatalyst and factors affecting the photocatalytic performance. The sacrificial H_2 evolution is a half reaction of photocatalytic water splitting into H_2 and O_2 in a 2:1 ratio. When a metal-sulfide photocatalyst absorbs light of which energy is larger than the bandgap, electrons are excited from a valence band to a conduction band. Here, the band levels and the bandgap mainly determined by constituent elements and the crystal structure are important. As the next step, the photogenerated electrons and holes migrate to the surface and react with water. Calcination temperature in the preparation affects the crystallinity, particle size, surface area, and defect formation, which are also significant positive and negative factors. The cocatalyst is often loaded on the surface for introducing active sites for water reduction to form H_2 by photogenerated electrons. Photogenerated holes should oxidize water to form O_2 in water splitting. However, the holes photogenerated in the valence band of a metal-sulfide photocatalyst cannot oxidize water because they oxidize a photocatalyst itself, resulting in photocorrosion. The photocorrosion is suppressed by a sacrificial reagent working as an electron donor instead of water. As mentioned above, the HER is a phenomenon influenced by both chemical composition and experimental conditions.

A number of metal-sulfide photocatalysts have been developed for hydrogen evolution from aqueous solutions containing sacrificial reagents under visible light irradiation.[28,31,32] In particular, Kudo, who is one of the collaborators of our study, and the Kudo group have developed many metal-sulfide photocatalysts such as $ZnS−CuInS_2−AgInS_2$,[33] $A_2^I−Zn−A^{IV}−S_4$(A^I=Cu, Ag; A^{IV}=Sn, Ge),[34] $ZnS−CuGaS_2$,[35] and Cu_3MS_4(M = V, Nb, Ta)[36] being active for sacrificial hydrogen evolution under visible light irradiation. It is necessary to develop a novel metal-sulfide photocatalyst for efficient hydrogen evolution in the present stage because the activities of the developed photocatalysts are still low. Therefore, MI is expected to efficiently develop novel photocatalytic materials. This research topic is challenging because numerous and reliable data on various photocatalysts are necessary to construct a prediction model utilizing MI. Here, we have a large amount of original experimental data on the activities of sacrificial hydrogen evolution over metal-sulfide photocatalysts. The HER over photocatalysts is strongly influenced by experimental conditions, such as synthesis conditions, reaction conditions, and chemical composition.[28,37,38] Specifically, their HER activities vary by 10^3 orders of magnitude depending on the experimental conditions, even if the chemical compositions of the photocatalysts are identical. This trend creates a situation in which there are very few optimal synthesis conditions and reaction conditions, and the distribution of HERs is often skewed toward lower activity. Therefore, predicting HER activity is known to be difficult.

As mentioned above, training a machine learning model on imbalanced data sets often reduces generalization to unseen data. In machine learning, a log transformation can be employed when the distribution of an objective variable is skewed. Because a log transformation helps to reduce skewness and stabilize variance, the data are more suitable for modeling. However, zero values cannot be transformed using the log function, and samples with large values of the objective variable tend to exhibit large discrepancies.

We propose a two-stage machine learning model aimed at handling imbalanced data sets without data thinning. This approach predicts the target by combining a first regression model that predicts the target quantitatively with a second classification model that determines the reliability of the values predicted by the first regression model. In other words, the second classification model defines the applicability scope of the first regression model. The reliability metrics of the two-stage machine learning model consider both the prediction accuracy of the prediction model and the metrics of the domain about whether the data itself is reliable or not. Reliability can be considered with residuals[39] or an approach that deals with domain applicability metrics,[40] as described in previous studies. In contrast to the previous studies, the data set we are working with has an imbalanced frequency distribution, with certain values appearing more frequently. In this study, we aimed to transform the regression problem of a data set with an imbalanced frequency distribution, characterized by the high occurrence of certain values, into a balanced binary classification problem to avoid data imbalance. Approaches combining models have been reported in the field of MI. Examples of these investigations are as follows. Sakaushi and coauthors have dealt with the loop between material synthesis and machine learning.[41] In this study, a seamless loop between prediction by machine learning and material development was achieved using Bayesian optimization to determine the subsequent experimental conditions for 11 composition-related variables. This study assumes that the properties are uniquely determined by the chemical composition. Talapatra and coauthors investigated bandgaps using density functional theory calculations.[42] Parts of the data with a certain threshold value were removed to obtain an accurate prediction model.

https://doi.org/10.1021/acsomega.4c06997
ACS Omega 2025, 10, 14626−14639

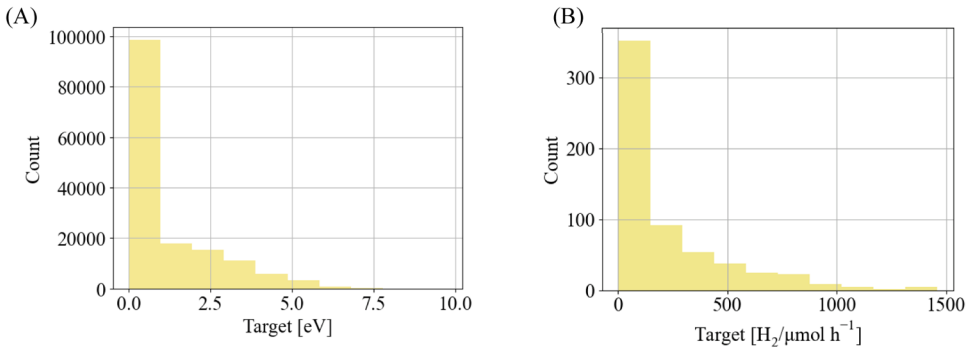


Figure 2. (A) Histogram of bandgaps in the Materials Project data set used in this study. (B) Histogram of HERs in the photocatalyst data set used in this study.

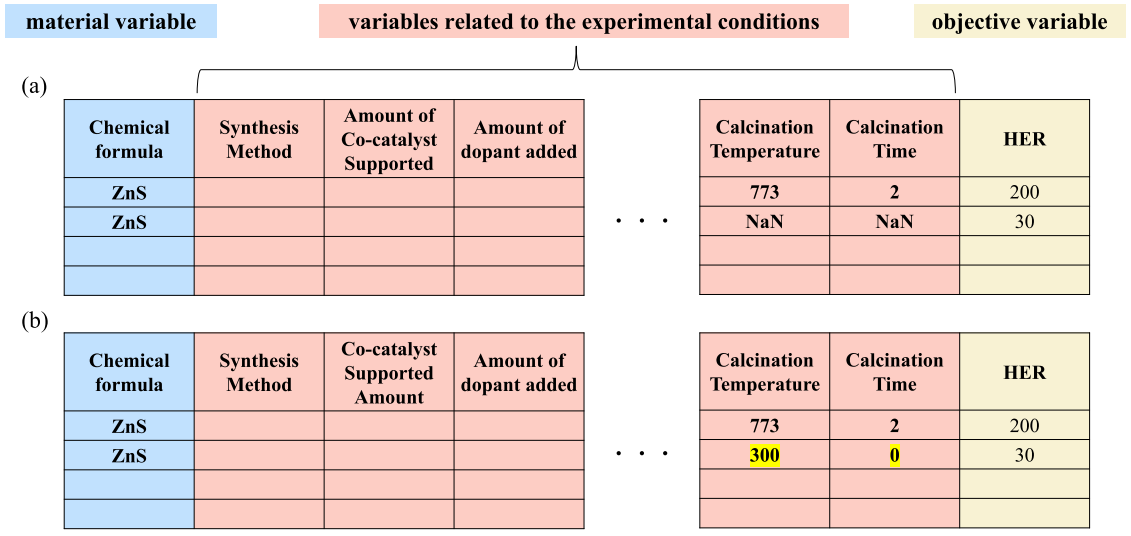


Figure 3. (a) Overview of the photocatalyst data set. (b) Example of the imputation of missing values.

Balachandran and coauthors also investigated perovskite compounds.[43] They classified perovskites or nonperovskites in the first step and cubic or noncubic in the second step. Here, we discuss our attempts to deal with imbalanced data sets without data thinning.

When searching for a photocatalyst with high HER activity using the constructed two-stage machine learning model, it is necessary to define variables related to the experimental conditions in addition to the chemical composition. Because the optimal set of variables for the experimental conditions for the chemical compositions to be explored is unknown, it is desirable to consider as many diverse sets of variables related to the experimental conditions as possible. Furthermore, the considered sets should be in a range in which the two-stage machine learning model can be driven well. We also propose a search scheme for variables related to the experimental conditions in anticipation of a photocatalyst search utilizing this two-stage machine learning model. This search scheme explicitly visualizes the importance of experimental conditions, as the activities can vary by 10^3 orders of magnitude even with identical chemical compositions.

2. DATA SET DESCRIPTION

In this study, we used two types of data sets that exhibit imbalance: the Materials Project data set (openly shared due to its public domain data) and the in-house metal-sulfide photocatalyst data set (not openly shared due to the confidentiality of experimental data). The former serves as a demonstration of public domain data, while the latter serves as a demonstration of our proprietary experimental data. The common characteristic imbalance in the two data sets is that data points are concentrated near zero, with the frequency decreasing as the values become larger. Detailed descriptions of each data set are provided below.

2.1. Materials Project Data Set. In this study, we used 153219 data collected from the Materials Project (version 2023.11.1) as public domain data. This data set consisted of a material variable (chemical composition) and an objective variable (bandgap). As shown in Figure 2A, the distribution of the bandgap was skewed toward smaller values. In this study, the material variables were converted into numerical feature vectors using Xenonpy[10] and converted into a format that could be processed by machine learning algorithms. The procedure is described in the following section. First, chemical compositions were converted into comp_dict, a dictionary of proportions by element, using Pymatgen.[6,7] Subsequently, 290 compositional features based on comp_dict were calculated from 58 features of each element (atomic number, bond radius, van der Waals radius, electronegativity, thermal conductivity, bandgap, polarizability, boiling point, melting point, etc.).[44] All of the chemical compositions used in this study were normalized. In this data set, there were no features with zero variance. Only duplicate features were removed, and

https://doi.org/10.1021/acsomega.4c06997
ACS Omega 2025, 10, 14626−14639

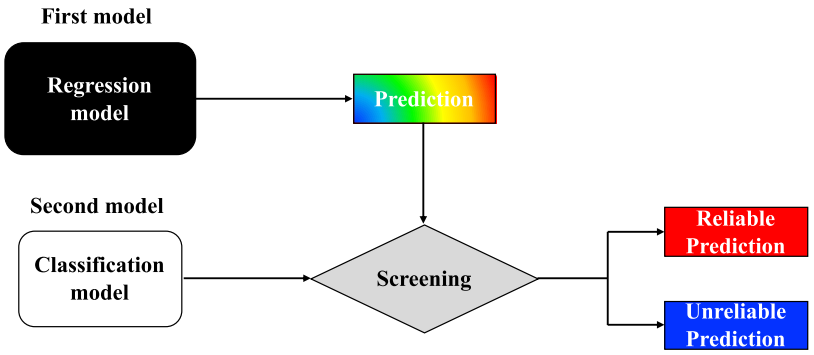


Figure 4. Pipeline of the two-stage machine learning model.

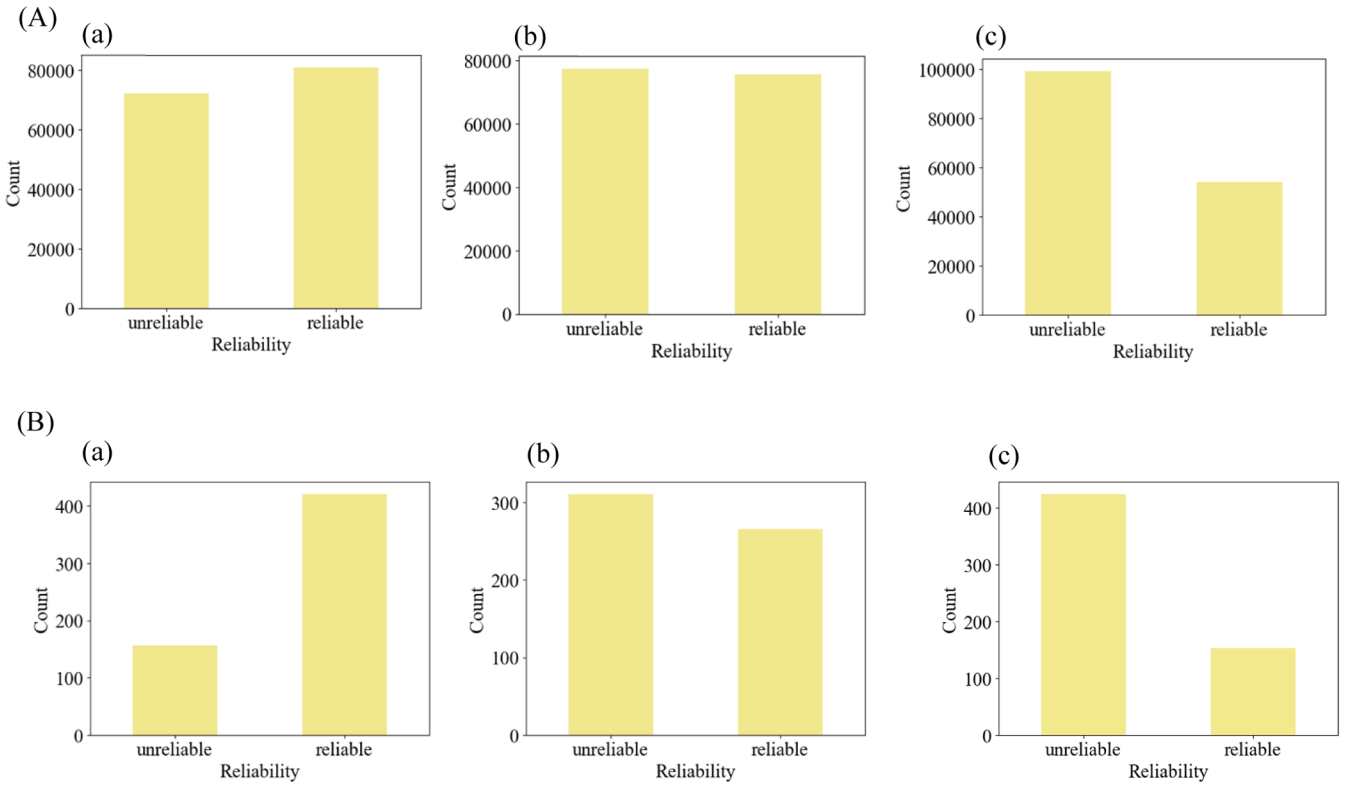


Figure 5. (A) Histogram of the Materials Project data set divided into reliable and unreliable categories: (a) cutoff = 0.0001 [eV], (b) cutoff = 0.1 [eV], and (c) cutoff = 1.0 [eV]. (B) Histogram of the photocatalyst data set divided into reliable and unreliable categories: (a) cutoff = 10 [H_2/μmol h^{-1}], (b) cutoff = 100 [H_2/μmol h^{-1}], and (c) cutoff = 300 [H_2/μmol h^{-1}].

232 features were used. Also, there were no missing values in this data set.

2.2. Photocatalyst Data Set. Kudo et al. have developed a number of metal-sulfide photocatalysts capable of producing hydrogen from an aqueous solution containing sacrificial reagents under visible light irradiation. For example, these data were used in this study.[32−36,45−53] 577 experimental data for these photocatalysts were used in this study. This data set consisted of a material variable (chemical composition), variables related to the experimental conditions (91 variables related to synthesis conditions and reaction conditions), and an objective variable (HER). In this data set, sacrificial hydrogen evolution was conducted under visible light irradiation ($\lambda \geq 420$ nm) using a 300 W Xe lamp attached with a cutoff filter (L42, HOYA). Figure 3a presents an overview of the data set. As shown in Figure 2B, the distribution of the HER was skewed toward smaller values. Similar to the Materials Project data set, the chemical compositions were converted into 290 compositional features. Compositional features with zero variance and duplicate features were removed, and 225 features were used.

If the values of a particular variable for the experimental conditions were not present, then it was assumed that these values were missing. An example of this case is shown in Figure 3a. Subsequently, a two-stage machine learning model was built using the data set containing these missing values, which are written as NaN (namely, Not a Number) in Figure 3a. In addition, Uniform Manifold Approximation and Projection (UMAP),[54,55] the dimensionality compression algorithm used in the present study, does not allow for the presence of missing values. The missing values were complemented as much as possible without compromising their intrinsic physical meaning. The yellow highlight in Figure 3b shows an example of missing value imputation. The missing values were comple-

https://doi.org/10.1021/acsomega.4c06997
ACS Omega 2025, 10, 14626−14639

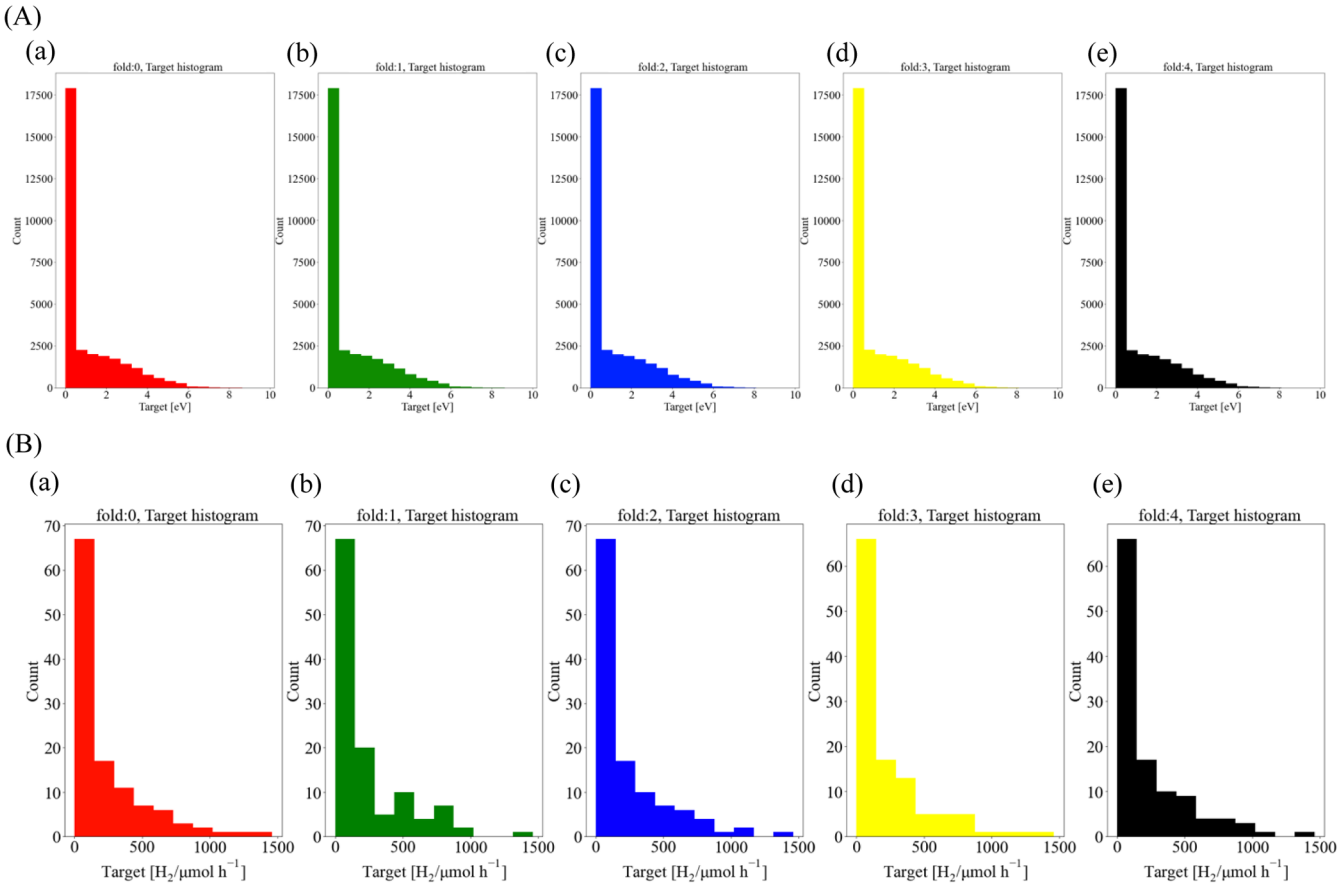


Figure 6. (A) Fold split results for the Materials Project data set: (a) fold:0, (b) fold:1, (c) fold:2, (d) fold:3, and (e) fold:4. (B) Fold split results for the photocatalyst data set: (a) fold:0, (b) fold:1, (c) fold:2, (d) fold:3, and (e) fold:4.

mented by replacing the calcination time with 0 h at room temperature (300 K). The missing values in the entire data set were imputed as follows: processing time, 0 h; processing temperature, 300 K (room temperature); and number of elements in the synthesis, 0.

3. TWO-STAGE MACHINE LEARNING MODEL

3.1. Method. Figure 4 shows the pipeline of the two-stage machine learning model proposed in this study. The two-stage model consisted of a first regression model and a second classification model. The second classification model determines the reliability of the values predicted by the first regression model. The second model yields a ″reliable″ judgment for the case of reliable, whereas an ″unreliable″ judgment is made in the case of unreliable.

The thresholds for the classification between reliable and unreliable in the second model were determined such that the numbers of reliable and unreliable values were balanced as much as possible. Here, the cutoff for determining reliability based on bandgap was set at 0.1 eV to balance the rates of reliable and unreliable data in the Materials Project data set. As shown in Figure 5A, the cutoff value was determined by testing various values to find the one that best balanced reliable and unreliable data. A data set is considered reliable if its bandgap is 0.1 [eV] or higher and unreliable if its bandgap is below 0.1 [eV]. Similarly, the cutoff for determining reliability based on HER was set at 100 [H_2/μmol h^{-1}] to balance the rates of reliable and unreliable data in the photocatalyst data set. As shown in Figure 5B, the cutoff value was determined by testing various values to find the one that best balanced reliable and unreliable data. Moreover, metal-sulfide photocatalysts with HER values of 100 [H_2/μmol h^{-1}] or higher are considered promising based on photocatalytic expertise. A data is considered reliable if its HER is 100 [H_2/μmol h^{-1}] or higher, and unreliable if its HER is below 100 [H_2/μmol h^{-1}]. In the Materials Project data set, the LightGBM[56,57] was used to construct the first and second models. In the photocatalyst data set, the LightGBM, the XGBoost,[58,59] the CatBoost,[60,61] the Random Forest,[62] the Gaussian Process,[63] and the Support Vector Machine[64] are used. The Operating System used in this method and the values of the hyperparameters are shown in the Appendix. The LightGBM, the XGBoost, and the CatBoost are based on a gradient-boosting algorithm that can be driven even in the presence of missing values. Scikit-learn[65,66] was used for the cross-validation, evaluation metrics, the Random Forest, the Gaussian Process, and the Support Vector Machine. The Random Forest, the Gaussian Process, and the Support Vector Machine do not allow for the presence of missing values. Missing values were imputed using the same method described in the UMAP section for the Photocatalyst Data Set. The default decision probability for the reliable and unreliable values of the second model was 0.5.

The same folds used for cross-validation were used for both models. This enabled the first and second model results to be combined during the performance evaluation. Specifically, they are combined via out-of-fold (OOF). OOF is the collection of all of the predictions of the machine learning model in the validation data for each fold during cross-validation, rearranged

https://doi.org/10.1021/acsomega.4c06997
ACS Omega 2025, 10, 14626−14639

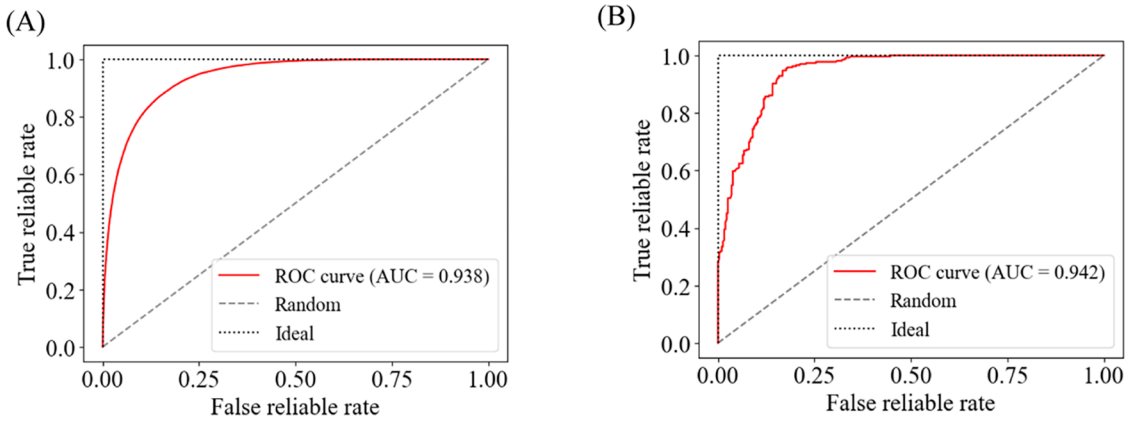


Figure 7. (A) ROC curves calculated using the OOF in the second model for the Materials Project data set. (B) ROC curves calculated using the OOF in the second model for the photocatalyst data set.

in the same order as the original training data. The cross-validation employed in this study was performed by dividing the objective variable into bins and stratifying the k-fold into 5-fold (stratified k-folds). The number of bins was determined using Sturge's law.[67] This validation design balances the difficulty of quantitatively validating data for each fold. This is expected to build models with good generalization performance.

Generally, when constructing a machine learning model, it is essential to identify which features (descriptors) are important for the target and to make physical sense of that importance. In this study, we calculated the average feature importance values using the get_score method with the weight importance type on the booster objects obtained from 5-fold cross-validation models.

3.2. Results and Discussion. *3.2.1. Materials Project Data Set.* Figure 5A(b) shows the histogram in which the data set was divided into reliable and unreliable data. Compared to Figure 2A, the imbalance of the data set has been eliminated by dividing it into two classes. Figure 6A shows the results of partitioning all data into 5-fold using stratified k-folds. The distributions of the partitioned data (Figure 6A (a−e)) almost resembled each other and the original data set shown in Figure 2A. This indicates that cross-validation can divide the data set into five partitions while retaining the imbalance in data sets.

The receiver operating characteristic (ROC) curve calculated using the OOF in the second model is shown in Figure 7A. The accuracy, F1 Score, and area under the ROC curve (AUC) are summarized in Table 1A. The accuracy, F1 score, and AUC were 0.861, 0.864, and 0.938, respectively. These results indicate that a second model with a good classification performance was constructed.

Figure 8A shows the effects of the combination of regression using the first model and classification using the second model on the prediction accuracy. The color of the plot in Figure 8A (a) indicates the probability value of reliable decisions in the second model. Figure 8A (a) shows a distorted shape with a shoulder near zero owing to the imbalanced bandgap distribution. Figure 8A (b) and (c) depicts the impact of the classification of Figure 8A (a) based on the score of the probability of reliability (the threshold was 0.5). In Figure 8A (b), the distortion is improved, and the plot points are closer to the ideal line indicated by the dotted line. However, Figure 8A (c) shows that the plot points were generally far from the ideal line. The coefficient of determination (R^2), root mean squared error (RMSE), and mean absolute error (MAE) indicated better accuracy for the data with reliable judgments and worse accuracy for the data with unreliable judgments compared to the case of all data. In conclusion, the proposed two-stage machine learning model was effective in improving the prediction accuracy of the bandgap.

Table 1. (A) Accuracy, F1 Score, and AUC in the Second Model on the Materials Project Data Set; (B) Accuracy, F1 Score, and AUC in the Second Model for Each Machine Learning Algorithm on the Photocatalyst Data Set

(A)		
accuracy	F1 Score	AUC
0.861	0.864	0.938

(B)			
algorithm	accuracy	F1 Score	AUC
LightGBM	0.858	0.854	0.929
XGBoost	0.875	0.870	0.942
CatBoost	0.873	0.868	0.933
Random Forest	0.868	0.864	0.933
Gaussian Process	0.858	0.852	0.912
Support Vector Machine	0.802	0.794	0.889

3.2.2. Photocatalyst Data Set. Figure 5B(b) shows the histogram in which the data set was divided into reliable and unreliable data. Compared to Figure 2B, the imbalance of the data set has been eliminated by dividing it into two classes. Figure 6B shows the results of partitioning all data into 5-fold using stratified k-folds. The distributions of the partitioned data (Figure 6B (a−e)) almost resembled each other, and the original data set is shown in Figure 2B. This indicates that cross-validation can divide the data set into five partitions while retaining the imbalance in data sets.

The accuracy, F1 score, and AUC of the second model for each machine learning algorithm are summarized in Table 1B. The results of XGBoost that achieved the highest accuracy are discussed in detail. The ROC curve calculated using the OOF in the second model with XGBoost is shown in Figure 7B. The accuracy, F1 score, and AUC of XGBoost were 0.875, 0.870, and 0.942, respectively. These results indicate that a second model with good classification performance was constructed.

The R^2, RMSE, and MAE of the first model for each machine learning algorithm are summarized in Table 2. The R^2, RMSE, and MAE of the two-stage machine learning model, based on the score of the probability of reliability (the

https://doi.org/10.1021/acsomega.4c06997
ACS Omega 2025, 10, 14626−14639

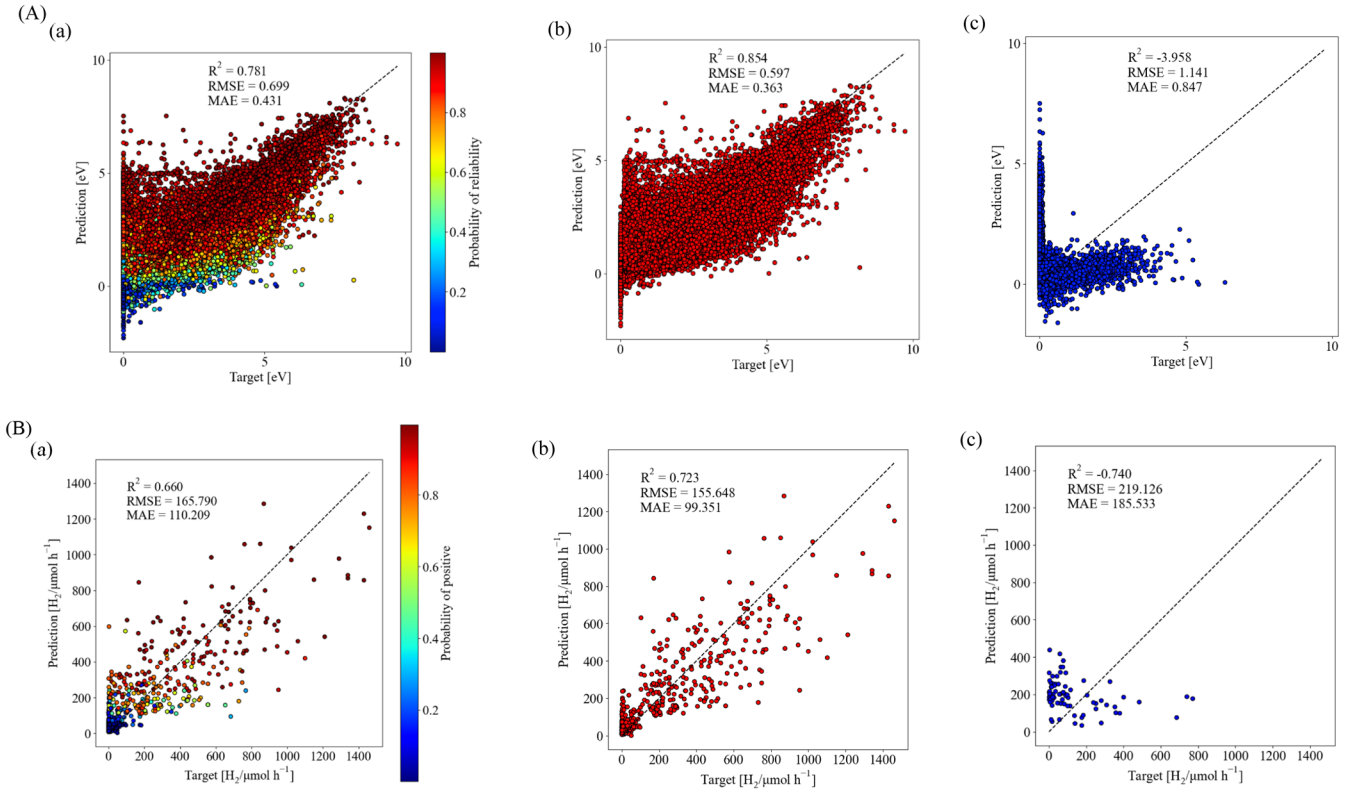


Figure 8. (A) (a) Plot of predicted versus measured values for the OOF of the first model on all data, (b) plot of predicted versus measured values on these for which the second model made reliable decisions, and (c) unreliable decisions for the Materials Project data set. (B) (a) Plot of predicted versus measured values for the OOF of the first model on all data, (b) plot of predicted versus measured values on these for which the second model made reliable decisions, and (c) unreliable decisions for the photocatalyst data set.

Table 2. R^2, RMSE, and MAE in the First Model for Each Machine Learning Algorithm on the Photocatalyst Data Set

algorithm	R^2	RMSE	MAE
LightGBM	0.589	182.382	121.048
XGBoost	0.66	165.79	110.209
CatBoost	0.648	168.793	113.643
Random Forest	0.602	179.374	125.222
Gaussian Process	0.604	178.917	127.09
Support Vector Machine	0.531	194.795	127.492

Table 3. R^2, RMSE, and MAE of the Two-Stage Machine Learning Model, Based on the Score of the Probability of Reliability (the Threshold Was 0.5), for Each Machine Learning Algorithm on the Photocatalyst Data Set

algorithm	$R_{(\text{reliable})}{}^2$	$\text{RMSE}_{(\text{reliable})}$	$\text{MAE}_{(\text{Reliable})}$
LightGBM	0.651	176.524	110.535
XGBoost	0.723	155.648	99.351
CatBoost	0.704	161.409	103.339
Random Forest	0.66	172.546	116.073
Gaussian Process	0.667	172.284	117.503
Support Vector Machine	0.604	192.014	117.592

threshold was 0.5), for each machine learning algorithm are summarized in Table 3. The results from Tables 2 and 3 confirmed an improvement in prediction accuracy for the two-stage machine learning model across all of the algorithms. This indicates that the use of the two-stage model is useful regardless of the algorithm. The results of XGBoost that achieved the highest accuracy are discussed in detail. Figure 8B shows the effects of the combination of regression using the first model and classification using the second model, both based on XGBoost, on the prediction accuracy. The color of the plot in Figure 8B (a) indicates the probability value of reliable decisions in the second model. Figure 8B (a) shows a distorted shape with a shoulder near zero owing to the imbalanced HER distribution. Figure 8B (b) and (c) depicts the impact of the classification of Figure 8B (a) based on the score of the probability of reliability (the threshold was 0.5). In Figure 8B (b), the distortion is improved, and the plot points are closer to the ideal line indicated by the dotted line. However, Figure 8B (c) shows that the plot points were generally far from the ideal line. The R^2, RMSE, and MAE

https://doi.org/10.1021/acsomega.4c06997
ACS Omega 2025, 10, 14626−14639

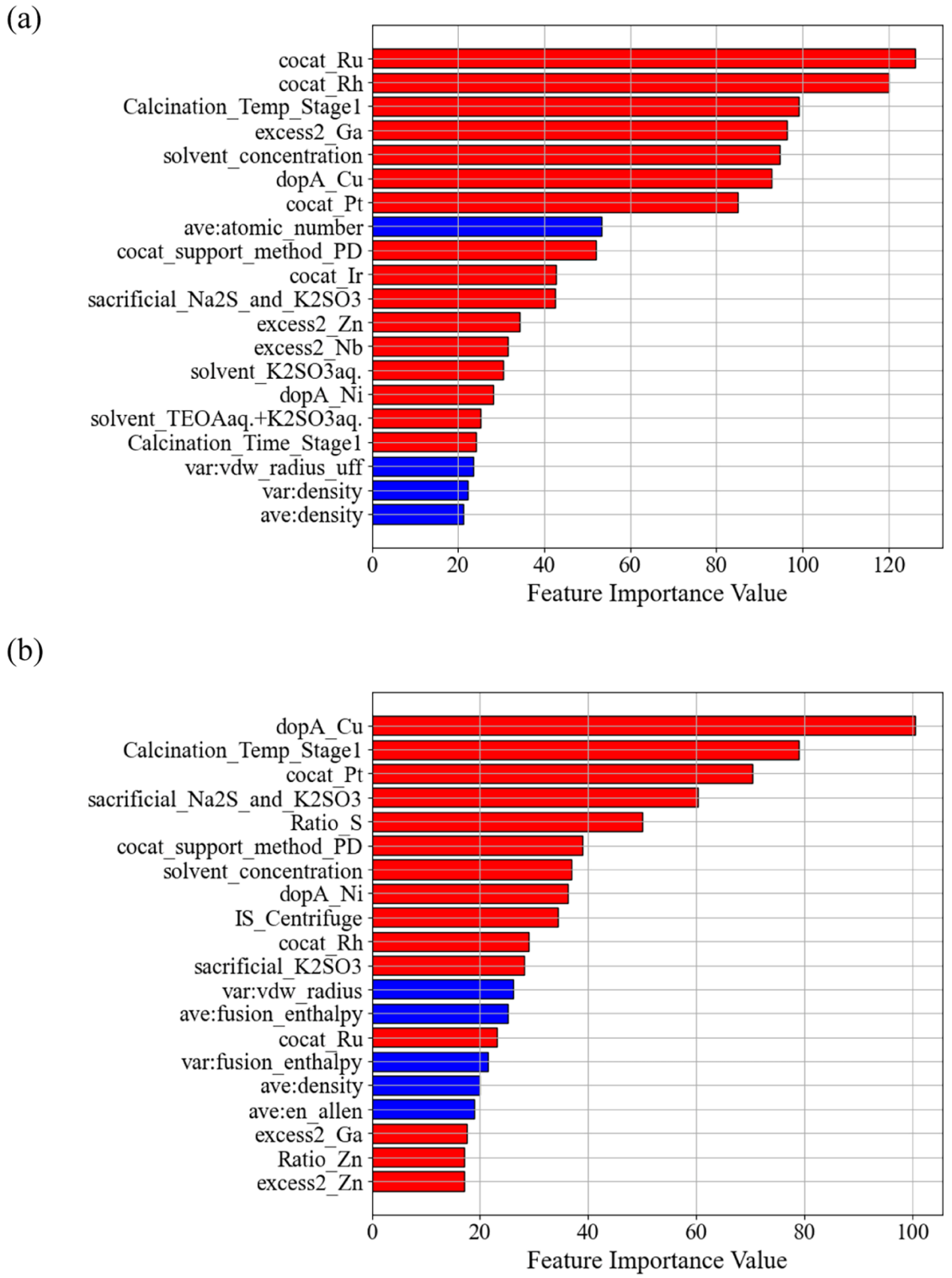


Figure 9. (a) Top 20 feature importance values for the first regression model; (b) top 20 feature importance values for the second classification model for the photocatalyst data set. Features related to experimental conditions are highlighted in red, and those derived from composition are highlighted in blue.

indicated better accuracy for the data with reliable judgments and worse accuracy for the data with unreliable judgments, compared to the case of all data. In conclusion, the proposed two-stage machine learning model was effective in improving the prediction accuracy of the HER.

The demonstrations using the Materials Project data set and the photocatalyst data set showed that the proposed two-step machine learning model is effective for handling imbalanced data, as illustrated in Figure 2.

3.2.3. Feature Importance. We calculated the feature importance of the two-stage machine learning model constructed using the photocatalyst data set with XGBoost because it achieved the highest prediction accuracy. Figure 9 shows the feature importance of the constructed two-stage

https://doi.org/10.1021/acsomega.4c06997
ACS Omega 2025, 10, 14626−14639

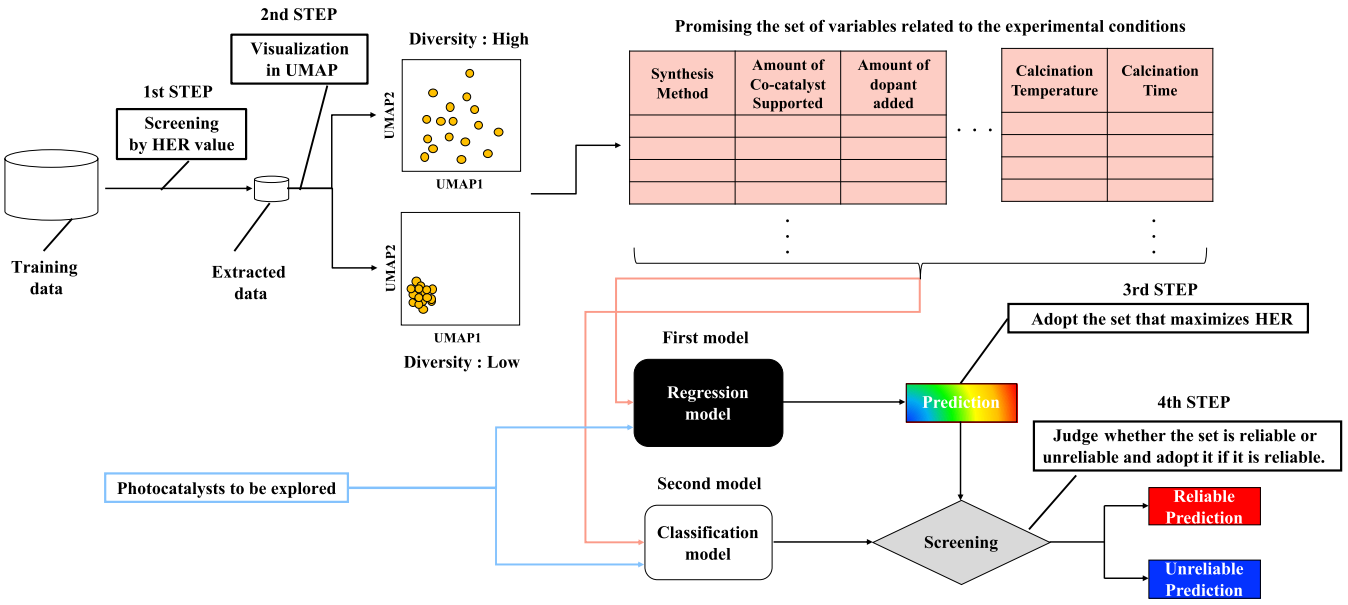


Figure 10. Procedures for the search scheme for variables related to the experimental conditions.

machine learning model, where features related to experimental conditions are highlighted in red and those related to chemical composition are highlighted in blue. Figure 9a,b shows the top 20 feature importance values for the first regression model and the second classification model, respectively. These results suggest that both experimental conditions and chemical composition play a significant role in predicting HER. Additionally, calcination temperature and cocatalyst were identified as more important features, both ranking among the top 3, in both the first regression model and the second classification model. This indicates that calcination temperature is closely related to HER, affecting crystallinity, particle size, surface area, and defect formation, and the cocatalyst is closely related to HER, introducing active sites for water reduction to form H_2 by photogenerated electrons, as explained previously in Figure 1. These findings are consistent with photocatalytic expertise.

4. SEARCH SCHEME FOR VARIABLES RELATED TO EXPERIMENTAL CONDITIONS

4.1. Method. Figure 10 illustrates the search scheme for variables related to the experimental conditions implemented in this study. This study proposes an approach to photocatalyst exploration. The actual exploration of photocatalysis is planned for future work. Among the four steps described below, the procedures from the third step onward correspond to the future work. In the first step, sets of variables for the experimental conditions were extracted from the training data under screening conditions. The screening conditions for the first step were set to satisfy the following conditions. A set of variables related to experimental conditions are extracted to satisfy that the HER is greater than or equal to 100 [H_2/μmol h^{-1}], which is the threshold of the second model of the two-stage machine learning model. For identical chemical compositions, the set of variables for the experimental conditions with the highest HERs was adopted. Because the sets were extracted from the training data, it is expected that the two-stage machine learning model would be well-driven in terms of the applicability domain of the model. In the second step, the sets were visualized in the principal 2D space to check the diversity, where UMAP was used to compress the dimensions of the variables for the experimental conditions. The operating systems used in this method and parameter values are presented in the Appendix. In the third step, a two-stage machine learning model is run for all combinations of the obtained sets for the photocatalyst to be explored. We then adopt the set of photocatalysts and variables for the experimental conditions that achieve the HER with the largest value in the first regression model. In the fourth step, we determine whether the photocatalyst with the set is reliable or unreliable; if it is reliable, we adopt the set.

4.2. Results and Discussion. Figure 11 shows the variables related to experimental conditions of the photocatalyst data set in the principal two-dimensional space constructed with UMAP. The orange circles indicate the

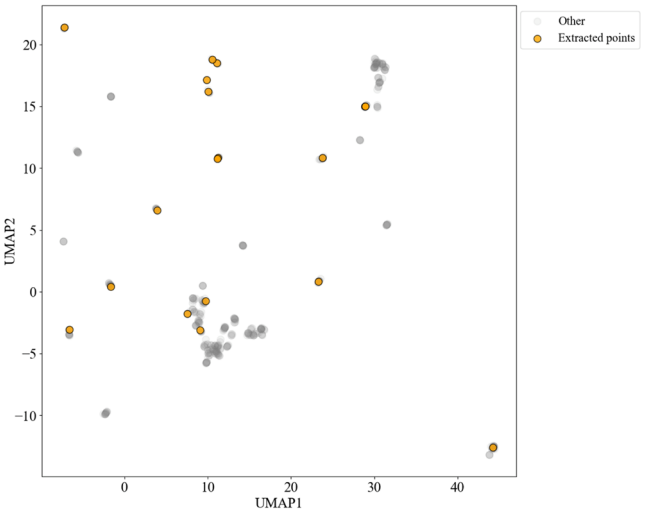


Figure 11. Variables related to experimental conditions of the photocatalyst data set in the principal two-dimensional space constructed with UMAP (focused on extracted points by the second step of the search scheme for variables related to experimental conditions).

https://doi.org/10.1021/acsomega.4c06997
ACS Omega 2025, 10, 14626−14639

extracted sets of variables for the experimental conditions in this scheme, which are scattered over space without clumping. The gray areas represent the other sets present in the training data. This confirmed the diversity in the promising sets of experimental condition variables obtained.

Figure 12 shows an additional visualization of UMAP. The overview of Figure 12 is the same as that of Figure 11 because

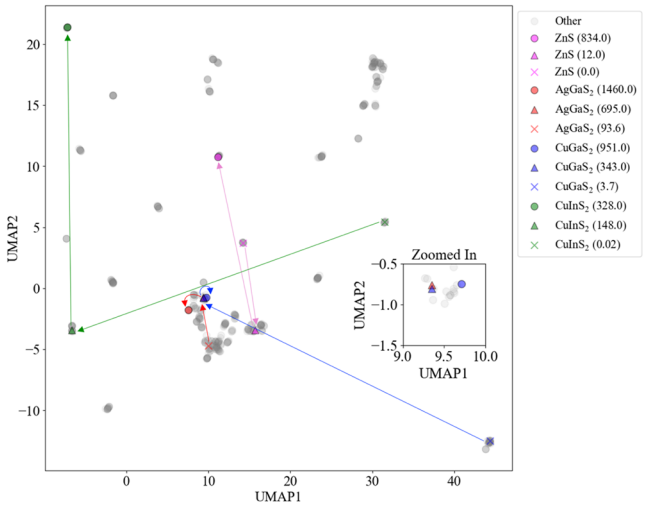


Figure 12. Variables related to experimental conditions of the photocatalyst data set in the principal two-dimensional space constructed with UMAP, including a zoomed-in view (focused on the ZnS, $AgGaS_2$, $CuGaS_2$, and $CuInS_2$). The values within parentheses are the HER values provided in $[H_2/\mu mol\ h^{-1}]$.

the plotting parameters are the same. However, there are additional roles to highlight the plots. Its highlights focused on the ZnS, $AgGaS_2$, $CuGaS_2$, and $CuInS_2$ are indicated by pink, red, blue, and green symbols, respectively. Specifically, we chose points related to experimental conditions corresponding to the minimum, median, and maximum HER values. These are plotted as crosses, triangles, and circles. The gray circles represent the other sets present in the training data. By focusing on the trend of the positions of the plots highlighted by the arrows with the corresponding colors, it can be recognized that the directions of the connected arrows in the order of minimum, median, and maximum were different from each other. The circles corresponding to the maximum of the HER of each photocatalyst are also located in different positions on the space of UMAP. Taking the plots of ZnS as an example, the arrow from the cross symbol to the triangular symbol is directed toward the lower-right side. In contrast, the arrow from the triangular symbol to the circular symbol was directed toward the upper left side. This trend is not similar to those of other representative materials ($AgGaS_2$, $CuGaS_2$, and $CuInS_2$). These discrepant trends indicate that the optimal variables for the experimental conditions depend on the composition of the photocatalyst. Additionally, in Figure 12, the values in parentheses are the HER values provided in $[H_2/\mu mol\ h^{-1}]$. These show that the minimum to maximum HER values of ZnS, $AgGaS_2$, $CuGaS_2$, and $CuInS_2$ each change significantly across 100 $[H_2/\mu mol\ h^{-1}]$, indicating that the HER varies significantly depending on the experimental conditions. In other words, for extracting information to optimize the variables related to experimental conditions of a specific composition, data sets corresponding to low HER values (less than 100 $[H_2/\mu mol\ h^{-1}]$) included important information to improve HER over the composition. If we do not use the data with HER less than 100 $[H_2/\mu mol\ h^{-1}]$, we could not use such information about the dependencies on the variables related to the experimental conditions and fail to understand them. Therefore, we confirmed the effectiveness of the approach presented in this study, which utilizes all of the data without thinning them.

5. CONCLUDING REMARKS

In the present study, a two-stage machine learning model was constructed for skewed data sets and applied to the data sets for bandgap values from the Materials Project and the in-house metal-sulfide photocatalysts. The results showed that the present machine learning model improved the prediction accuracy of the targets (bandgap and HER) compared with that of the one-stage regression model. In the photocatalyst data set, results indicate that the use of the two-stage model is useful regardless of the algorithm for skewed data sets. The proposed two-stage machine learning model is available with various regression models and can be applied to a wide range of tasks. We calculated the feature importance of the two-stage machine learning model constructed using the photocatalyst data set. Findings derived from feature importance were consistent with photocatalytic expertise.

In addition, a search scheme for variables related to the experimental conditions was constructed in anticipation of a photocatalyst search utilizing this two-stage machine learning model. This scheme allowed the preparation of diverse and promising sets of experimental condition variables for photocatalyst exploration in future work. It also confirmed the effectiveness of the approach presented in this study of using all of the data without thinning out.

The two-stage machine learning model and search scheme allow us to develop novel materials for photocatalysts that consider imbalance in data sets and experimental conditions. We believe that the present approach is effective not only for photocatalysts but also for various materials for which the experimental conditions should be considered.

APPENDIX

The Operating System used in this study

Linux (Ubuntu 20.04.6 LTS)

The parameters of the first regression model (LightGBM)

```
import lightgbm as lgbm

lgbm_params_reg = {
    "objective": "regression",
    "learning_rate": 0.01,
    "reg_alpha": 0.1,
    "reg_lambda": 0.1,
    "random_state": 42,
    "max_depth": 5,
    "n_estimators": 1000000,
    "colsample_bytree": 0.7,
}

model_reg = lgbm.LGBMRegressor(**lgbm_params_reg)

model_reg.fit(
    X_train,
    y_train,
    eval_set=[(X_valid, y_valid)],
    eval_metric="rmse",
    callbacks=[
        lgbm.early_stopping(stopping_rounds=10, verbose=True),
        lgbm.log_evaluation(200),
    ],
)

pred = model_reg.predict(X_valid)
```


https://doi.org/10.1021/acsomega.4c06997
ACS Omega 2025, 10, 14626−14639

(XGBoost)

```
import xgboost as xgb

xgb_params_reg = {
    "objective": "reg:squarederror",
    "learning_rate": 0.05,
    "alpha": 0.1,
    "lambda": 0.1,
    "random_state": 42,
    "max_depth": 5,
    "n_estimators": 1000000,
    "colsample_bytree": 0.7,
}

model_reg = xgb.XGBRegressor(**xgb_params_reg)

model_reg.fit(
    X_train,
    y_train,
    eval_set=[(X_valid, y_valid)],
    early_stopping_rounds=10,
    verbose=200,
)

pred = model_reg.predict(X_valid)
```

(CatBoost)

```
from catboost import CatBoostRegressor

catboost_params_reg = {
    "learning_rate": 0.03,
    "l2_leaf_reg": 0.1,
    "random_seed": 42,
    "depth": 5,
    "iterations": 1000000,
    "colsample_bylevel": 0.7,
    "eval_metric": "RMSE",
    "early_stopping_rounds": 10,
}

model_reg = CatBoostRegressor(**catboost_params_reg)

model_reg.fit(
    X_train, y_train, eval_set=(X_valid, y_valid), use_best_model=True, verbose=200
)

pred = model_reg.predict(X_valid)
```

(Random Forest)

```
from sklearn.ensemble import RandomForestRegressor

rf_params_reg = {
    "random_state": 42,
    "max_depth": 5,
    "n_estimators": 1000,
    "max_features": 0.7,
}

model_reg = RandomForestRegressor(**rf_params_reg)

model_reg.fit(X_train, y_train)

pred = model_reg.predict(X_valid)
```

(Gaussian Process)

```
from sklearn.gaussian_process import GaussianProcessRegressor
from sklearn.gaussian_process.kernels import RBF, ConstantKernel, WhiteKernel
from sklearn.preprocessing import StandardScaler

scaler_X = StandardScaler()
scaler_y = StandardScaler()
X_train = scaler_X.fit_transform(X_train)
X_valid = scaler_X.transform(X_valid)
y_train = scaler_y.fit_transform(y_train.reshape(-1, 1)).ravel()
y_valid = scaler_y.transform(y_valid.reshape(-1, 1)).ravel()

model_reg = GaussianProcessRegressor(
    kernel=ConstantKernel() * RBF() + WhiteKernel(), alpha=0, random_state=42
)

model_reg.fit(X_train, y_train)

pred = model_reg.predict(X_valid)
pred_original = scaler_y.inverse_transform(pred.reshape(-1, 1)).ravel()
```

(Support Vector Machine)

```
from sklearn.preprocessing import StandardScaler
from sklearn.svm import SVR

scaler_X = StandardScaler()
scaler_y = StandardScaler()
X_train = scaler_X.fit_transform(X_train)
X_valid = scaler_X.transform(X_valid)
y_train = scaler_y.fit_transform(y_train.reshape(-1, 1)).ravel()
y_valid = scaler_y.transform(y_valid.reshape(-1, 1)).ravel()

model_reg = SVR()

model_reg.fit(X_train, y_train)

pred = model_reg.predict(X_valid)
pred_original = scaler_y.inverse_transform(pred.reshape(-1, 1)).ravel()
```

The parameters of the second classification model

(LightGBM)

```
import lightgbm as lgbm

lgbm_params_clf = {
    "objective": "binary",
    "learning_rate": 0.01,
    "reg_alpha": 0.1,
    "reg_lambda": 0.1,
    "random_state": 42,
    "max_depth": 5,
    "n_estimators": 1000000,
    "colsample_bytree": 0.7,
    "is_unbalance": True,
}

model_clf = lgbm.LGBMClassifier(**lgbm_params_clf)

model_clf.fit(
    X_train,
    y_train,
    eval_set=[(X_valid, y_valid)],
    eval_metric="binary_logloss",
    callbacks=[
        lgbm.early_stopping(stopping_rounds=10, verbose=True),
        lgbm.log_evaluation(200),
    ],
)

pred_proba = model_clf.predict_proba(X_valid)
pred = model_clf.predict(X_valid)
```

(XGBoost)

```
import xgboost as xgb

# reliable_data_count : Number of reliable data
# unreliable_data_count : Number of unreliable data

xgb_params_clf = {
    "objective": "binary:logistic",
    "learning_rate": 0.05,
    "alpha": 0.1,
    "lambda": 0.1,
    "random_state": 42,
    "max_depth": 5,
    "n_estimators": 1000000,
    "colsample_bytree": 0.7,
    "scale_pos_weight": unreliable_data_count / reliable_data_count,
}

model_clf = xgb.XGBClassifier(**xgb_params_clf)

model_clf.fit(
    X_train,
    y_train,
    eval_set=[(X_valid, y_valid)],
    early_stopping_rounds=10,
    verbose=200,
)

pred_proba = model_clf.predict_proba(X_valid)
pred = model_clf.predict(X_valid)
```

(CatBoost)

https://doi.org/10.1021/acsomega.4c06997
ACS Omega 2025, 10, 14626−14639

```
from catboost import CatBoostClassifier

catboost_params_clf = {
    "learning_rate": 0.05,
    "l2_leaf_reg ": 0.1,
    "random_state": 42,
    "depth": 5,
    "iterations": 1000000,
    " early_stopping_rounds": 10,
    "auto_class_weights": "Balanced",
}

model_clf = CatBoostClassifier(**catboost_params_clf)

model_clf.fit(
    X_train, y_train, eval_set=(X_valid, y_valid), use_best_model=True, verbose=200
)

pred_proba = model_clf.predict_proba(X_valid)
pred = model_clf.predict(X_valid)
```

(Random Forest)

```
from sklearn.ensemble import RandomForestClassifier

rf_params_clf = {
    "random_state": 42,
    "max_depth": 5,
    "n_estimators": 1000,
    "max_features": 0.7,
    "class_weight": "balanced",
}

model_clf = RandomForestClassifier(**rf_params_clf)

model_clf.fit(X_train, y_train)

pred_proba = model_clf.predict_proba(X_valid)
pred = model_clf.predict(X_valid)
```

(Gaussian Process)

```
from sklearn.gaussian_process import GaussianProcessClassifier
from sklearn.gaussian_process.kernels import RBF, ConstantKernel, WhiteKernel
from sklearn.preprocessing import StandardScaler

scaler_X = StandardScaler()
X_train = scaler_X.fit_transform(X_train)
X_valid = scaler_X.transform(X_valid)

model_clf = GaussianProcessClassifier(
    kernel=ConstantKernel() * RBF() + WhiteKernel(), random_state=42
)

model_clf.fit(X_train, y_train)

pred_proba = model_clf.predict_proba(X_valid)
pred = model_clf.predict(X_valid)
```

(Support Vector Machine)

```
from sklearn.preprocessing import StandardScaler
from sklearn.svm import SVC

scaler_X = StandardScaler()
X_train = scaler_X.fit_transform(X_train)
X_valid = scaler_X.transform(X_valid)

model_clf = SVC(probability=True, random_state=42)

model_clf.fit(X_train, y_train)

pred_proba = model_clf.predict_proba(X_valid)
pred = model_clf.predict(X_valid)
```

The parameters of the UMAP

```
import umap

reducer = umap.UMAP(n_components=2, random_state=42, n_neighbors=15)
```

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acsomega.4c06997.

The Materials Project dataset (ZIP)

■ AUTHOR INFORMATION

Corresponding Authors

Yosuke Harashima − *Division of Materials Science, Nara Institute of Science and Technology, Ikoma-shi, Nara-ken 630-0192, Japan; Data Science Center, Nara Institute of Science and Technology, Ikoma-shi, Nara-ken 630-0192, Japan*; Email: harashima.yosuke@ms.naist.jp

Akihiko Kudo − *Department of Applied Chemistry, Faculty of Science, Tokyo University of Science, Shinjuku-ku, Tokyo 162-8601, Japan; Carbon Value Research Center, Research Institute for Science & Technology, Tokyo University of Science, Noda-shi, Chiba-ken 278-8510, Japan*; orcid.org/0000-0002-5665-5482; Email: a-kudo@rs.tus.ac.jp

Mikiya Fujii − *Division of Materials Science, Nara Institute of Science and Technology, Ikoma-shi, Nara-ken 630-0192, Japan; Data Science Center and Center for Material Research Platform, Nara Institute of Science and Technology, Ikoma-shi, Nara-ken 630-0192, Japan*; orcid.org/0000-0002-3728-3097; Email: fujii.mikiya@ms.naist.jp

Authors

Wataru Takahara − *Division of Materials Science, Nara Institute of Science and Technology, Ikoma-shi, Nara-ken 630-0192, Japan*; orcid.org/0009-0007-3442-8338

Ryuto Baba − *Division of Materials Science, Nara Institute of Science and Technology, Ikoma-shi, Nara-ken 630-0192, Japan*

Tomoaki Takayama − *Division of Materials Science, Nara Institute of Science and Technology, Ikoma-shi, Nara-ken 630-0192, Japan; Data Science Center, Nara Institute of Science and Technology, Ikoma-shi, Nara-ken 630-0192, Japan*; orcid.org/0000-0002-4736-7454

Shogo Takasuka − *Division of Materials Science, Nara Institute of Science and Technology, Ikoma-shi, Nara-ken 630-0192, Japan*

Yuichi Yamaguchi − *Department of Applied Chemistry, Faculty of Science, Tokyo University of Science, Shinjuku-ku, Tokyo 162-8601, Japan; Carbon Value Research Center, Research Institute for Science & Technology, Tokyo University of Science, Noda-shi, Chiba-ken 278-8510, Japan*

Complete contact information is available at: https://pubs.acs.org/10.1021/acsomega.4c06997

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This work was supported by the grant JSPS KAKENHI (Grant Numbers 22K03449 and 23H00248) and by MEXT as "Program for Promoting Research on the Supercomputer Fugaku" (realization of innovative light energy conversion materials utilizing the supercomputer Fugaku, Grant Number JPMXP1020210317). The computation was partly conducted using the facilities of the Supercomputer Center at the Institute for Solid State Physics at the University of Tokyo.

■ REFERENCES

(1) Agrawal, A.; Choudhary, A. Perspective: Materials Informatics and Big Data: Realization of the "Fourth Paradigm" of Science in Materials Science. *APL Mater.* **2016**, *4* (5), No. 053208.

(2) Liu, Y.; Zhao, T.; Ju, W.; Shi, S. Materials Discovery and Design Using Machine Learning. *J. Materiomics* **2017**, *3* (3), 159−177.

(3) Ramprasad, R.; Batra, R.; Pilania, G.; Mannodi-Kanakkithodi, A.; Kim, C. Machine Learning in Materials Informatics: Recent Applications and Prospects. *npj Comput. Mater.* **2017**, *3* (1), No. 54.
(4) Lee, S.; Byun, H.; Cheon, M.; Kim, J.; Lee, J. H. Machine Learning-Based Discovery of Molecules, Crystals, and Composites: A Perspective Review. *Korean J. Chem. Eng.* **2021**, *38* (10), 1971−1982.
(5) Choudhary, K.; DeCost, B.; Chen, C.; Jain, A.; Tavazza, F.; Cohn, R.; Park, C. W.; Choudhary, A.; Agrawal, A.; Billinge, S. J. L.; Holm, E.; Ong, S. P.; Wolverton, C. Recent Advances and Applications of Deep Learning Methods in Materials Science. *npj Comput. Mater.* **2022**, *8* (1), No. 59.
(6) Ong, S. P.; Richards, W. D.; Jain, A.; Hautier, G.; Kocher, M.; Cholia, S.; Gunter, D.; Chevrier, V. L.; Persson, K. A.; Ceder, G. Python Materials Genomics (Pymatgen): A Robust, Open-Source Python Library for Materials Analysis. *Comput. Mater. Sci.* **2013**, *68*, 314−319.
(7) materialsproject/pymatgen. *Python Materials Genomics (pymatgen) is a Robust Materials Analysis Code That Defines Classes for Structures and Molecules with Support for Many Electronic Structure Codes. It Powers the Materials Project*, https://github.com/materialsproject/pymatgen (accessed May 14, 2024).
(8) Ward, L.; Dunn, A.; Faghaninia, A.; Zimmermann, N. E. R.; Bajaj, S.; Wang, Q.; Montoya, J.; Chen, J.; Bystrom, K.; Dylla, M.; Chard, K.; Asta, M.; Persson, K. A.; Snyder, G. J.; Foster, I.; Jain, A. Matminer: An Open Source Toolkit for Materials Data Mining. *Comput. Mater. Sci.* **2018**, *152*, 60−69.
(9) hackingmaterials/matminer. *Data Mining for Materials Science*, https://github.com/hackingmaterials/matminer (accessed May 14, 2024).
(10) yoshida-lab/XenonPy. *XenonPy is a Python Software for Materials Informatics*, https://github.com/yoshida-lab/XenonPy (accessed May 14, 2024).
(11) rdkit/rdkit. *Official Sources for the RDKit Library*, https://github.com/rdkit/rdkit (accessed May 14, 2024).
(12) Moriwaki, H.; Tian, Y. S.; Kawashita, N.; Takagi, T. Mordred: A Molecular Descriptor Calculator. *J. Cheminform.* **2018**, *10* (1), No. 4.
(13) mordred-descriptor/mordred. *A Molecular Descriptor Calculator*, https://github.com/mordred-descriptor/mordred (accessed May 14, 2024).
(14) Bolton, E. E.; Wang, Y.; Thiessen, P. A.; Bryant, S. H. PubChem: Integrated Platform of Small Molecules and Biological Activities. *Annu. Rep. Comput. Chem.* **2008**, *4*, 217−241.
(15) Kim, S.; Chen, J.; Cheng, T.; Gindulyte, A.; He, J.; He, S.; Li, Q.; Shoemaker, B. A.; Thiessen, P. A.; Yu, B.; Zaslavsky, L.; Zhang, J.; Bolton, E. E. PubChem 2023 Update. *Nucleic Acids Res.* **2023**, *51* (D1), D1373−D1380.
(16) PubChem https://pubchem.ncbi.nlm.nih.gov/ (accessed May 30, 2024).
(17) Jain, A.; Ong, S. P.; Hautier, G.; Chen, W.; Richards, W. D.; Dacek, S.; Cholia, S.; Gunter, D.; Skinner, D.; Ceder, G.; Persson, K. A. Commentary: The Materials Project: A Materials Genome Approach to Accelerating Materials Innovation. *APL Mater.* **2013**, *1* (1), No. 011002.
(18) Materials Project - Home https://next-gen.materialsproject.org/ (accessed May 30, 2024).
(19) Ramakrishnan, R.; Dral, P. O.; Rupp, M.; von Lilienfeld, O. A. Quantum Chemistry Structures and Properties of 134 Kilo Molecules. *Sci. Data* **2014**, *1*, No. 140022.
(20) Quantum-Machine.org: Home http://www.quantum-machine.org/ (accessed May 30, 2024).
(21) Irwin, J. J.; Shoichet, B. K. ZINC-A Free Database of Commercially Available Compounds for Virtual Screening. *J. Chem. Inf. Model.* **2005**, *45* (1), 177−182.
(22) Irwin, J. J.; Sterling, T.; Mysinger, M. M.; Bolstad, E. S.; Coleman, R. G. ZINC: A Free Tool to Discover Chemistry for Biology. *J. Chem. Inf. Model.* **2012**, *52* (7), 1757−1768.
(23) Sterling, T.; Irwin, J. J. ZINC 15 - Ligand Discovery for Everyone. *J. Chem. Inf. Model.* **2015**, *55* (11), 2324−2337.
(24) Irwin, J. J.; Tang, K. G.; Young, J.; Dandarchuluun, C.; Wong, B. R.; Khurelbaatar, M.; Moroz, Y. S.; Mayfield, J.; Sayle, R. A. ZINC20 - A Free Ultralarge-Scale Chemical Database for Ligand Discovery. *J. Chem. Inf. Model.* **2020**, *60* (12), 6065−6073.
(25) ZINC https://zinc.docking.org/ (accessed May 30, 2024).
(26) Takahara, W.; Kobayashi, Y.; Morita, M.; Okuyama, K.; Kawamura, N. Building Machine Learning Models on Thermosetting Resin Composite Materials toward the Prediction of Physical Characteristics. *J. Comput. Chem. Jpn.* **2021**, *20* (1), 14−21.
(27) Okuyama, M.; Nakazawa, Y.; Funatsu, K. A Data-Driven Scheme to Search for Alternative Composite Materials. *Sci. Technol. Adv. Mater. Methods* **2022**, *2* (1), 106−118.
(28) Kudo, A.; Miseki, Y. Heterogeneous Photocatalyst Materials for Water Splitting. *Chem. Soc. Rev.* **2009**, *38* (1), 253−278.
(29) Nishiyama, H.; Yamada, T.; Nakabayashi, M.; Maehara, Y.; Yamaguchi, M.; Kuromiya, Y.; Nagatsuma, Y.; Tokudome, H.; Akiyama, S.; Watanabe, T.; Narushima, R.; Okunaka, S.; Shibata, N.; Takata, T.; Hisatomi, T.; Domen, K. Photocatalytic Solar Hydrogen Production from Water on a 100-m^2 Scale. *Nature* **2021**, *598* (7880), 304−307.
(30) Nandy, S.; Hisatomi, T.; Takata, T.; Setoyama, T.; Domen, K. Recent Advances in Photocatalyst Sheet Development and Challenges for Cost-Effective Solar Hydrogen Production. *J. Mater. Chem. A* **2023**, *11* (38), 20470−20479.
(31) Zhang, K.; Guo, L. Metal Sulphide Semiconductors for Photocatalytic Hydrogen Production. *Catal. Sci. Technol.* **2013**, *3* (7), 1672−1690.
(32) Takayama, T.; Tsuji, I.; Aono, N.; Harada, M.; Okuda, T.; Iwase, A.; Kato, H.; Kudo, A. Development of Various Metal Sulfide Photocatalysts Consisting of d^0, d^5, and d^{10} Metal Ions for Sacrificial H_2 Evolution under Visible Light Irradiation. *Chem. Lett.* **2017**, *46* (4), 616−619.
(33) Tsuji, I.; Kato, H.; Kudo, A. Visible-Light-Induced H_2 Evolution from an Aqueous Solution Containing Sulfide and Sulfite over a ZnS-$CuInS_2$-$AgInS_2$ Solid-Solution Photocatalyst. *Angew. Chem., Int. Ed.* **2005**, *44* (23), 3565−3568.
(34) Tsuji, I.; Shimodaira, Y.; Kato, H.; Kobayashi, H.; Kudo, A. Novel Stannite-Type Complex Sulfide Photocatalysts A^I_2-Zn-A^{IV}-S_4 (A^I = Cu and Ag; A^{IV} = Sn and Ge) for Hydrogen Evolution under Visible-Light Irradiation. *Chem. Mater.* **2010**, *22* (4), 1402−1409.
(35) Kato, T.; Hakari, Y.; Ikeda, S.; Jia, Q.; Iwase, A.; Kudo, A. Utilization of Metal Sulfide Material of $(CuGa)_{1-x}Zn_{2\ x}S_2$ Solid Solution with Visible Light Response in Photocatalytic and Photoelectrochemical Solar Water Splitting Systems. *J. Phys. Chem. Lett.* **2015**, *6* (6), 1042−1047.
(36) Ikeda, S.; Aono, N.; Iwase, A.; Kobayashi, H.; Kudo, A. Cu_3MS_4 (M = V, Nb, Ta) and Its Solid Solutions with Sulvanite Structure for Photocatalytic and Photoelectrochemical H_2 Evolution under Visible-Light Irradiation. *ChemSusChem* **2019**, *12* (9), 1977−1983.
(37) Osterloh, F. E. Inorganic Materials as Catalysts for Photochemical Splitting of Water. *Chem. Mater.* **2008**, *20* (1), 35−54.
(38) Wang, Q.; Domen, K. Particulate Photocatalysts for Light-Driven Water Splitting: Mechanisms, Challenges, and Design Strategies. *Chem. Rev.* **2020**, *120* (2), 919−985.
(39) Guha, R.; Jurs, P. C. Determining the Validity of a QSAR Model - A Classification Approach. *J. Chem. Inf. Model.* **2005**, *45* (1), 65−73.
(40) Sheridan, R. P. Using Random Forest to Model the Domain Applicability of Another Random Forest Model. *J. Chem. Inf. Model.* **2013**, *53* (11), 2837−2850.
(41) Sakaushi, K.; Hoisang, W.; Tamura, R. Human-Machine Collaboration for Accelerated Discovery of Promising Oxygen Evolution Electrocatalysts with On-Demand Elements. *ACS Cent. Sci.* **2023**, *9* (12), 2216−2224.
(42) Talapatra, A.; Uberuaga, B. P.; Stanek, C. R.; Pilania, G. Band Gap Predictions of Double Perovskite Oxides Using Machine Learning. *Commun. Mater.* **2023**, *4* (1), No. 46.
(43) Balachandran, P. V.; Emery, A. A.; Gubernatis, J. E.; Lookman, T.; Wolverton, C.; Zunger, A. Predictions of New ABO_3 Perovskite

https://doi.org/10.1021/acsomega.4c06997
ACS Omega 2025, 10, 14626−14639

Compounds by Combining Machine Learning and Density Functional Theory. *Phys. Rev. Mater.* **2018**, *2* (4), No. 043802.
(44) Liu, C.; Fujita, E.; Katsura, Y.; Inada, Y.; Ishikawa, A.; Tamura, R.; Kimura, K.; Yoshida, R. Machine Learning to Predict Quasicrystals from Chemical Compositions. *Adv. Mater.* **2021**, *33* (36), No. 2102507.
(45) Kudo, A.; Sekizawa, M. Photocatalytic H_2 Evolution under Visible Light Irradiation on $Zn_{1-x}Cu_xS$ Solid Solution. *Catal. Lett.* **1999**, *58*, 241−243.
(46) Kudo, A.; Sekizawa, M. Photocatalytic H_2 Evolution under Visible Light Irradiation on Ni-Doped ZnS Photocatalyst. *Chem. Commun.* **2000**, No. 15, 1371−1372.
(47) Tsuji, I.; Kudo, A. H_2 Evolution from Aqueous Sulfite under Visible-Light Irradiation over Pb and Halogen-Codoped ZnS Photocatalysts. *J. Photochem. Photobiol., A* **2003**, *156* (1−3), 249−252.
(48) Kudo, A.; Nagane, A.; Tsuji, I.; Kato, H. H_2 Evolution from Aqueous Potassium Sulfite Solutions under Visible Light Irradiation over a Novel Sulfide Photocatalyst $NaInS_2$ with a Layered Structure. *Chem. Lett.* **2002**, *31* (9), 882−883.
(49) Hayashi, T.; Niishiro, R.; Ishihara, H.; Yamaguchi, M.; Jia, Q.; Kuang, Y.; Higashi, T.; Iwase, A.; Minegishi, T.; Yamada, T.; Domen, K.; Kudo, A. Powder-Based $(CuGa_{1-y}In_y)_{1-x}Zn_{2x}S_2$ Solid Solution Photocathodes with a Largely Positive Onset Potential for Solar Water Splitting. *Sustainable Energy Fuels* **2018**, *2* (9), 2016−2024.
(50) Tsuji, I.; Kato, H.; Kobayashi, H.; Kudo, A. Photocatalytic H_2 Evolution under Visible-Light Irradiation over Band-Structure-Controlled $(CuIn)_xZn_{2(1-x)}S_2$ Solid Solutions. *J. Phys. Chem. B* **2005**, *109* (15), 7323−7329.
(51) Tsuji, I.; Kato, H.; Kobayashi, H.; Kudo, A. Photocatalytic H_2 Evolution Reaction from Aqueous Solutions over Band Structure-Controlled $(AgIn)_xZn_{2(1-x)}S_2$ Solid Solution Photocatalysts with Visible-Light Response and Their Surface Nanostructures. *J. Am. Chem. Soc.* **2004**, *126* (41), 13406−13413.
(52) Kaga, H.; Kudo, A. Cosubstituting Effects of Copper(I) and Gallium(III) for $ZnGa_2S_4$ with Defect Chalcopyrite Structure on Photocatalytic Activity for Hydrogen Evolution. *J. Catal.* **2014**, *310*, 31−36.
(53) Yamato, K.; Iwase, A.; Kudo, A. Photocatalysis Using a Wide Range of the Visible Light Spectrum: Hydrogen Evolution from Doped $AgGaS_2$. *ChemSusChem* **2015**, *8* (17), 2902−2906.
(54) McInnes, L.; Healy, J.; Saul, N.; Großberger, L. UMAP: Uniform Manifold Approximation and Projection. *J. Open Source Software* **2018**, *3* (29), No. 861.
(55) lmcinnes/umap. *Uniform Manifold Approximation and Projection*, https://github.com/lmcinnes/umap (accessed May 15, 2024).
(56) Ke, G.; Meng, Q.; Finley, T.; Wang, T.; Chen, W.; Ma, W.; Ye, Q.; Liu, T.-Y. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. In *Adv. Neural Inf. Process. Syst.*; Guyon, I.; von Luxburg, U.; Bengio, S.; Wallach, H.; Fergus, R.; Vishwanathan, S.; Garnett, R., Eds.; Curran Associates, Inc., 2017; Vol. *30*.
(57) microsoft/LightGBM. *A fast, distributed, high performance gradient boosting (GBT, GBDT, GBRT, GBM or MART) framework based on decision tree algorithms, used for ranking, classification and many other machine learning tasks*, https://github.com/microsoft/LightGBM (accessed May 15, 2024).
(58) Chen, T.; Guestrin, C. In *XGBoost: A Scalable Tree Boosting System*, Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 13-17-August, 2016; pp 785−794 .
(59) dmlc/xgboost. *Scalable, Portable and Distributed Gradient Boosting (GBDT, GBRT or GBM) Library, for Python, R, Java, Scala, C++ and more. Runs on single machine, Hadoop, Spark, Dask, Flink and DataFlow*, https://github.com/dmlc/xgboost (accessed Dec 21, 2024).
(60) Prokhorenkova, L.; Gusev, G.; Vorobev, A.; Dorogush, A. V.; Gulin, A. CatBoost: Unbiased Boosting with Categorical Features. In *Advances in Neural Information Processing Systems* 2018; Vol. *31*.
(61) catboost/catboost. *A fast, scalable, high performance Gradient Boosting on Decision Trees library, used for ranking, classification, regression and other machine learning tasks for Python, R, Java, C++. Supports computation on CPU and GPU*, https://github.com/catboost/catboost (accessed Dec 21, 2024).
(62) Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45* (1), 5−32.
(63) Rasmussen, C. E.; Williams, C. K. I. *Gaussian Processes for Machine Learning*; MIT Press: Cambridge, MA, 2006.
(64) Cortes, C.; Vapnik, V. Support-Vector Networks. *Mach. Learn.* **1995**, *20* (3), 273−297.
(65) scikit-learn/scikit-learn. *scikit-learn: machine learning in Python*, https://github.com/scikit-learn/scikit-learn (accessed May 15, 2024).
(66) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; Duchesnay, E. Scikit-Learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12* (85), 2825−2830.
(67) Sturges, H. A. The Choice of a Class Interval. *J. Am. Stat. Assoc.* **1926**, *21* (153), 65−66.

https://doi.org/10.1021/acsomega.4c06997
ACS Omega 2025, 10, 14626−14639