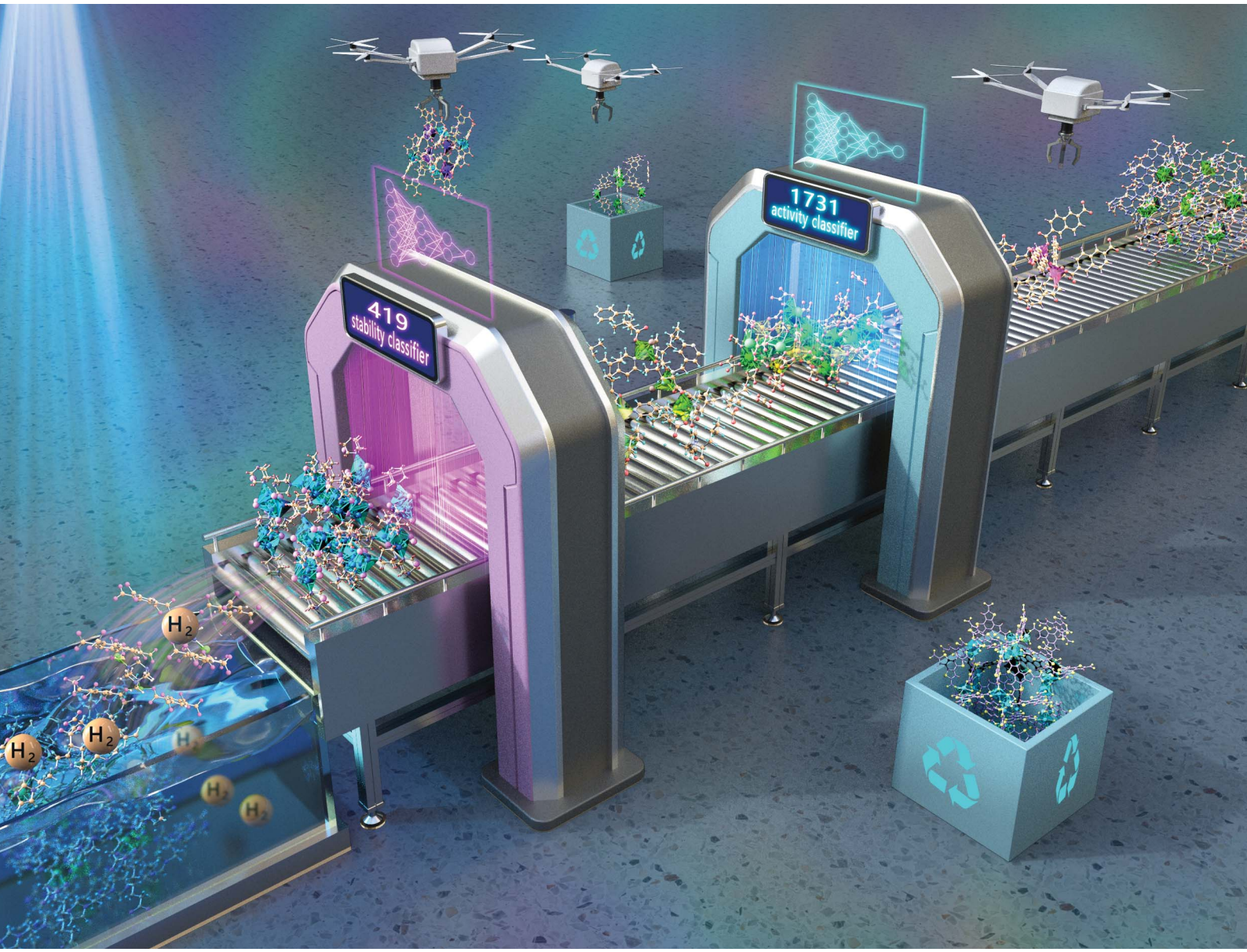
Volume 17
Number 11
18 March 2026
Pages 5283–5762

Chemical Science

rsc.li/chemical-science



ISSN 2041-6539

ROYAL SOCIETY OF CHEMISTRY

EDGE ARTICLE

Xiao Niu *et al.*

Machine learning guided discovery of water stable metal–organic frameworks for photocatalytic hydrogen production

Chemical Science

ROYAL SOCIETY OF CHEMISTRY

EDGE ARTICLE

View Article Online
View Journal | View Issue

Cite this: *Chem. Sci.*, 2026, **17**, 5376

All publication charges for this article have been paid for by the Royal Society of Chemistry

Machine learning guided discovery of water stable metal–organic frameworks for photocatalytic hydrogen production

Xiao Niu,[ab] Zhiming Zhang,[b] Xiaoyu Wu,[b] Yan Liu,[a] Yong Cui *[a] and Jianwen Jiang *[b]

Received 26th October 2025
Accepted 8th February 2026

DOI: 10.1039/d5sc08277c

rsc.li/chemical-science

With remarkably tunable porosity and modular chemistry, metal–organic frameworks (MOFs) present a versatile platform for photocatalytic hydrogen (H_2) production. However, identifying high-performing and water stable MOFs from the vast design space is challenging. In this study, we develop a hierarchical screening strategy to accelerate the discovery of photocatalytically active MOFs with robust water stability. First, machine learning (ML) classifiers are trained on experimental H_2 production data to predict photocatalytic performance, achieving high accuracy and excellent transferability. Then, starting from 11660 structures in the CoRE-MOF database, 1731 are shortlisted to be photocatalytically active. Detailed structure–performance analyses reveal that linker flexibility and aliphatic character positively correlate with H_2 evolution activity, while excessive aromaticity and rigidity are detrimental. Finally, a water stability classifier is applied to further identify 419 MOFs to be simultaneously photocatalytically promising and water stable. The ML-guided strategy provides a quantitative and interpretable path toward the discovery of new MOFs as photocatalysts, and it would facilitate future experimental exploration for efficient photocatalytic H_2 production.

1. Introduction

The rapid industrialization and population growth have intensified the consumption of conventional fossil fuels, causing severe global warming and environmental deterioration.[1,2] Hydrogen (H_2) is considered a clean fuel towards carbon neutrality.[3] Among various production routes for H_2, photocatalytic water splitting offers a sustainable pathway by utilizing readily available sunlight and water.[4] However, most existing photocatalysts fail to meet practical requirements due to insufficient activity, poor visible-light response, and, especially, limited water stability.[5] There is an urgent need to develop high-performance and water-stable photocatalysts for water splitting.

With a remarkable tunability of building blocks and modular architectures, metal–organic frameworks (MOFs) have attracted significant attention as photocatalysts for hydrogen evolution reaction (HER).[6,7] By independently modulating metal nodes and organic linkers, MOFs allow for precise tuning of electronic band gap, redox potential, and charge transport pathway, which are essential for achieving suitable band alignment and efficient carrier dynamics in photocatalytic HER.[8] A handful of MOFs such as UiO-66(Zr),[9] MIL-125(Ti)[10] and MIL-53(Al)[11] have been investigated for HER. However, their photocatalytic performance is significantly below the requirements for practical applications. This is primarily due to their low solar-to-hydrogen (STH) conversion efficiency,[12,13] strong dependence on sacrificial agent, and severe electron–hole recombination. It is highly desired to develop new MOF-based photocatalysts that are highly efficient, water stable and readily responsive to visible light, thereby advancing solar-driven H_2 production. At present, over 120 000 MOFs have been experimentally synthesized and thus it is not feasible to identify promising MOFs from such a large chemical space by conventional trial-and-error methods.

In the past several years, machine learning (ML) has emerged as a transformative data-driven tool, playing an increasingly important role in structure prediction, property evaluation, and materials discovery.[14,15] By learning complex structure–property relationships from a data set, ML can accelerate the screening and design of new materials. In the field of MOFs, ML has been successfully applied to predict gas adsorption and diffusion, mechanical strength, and water stability.[16–18] There were also a few ML studies for photocatalytic water splitting. By combining density functional theory (DFT) calculations and ML, Wang *et al.* screened over 20 375 MOFs in the Quantum-MOF (QMOF) database and identified 14 MOFs with superior overall water splitting potential.[19] Similarly, Mourino *et al.* also applied DFT and ML to evaluate 314 MOFs,

[a]*State Key Laboratory of Synergistic Chem-Bio Synthesis, School of Chemistry and Chemical Engineering, Shanghai Jiao Tong University, Shanghai 200240, China. E-mail: yongcui@sjtu.edu.cn*

[b]*Department of Chemical and Biomolecular Engineering, National University of Singapore, 117576, Singapore. E-mail: chejj@nus.edu.sg*

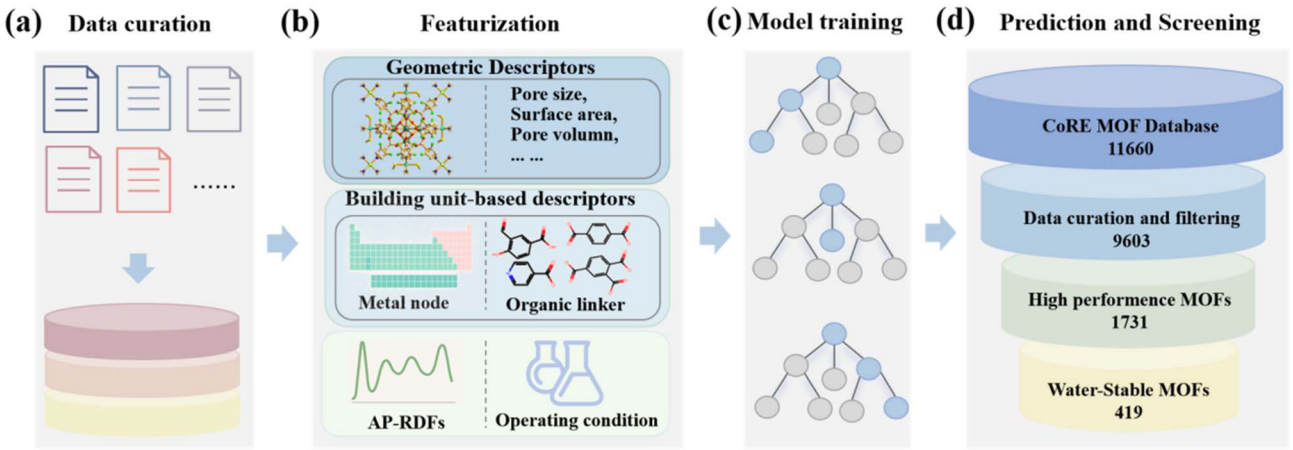


Fig. 1 Workflow to predict photocatalytic performance of MOFs. (a) Data curation, (b) featurization, (c) model training, and (d) prediction and screening.

highlighted the role of band gap, band alignment and charge separation in photocatalytic activity, and identified promising structural motifs such as Ti clusters and rod-shaped metal nodes.[20] Despite demonstrating the robustness of ML, these studies utilized the data sets from DFT calculations, primarily emphasized electronic descriptors, and did not take the stability of MOFs into account.

To bridge the gap, in this study, we curated an experimental data set with MOFs as photocatalysts for water splitting and featurized MOFs with multi-level descriptors, then developed ML models for photocatalytic performance prediction, and, finally, water stability determination was incorporated. This approach enables high-throughput screening of MOFs with both high photocatalytic performance and practical water stability, thus accelerating the discovery of MOFs for H_2 production. Fig. 1 illustrates the overall workflow. Specifically, a data set with 92 MOFs was curated with experimentally measured H_2 evolution rates. The MOFs were featurized using a combination of geometric descriptors, building units (metals and linkers), and atomic property-weighted radial distribution functions (AP-RDFs). Different ML classifiers were trained and interpreted through feature importance analysis. To evaluate transferability, the classifiers were tested on 12 recent data points of 10 MOFs, including both unseen structures and new operating conditions. Next, the best classifier was applied to predict the photocatalytic performance of 9603 MOFs from the CoRE MOF database, identifying a set of top-performing MOFs. Retrospective comparison with recent experimental data confirmed the predictive accuracy. Finally, water stability of top-performing MOFs was evaluated *via* a recently developed ML model, resulting in the identification of MOFs with both high potential in HER and robust water tolerance.

2. Results and discussion

2.1. Structure–performance relationships

When acting as photocatalysts for HER, MOFs can facilitate charge separation and transfer, which is influenced by the metal–ligand coordination, pore environment and electronic properties of MOFs.[21] The photocatalytic efficiency of MOFs is closely tied to their ability to absorb visible light, generate photoexcited charge carriers, and catalyze redox reactions at active sites. In particular, the complex interplay among building units such as catalytic metal nodes and functionalized linkers is crucial in determining photocatalytic activity. In our study, 92 experimentally examined MOFs were compiled and classified based on their HER performance. Thus, it is instructive to explore quantitative structure–performance relationships.

Fig. S1a shows the correlation matrix of Pearson, Spearman and Kendall coefficients between geometric descriptors and H_2 production rate. The geometric descriptors including the largest cavity diameter (LCD), pore limiting diameter (PLD), largest free path diameter (LFPD), density, volumetric surface area (VSA), gravimetric surface area (GSA), void fraction (VF) and pore volume (PV) were estimated by using Zeo++.[22] Among these, VSA and GSA exhibit the most notable positive correlations with H_2 production rate. This is further confirmed by the boxplot comparison between low- and high-performing MOFs in Fig. S1b, where high-performing MOFs generally possess larger surface areas. These findings suggest that surface accessibility and availability of active sites play a more significant role than pore diameter or volume alone in governing photocatalytic activity.[8] A recent study highlighted that a material with a larger surface area and a higher porosity would tend to exhibit superior catalytic performance,[23] which is consistent with our analysis based on the geometric descriptors.

It is also essential to evaluate structure–performance relationships at the level of building units. Fig. 2a illustrates the probability histogram of metal types in 92 collected MOFs. There is a significant imbalance, especially with Ti, Co and Zr dominating the data set. We observe considerable variation in the H_2 production rate *versus* metal type. As shown in Fig. 2b, lanthanide-based MOFs (*e.g.*, those containing Ho and Yb) exhibit a remarkably high median H_2 production rate. This is attributed to the unique 4f electronic structures and large ionic radii of lanthanide metals, thus facilitating favorable photo-induced electron transfer. Although widely used in MOF synthesis, common transition metals such as Zn, Co and Zr are

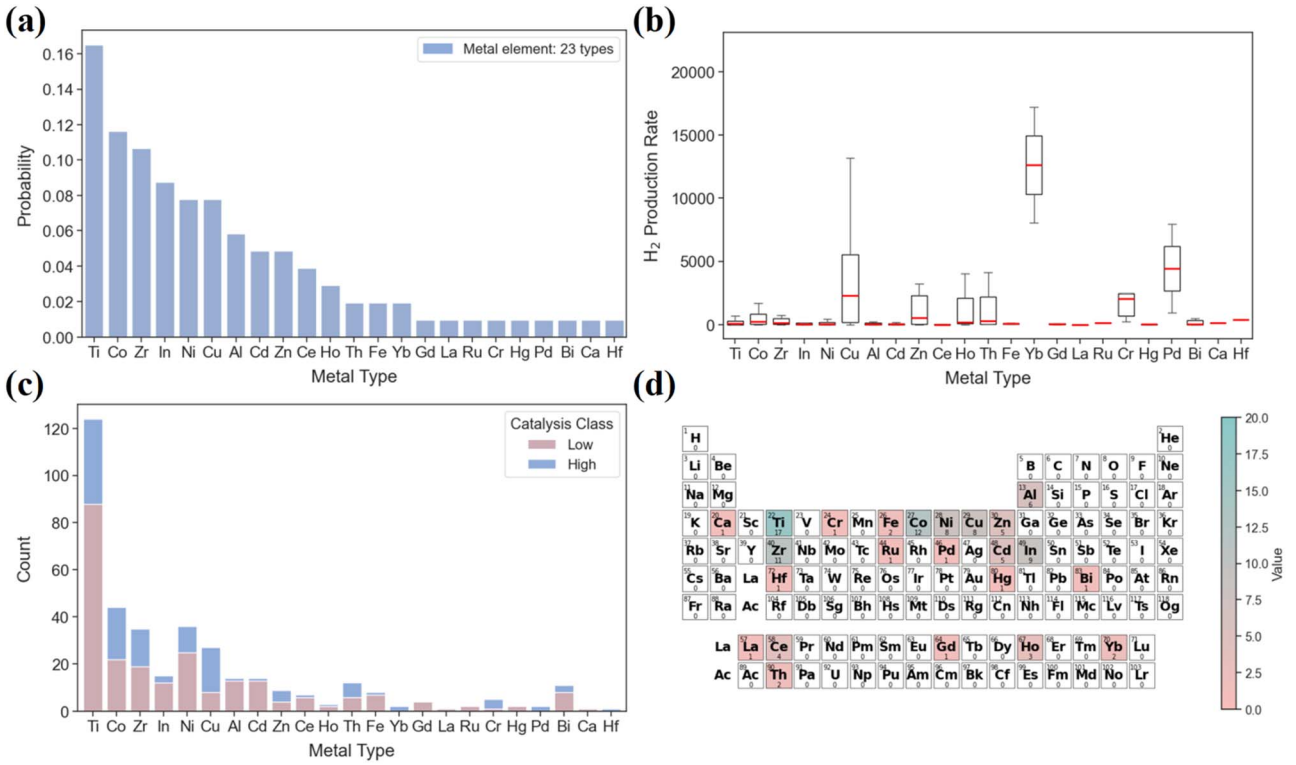


Fig. 2 (a) Probability histogram of metal types in 92 collected MOFs. (b) H_2 production rate *versus* metal type. (c) Count of metal types in low- and high-performing MOFs. (d) Periodic table heatmap for the diversity of metal types.

predominantly associated with low H_2 production rates. Fig. 2c shows the count of metal types in high- and low-performing MOFs. Metals like Cu and Yb appear more frequently in high-performing MOFs, whereas Zr and Ti, despite their prevalence, primarily exist in low-performing MOFs. To further visualize the diversity of metal types, a periodic table heatmap is plotted in Fig. 2d, highlighting the localized clustering of catalytically active metals in the d-block and f-block regions. These findings underscore the importance of metal selection in the design of photocatalytic MOFs, suggesting that rare-earth and certain transition metals may offer enhanced activity due to their intrinsic electronic and coordination properties. However, the imbalance in metal types suggests a bias in the reported MOFs, with many underexplored metal candidates possibly offering untapped catalytic potential. Future research should therefore target less-explored metals, particularly those observed in high-performing outliers, to diversify the chemical space and improve the generalizability of predictive ML models.

To elucidate the underlying physicochemical factors, we further explored the correlation between key metal properties and photocatalytic performance. As shown in Fig. S2, high-performing MOFs generally contain metals with higher electronegativity and larger ionization energy, implying stronger electron-withdrawing ability and greater oxidation stability in promoting efficient charge separation. Metals with smaller atomic radii also tend to enhance activity due to a more compact coordination environment, whereas molecular weight shows a negligible correlation, confirming that electronic rather than mass-related factors dominate the photocatalytic behavior of MOFs.

Subsequently, we analyzed the relationship between organic linkers and photocatalytic performance. Representative linkers commonly employed in MOF construction were selected, mostly containing carboxylates as coordination sites, along with polar functional groups such as hydroxyl, amine and heterocycles (Fig. S3). These functional groups can enhance interactions with water through hydrogen bonding and facilitate visible-light absorption by modulating the local electronic environment.[24,25] We calculated six RDKit-based molecular descriptors (RDKitDP), including molecular weight, partition coefficient ($\log P$), topological polar surface area (TPSA), aromatic ring count, and hydrogen bond (H-bond) donor and acceptor counts, for organic linkers. As evidenced in Fig. S2, high-performing MOFs exhibit larger TPSA and a greater H-bond acceptor count, both of which are positively correlated with Pearson, Spearman and Kendall coefficients. These properties of organic linkers improve surface hydrophilicity and strengthen hydrogen bonding with water, which in turn facilitates proton-coupled electron transfer during HER. Meanwhile, the H-bond donor count also exhibits a moderate positive trend, highlighting the importance of hydrogen bonding in forming a catalytically active microenvironment. Conversely, aromatic ring count and $\log P$ display negative correlations with performance, as excessive aromaticity increases linker rigidity and promotes π–π stacking, thus hindering charge mobility and restricting reactant accessibility. Similarly, more hydrophobic linkers (*i.e.*, higher $\log P$) reduce surface wettability, impeding water

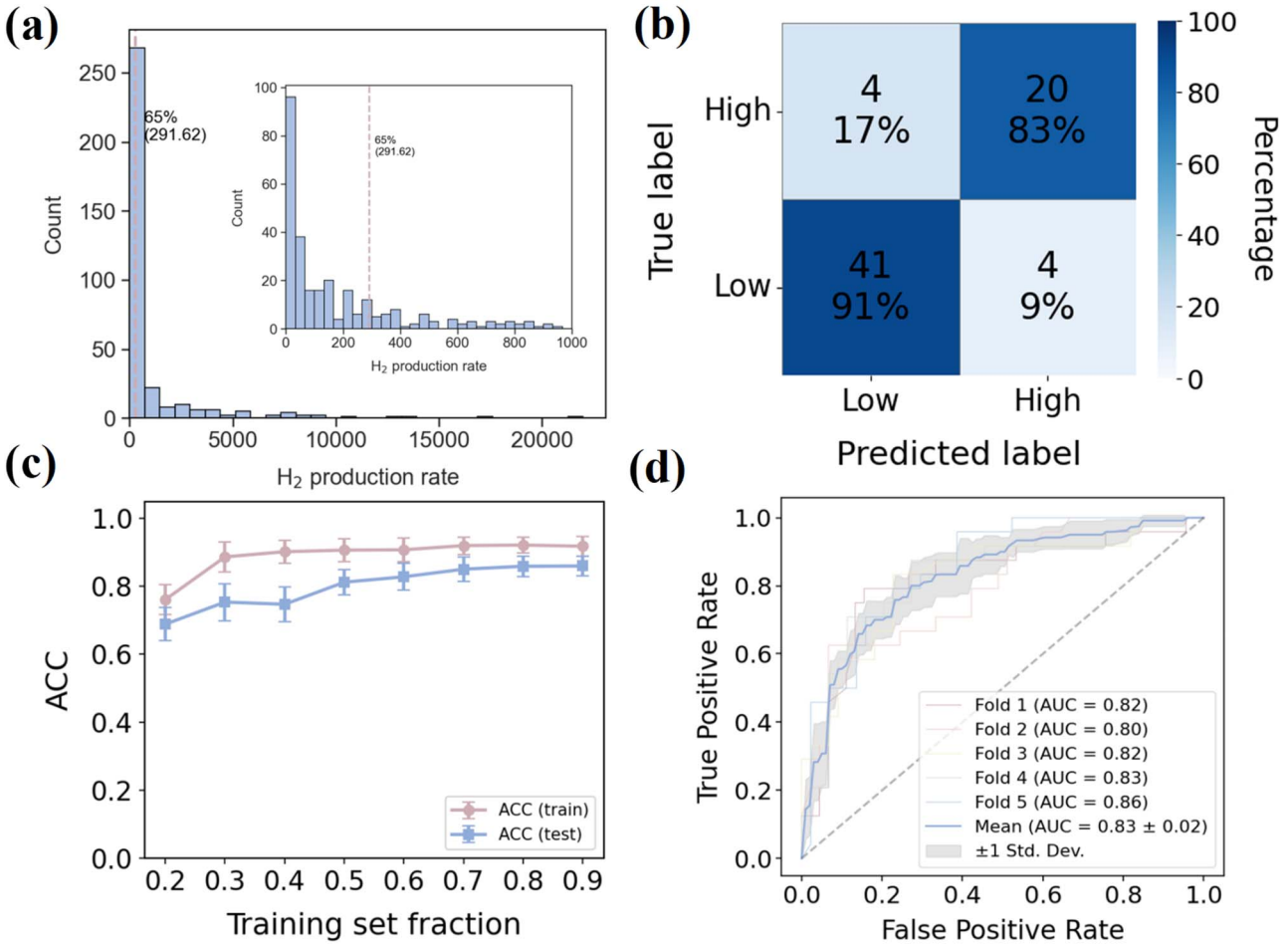


Fig. 3 (a) Distribution of experimental H_2 production rates. (b) Normalized confusion matrix of the LightGBM classifier based on the RFE-processed set. (c) ACC curves of the LightGBM classifier based on the RFE-processed set. Error bars represent standard deviation over 10 random splits of training/test sets. (d) ROC curves with five-fold cross validation based on the RFE-processed set.

adsorption and interaction with active sites. Among the top 20 Molecular ACCess System (MACCS) fingerprints most correlated with photocatalytic activity (Fig. S4), several fragments show positive correlations. In particular, alkyl and amine linkages and polar functional groups possess the strongest correlations, suggesting that linker flexibility and hydrogen-bonding capability favor efficient charge transfer and water interaction. In contrast, oxygen-bridged or highly aromatic motifs possess weaker correlations, implying that excessive rigidity or hydrophobicity may limit catalytic efficiency. Overall, high-performing MOFs tend to integrate metal nodes with strong electron-withdrawing ability and organic linkers with balanced polarity, moderate hydrogen-bonding capacity, and limited aromatic rigidity, which collectively enhance charge separation, water adsorption, and photocatalytic turnover.

2.2. ML classifiers

From the above analysis, given the complexity and nonlinearity of structure–performance relationships, traditional heuristics are apparently insufficient to capture the effects of different factors. Therefore, we employ ML as a data-driven approach to quantitatively examine these complex relationships. The experimental H_2 evolution rates were classified into high- and low-performing classes. As shown in Fig. 3a, rates above the 65th percentile (291.62 μmol g^{-1} h^{-1}) were high-performing, while the remaining 65th were low-performing. The threshold of 291.62 μmol g^{-1} h^{-1} was chosen to account for the skewed distribution of H_2 evolution rates in the data set, ensuring sufficient high-performing data points while maintaining class balance for the binary classification. To justify the threshold selection and evaluate its sensitivity, the performance of the ML classifier was also evaluated at the 60th (232.60 μmol g^{-1} h^{-1}) and 70th (386.40 μmol g^{-1} h^{-1}) percentiles, respectively. As shown in Table S1, a classifier based on Light Gradient Boosting Machine (LightGBM) exhibits exceptional robustness across different thresholds, with average accuracy remaining between 0.84 and 0.88 and the area under the curve (AUC) stabilizing between 0.83 and 0.88. Furthermore, the feature importance analysis (Fig. S5) indicates that despite variation in the performance threshold, the relative contributions of key descriptors such as C_Fermi_Level_aver (Fermi level of the cocatalyst) and pH value remain highly consistent. This stability in feature weights strongly demonstrates that the classifier captures

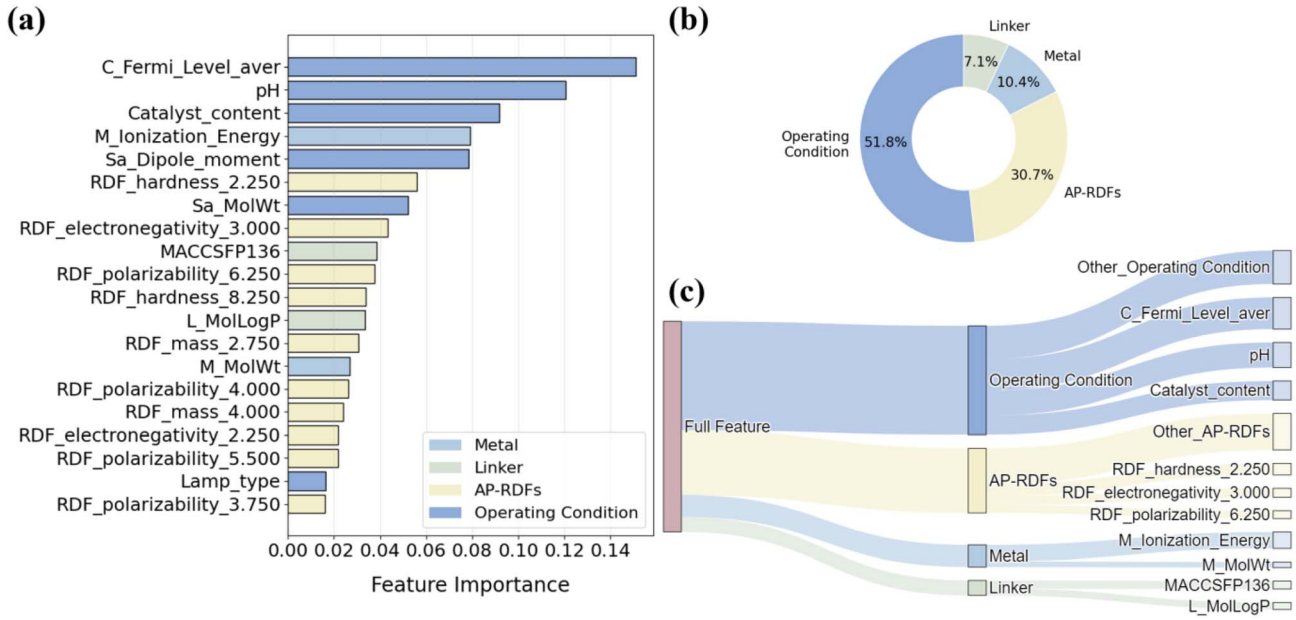


Fig. 4 (a) All features based on average importance with five-fold cross validation. (b) Cumulative feature importance. (c) Sankey diagram illustrating the breakdown of permutation feature importance of different feature types. The node width for each feature component reflects its proportional contribution to the total feature importance.

intrinsic physicochemical mapping relationships rather than being constrained by specific numerical definitions.

Different algorithms, including LightGBM, Random Forest (RF), Gradient Boosting (GB), Extremely Randomized Trees (ET), eXtreme Gradient Boosting (XGBoost) and Categorical Boosting (CatBoost) were applied to train ML classifiers. Among them, LightGBM was found to achieve the highest accuracy in terms of receiver operating characteristic (ROC) curves (Fig. S6). Then with the LightGBM classifier, various descriptor sets were examined. As indicated in Table S2, the combination of metal-based descriptors and operating conditions turned out to show the best predictive performance. Therefore, subsequent analyses were conducted using the LightGBM classifier, unless otherwise stated.

To improve classifier performance and eliminate redundant features, recursive feature elimination (RFE) was applied, reducing the feature dimensionality from 768 to 21 (Table S3). Based on the RFE-processed set, the normalized confusion matrix was predicted by the LightGBM classifier. As shown in Fig. 3b, strong classification performance is observed, with a true positive rate of 83% for the high-performing class and a true negative rate of 91% for the low-performing class. The average accuracy (ACC), positive predictive value (PPV), true positive rate (TPR), and F1 score reach 0.88, 0.83, 0.83, and 0.83, respectively (Table S4). The learning curves of the LightGBM classifier from 10 random splits of training/test sets demonstrate that the metric difference is reduced (Fig. 3c and S7), suggesting improved classifier generalizability. In addition, ROC curves with five-fold cross validation are presented in Fig. 3d. The mean area under the ROC curve (*i.e.*, the mean AUC) reaches 0.83, further validating the robustness and predictive reliability of the LightGBM classifier.

It is instructive to quantify the effects of different features and interpret the learned classifier. Fig. 4a displays all the features based on average importance with five-fold cross validation. Among the features, operating conditions such as pH and C_Fermi_Level_aver emerge as the most influential ones, indicating that operating conditions play a dominant role in determining photocatalytic activity. As shown in Fig. 4b, operating conditions, AP-RDFs, linkers and metals cumulatively account for 51.8%, 30.7%, 10.4% and 7.1%, respectively, of the feature importance, suggesting that both operating conditions and structural descriptors jointly govern the prediction of the LightGBM classifier. This observation is further supported by SHapley Additive exPlanations (SHAP) analysis (Fig. S8), which reveals how individual descriptors influence the prediction direction and magnitude. For instance, a high log P negatively shifts the prediction, implying that a hydrophobic linker would reduce photocatalytic activity. Several AP-RDFs capture the radial distributions of atomic-pair reactivity, reflecting how spatial organization and electronic environment affect light harvesting, carrier transport, and exciton migration. Additionally, M_Ionization_Energy (first ionization energy of a metal), a representative metal descriptor, exhibits a moderate level of importance and provides insight into the intrinsic redox behavior of a metal node. A metal with a lower ionization energy is generally more prone to participate in redox-mediated catalytic cycles, such as single-electron transfer (SET) or coordination-driven hydrogen evolution process. This characteristic is also correlated with the ability of metal center to stabilize radical or charged intermediates, thus modulating the energy landscape of photocatalytic reaction. Fig. 4c integrates these insights into a hierarchical Sankey diagram that illustrates the breakdown of permutation feature importance of

Table 1 Validation performance

Accuracy	Precision	TPR	F1 score	AUC
0.83	1.00	0.78	0.88	0.85

different feature types. The diagram visually highlights the dominant contributions of operating conditions and AP-RDF descriptors, while metal and linker descriptors provide complementary structural and electronic information. Overall, these results underscore that optimal photocatalysts tend to balance the electronic structure, surface polarity, and catalytic site exposure of MOFs, while avoiding excessive hydrophobicity or overly delocalized electron density. The combination of interpretable features with domain-relevant chemistry provides useful guidance for future rational design of MOFs for photocatalytic H_2 production.

The partial-dependence plots of H_2 production rate on key descriptors are illustrated in Fig. S9. The rate increases sharply at a low C_Fermi_Level_aver or pH (Fig. S9a and b), suggesting strong enhancement of photocatalytic activity under acidic conditions and at a low Fermi level. Physicochemically, these descriptors promote water activation and charge separation at the MOF/solution interface. Similarly, Catalyst_content and Sa_Dipole_moment (*i.e.*, the dipole moment of sacrificial agent) display strong positive effects (Fig. S9c and d), implying that increasing the density of active centers and the polarity of sacrificial agent facilitates reaction turnover. Additionally, there exists a distinct performance drop when L_MolLog$P$ ($\log P$ of organic linker) exceeds ~1.0 (Fig. S9e), indicating that an overly hydrophobic linker may impede water accessibility and interfacial proton transfer. Relatively, MACCSFP136 has a weaker effect on photocatalytic activity (Fig. S9f).

2.3. Validation and transferability of the ML classifier

To validate and examine the transferability of the LightGBM classifier for photocatalytic water splitting, we conducted out-of-sample validation using a curated set of 10 MOFs, comprising

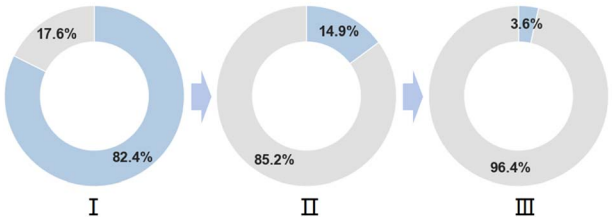


Fig. 6 Statistics in step (I–III) of the screening workflow. The blue and gray regions denote the percentages of shortlisted and discarded MOFs, respectively.

12 experimentally reported data points published between 2024 and 2025. Among these, 8 MOFs were entirely absent from the training set, representing structure-level out-of-samples. Despite not being seen in the training classifier, as shown in Table 1 and Fig. 5, these MOFs were accurately classified with respect to their H_2 evolution performance. This reveals the good transferability of the LightGBM classifier beyond the original training set. The remaining 2 MOFs were included in the training set, but also in the validation set under different operating conditions such as variations in pH and co-catalyst usage. While these conditions were part of the feature space, they had not been previously paired with the same MOFs. The LightGBM classifier still gives consistent and accurate predictions for these new structure-condition combinations, indicating its ability to make reliable predictions within a similar experimental landscape. To exclude potential optimistic bias arising from structural overlap, a stricter re-analysis was conducted by completely removing the 2 MOFs from the out-of-sample validation set and retaining only the 8 MOFs with entirely new structures. In this stricter setting, the LightGBM classifier again exhibits robust performance (Table S5 and Fig. S10). Specifically, it correctly identifies 5 out of 7 high-performing MOFs without generating any false positives. Collectively, the results validate the robustness of the LightGBM classifier in both extrapolating to unseen structures and transferring across varying operating conditions, supporting its

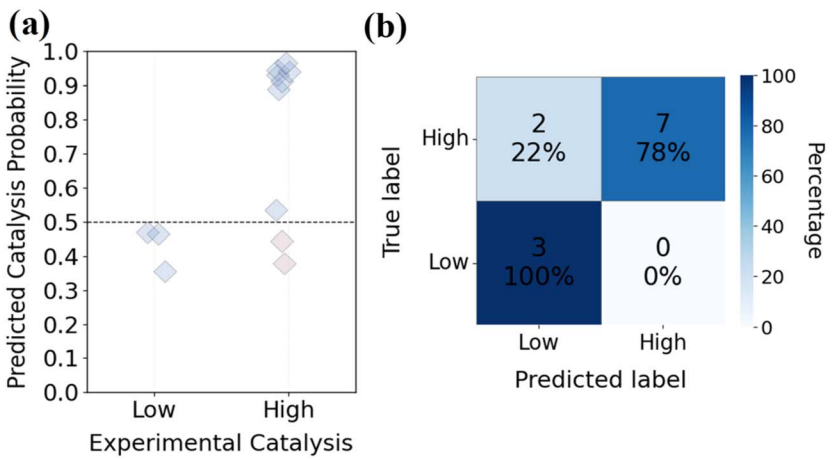
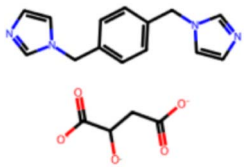
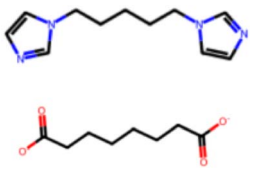
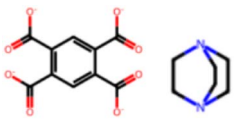
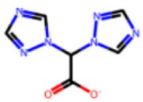
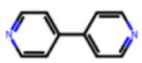


Fig. 5 (a) Predicted catalysis probability *versus* experimental catalysis label. Data points are represented as translucent hexagons to depict data density and colored by the classification correctness: correct (blue) and incorrect (red). (b) Confusion matrix of the classification results.

Table 2 Top five MOFs

MOF	Predicted probability	Metal	Linker
CUYQIX_clean	0.99	Zn	
IYEQUA_clean	0.99	Cd	
MURCAD_clean	0.99	Cu	
QEYWUN_clean	0.99	Cd	
UQULAU_clean	0.99	Co	

utility for high-throughput screening of MOFs for photocatalysis.

2.4. Top CoRE-MOFs

The validated LightGBM classifier was applied to predict the photocatalytic performance of CoRE-MOFs and screen out the top ones. Table S6 illustrates the hierarchical screening workflow. (I) Starting from 11 660 structures in the CoRE-MOF database, we performed structural and descriptor-based filtering, resulting in a reduced set of 9603 MOFs. (II) The LightGBM classifier was employed to predict the photocatalytic performance of these MOFs and top-performing ones were shortlisted. (III) A recently developed water stability classifier for MOFs[26] was subsequently utilized to identify stable structures among the top-performing MOFs. As summarized in Fig. 6, the number of MOFs was progressively reduced to 82.4%, 14.9%, and 3.6% at each step, respectively, ultimately yielding 419 water-stable top-performing MOFs. Notably, in step II, the number of MOFs was reduced significantly from 82.4% to 14.9%, underscoring the strong capability of the LightGBM classifier in efficiently narrowing down the candidate space. Among the 419 MOFs, Table 2 lists the top five structures, which possess the highest priority for experimental exploration. To assess the robustness of screening, the consistency of top-performing MOFs was examined across different quantile thresholds (Table S7). For the top 5 MOFs, the overlap rate reaches as high as 0.80. Such a high consistency for the top 5 MOFs is significant and it underscores that the top MOFs identified by the LightGBM classifier remain reliable regardless of the fine-tuning of performance boundaries, thereby ensuring the credibility of screened MOFs.

To further evaluate the practical applicability of our ML classifier, we performed a retrospective validation by cross-referencing the predicted top MOFs with recent experimental data. As listed in Table 3, several MOFs including HAKSEU and HAKSOE were reported to exhibit excellent H_2 evolution activity under visible light, with production rates exceeding 1000 μmol g^{-1} h^{-1}.[27] Although the operating conditions are different from those used in our predictions, the high predictive probability (0.92–0.93) highlights the robustness and generalizability of the LightGBM classifier across different structure-condition combinations. Notably, these MOFs also appear in our out-of-sample validation set but under different operating conditions. As the LightGBM classifier incorporates both structural and operating conditions as input features, the corresponding

Table 3 Retrospective validation of predictions against recent experimental data

MOF	Predictive probability	Experimental H_2 rate (μmol g^{-1} h^{-1})[27]
HAKSEU	0.93	1005.40
HAKSOE	0.92	1478.75

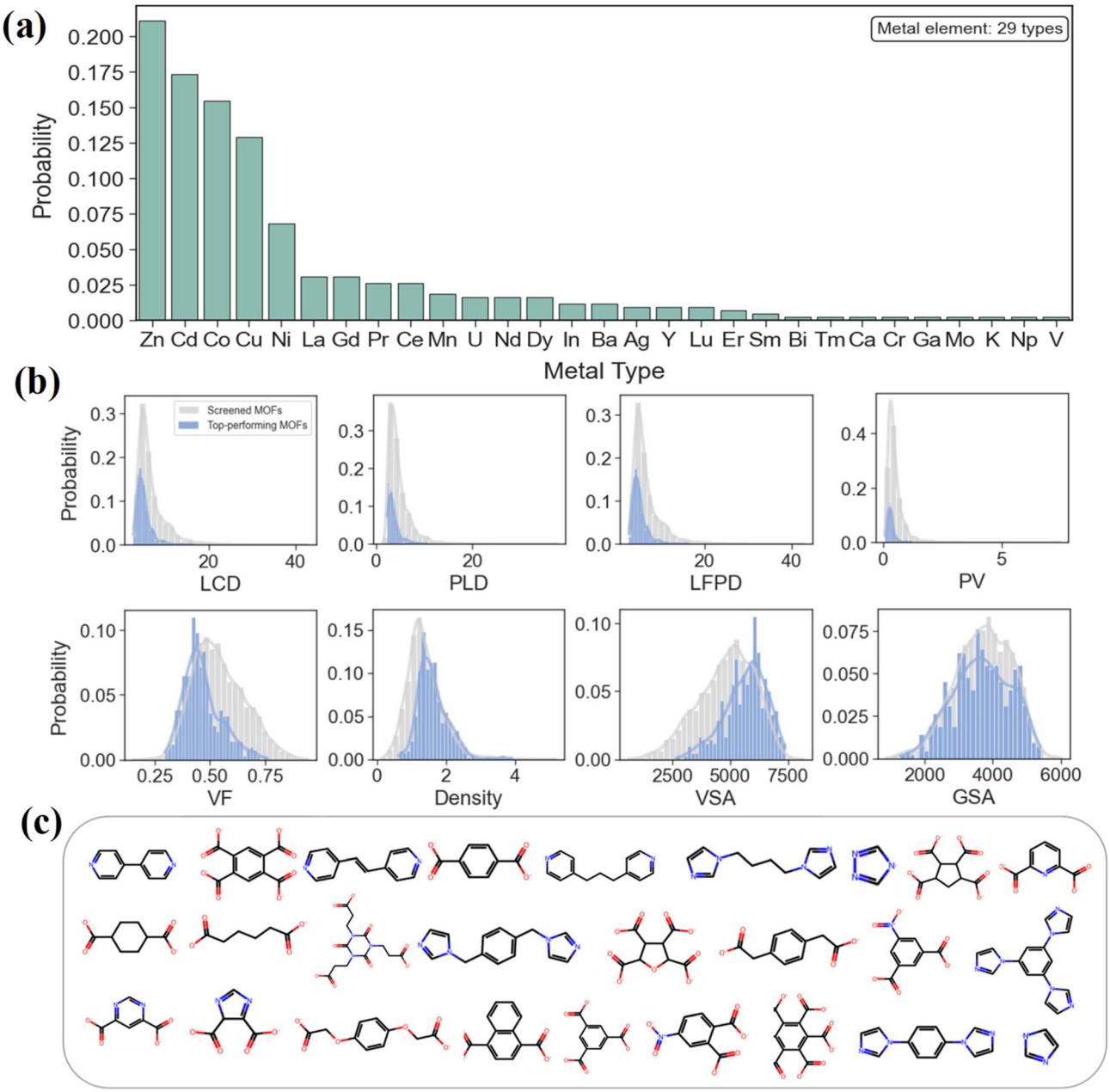


Fig. 7 Predicted 419 top-performing MOFs. (a) Probability of metal types. (b) Comparison of geometric properties between top-performing MOFs and the screened 9603 MOFs. (c) Representative linkers in top-performing MOFs.

predictions reflect distinct structure-condition pairs and thus do not compromise the independence of validation process.

Subsequently, we performed a comprehensive structural analysis of the 419 top-performing MOFs. As shown in Fig. 7a, the metal nodes in these MOFs are dominated by transition metals, such as Zn, Cd, Co and Cu. Fig. 7b displays the key geometric descriptors in these MOFs. Compared to the initial screened data set of 9603 MOFs, the top-performing MOFs tend to possess moderate pore sizes (LCD, PLD, and LFPD primarily in the range of 10–20 Å), higher void fractions, and larger accessible surface areas (VSA and GSA). These structural characteristics are favorable for enhancing mass transport, increasing the accessibility of catalytic sites, and creating a microenvironment conducive to photocatalytic reactions. For the organic linkers, representative structures shown in Fig. 7c reveal the prevalence of aromatic backbones (*e.g.*, phenyl and pyridyl groups) and polar functional groups (*e.g.*, carboxylates, amides and ethers). These functionalities not only facilitate stable coordination with metal nodes, but also enhance water adsorption and charge transfer by tuning the electronic and polar properties of MOFs. Additionally, the t-distributed stochastic neighbor embedding (t-SNE)[28] map in Fig. S11 illustrates the distribution of top-performing MOFs in the overall feature space. While the top MOFs are broadly distributed, they exhibit discernible clustering patterns, suggesting that they share critical structural motifs despite their diversity. This

reflects the ability of the LightGBM classifier to capture relevant features within a complex chemical landscape. The hierarchical screening significantly improves the efficiency of candidate selection while preserving structural diversity and chemical plausibility.

3. Conclusion

In this study, we have developed a hierarchical screening strategy that integrates ML-based photocatalytic performance prediction, water stability classification, and structure–performance analysis to accelerate the discovery of MOFs for H_2 production. A dedicated data set with experimentally reported H_2 evolution rates is curated to train ML classifiers tailored for photocatalytic water splitting. The LightGBM classifier demonstrates high accuracy and strong transferability in out-of-sample validation. Structural analysis reveals that top-performing MOFs are rich in catalytically favorable transition metals (*e.g.*, Zn, Cd, Co and Cu) and exhibit moderate pore sizes, high void fractions and aromatic linkers bearing polar functional groups such as carboxylates and pyridyls. These features contribute to improved charge separation, mass transport, and aqueous compatibility. From the CoRE-MOF database, structure-based and descriptor-based filtering is performed to narrow the structures down from 11 660 to 9603. Subsequently, 1731 structures are identified to be photocatalytically active, of which 419 are predicted to be water stable. Retrospective validation against experiments further supports the accuracy of our ML model. Overall, this study highlights the practical utilization of ML in discovering high-performing photocatalytic MOFs and provides a solid data-driven foundation for interpreting structure–performance relationships, and it would be insightful to guide the future rational design of MOFs for efficient photocatalytic H_2 production.

It is worthwhile to note that the encoded structural and chemical descriptors in our ML model, such as metal node properties, linker polarity and pore architecture, may also hold relevance beyond photocatalysis, *e.g.*, for electrocatalytic H_2 production. However, inherent differences between photocatalytic and electrocatalytic reactions, including the presence of electrolytes, applied potentials and electrode interfaces, could affect model performance and limit direct transferability. Therefore, while our model may serve as a preliminary predictive tool for electrocatalytic H_2 production, experimental validation and model refinement under relevant reaction conditions are essential to ensure predictive accuracy.

4. Methodology

4.1. Data curation

The experimental photocatalysis data were curated from peer-reviewed manuscripts reported in the literature. The manuscripts were identified *via* a targeted search in the Web of Science using keywords "metal–organic framework" AND "photocat*" AND "hydrogen" OR "wat*", where the wildcard symbol "*" was employed to encompass keyword variants such as "photocatalysis", "photocatalytic", "water", and "watery". A total of 100 MOFs with experimentally reported H_2 evolution rates were collected. In addition to rates, operating conditions were also extracted when available, including catalyst and cocatalyst loadings, substrate type, sacrificial agent, reaction temperature, pH, irradiation, and lamp type.

4.2. Featurization

The collected MOFs were encoded to machine-readable descriptors including geometric, building unit- and atom-based descriptors as shown in Table S8. First, the crystallographic information files (CIFs) from the Cambridge Structural Database (CSD) were preprocessed using a solvent-removal script.[29] Then, eight geometric descriptors reflecting overall crystalline architectures were calculated by using Zeo++.[22] They were the largest cavity diameter (LCD, Å), pore limited diameter (PLD, Å), largest free path diameter (LFPD, Å), volumetric surface area (VSA, $m^2\ cm^{-3}$), gravimetric surface area (GSA, $m^2\ g^{-1}$), void fraction (VF), pore volume (PV, $cm^3\ g^{-1}$), and density (kg m^{-3}). Fig. S12 and S13 show the distributions and correlation matrix of geometric descriptors in the collected MOFs. For the building unit-based descriptors, MOFs were deconstructed into different metal atoms and organic linkers (Table S8) *via* the MOFid toolkit.[22,30] The metal atoms were described by physiochemical properties such as atomic mass, radius and electronegativity. In MOFs with mixed metals, the averaged properties were estimated based on the ratio of metals. The organic linkers were represented by the simplified molecular-input line-entry system (SMILES) and calculated using RDKitDP[31] and MACCS fingerprints,[32] respectively. In addition, atom-based descriptors were incorporated by the atomic-property-weighted radial distribution functions (AP-RDFs), which encompass the global description of atomic environment.[33] To capture the physicochemical landscape relevant to photocatalytic HER, experimental operating conditions were included as features (Table S9). Solvent was characterized using its physicochemical parameters such as dielectric constant and acid dissociation constant (pK_a). For cocatalyst, band gap and Fermi level were used to quantify its electronic properties. Sacrificial agent was characterized by its ionization energy. These operating conditions are critical to determine photocatalytic HER performance. All the continuous features, including 565 AP-RDF descriptors, were standardized using Z-score normalization.

4.3. Machine learning

The curated data set was employed to train and validate ML classifiers. Following the timeline, the data set was split into two classes: (1) 92 MOFs published before 2024 were used for training, representing established knowledge; (2) 10 MOFs published after 2024 were reserved for out-of-sample validation, among which two existed in the training set but under different operating conditions. Different algorithms were adopted for training, including Random Forest (RF), Gradient Boosting (GB), Extremely Randomized Trees (ET), Light Gradient Boosting Machine (LightGBM), eXtreme Gradient Boosting (XGBoost), and Categorical Boosting (CatBoost) as implemented in the scikit-learn toolkit.[34,35] The performance of each

classifier was optimized by tuning its respective hyperparameters (Table S10) and quantified using the receiver operating characteristic (ROC) curve and area under the ROC curve (AUC). The ROC compares a true positive value against a false positive value, and the AUC measures the overall ability of a classifier to discriminate between classes, thus ensuring a comprehensive evaluation of the classifier. For each loop during training, the data set was split into a training set and a test set. The descriptors were centered to zero and scaled using the mean and standard deviation. To mitigate overfitting, grid search was employed with cross validation for hyperparameter tuning, feature engineering, and comprehensive classifier evaluation. First, five-fold cross validation was adopted to optimize hyperparameters, which involved dividing the data set into five subsets and iteratively training the classifier on four subsets while validating the remaining subset. Hyperparameters were tuned by grid search for comprehensive and systematic exploration across the whole hyperparameter space, like n_estimators, max_depth, and max_features. For all the classifiers, Table S10 lists the search space. Specifically, the exact optimized hyperparameters for the final LightGBM classifier are provided in Table S11. Then, recursive feature elimination (RFE) was implemented to further enhance classifier robustness. By iteratively removing less relevant descriptors, RFE helped focus on the most informative descriptors. Lastly, a thorough evaluation of classifier performance and detection of potential overfitting were conducted using a series of metrics including ACC, PPV, TPR, F1 and AUC.

$$\mathrm{ACC} = \frac{\text{true positives} + \text{true negatives}}{\text{total predictions}} \quad (1)$$

$$\mathrm{PPV} = \frac{\text{true positives}}{\text{true positives} + \text{false positives}} \quad (2)$$

$$\mathrm{TPR} = \frac{\text{true positives}}{\text{true positives} + \text{false negatives}} \quad (3)$$

$$\mathrm{F1} = \frac{2 \times \mathrm{PPV} \times \mathrm{TPR}}{\mathrm{PPV} + \mathrm{TPR}} \quad (4)$$

where ACC is the overall accuracy, PPV is the positive predictive value, TPR is the true positive rate, and F1 is the harmonic mean of PPV and TPR. In addition, Pearson, Spearman and Kendall correlation coefficients between descriptors and targets were calculated. The importance of descriptors was quantified using the mean information gain and the SHapley Additive exPlanations (SHAP) to interpret ML classifiers.[34,36]

5. Predictions

After validation, the best-performing LightGBM classifier was applied to predict the photocatalytic performance of MOFs in the CoRE 2019 database.[37] Derived from experimentally synthesized structures, the CoRE MOF database contains 11 660 computation-ready CIF files. To ensure data quality, we first performed curation and structural filtering to remove entries with missing atoms or unreasonable geometries. Given that most of the 92 collected MOFs were sourced from the CoRE 2019 database and the CSD database, duplicate entries were identified and excluded. Finally, 9603 structures (~80%) were retained for featurization using the same procedure described above and used for predictions. Although our collected data set only counts a small proportion of 9603 CoRE MOFs, it fairly well samples the chemical space of CoRE MOFs as shown in the t-SNE map (Fig. S14). All the descriptors were normalized and standardized by using the same parameters used to train ML classifiers. To ensure consistency, a standardized set of experimental operating conditions was applied, including representative values for catalyst concentration, cocatalyst loading, temperature, pH, and other key parameters (Table S12). By using these values, we aimed to ensure consistent and interpretable inputs when applying the classifier for new predictions.

Author contributions

X. N.: investigation, formal analysis, data curation, validation, writing – original draft. Z. Z.: investigation, writing – review & editing, supervision. X. W.: formal analysis, validation. Y. L.: formal analysis, visualization. Y. C.: funding acquisition, writing – review & editing, supervision. J. J.: funding acquisition, writing – review & editing, supervision.

Conflicts of interest

The authors declare there is no competing interest.

Data availability

Data and codes related to this study are provided on GitHub (https://github.com/NXCCC/StableMOF_PhotoH2).

Supplementary information (SI): structure-performance relationships; ML classifier performance; predictions for CoRE-MOFs; featurization. See DOI: https://doi.org/10.1039/d5sc08277c.

Acknowledgements

This work was supported by the National Key R&D Program of China (2021YFA1200402, 2021YFA1501501, 2022YFA1503302, and 2021YFA1200302), the National Natural Science Foundation of China (Grants 22331007, 22225111, and 22174125), the Key Project of Basic Research of Shanghai (22JC1402000), and the National Research Foundation Singapore (NRF-CRP26-2021RS-0002). X. N. gratefully acknowledges the support from the China Scholarship Council (CSC).

References

1 J. Khan and M. H. Arsalan, *Renewable Sustainable Energy Rev.*, 2016, **55**, 414–425.
2 S. J. Davis, K. Caldeira and H. D. Matthews, *Science*, 2010, **329**, 1330–1333.

3 L. V. Mattos, G. Jacobs, B. H. Davis and F. B. Noronha, *Chem. Rev.*, 2012, **112**, 4094–4123.
4 C. Acar, I. Dincer and G. F. Naterer, *Int. J. Energy Res.*, 2016, **40**, 1449–1473.
5 M. Tahir, S. Tasleem and B. Tahir, *Int. J. Hydrogen Energy*, 2020, **45**, 15985–16038.
6 H. Wang, X. Zhang, W. Zhang, M. Zhou and H. L. Jiang, *Angew. Chem., Int. Ed.*, 2024, **63**, e202401443.
7 X. Lu, X. Yu, B. Li, X. Sun, L. Cheng, Y. Kai, H. Zhou, Y. Tian and D. Li, *Adv. Sci.*, 2024, **11**, e2405643.
8 D. Li, M. Kassymova, X. Cai, S. Q. Zang and H. L. Jiang, *Coord. Chem. Rev.*, 2020, **412**, 213262.
9 J. Zhang, T. Bai, H. Huang, M. H. Yu, X. Fan, Z. Chang and X. H. Bu, *Adv. Mater.*, 2020, **32**, e2004747.
10 M. Cabrero-Antonino, J. Albero, C. Garcia-Valles, M. Alvaro, S. Navalon and H. Garcia, *Chem.–Eur. J.*, 2020, **26**, 15682–15689.
11 Y. An, Y. Liu, P. An, J. Dong, B. Xu, Y. Dai, X. Qin, X. Zhang, M. H. Whangbo and B. Huang, *Angew. Chem., Int. Ed.*, 2017, **56**, 3036–3040.
12 P. Salcedo-Abraira, A. A. Babaryk, E. Montero-Lanzuela, O. R. Contreras-Almengor, M. Cabrero-Antonino, E. S. Grape, T. Willhammar, S. Navalon, E. Elkaim, H. Garcia and P. Horcajada, *Adv. Mater.*, 2021, **33**, e2106627.
13 K. Sun, Y. Qian and H. L. Jiang, *Angew. Chem., Int. Ed.*, 2023, **62**, e202217565.
14 M. I. Jordan and T. M. Mitchell, *Science*, 2015, **349**, 255–260.
15 K. M. Jablonka, D. Ongari, S. M. Moosavi and B. Smit, *Chem. Rev.*, 2020, **120**, 8066–8129.
16 X. Wu, R. Zheng and J. Jiang, *J. Chem. Theory Comput.*, 2025, **21**, 900–911.
17 Z. Zhang, H. Tang, M. Wang, B. Lyu, Z. Jiang and J. Jiang, *ACS Sustain. Chem. Eng.*, 2023, **11**, 8148–8160.
18 Z. Zhang, A. S. Palakkal, X. Wu, J. Jiang and Z. Jiang, *Environ. Sci. Technol.*, 2025, **59**, 9123–9133.
19 C. Wang, Y. Wan, S. Yang, Y. Xie, S. Chu, Y. Chen and X. Yan, *Adv. Funct. Mater.*, 2024, **34**, 2313596.
20 B. Mourino, S. Majumdar, X. Jin, F. McIlwaine, J. Van Herck, A. Ortega-Guerrero, S. Garcia and B. Smit, *Chem. Sci.*, 2025, **16**, 11434–11446.
21 A. Meng, L. Zhang, B. Cheng and J. Yu, *Adv. Mater.*, 2019, **31**, 1807660.
22 T. F. Willems, C. H. Rycroft, M. Kazi, J. C. Meza and M. Haranczyk, *Microporous Mesoporous Mater.*, 2012, **149**, 134–141.
23 Y. Zhang, D. Ma, J. Li, C. Zhi, Y. Zhang, L. Liang, S. Mao and J. W. Shi, *Coord. Chem. Rev.*, 2024, **517**, 215995.
24 J. Y. Zeng, X. S. Wang, B. R. Xie, Q. R. Li and X. Z. Zhang, *J. Am. Chem. Soc.*, 2022, **144**, 1218–1231.
25 A. Cadiau, N. Kolobov, S. Srinivasan, M. G. Goesten, H. Haspel, A. V. Bavykina, M. R. Tchalala, P. Maity, A. Goryachev and A. S. Poryvaev, *Angew. Chem., Int. Ed.*, 2020, **132**, 13570–13574.
26 Z. Zhang, F. Pan, S. A. Mohamed, C. Ji, K. Zhang, J. Jiang and Z. Jiang, *Small*, 2024, **20**, 2405087.
27 T. Chen, C. Lu, J. Wang, Y. Kong, T. Liu, S. Ying, X. Ma and F. Y. Yi, *Electrochim. Acta*, 2024, **480**, 143927.
28 L. Van der Maaten and G. Hinton, *J. Mach. Learn. Res.*, 2008, **9**, 2579–2605.
29 C. R. Groom, I. J. Bruno, M. P. Lightfoot and S. C. Ward, *Acta Crystallogr., Sect. B*, 2016, **72**, 171–179.
30 B. J. Bucior, A. S. Rosen, M. Haranczyk, Z. Yao, M. E. Ziebel, O. K. Farha, J. T. Hupp, J. I. Siepmann, A. Aspuru-Guzik and R. Q. Snurr, *Cryst. Growth Des.*, 2019, **19**, 6682–6697.
31 G. Landrum, *RDKit, Open-Source Cheminformatics Software*, 2016, http://www.rdkit.org/.
32 J. L. Durant, B. A. Leland, D. R. Henry and J. G. Nourse, *J. Chem. Inf. Comput. Sci.*, 2002, **42**, 1273–1280.
33 M. Fernandez, N. R. Trefiak and T. K. Woo, *J. Phys. Chem. C*, 2013, **117**, 14095–14105.
34 F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot and É. Duchesnay, *J. Mach. Learn. Res.*, 2011, **12**, 2825–2830.
35 L. Buitinck, G. Louppe, M. Blondel, F. Pedregosa, A. Mueller, O. Grisel, V. Niculae, P. Prettenhofer, A. Gramfort and J. Grobler, *arXiv*, 2013, preprint, arXiv:1309.0238, DOI: 10.48550/arxiv.1309.0238.
36 S. M. Lundberg and S. I. Lee, *Adv. Neural Inf. Process. Syst.*, 2017, **30**, 4765–4774.
37 Y. G. Chung, E. Haldoupis, B. J. Bucior, M. Haranczyk, S. Lee, H. Zhang, K. D. Vogiatzis, M. Milisavljevic, S. Ling, J. S. Camp, B. Slater, J. I. Siepmann, D. S. Sholl and R. Q. Snurr, *J. Chem. Eng. Data*, 2019, **64**, 5985–5998.