Supplementary information

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

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

[a]*State Key Laboratory of Synergistic Chem-Bio Synthesis, School of Chemistry and Chemical Engineering, Shanghai Jiao Tong University, Shanghai 200240, China*

[b]*Department of Chemical and Biomolecular Engineering, National University of Singapore, 117576 Singapore*

Corresponding authors: yongcui@sjtu.edu.cn (Y. Cui), chejj@nus.edu.sg (J. Jiang)

Table of Contents

S1. Structure-performance relationships

S2. ML classifier performance

S3. Predictions for CoRE-MOFs

S4. Featurization

S1. Structure-performance relationships

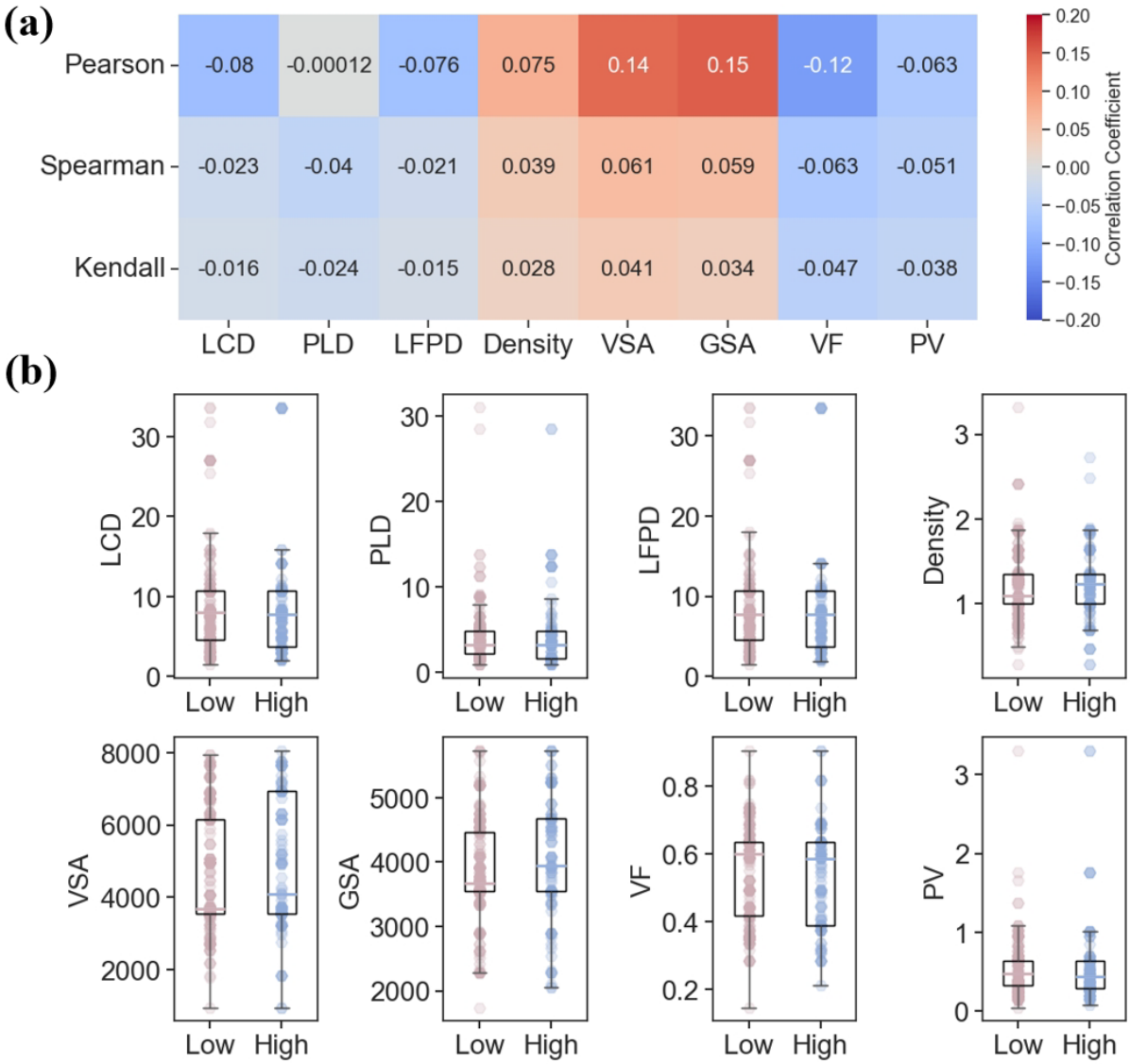

Fig. S1. (a) Correlation matrix of Pearson, Spearman and Kendall coefficients between geometric descriptors and H_2 production rate. (b) Boxplot distributions of geometric descriptors in low- and high-performing MOFs.

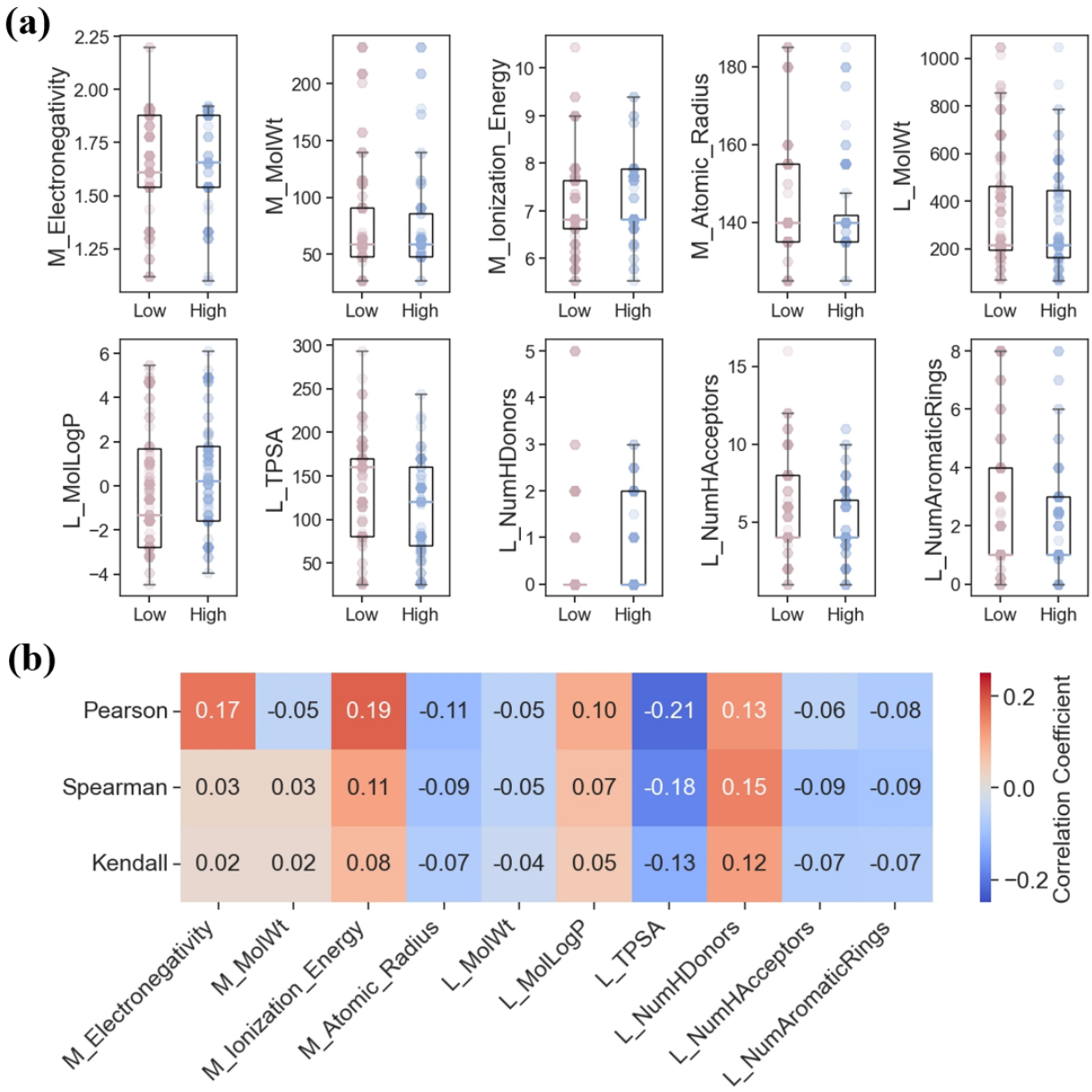

Fig. S2. (a) Boxplot distributions of metal and linker descriptors in low- and high-performing MOFs. (b) Correlation matrix of Pearson, Spearman and Kendall coefficients between metal, linker and H_2 production rate.

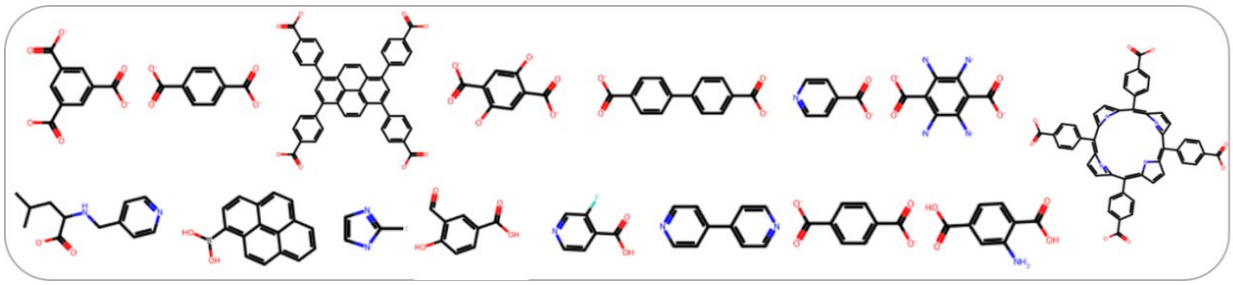

Fig. S3. Representative organic linkers in the collected MOFs.

Fig. S4. Correlation matrix of Pearson, Spearman and Kendall coefficients for the top 20 MACCS fingerprints of linkers most correlated with H_2 production rate.

S2. ML classifier performance

Table S1. Performance of LightGBM classifier under different percentile thresholds.

Percentile	Threshold Value ($\mu mol \cdot g^{-1} \cdot h^{-1}$)	Accuracy	Precision	TPR	F1 Score	AUC
60th	232.60	0.84	0.81	0.77	0.79	0.83
70th	386.40	0.88	0.84	0.83	0.82	0.88

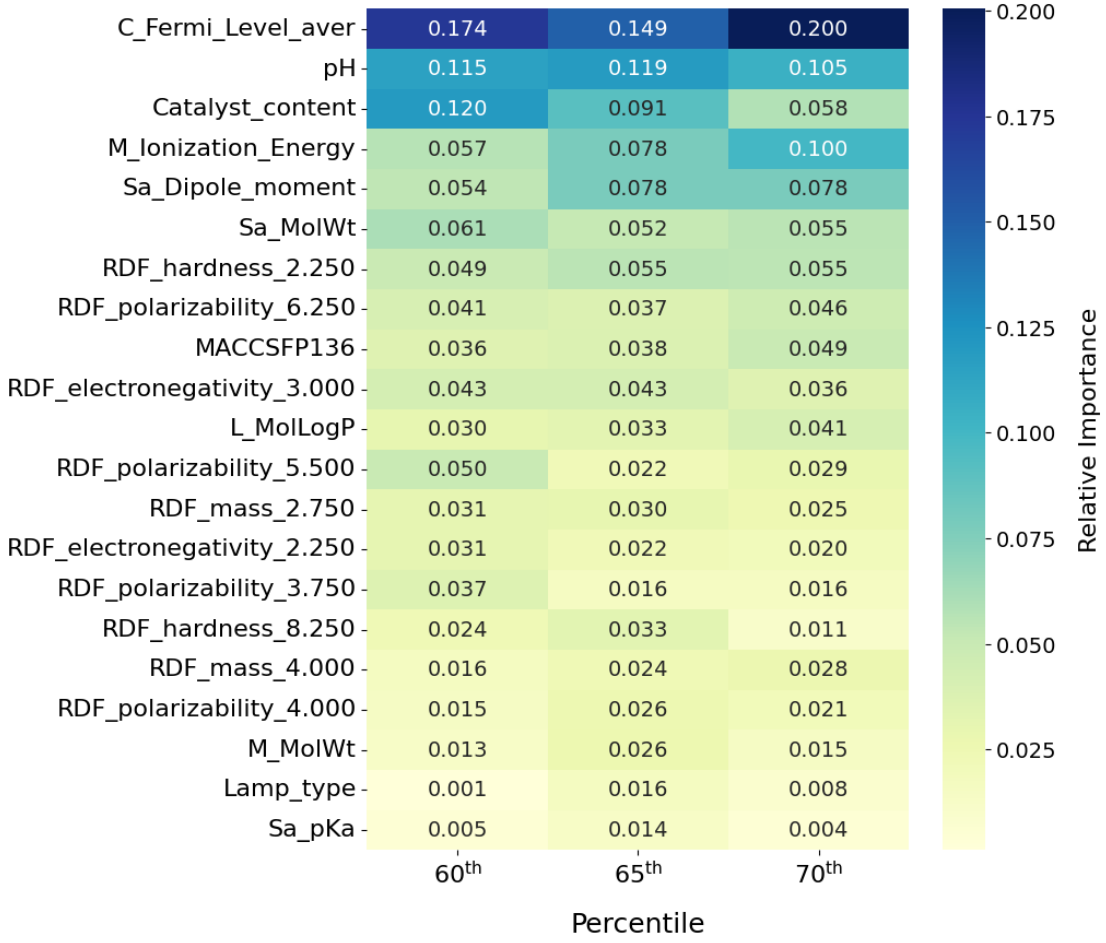

Fig. S5. Heatmap of relative feature importance across 60th, 65th, and 70th percentile thresholds.

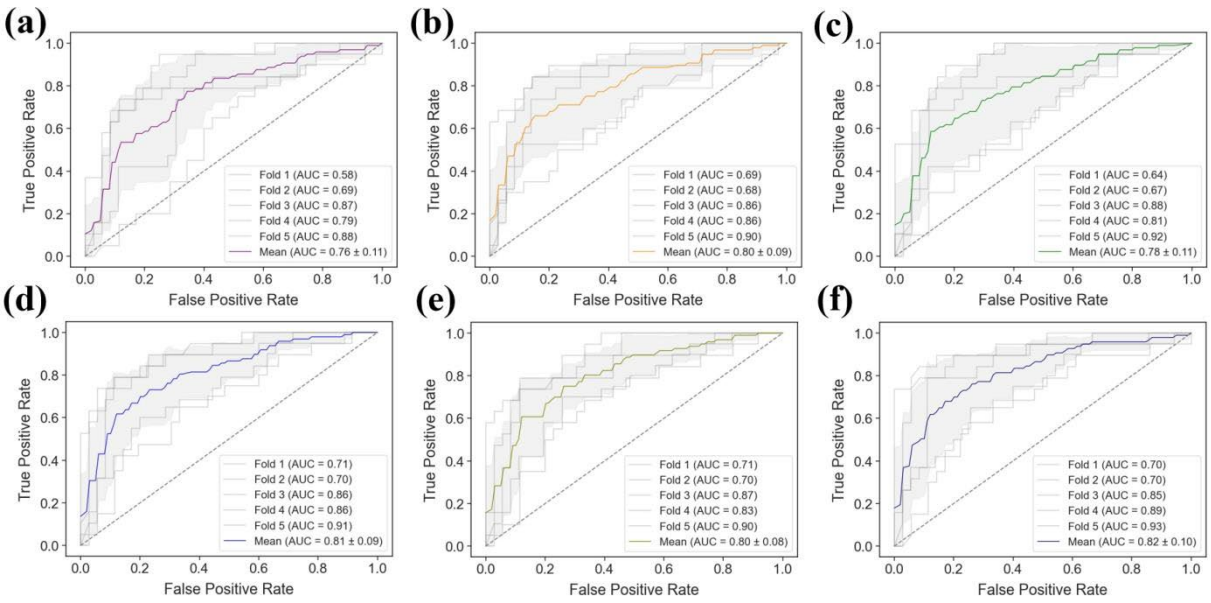

Fig. S6. Receiver operating characteristic (ROC) curves of (a) Extremely Randomized Trees, (b) Gradient Boosting, (c) Random Forest, (d) XGBoost, (e) CatBoost, and (f) LightGBM classifiers from five-fold cross-validation in binary classification based on original data set.

Table S2. Performance of LightGBM classifier using different descriptor sets based on original data.

Descriptor set	ACC	PPV	TPR	F1	AUC
Geometric + Operating Conditions	0.86	0.79	0.79	0.79	0.90
Metal + Operating Conditions	0.87	0.78	0.88	0.82	0.92
Linker + Operating Conditions	0.84	0.78	0.75	0.77	0.91
AP-RDFs + Operating Conditions	0.86	0.79	0.79	0.79	0.88

ACC: the overall accuracy, PPV: the positive predictive value, TPR: the true positive rate, F1: the harmonic mean of PPV and TPR, and AUC: the area under receiver operating characteristic (ROC) curve.

Table S3. RFE-processed descriptors and their interpretation. Descriptors starting with 'Sa', 'C', 'M' and 'L' represent sacrificial agent, cocatalyst, metal and linker properties, respectively. Numeric descriptors are MACCS keys.

Descriptor	Description
Catalyst_content (mg)	Amount of catalyst
pH	pH of reaction solution
Lamp_type	Type of lamp
Sa_pKa	Acidity constant (pK_a) of sacrificial agent
C_Fermi_Level_aver (eV)	Fermi level of cocatalyst
Sa_Dipole_moment (D)	Dipole moment of sacrificial agent
Sa_Ionization_energy (eV)	Ionization energy of sacrificial agent
Sa_MolWt (g/mol)	Molar mass of sacrificial agent
M_MolWt (g/mol)	Atomic weight of metal
M_Ionization_Energy (eV)	First ionization energy of metal
L_MolLogP	Molecular LogP of organic linker
MACCSFP136	O=A > 1
RDF_electronegativity_2.25	RDF of atomic electronegativity at 2.25 Å
RDF_electronegativity_3.00	RDF of atomic electronegativity at 3.00 Å
RDF_hardness_2.25	RDF weighted by atomic hardness (η) within 2.25 Å radius
RDF_hardness_8.25	RDF weighted by atomic hardness (η) within 8.25 Å radius
RDF_polarizability_3.75	RDF weighted by atomic polarizability at 3.75 Å
RDF_polarizability_4.00	RDF weighted by atomic polarizability at 4.00 Å
RDF_polarizability_5.50	RDF weighted by atomic polarizability at 5.50 Å
RDF_polarizability_6.25	RDF weighted by atomic polarizability at 6.25 Å
RDF_mass_4.00	RDF weighted by atomic mass at 4.00 Å

In MACCS keys, (1) atom symbol: A (any valid periodic table element), Q (hetro atoms; any non-C or non-H atom), X (halogen), Z (other than H, C, N, O, Si, P, S, F, Cl, Br, I); (2) bond type: - (single), = (double), T (triple), # (triple), ~ (single or double bond), % (aromatic bond); $ (ring), ! (chain or non-ring bond), @ (ring linkage and number following it specifies atom position).

Table S4. Performance of RFE-processed dataset.

Accuracy	Precision	TPR	F1 Score
0.88	0.83	0.83	0.83

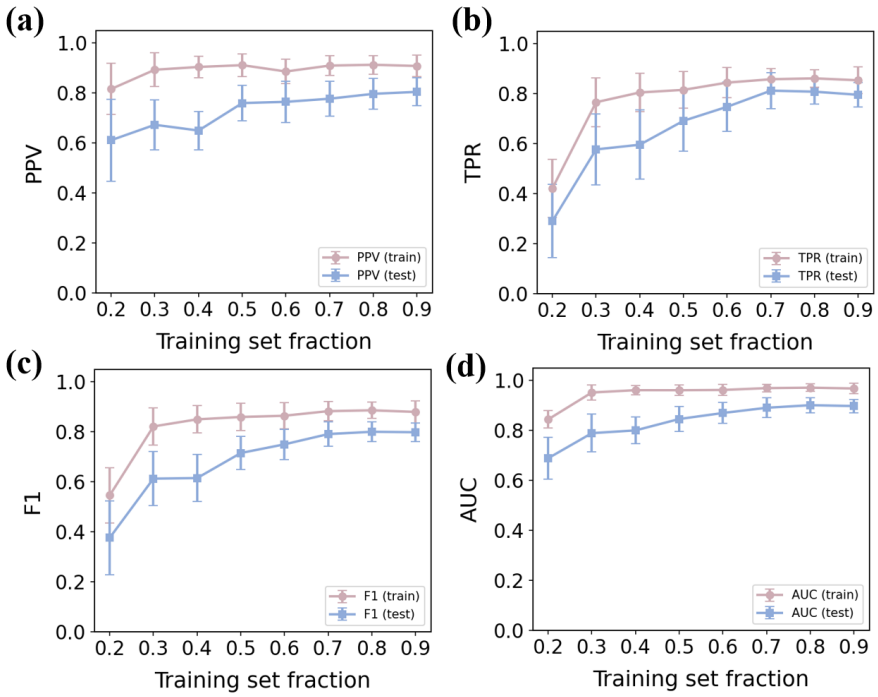

Fig. S7. Learning curves of LightGBM classifier in binary classification. The average values of PPV, TPR, F1 and AUC were generated using 10 random splits of training/test sets.

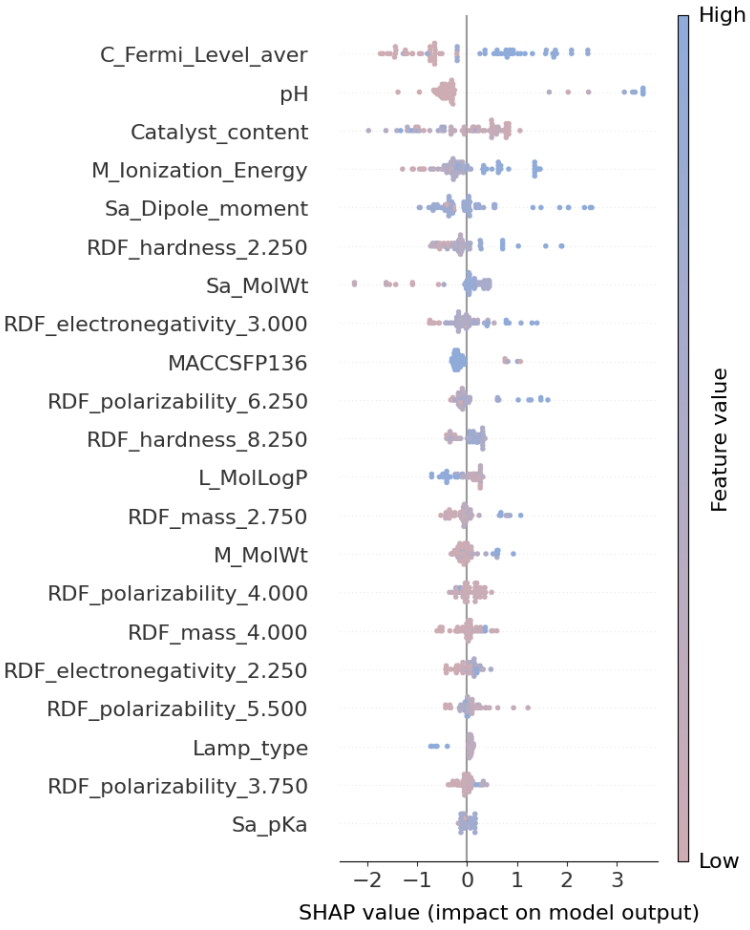

Fig. S8. SHAP summary plot of all the features for high photocatalytic performance.

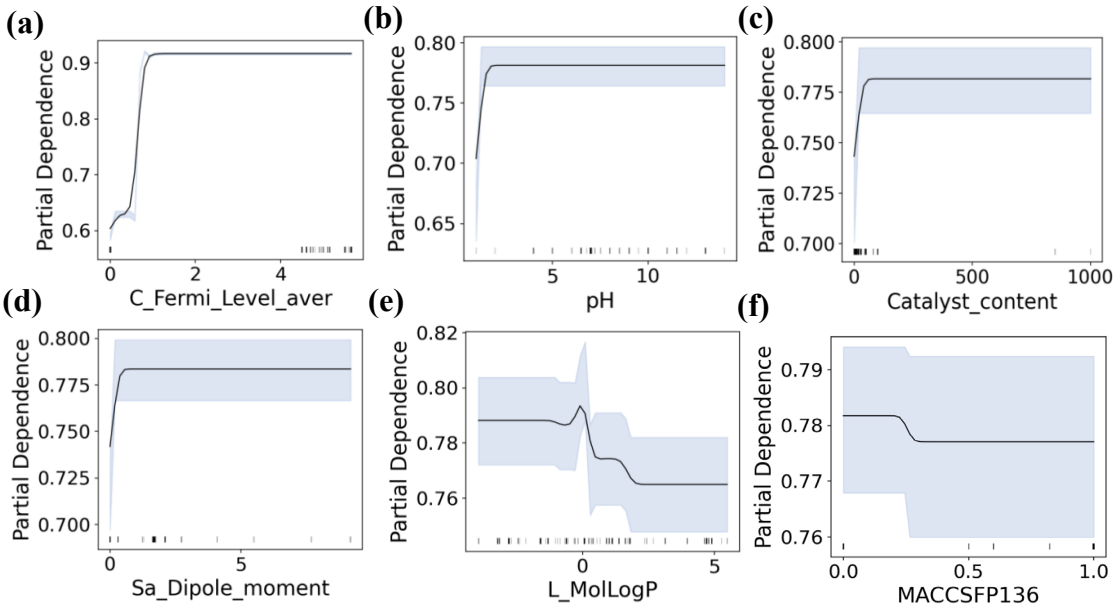

Fig. S9. Partial-dependence plots of H_2 production rate on (a) Fermi level of cocatalyst (C_Fermi_Level_aver), (b) pH, (c) catalyst content, (d) dipole moment of sacrificial agent (Sa_Dipole_moment), (e) molecular LogP of organic linker (L_MolLogP), (f) MACCSFP136.

Table S5. Validation performance for 8 MOFs with entirely new structures.

Accuracy	Precision	TPR	F1 Score	AUC
0.75	1.00	0.71	0.83	0.71

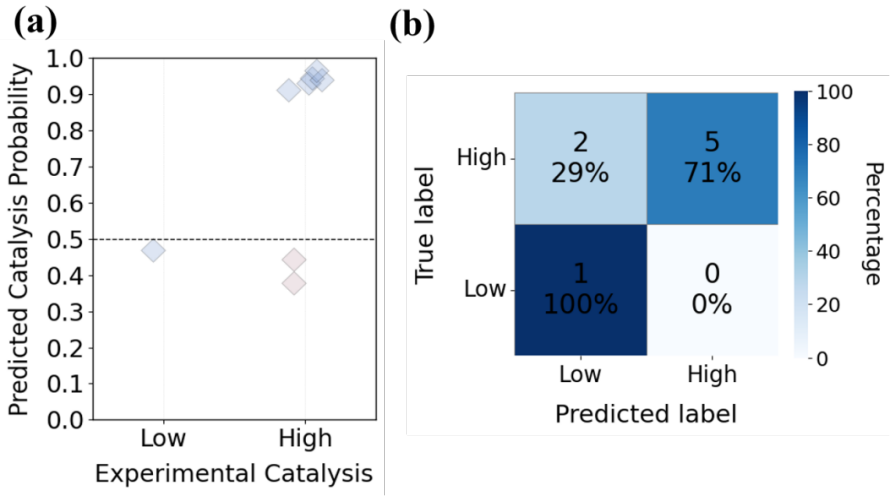

Fig. S10. Classification performance on the stricter out-of-sample validation set for 8 MOFs with entirely new structures. (a) Predicted catalysis probability versus experimental catalysis label. Data points are denoted as translucent hexagons to depict data density and colored by the classification correctness: correct (blue) and incorrect (red). (b) Confusion matrix of the classification results.

S3. Predictions for CoRE-MOFs

Table S6. Hierarchical screening workflow.

Step	Condition	Number of MOFs
Initial	Original data	11660
I	Data curation and filtering	9603
II	Predicted top-performing MOFs	1731
III	Predicted water stable top-performing MOFs	419

Table S7. Consistency analysis of top-performing MOFs across different thresholds relative to the 65th percentile.

Percentile	Threshold value ($\mu mol \cdot g^{-1} \cdot h^{-1}$)	Count	Overlap Count	Overlap Rate
60th	232.60	5	4	0.80
		100	79	0.79
70th	386.40	5	4	0.80
		100	63	0.63

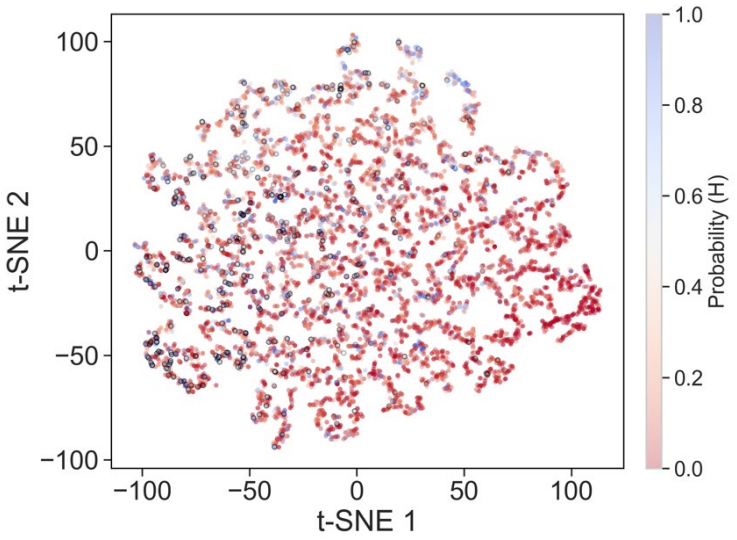

Fig. S11. t-SNE map of predictions by LightGBM classifier. H is the predicted probability of photocatalysis. The marked points represent 419 water stable top-performing MOFs.

S4. Featurization

Table S8. Descriptors.

Descriptor (Number)	Meaning	Number of featurizable MOFs
Geometric (8)	Largest cavity diameter (LCD, Å), pore limited diameter (PLD, Å), largest free path diameter (LFPD, Å)	100
	volumetric surface area (VSA, m^2/cm^3), gravimetric surface area (GSA, m^2/g), void fraction (VF), pore volume (PV, cm^3/g), density (kg/m^3)	100
Linker (173)	RDKit descriptors (6): molecular weight (mol/g), aromatic ring number, ...	100
	Molecular ACCess System (MACCS) fingerprint (167)	
Metal (4)	Atomic radius (Å), electronegativity, ionization energy (eV), atomic mass (g/mol)	100
AP-RDFs (565)	Electronegativity (EN-) (113), van der Waals volume (V-) (113), polarizability (α-) (113), hardness (H-) (113), mass (M-) (113)	100

Table S9. Experimental operating conditions.

Operating Condition	Remark
Temperature	Reaction temperature (°C)
Sacrificial agent	Reaction sacrificial agent
Catalyst	Catalyst content (mg)
pH	pH of reaction solution
Cocatalyst	Cocatalyst type and loading (wt%)
Substrate	Reaction substrate type and volume ratio (v/v)
Irradiation	Type of light irradiation
Lamp	Type of lamp
H_2 production rate	Amount of H_2 per mole of catalyst per unit of time ($\mu mol\ g^{-1}\ h^{-1}$)

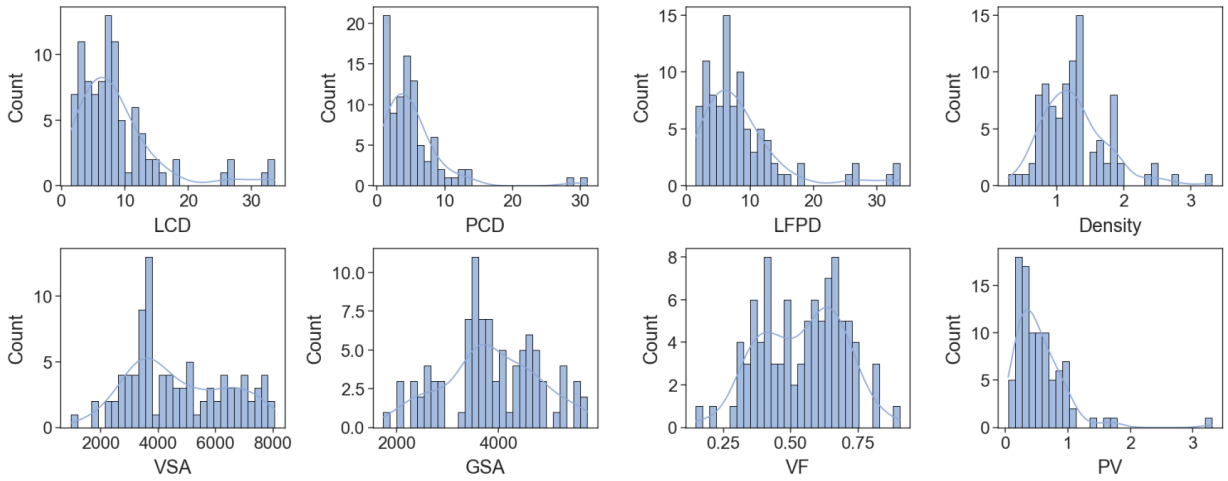

Fig. S12. Distributions of eight geometric descriptors in 92 collected MOFs. The lines indicate kernel density estimations.

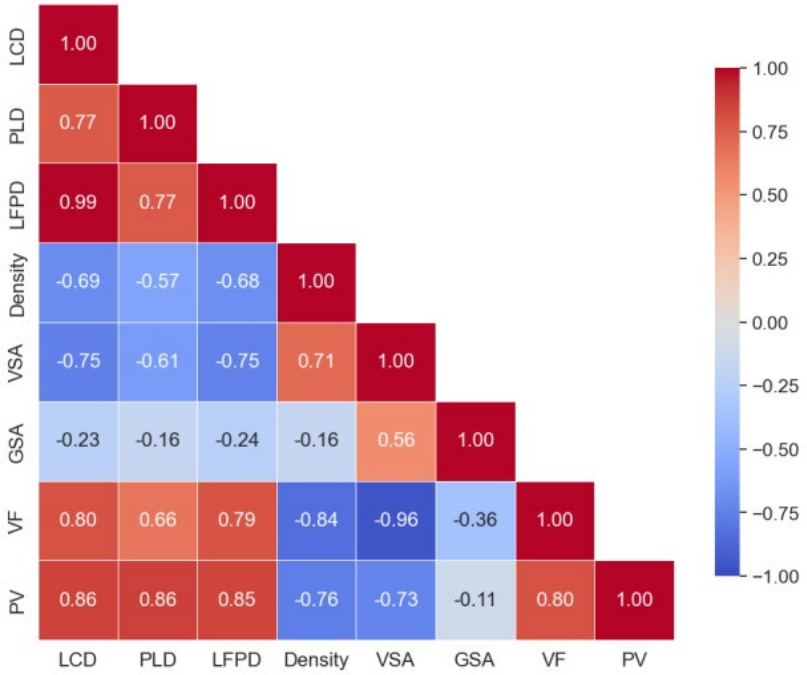

Fig. S13. Pearson correlation coefficient (PCC) matrix among eight geometric descriptors. The color scale indicates PCC value.

Table S10. Hyperparameter grids for six classifiers.

Classifier	Hyperparameter
Extremely Randomized Trees	'n_estimators': [100, 150, 200, 250], 'max_depth': [6, 7, 9, 10, 15, 20]
Gradient Boosting	'n_estimators': [100, 150, 200, 250], 'learning_rate': [0.02, 0.05, 0.1], 'max_depth': [3, 5, 7, 9, 11]
XGBoost	'n_estimators': [100, 150, 200, 250], 'max_depth': [10, 15], 'learning_rate': [0.2, 0.1, 0.15], 'subsample': [0.8, 1.0], 'colsample_bytree': [0.8, 1.0]
Random Forest	'n_estimators': [50,100,150,200], 'max_depth': [5,8,10,13]
CatBoost	'iterations': [100, 1000], 'depth': [3, 5, 7, 10], 'learning_rate': [0.02, 0.1, 0.15]
LightGBM	'n_estimators': [50, 100, 150], 'max_depth': [5, 6, 7, 10], 'learning_rate': [0.05, 0.1, 0.12]

Table S11. Optimized hyperparameters of LightGBM classifier.

Hyperparameter	Optimized Value
n_estimators	150
max_depth	10
learning_rate	0.12

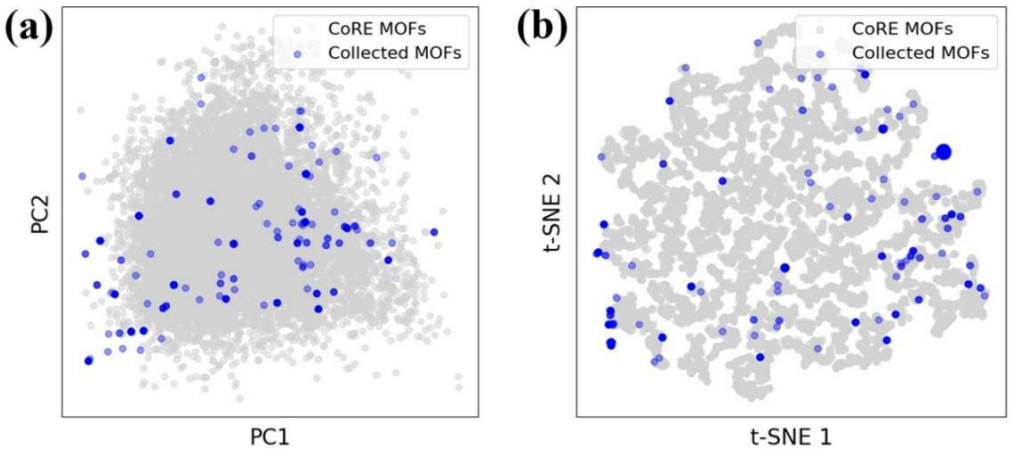

Fig. S14. (a) PCA and (b) t-SNE map. The gray points indicate the overall feature space of 9603 CoRE-MOFs and the blue points indicate 92 collected MOFs.

Table S12. Standardized input conditions used in screening.

Parameter Type	Descriptor	Standardization	Value
Continuous	Catalyst concentration	Median	10 mg
	Cocatalyst loading	Median	0 wt%
	Temperature	Median	25 °C
	pH	Median	7
Categorical	Sacrificial agent	Mode	TEOA
	Substrate	Mode	H_2O
	Cocatalyst	Mode	No
	Light	Mode	Vis
	Lamp_type	Mode	300 W Xe lamp