Electronic Supplementary Material (ESI) for EES Catalysis.

Predicting the rates of photocatalytic hydrogen evolution over cocatalyst-deposited TiO_2 using machine learning with active photon flux as unifying feature

Supplementary Information

Yousof Haghshenas,[a] Wei Ping Wong,[b] Denny Gunawan,[a] Alireza Khataee,[c] Ramazan Keyikoğlu,[c] Amir Razmjou,[d] Priyank Kumar,[a] Cui Ying Toe,[a,e] Hassan Masood,[a] Rose Amal,[a] Vidhyasaharan Sethu,[f] Wey Yang Teoh[a,b,*]

[a]School of Chemical Engineering, The University of New South Wales, NSW 2052, Australia

[b]Department of Chemical Engineering, Sustainable Process Engineering Centre (SPEC), Universiti Malaya, 50603 Kuala Lumpur, Malaysia

[c]Department of Environmental Engineering, Gebze Technical University, 41400 Gebze, Turkey

[d]Mineral Recovery Research Center (MRRC), School of Engineering, Edith Cowan University, Joondalup, Perth, WA, 6027, Australia

[e]School of Engineering, The University of Newcastle, Callaghan, New South Wales 2038, Australia

[f]School of Electrical Engineering and Telecommunications, The University of New South Wales, NSW 2052, Australia

*Corresponding author. E-mail: wy.teoh@um.edu.my

All data have been provided with code in the GitHub repository.

https://github.com/yousof96/ligh_intensity_paper

Acronyms

a: Lattice parameter

*AC: A*lcohol concentration

AcP: Active photons flux

AMW: Alcohol molecular weight

ApP: Apparent photon flux

AT: Alcohol type

ATI: Alcohol type indicator

CAN: Cocatalyst atomic number

CCT: Cocatalyst type

CEN: Cocatalyst electronegativity

CL: Cocatalyst loading

CWF: Cocatalyst work function

D: Cocatalyst dispersion

d_{anatase}*:* Size of anatase crystal

d_{rutile}*:* Size of anatase crystal

E_{g}*:* Bandgap

g_{catalyst}: Photocatalyst loading

HY: Hydrogen yield

LI: Light intensity

N_{active}*:* Total amount of cocatalyst active sites

RMSE: Root mean squared error

S_{m}*:* Surface area per atom

SSA: Specific surface area

$t_{\mathrm{calcination}}$*:* Calcination time

$T_{\mathrm{calcination}}$*:* Calcination temperature

V_{m}*:* atomic volume

Rate: Hydrogen evolution rate

X_{anatase}*:* Fraction of anatase phase

X_{rutile}*:* Fraction of rutile phase

Dataset analysis

Following features were extracted from the published papers: Hydrogen evolution rate, activity, cocatalyst type, cocatalyst loading, cocatalyst atomic number, cocatalyst electronegativity, cocatalyst work function, organic doner type, alcohol concentration, alcohol molecular weight, number of hydroxyl group, number of hydrogen atoms, number of alpha hydrogens, polarity, liquid solvent volume, reaction time, photocatalyst type, photocatalyst loading, photocatalyst bandgap, specific surface area, rutile phase fraction, anatase phase fraction, brookite phase fraction, rutile crystal size, anatase crystal size, space group number, space group symbol, synthesises method name, synthesises method description, calcination time, calcination temperature, activity promotion method, activity promotion method description, pore volume, pore diameter, lamp type, lamp power, light intensity, light intensity in Einstein, irradiation area, irradiation distance, photonic photon, apparent photon flux, active photon flux, photonic efficiency, and apparent quantum efficiency.

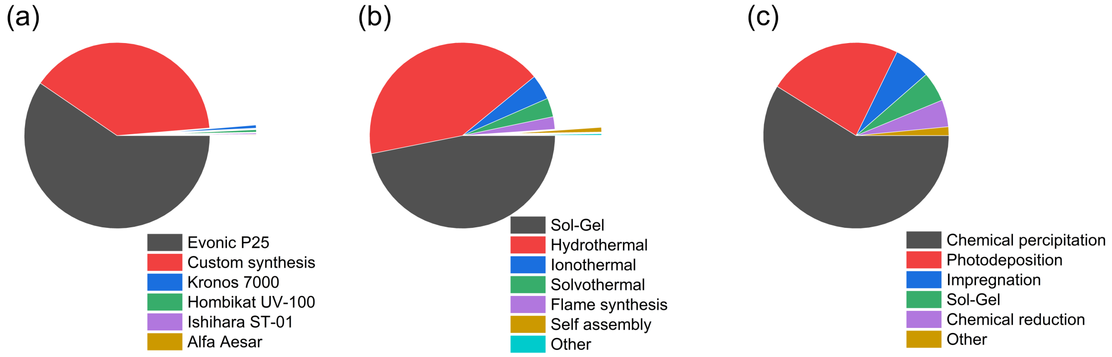


Figure S1. Breakdown of (a) types of TiO_2 photocatalysts, (b) synthesis methods of TiO_2, (c) cocatalyst deposition methods among the 946 entries of the literature-extracted dataset.

To prepare the database for statistical analysis, A set of techniques including missing value replacement, categorical data handling, scaling, and feature selection must apply to data. In this study, data from all designed experiments had no problem with missing values but extracted data from literature was highly inconsistent and 42 percent of information was not reported. To avoid computational issues, samples with missing information were removed. After dealing with categorical features, data was scaled to range between [-1, 1] to avoid scaling effects for further analysis.

Feature Engineering

Calculation of apparent and active photon fluxes

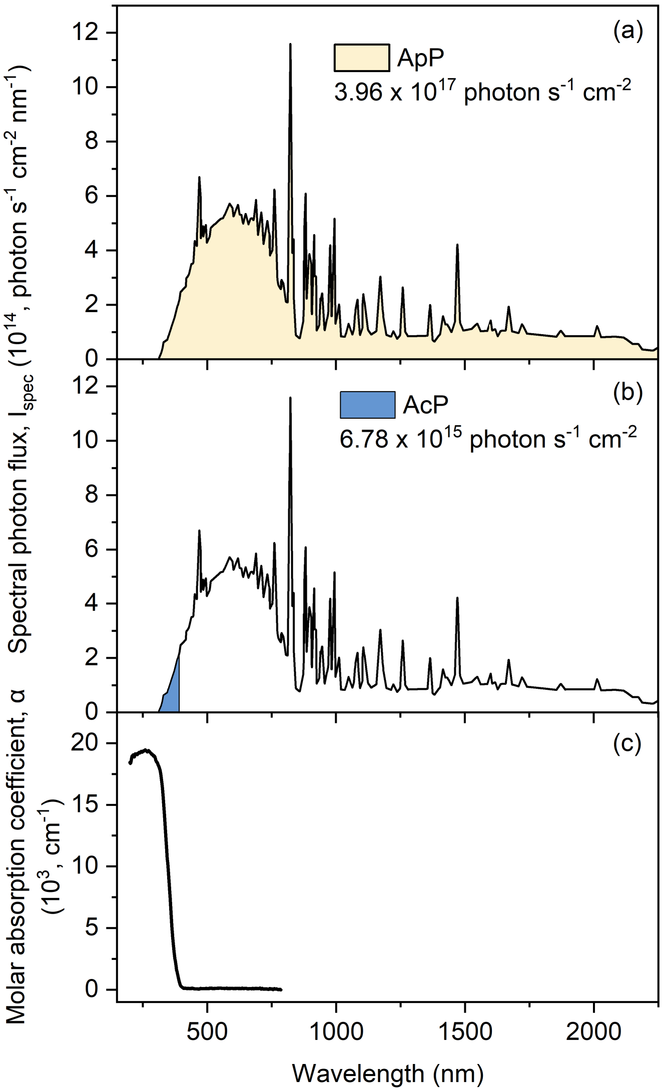


Figure S2. Simulated solar A.M. 1.5 G (Newport, Sol3A) spectrum at 1 sun (100 mW cm^{-2}) intensity with the shaded areas representing (a) apparent photon flux, $ApP(\lambda, I)$ and (b) the active photon flux, $AcP(\lambda, I)$, that can be absorbed by the TiO_2. (c) The molar absorption coefficient of TiO_2 showing the photoabsorption range for the calculation of AcP.

For any defined light source of known spectrum and intensity, the apparent photon flux, $ApP(\lambda, I)$ can be calculated by integrating the whole light spectrum. For example, the $ApP(\lambda, I)$ of simulated solar spectrum (A.M. 1.5 G, 1 sun):

$$ApP(\lambda, I) = \int_0^{\infty} \frac{I_{spec}}{E_\lambda} d\lambda = 3.96 \times 10^{17}\, photon\, s^{-1}\, cm^{-2} \qquad \textbf{Eq. S1}$$

Besides, the available active photon flux, $AcP(\lambda, I)$, can be calculated by integrating at and below the absorption threshold. For example, in the case of anatase TiO_2 suspension of bandgap 3.2 eV (absorption threshold 388 nm) irradiated under simulated solar spectrum (A.M. 1.5 G, 1 sun):

$$AcP(\lambda, I) = \int_0^{388\, nm} \frac{I_{spec}}{E_\lambda} d\lambda = 6.78 \times 10^{15}\, photon\, s^{-1}\, cm^{-2} \qquad \textbf{Eq. S1}$$

where λ is wavelength in nm, I_{spec} is spectral irradiance in mW cm^{-2} nm^{-1}, E_λ is the energy of a photon at the specific wavelength in mJ.

The total photon flux absorbed by the TiO_2, or active photon flux (AcP) can be calculated from

$$N_{Abs}(\lambda, I, \alpha, t) = \int_0^{388\, nm} \frac{I_{spec} \times (1 - 10^{-A})}{E_\lambda} d\lambda \qquad \textbf{Eq. S2}$$

where *A* is the absorbance of TiO_2 given by,

$$A\,(\lambda, \alpha, t) = \frac{\alpha(\lambda)(cm^{-1}) \times C\,(\mathrm{g\,L^{-1}}) \times t_{pen}(cm)}{\rho\,(g\, cm^{-3}) \times 1000} \qquad \textbf{Eq. S3}$$

where α is the molar absorption coefficient of TiO_2 (see Figure S2b), *C* is the concentration of TiO_2, t_{pen} is the light penetration depth into the suspension, ρ is the bulk density of TiO_2. By integrating Eq. S2 at C = 1 g L^{-1} of TiO_2, $\rho = 4.26\, g\, cm^{-3}$ and at $t_{pen} = 4.3\, cm$, the $N_{Abs}(\lambda, I)$ was calculated to be 6.71 x 10^{15} photon s^{-1} cm^{-2}.

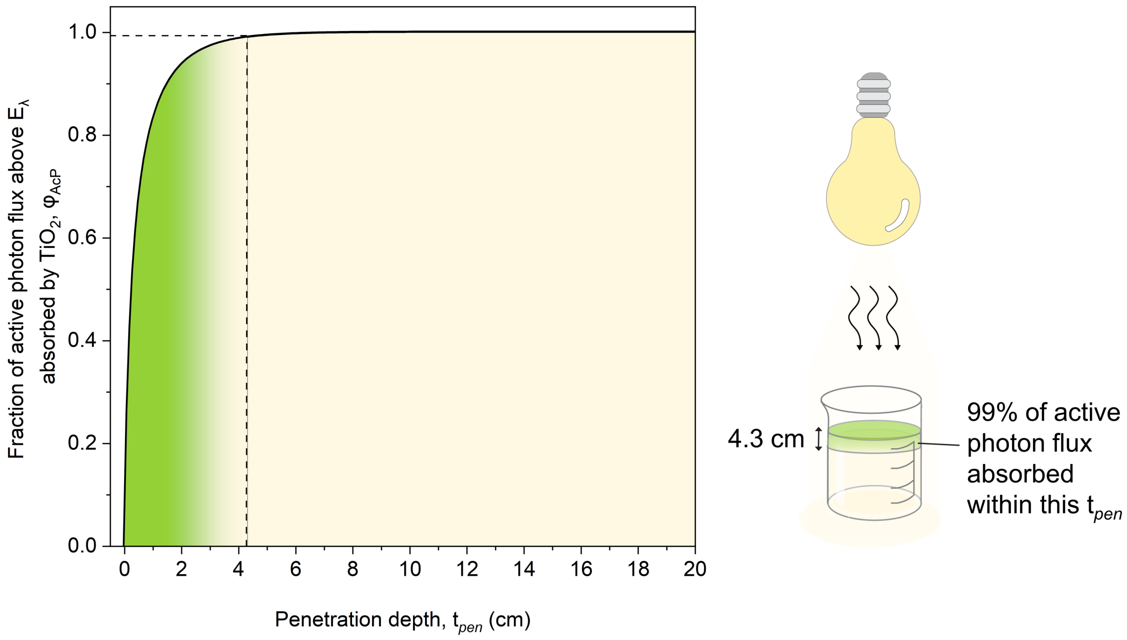


Figure S3. Fraction of active photon flux above E_λ absorbed by TiO_2 along the penetration depth. The following parameters were used in the calculation: the concentration of TiO_2 = 1 g L^{-1} and the density of TiO_2 = 4.26 g cm^{-3}

The fraction of photon absorbed to that available at $t_{pen} = 4.3\ cm$ can thus be given as,

$$\varphi_{AcP} = \frac{N_{Abs}(\lambda, I, \alpha, t)}{AcP(\lambda, I)} = \frac{6.71\ \times 10^{15}\ photon\ s^{-1}\ cm^{-2}}{6.78\ \times 10^{15}\ photon\ s^{-1}\ cm^{-2}} = 0.99 \qquad \textbf{Eq. S4}$$

From Figure S3, it can be seen that 99% of the available active photons are absorbed within 4.3 cm of the TiO_2 suspension, thus representing a typically small fraction of the suspension volume near the surface where the photocatalyst can be activated.

Categorical feature representation (CT and AT)

The effect of different types of cocatalysts and alcohol in the machine learning model was considered. These features are categorical and need encodings before using for modeling. Common methods of encoding such as one hot encoding are not suitable. Since various types of cocatalysts and alcohol are available in the current database and if one hot encoding is used, the number of features increases significantly which is not consistent with a limited number of data. Here cocatalyst atomic number (CAN), cocatalyst electronegativity (CEN), and cocatalyst work function (CWF) were used instead of categorical feature for cocatalyst type and alcohol molecular weight (AMW) instead of categorical feature for alcohol type.

Table S1 shows the initial statistical information of the extracted data from the literature.

Table S1. List of all features from literature-extracted dataset and their statistical ranges. A high standard deviation (std dev) is an indicator of a wide coverage range of input data.

Metric	Count	Mean	Std	Min	Q1	Q2	Q3	Max	Skewness	Kurtosis	Entropy 1	Entropy 2	Gini index	Excluding criteria
Photocatalyst features														
Pore volume	275	-	-	-	-	-	-	-	7.2	58.2	6.4	2.98	0.97	4, 6
Pore diameter	174	10.64	8.39	1.58	5.41	6.76	14.4	35.2	3.39	10.82	4.75	2.73	0.93	4, 6
SSA	833	71.6	66.1	1	43.3	47	75.8	600	-5.43	37.4	8.4	3.47	0.85	
E_g	726	31.4	0.11	2.25	3.1	3.15	3.2	3.54	-4.35	32.1	8.41	2.3	0.71	
$X_{anatase}$	879	0.136	0.19	0	0	0.1431	0.16	1	8.38	93.1	8.29	1.9	0.66	
$T_{calcination}$	934	374	218	0	300	400	500	1000	-0.23	2.04	8.29	1.68	0.62	
Space group number	781	-	-	-	-	-	-	-	1.36	0.34	1.75	2.07	0.6	1
Space group symbol	781	-	-	-	-	-	-	-	1.36	0.34	1.75	2.07	0.6	1
Synthesis method	927	-	-	-	-	-	-	-	1.93	2.42	1.63	1.71	0.57	1
d_{rutile}	506	54.4	70.1	0	43	50	50	1090	7.73	63.5	8.33	1.54	0.53	
Photocatalyst	924	-	-	-	-	-	-	-	0.89	-0.98	1.07	1.38	0.49	1, 3
Brookite phase fraction	37	0.78	0.4	0	1	1	1	1	0.57	-1.5	0.87	1.82	0.34	3, 6
$d_{anatase}$	763	23.8	16.2	3.9	13.6	25	25	249	11.5	139	8.28	0.98	0.31	3
$t_{calcination}$	930	2.41	4.27	0	2	2	2	80	-2.29	10.8	8.3	0.44	0.14	3
Cocatalyst features														
CT	946	-	-	-	-	-	-	-	0.93	-0.52	2.55	2.36	0.79	1
Promotion method	790	-	-	-	-	-	-	-	0.75	-1.11	2.31	2.42	0.76	1
CL	945	1.14×10^{-2}	1.33×10^{-2}	0	5×10^{-3}	1×10^{-2}	2×10^{-2}	0.2	3.56	3.08	8.02	2.41	0.75	
CAN	945	52.2	30.4	0	28	78	79	79	0.193	-1.53	8.17	2.05	0.71	
CEN	945	1.94	0.82	0	1.91	2.28	2.54	2.68	-2.38	7.41	8.33	2.05	0.71	
CWF	945	4.35	1.83	0	4.53	5.06	5.54	5.54	-4.16	16.2	8.34	1.82	0.67	
Light source features														
Lamp type	927	-	-	-	-	-	-	-	3.84	15.34	3.56	2.25	0.84	1, 4
Lamp power	904	-	-	-	-	-	-	-	2.08	2.89	3.14	2.19	0.84	1, 4
AcP	499	1.63×10^{17}	1.21×10^{18}	1.70×10^{16}	6.22×10^{16}	6.32×10^{16}	6.32×10^{16}	1.92×10^{19}	-2.89	6.75	8.4	2.16	0.69	
LI in Einstein	43	2.06×10^{-6}	2.01×10^{-6}	6×10^{-8}	1.67×10^{-7}	3.79×10^{-7}	4.2×10^{-6}	4.2×10^{-6}	0.38	-1.42	1.75	2.9	0.66	6
ApP	523	1.06×10^{19}	1.65×10^{20}	2.28×10^{16}	2.4×10^{17}	2.4×10^{17}	2.4×10^{17}	2.67×10^{21}	2.8	5.95	1.26	1.21	0.35	3, 4
LI	830	14.6	32.1	2.2	6.5	6.5	6.5	250	-2.88	9.22	8.39	0.89	0.33	3, 4
Irradiation area	504	11.95	3.21	2.85	12.88	12.88	12.88	20	0.68	-1.5	0.59	1.24	0.21	3, 4

Organic substrate features													
Polarity	10	0.62	0.24	0.11	0.55	0.63	0.78	1	0	-3	3.32	3.32	0.9	6
No. of alpha hydrogens	10	2	1.63	0	1	2	2.75	5	0.62	-0.96	2.44	3.14	0.88	6
AC	933	0.28	0.29	0	0.1	0.1	0.4	1	0.59	-1.26	7.73	3.67	0.88	
AMW	946	51.1	25.9	18	32	46.1	62.1	149	0.82	-0.52	8.32	2.05	0.71	
AT	945	-	-	-	-	-	-	-	1.52	0.97	2.1	2.02	0.7	1
ATI	927	0.11	1.03	-2.15	-0.17	0.21	0.96	0.96	1.24	0.14	2.09	2.19	0.7	
No. of hydrogen atoms	10	6.8	2.52	2	6	7	8	10	0	-1.75	2.17	3.1	0.7	6
No. of hydroxyl groups	10	1.2	0.78	0	1	1	1	3	1.15	-0.66	1.35	2.25	0.48	6
Reaction features													
HY	923	1.24×10^{4}	2.25×10^{4}	0	1800	8000	1.85×10^{4}	0.554	0.35	-0.25	8.08	7.63	0.99	2
Rate	885	4.4×10^{2}	1.2×10^{3}	0	57.8	1.31×10^{2}	3.3×10^{2}	2.77×10^{3}	2.14	4.32	7.62	3.27	0.99	
Photonic efficiency	183	0.27	0.19	1.82×10^{-4}	0.12	0.26	0.38	1.04	3.18	8.15	7.37	3.3	0.99	2
AQE	184	7.26×10^{-2}	5.19×10^{-2}	0	3.12×10^{-2}	6.9×10^{-2}	0.1	0.27	2.25	4.34	7.15	3.27	0.99	2
Solution volume	781	0.34	2	0.01	0.02	0.07	0.2	18.75	3.47	11.8	2.97	2.15	0.77	4
Reaction time	159	-	-	-	-	-	-	-	2.11	3.6	2.57	2.47	0.75	4
$g_{catalyst}$	904	0.1	0.2	5×10^{-3}	6.5×10^{-3}	1.2×10^{-2}	0.1	2	3.65	12.7	2.65	1.94	0.71	4

Excluding criteria: (1) Descriptive feature, (2) Need Rate for calculations, (3) Gini index lower than 0.5, (4) High positive kurtosis/high skewness, (5) Low standard deviation, (6) High missing value

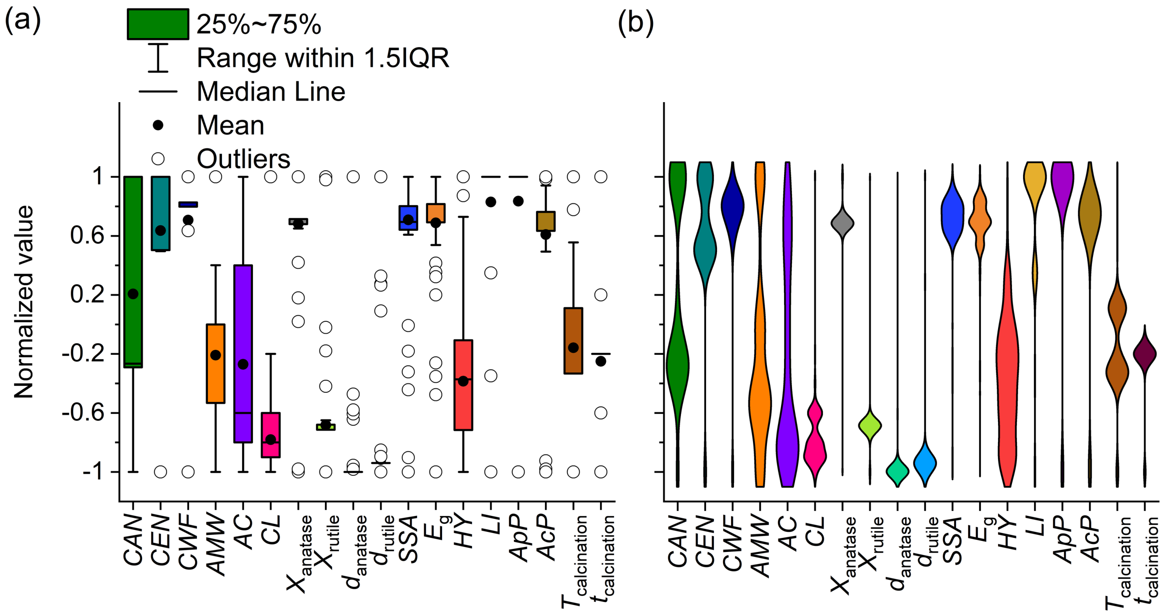


Figure S4. Statistical analysis of numerical features in the literature-extracted dataset using (a) box plot and (b) distribution. The high number of outliers and skewed distribution shows inconsistency in the dataset.

To narrow the selection of features from Table S1, following criteria were considered. First, we excluded all descriptive features (i.e., lamp type, synthesis method description, promotion methods, and so on). Next, all features that require Rate for calculations were removed since Rate is the target and supposed to be unknown during prediction (i.e., AQE, photonic efficiency). Features with Gini Index lower than 0.5 were excluded (i.e., irradiation area). We also removed features with a lot of missing values (i.e., LI). For the features belonging to the same group such as CWF, CAN, and CEN that represent the cocatalyst type information, only one representative feature is considered for modelling as shown in Figure S5. In the case of alcohol type, ATI is formulated as a representative feature. Because the ATI was calculated using related features of alcohol, e.g., AMW, No. of alpha hydrogens, No. of hydroxyl groups, these terms naturally become redundant. The advantage of ATI, compared to individual features that only partially describe the molecule itself, is better standard deviation, lower skewness and kurtosis (as elaborated in the Alcohol type indicator section). For features within the same group, the one with more negative kurtosis and/or higher standard deviation was selected. Features with large positive kurtosis, e.g., pore diameter and pole volume, were excluded due to the large number of outliers. The outliers of selected features are readily shown in the Figure S4. As shown in Figure S4, the distribution of the extracted data from the literature includes many outliers which made challenges in the model training step. Some features (e.g., CL and AC) have a normal distribution, while others (e.g., d_{anatase}, d_{rutile}, and SSA) include skewed distribution with many outliers.

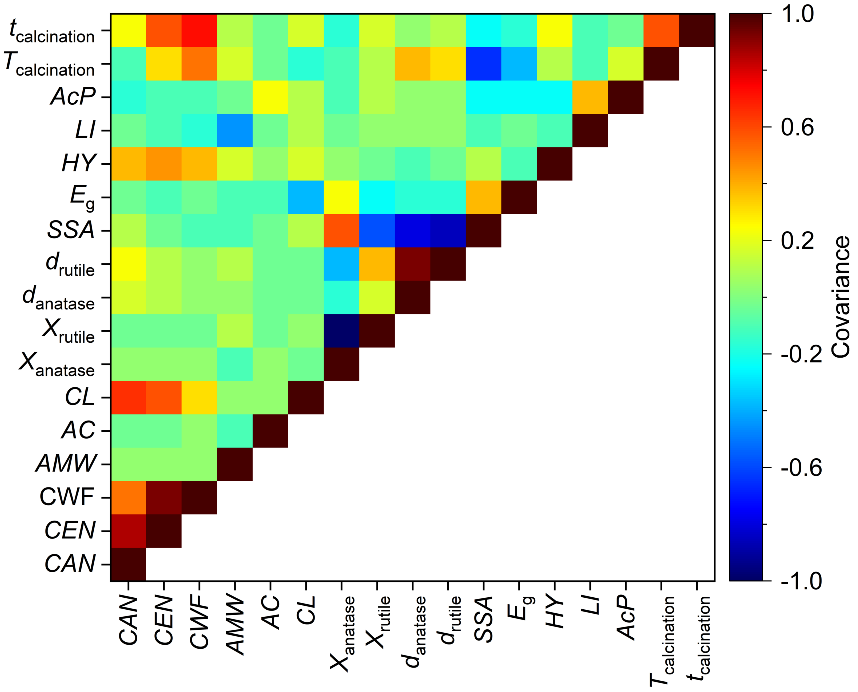


Figure S5. The covariance (linear relationship) of the numerical feature of the literature-extracted dataset. Extreme colors show a strong correlation between every two features.

The correlation coefficient of all numerical features of the extracted data from the literature is shown in Figure S5. A fair correlation between all features is visible from the figure which indicates the independency of the feature and the highly non-linear nature of the problem. Figure S6 shows the fuzzy curves results, where from fuzzy surfaces analysis the removal of AMW and LI was recommended.

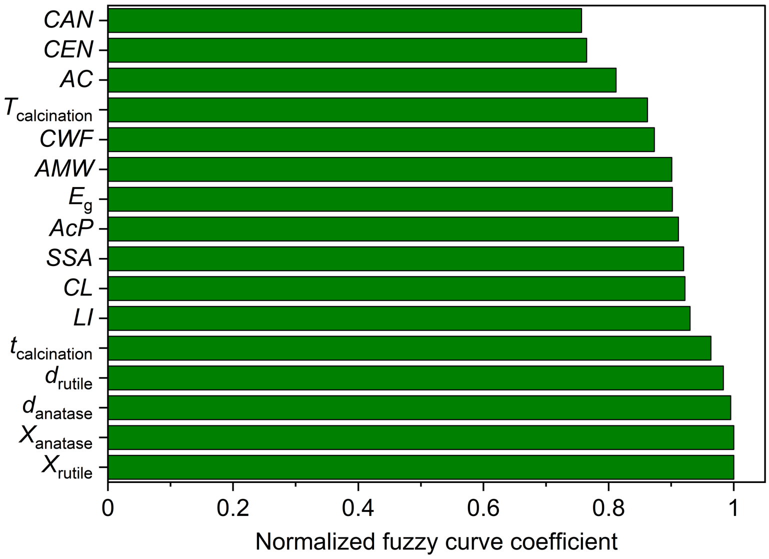
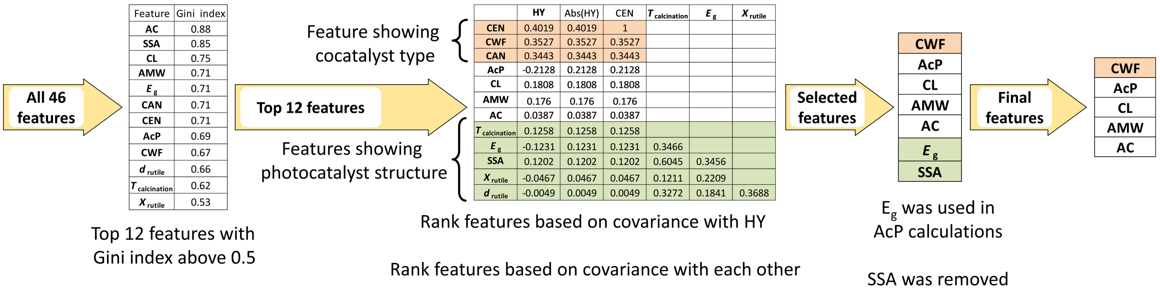


Figure S6. Normalized fuzzy curve coefficient for the numerical features of the literature-extracted dataset. The high fuzzy curve coefficient shows the nonlinear relationship of feature with Rate.



All 46 features
Top 12 features
Feature showing cocatalyst type
Features showing photocatalyst structure
Selected features
Final features

Feature	Gini index
AC	0.88
SSA	0.85
CL	0.75
AMW	0.71
E_g	0.71
CAN	0.71
CEN	0.71
AcP	0.69
CWF	0.67
d_{rutile}	0.66
$T_{calcination}$	0.62
X_{rutile}	0.53

Top 12 features with Gini index above 0.5

	HY	Abs(HY)	CEN	$T_{calcination}$	E_g	X_{rutile}
CEN	0.4019	0.4019	1			
CWF	0.3527	0.3527	0.3527			
CAN	0.3443	0.3443	0.3443			
AcP	-0.2128	0.2128	0.2128			
CL	0.1808	0.1808	0.1808			
AMW	0.176	0.176	0.176			
AC	0.0387	0.0387	0.0387			
$T_{calcination}$	0.1258	0.1258	0.1258			
E_g	-0.1231	0.1231	0.1231	0.3466		
SSA	0.1202	0.1202	0.1202	0.6045	0.3456	
X_{rutile}	-0.0467	0.0467	0.0467	0.1211	0.2209	
d_{rutile}	-0.0049	0.0049	0.0049	0.3272	0.1841	0.3688

Rank features based on covariance with HY

Rank features based on covariance with each other

Selected features
CWF
AcP
CL
AMW
AC
E_g
SSA

E_g was used in AcP calculations

SSA was removed

Final features
CWF
AcP
CL
AMW
AC

Figure S7. Features selection based on the Gini index and covariance of input features with HER rate. Highlighted rows in orange are features related to the type of the cocatalyst and while green rows are related to the TiO_2 structure.

Due to high number of missing information for Rate, we performed features selection analysis using HY, which has a high linear correlation with Rate. Doing so allows us to access a larger dataset despite the high interchangeability of the two. Features X_{rutile}, $X_{anatase}$, $d_{anatase}$ and d_{rutile} are related to the structural information of photocatalysts, while $T_{calcination}$ and $t_{calcination}$ directly impacts the photocatalysts structure. Since only TiO_2 photocatalyst was used and E_g is an abstract representation of the structural effects of photocatalyst, X_{rutile}, $X_{anatase}$, $d_{anatase}$, d_{rutile}, $T_{calcination}$ and $t_{calcination}$ can be ignored against E_g. Moreover, these features are highly linearly correlated in binary form (e.g., $d_{anatase}$ with d_{rutile}, $X_{anatase}$ with X_{rutile}, and

$T_{\mathrm{calcination}}$ with $t_{\mathrm{calcination}}$). Hence, there exists redundancies if all are added to the model input. In addition, based on statistical information on these features in Table S1 and their box plot in Figure S4, substantial amount of outliers for these features were seen. Hence, their distributions are not very helpful as specific descriptors compared with E_g. Finally, since E_g is embedded in the AcP calculations, it does not need to be an explicit feature.

Among descriptors for the type of cocatalyst, CEN, CAN, and CWF were used. Based on the statistical information (Table S1) a better distribution for CAN and CWF in comparison with CEN was seen. Moreover, in the box plot (Figure S4a), a better distribution for CAN in comparison with CEN was seen. Fuzzy curve analysis (Figure S6) did not show significant differences between CEN and CAN. Since CWF is a direct quantity describing the all-important Schottky barrier that dictates the efficiencies of charge separation, it is the favoured choice among these three features (as the cocatalyst type indicator). The effects of CAN in comparison with CWF were checked as well.

Alcohol type indicator

Figure S8 shows the ATI value for each alcohol structure. In Figure S8, the x-axis shows the smiles of alcohol, and the y-axis shows the ATI value. There are two advantages of ATI in comparison with single alcohol property as a feature, first, ATI gives different values for different alcohol and different structures for example AMW is the same for 1-propanol (CCCO) and 2-propanol (isopropanol: CCOC) but ATI is not. Second, there is an overall trend in ATI which can be used for qualitative analysis. The complete data of PCA analysis for ATI was provided with other databases.

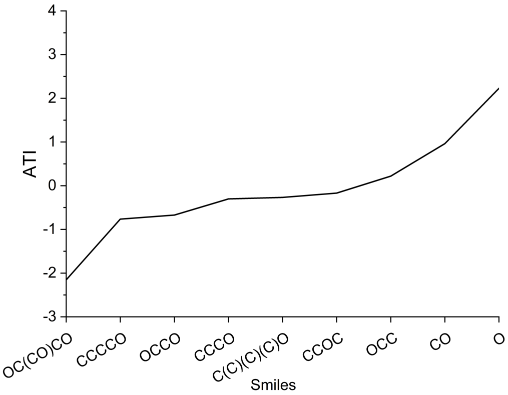


Figure S8. ATI values for different organic substances. Smiles were used for the x-axis since ATI is sensitive to Smiles.

Forward Selection

In this process, different models are trained multiple times starting with one feature and adding one feature in every step until the last step which is a model trained with all features. In Figure S9a, the results of the forward selection are shown. In Figure S9a, the y-axis shows the added feature, and the x-axis shows the R squared of the model on the same test data. The first model was trained using AcP, for the second model, CL was added to AcP for training, and so on. It is obvious that after adding more features, the model must have a higher R squared but this trend is not always incremental. As it is clear in Figure 9a, adding some features (e.g., SSA) did not make a significant change in the model metric while adding others (e.g., CL) showed distinguished improvement in the model metric. These results can tell the effects of adding a feature. Here it could be concluded that adding SSA was not necessary from the data-driven point of view while having CL was required.

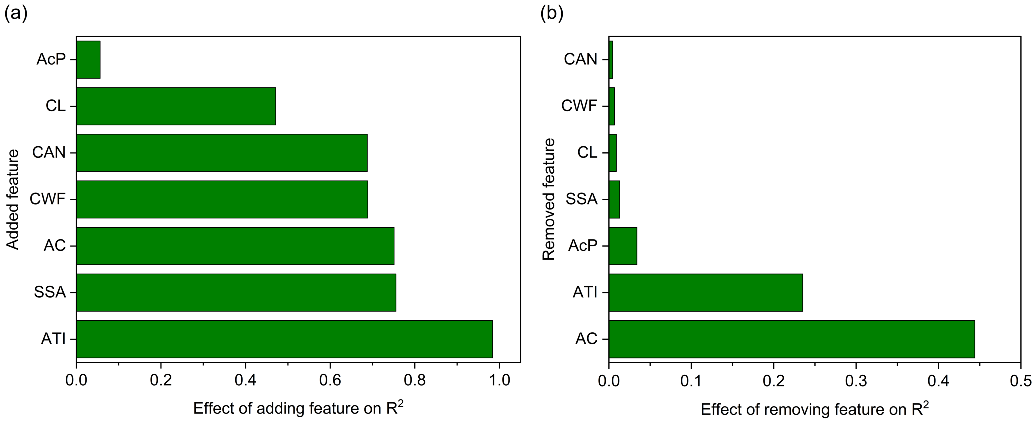


Figure S9. Share of feature contribution using (a) forward selection and (b) backward elimination analysis. This result shows the nonlinear relationships between input features and Rate.

Backward Elimination

This method is against forwarding selection. Here the first model is trained with all features and in every step, one feature is eliminated from the input feature set, and the model is trained with a new feature set. This process continues until multiple models are trained in a way that each of them missed only one feature. In Figure S9b results of backward elimination are shown. In this plot, the y-axis shows the name of the feature that was missed in model training and the x-axis shows the difference of R squared of the model with all features and the model with the missed feature for the same test data. As it is clear in Figure S9b, removing AC from the feature set had the highest effect on the model metric and makes it less accurate. Removing CWF had a slightly more adverse effect on the model metric in comparison with CAN. Moreover, removing SSA did not make a significant change in the model metrics.

It must be noted that forward selection and backward elimination results depend on the model performance. However, when the performance of the model is acceptable from a statistical point of view ($R^2 \sim 0.91$ on test data), the results of forwarding selection and backward elimination under the supervision of expert knowledge are reliable.

In conclusion, SSA and CAN can be removed. For CAN, it was shown that CWF had enough information about the type of cocatalyst for the rate/activity model, and by summing up the results of statistical decisions with forward selection and backward elimination the removal of CAN can be verified. For SSA, statistical information gave some suggestions to remove it but not enough. However, in the forward selection and backward elimination it was shown that the removal of SSA cannot make any adverse effects. These were reasons for the data-driven part. However, from the photocatalyst domain knowledge part, for titania-loaded materials, the cocatalyst loading amount has enough information about the surface of the material. Therefore, if there are any differences in surface, it is included in the CL. By these justifications, removing SSA was reliable from the photocatalyst domain knowledge point of view.

Active learning

As was described in the main manuscript, since a limited database was used, data splitting must be done smartly. For this reason, an active learning approach was used. In this approach, training started with 10 randomly selected samples, and a model was trained using this data. The rest of the data was assumed as the test dataset. A random forest and a Gaussian process regression with a combination of radial basis function (RBF) and the constant kernel were used as models for active learning. A 5-fold cross-validation approach was used to avoid overfitting. For the sake of consistency with the main predictor model, AC, ATI, CL, CWF, and AcP were used as the input features and Rate as the target feature. After training the first model, it was used to predict the Rate for the rest of the data (test dataset). After calculating the error and variance of prediction, a trade-off between exploration and exploitation strategies was used to select the next 10 samples to be added to the training dataset for the next iteration of the model training. It was shown that the trade-off between exploration and exploitation looks better in comparison with only focusing on them separately.[1] This process continued until covering ~85 percent of data as the training set and 15 percent as the test set. At the end of the active learning iterative process, 420 samples as the training set and 69 samples as the test were selected. These datasets were used to train and optimize the main Rate predictor model. Figure S10 shows the results of the active learning process. In this figure, the x-axis shows the iteration number in the active learning process and the y-axis shows the RMSE of each model on the test set for each iteration. In this figure, it is obvious that the model was improved over each iteration, and the data with the most information was selected.

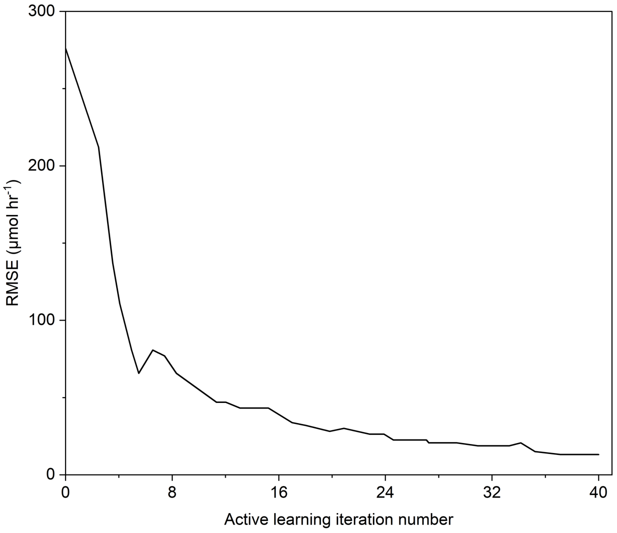


Figure S10. Changes in the RMSE of the random forest model during active learning iterations. The RMSE show the accuracy of the random forest model.

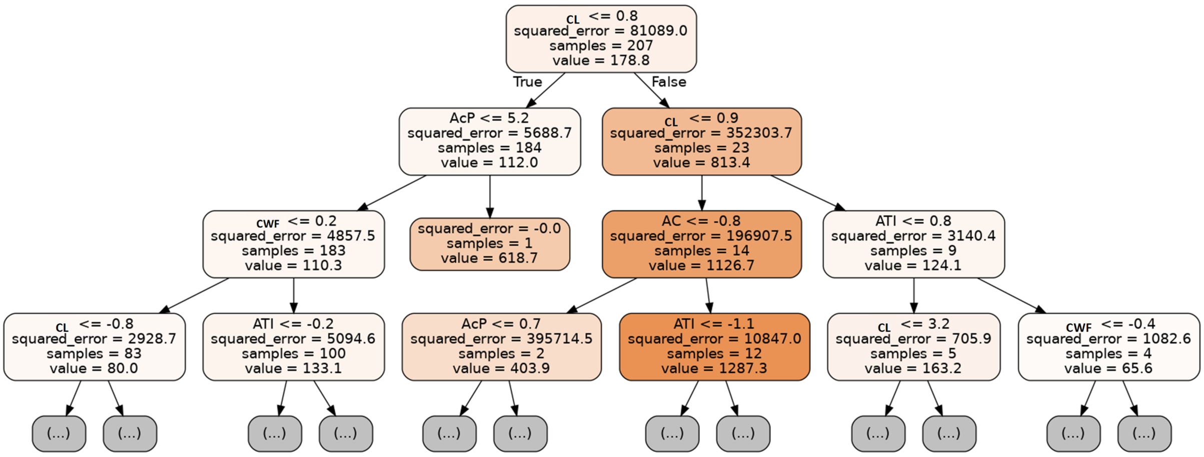


Figure S11. Visualization of the first tree in the random forest model used for the active learning process. The sequence of features in the tree shows their contributions on the model prediction during active learning iterations. the higher the feature is in the tree, the more contribution in the prediction.

It should be noted that in the active learning process, the compatibility of the model with domain knowledge must be checked. For this reason, the first tree of the random forest model in active learning was visualized to see the relative importance of each feature in decision-making using trees in Figure S11. As it is clear in this figure, CL is at the top of the tree which shows the high importance of this feature in the prediction of Rate. After CL, AcP, and CWF. This sequence of features is almost compatible with domain knowledge of the photocatalyst reaction and reported literature. It should be noted that the models used for active learning were not optimized, and the purpose of those models was not to achieve the highest accuracy in the prediction of Rate.

Training main Rate predictor model with TPOT

So far, the training and test datasets were successfully selected. Since a small database was used, ANN was not considered in the model selection. Even with ignoring artificial neural networks, many models can be used and each of them has multiple hyperparameters that should be tuned. Therefore, choosing the best model with the best and most optimized hyperparameters could be a challenging task. To achieve the best-optimized model, TPOT was used. As was described in the main manuscript, TPOT searches for the best model with optimized hyperparameters using a genetic algorithm. In the TPOT pipeline, 5 generations with a population of 100 candidates were used which were repeated 100 time for each combination of hyperparameters. To avoid overfitting, 10-times repeated 10-fold cross-validation was used during the TPOT training process. Details of model parameters were shown in Table S2, and details of model error were shown in Table S3. Detailed results of Rate prediction with an optimized model were shown in Figure S12. Figure S12a and b shows the results for the training dataset and Figure S12c, and d shows the results for the test dataset.

Table S2. Details of model parameters optimized in TPOT process.

Model	**Parameters**

Random Forest Regressor	Number of estimators = 100
	Max_features = 0.75
	Min_samples_leaf = 2
	Min_samples_split =10

RidgeCV	Alphas = array ([0.1, 1., 10.])

KNeighborsRegressor	N_neighbors = 8
	Weights = distance
	Metric = minkowski
	Leaf size = 30

Detailed error analysis for every cocatalyst showed that all the samples with Pt cocatalyst fell into the training dataset which slightly led the model toward accurate predictions for Pt cocatalyst within the training dataset. However, for every cocatalyst in the training and test dataset, error was normally distributed with a mean close to zero across the whole range of HER, for Au, Cu, and Ni in the test dataset, the model tends to slightly underestimate the Rate for very low values of Rate and overestimate for higher values of Rate. This skewness was affected by the abnormal distribution of values of Rate in the test dataset in which 75% of Rates were less than the average of all samples.

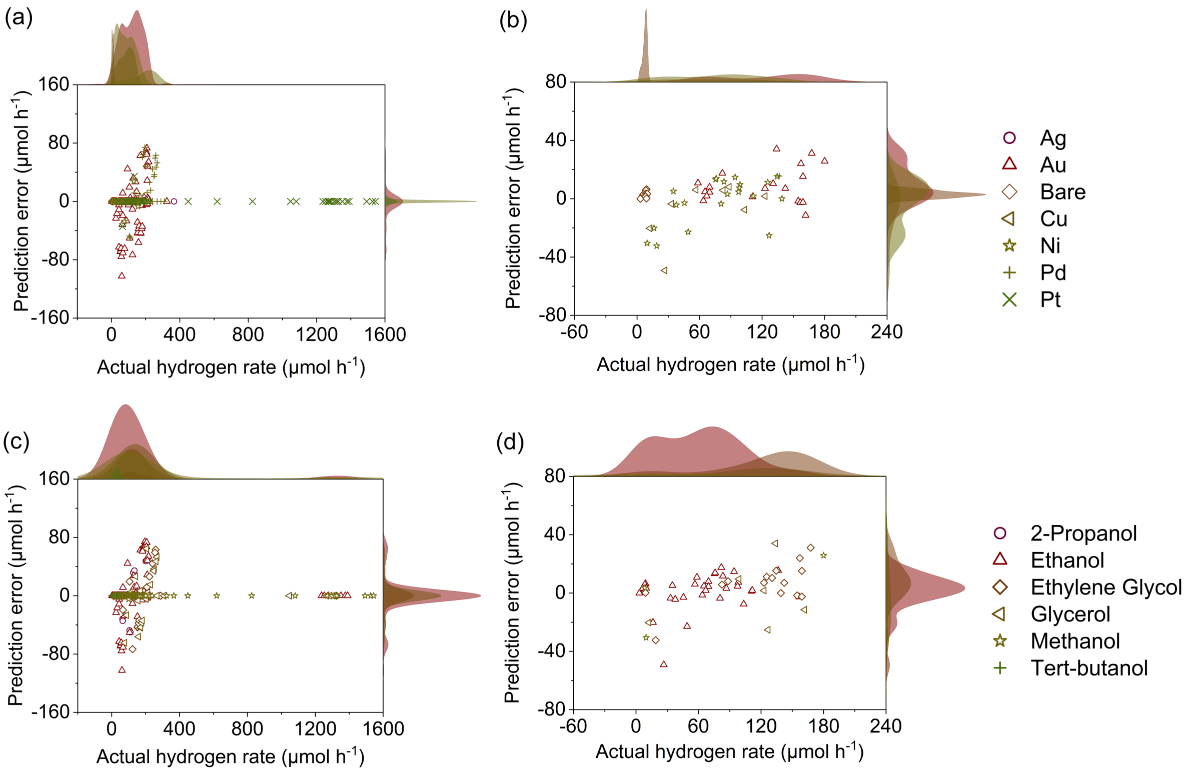


Figure S12. Absolute error analysis for (a) cocatalysts in training dataset, (b) cocatalysts in test dataset, (c) organic substrates in training dataset, and (d) organic substrates in test dataset.

The same analysis for organic substrates showed that a higher portion of samples with methanol fell into the training dataset compared to the test dataset which described the lower error for those samples in the latter. On the other hand, multiple samples with ethanol fell into the test dataset which decreased the number of observations that the model experienced during training for ethanol. In addition, in the training dataset for every organic substrate, the error was normally distributed with a mean close to zero which indicates there was no bias in the prediction error within any single group of cocatalysts and organic substrates. Although, for Ethylene glycol and Glycerol organic substrates, the model tends to slightly underestimate the Rate for very low values of Rate and overestimate for higher values of Rate.

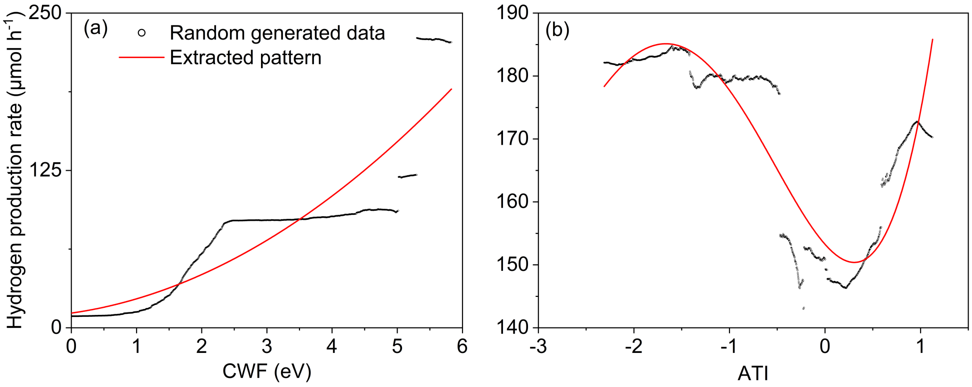


Figure S13. The effect of (a) CWF and (b) type of organic substrate on average predicted hydrogen rate by the model for random generated data

The detailed error analysis indicated that the training strategy and the use of active learning to split data into training and test sets in a smart way decreased the adverse effects of the imbalanced distribution of data in the initial dataset. However, it was not possible to completely overcome issues with the imbalance distribution of data without missing information from some experiments, the currently developed model reflects the initial distribution of data within its prediction while having a reasonably accurate prediction for rare experimental conditions.

Importance analysis of CWF (Figure S13a) and alcohol type (Figure S13b) is shown that an incremental relationship between CWF and the average predicted hydrogen rate is clear. Moreover, ATI has a nonlinear impact on the average predicted hydrogen rate.

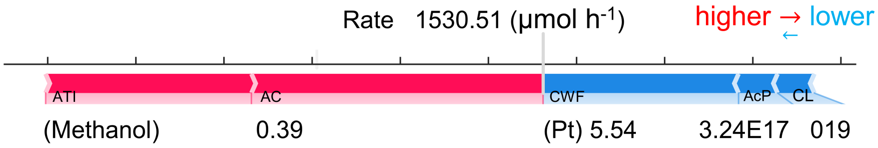


Figure S14. An example of local model interpretation using SHAP analysis for Methanol reforming on Pt-TiO$_2$. The red colour of AC shows that the value of the related feature had a positive impact on the hydrogen rate and pushed it to a higher value. However, the blue colour shows the opposite effect for CL (CWF: cocatalyst work function, AcP: active photon flux, CL: cocatalyst loading, AC: alcohol concentration)

Local interpretation using SHAP analysis for a random sample entry showed the experimental design of AC and ATI in that sample had negative effects while AcP, CL, and CWF had positive effects on the Rate. Moreover, AC contributed higher than other features in the prediction whereas CL was the least contributive feature in the value of Rate for that specific sample (Figure S14). This analysis provided a qualitative analysis for every sample using the developed model.

Table S3. The top 5 candidates of search space for Cond. 1 Bayesian optimization (2-Propanol). The first row is the experimentally verified candidate with highest Rate.

MT	AT	ML	AcP	AC	Rate
Pt	2-propanol	1.54×10^{-2}	8.46×10^{17}	0.40	69.91
Pt	2-propanol	2.23×10^{-2}	3.88×10^{17}	0.47	68.12
Pt	2-propanol	5.27×10^{-2}	3.76×10^{17}	0.62	62.92
Pt	2-propanol	3.09×10^{-2}	3.75×10^{17}	0.43	59.34
Pt	2-propanol	2.37×10^{-2}	4.58×10^{17}	0.58	58.53

Table S4. The top 5 candidates of search space for Cond. 2 Bayesian optimization (Ethylene glycol). The first row is the experimentally verified candidate with highest Rate.

MT	AT	ML	AcP	AC	Rate
Pt	Ethylene glycol	2.11×10^{-2}	3.46×10^{17}	0.38	75.12
Pt	Ethylene glycol	5.27×10^{-2}	3.76×10^{17}	0.62	64.24
Pt	Ethylene glycol	3.07×10^{-2}	3.74×10^{17}	0.42	59.53
Pt	Ethylene glycol	7.06×10^{-2}	3.78×10^{17}	0.40	58.75
Pt	Ethylene glycol	4.11×10^{-2}	4.32×10^{17}	0.45	56.72

Comparing the model accuracy with and model without AcP as input features

As was stated in the main manuscript, the AcP is essential to be used as an input feature from the domain knowledge point of view as well as statistical reasons (better model metrics). Here it is shown what happens if one does not use AcP as an input feature in the model. For this reason, using the same database, another TPOT-optimized model was trained which predicted Rate with CL, CWF, AC, and ATI as input features.

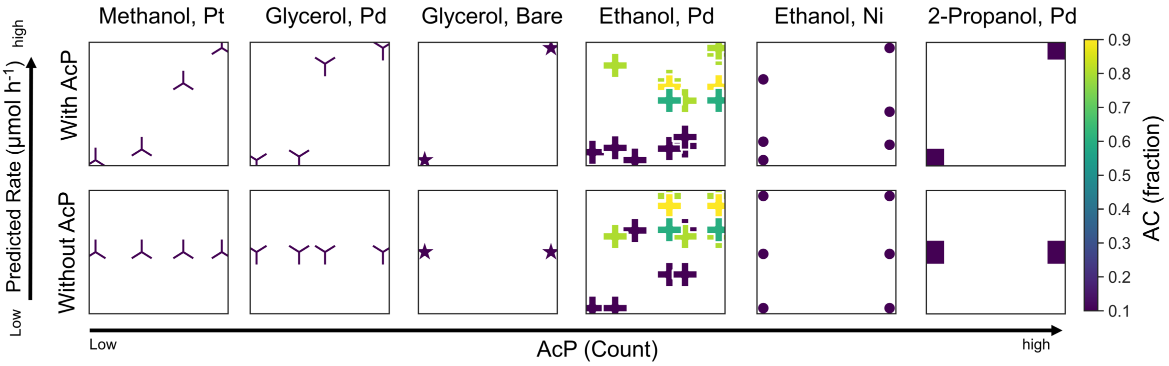


Figure S15. Comparison of samples with different AcP and Rate for (top row) model including AcP and (bottom row) model without AcP in input features. The different combinations of alcohol and cocatalyst were shown in different columns. Since alcohol concentration has a nonlinear effect, different colours were assigned to different AC values. A visible sensitivity to the effect of AcP is clear in the model with AcP while there are not any differences in the model without AcP.

Since overall metrics such as RMSE cannot tell everything, one needs to see what happens if only AcP changes while other features are constant. For this purpose, samples in the database which only are different in values of AcP, and Rate were extracted. The hydrogen rate with both models (model with AcP and model without AcP in input features) was predicted for them. Based on the real data, when everything is constant and only the AcP changes, the hydrogen rate also changes in the same direction (the higher AcP, the higher the Rate). By plotting these samples in Figure S15, it was shown that the model without AcP cannot see the effect of light in predictions. In this figure, every column presents data for one combination of alcohol and cocatalyst. The top row is the results of the prediction for the model with AcP in input features and the bottom row is for the model without AcP in input features. Every plot shows the scatter of model prediction (y-axis) vs AcP (x-axis) and different colours represent different values of AC. it is clear that in the top row, the model predictions are sensitive to AcP and an overall incremental trend between Rate and AcP can be captured. However, in the bottom row, the horizontal trend of model predictions confirms there is no sensitivity to AcP. In conclusion, one can say that AcP as an input feature allows to make the comparison of results between different literature with different light sources and make the model sensitive to light characteristics.

Figure S16 shows a brief analysis of the model without AcP in input features and its compatibility with domain knowledge. Figure S16a shows the SHAP analysis for feature importance of the model missing AcP in input features in which the x-axis shows the importance value and the y-axis shows the name of the feature. Here CWF is the least important feature which is not true, and the type of cocatalyst is Definity more important than the type of alcohol.[2] Moreover, AC is less important than CL due to the surface effects of CL. Therefore, the feature importance of the model without AcP shows less compatibility with domain knowledge of photocatalyst reaction and reported literature while in the model with AcP in the input features, there was more support for domain knowledge and literature.

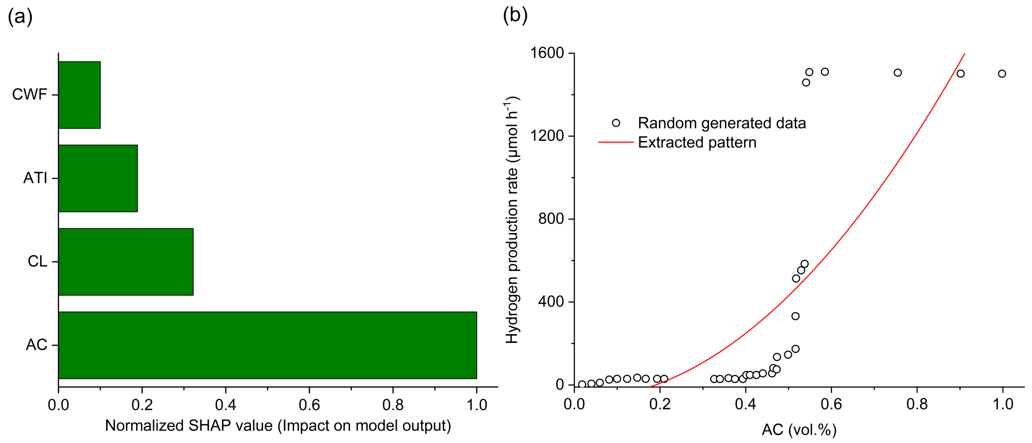


Figure S16. (a) Normalized SHAP values of the four most influential features for predicting the photocatalytic hydrogen evolution rates in a model without AcP among the input features. (b) Trends of predicted hydrogen evolution rates as a function of alcohol concentration (AC) in the model without AcP among the input features. The irrational order in the level of contribution between different features and the meaningless pattern between AC and hydrogen rate is an example of weak fitting of the model which happens in the absence of AcP.

Figure S16b also shows the relationship between the concentration of alcohol and the Rate for the model without AcP in the input feature. In this plot, the x-axis shows the concentration of alcohol, and the y-axis shows the Rate. As it is clear, the pattern that this model learned states there is a jump in AC = 0.5 on hydrogen rate, while experimental analysis did not confirm this trend. Moreover, this plot says that after AC ~ 0.5, increasing alcohol concentration does not have any effect on the Rate, while this trend is not also accepted.[3]

As a result, using AcP as an input feature can increase the accuracy of the model as well as make the model more compatible with experimental observations. This is because of the highly important role of AcP and its interactions with other features such as AC and CL. Therefore, using AcP in an activity/rate model is highly recommended referring to its positive effect on accuracy as well as its role in the physics of the photocatalytic reaction.

Turnover frequency (TOF) of cocatalyst

TOF is a measure of number of reactions per surface active site. Since the hydrogen evolution occurs exclusively on the cocatalyst, it is possible to calculate the TOF if one knows the dispersion (*D*, i.e., ratio of surface to bulk atoms) of the cocatalyst. This is in turn possible to deduce from the cocatalyst deposit size (*d*):

$$D = \frac{6\,V_m}{S_m d} \qquad \textbf{Eq. S5}$$

where V_m is the atomic volume, while S_m is the surface area per atom. Both parameters can be easily calculated from the unit cell dimension of the cocatalyst. In the case of Au cocatalyst (lattice parameter, a = 0.4079 nm, face-cubic-centre), the V_m and S_m can be calculated as 0.01697 nm^3 atom^{-1} and 0.08319 nm^2 atom^{-1}, respectively. Likewise, in the case of Pt cocatalyst (lattice parameter, a = 0.3912 nm, face-cubic-centre), the V_m and S_m can be calculated as 0.01497 nm^3 atom^{-1} and 0.07652 nm^2 atom^{-1}, respectively. Figure S17a, b shows the particle size distributions of Au and Pt cocatalysts on TiO_2,[4,5] at different cocatalyst loadings prepared by chemical precipitation, the most common deposition method in the collected dataset. The data on particle size distributions were best fitted to log-normal distribution, from which the mean particle sizes a function of cocatalyst loadings. As shown in the insets of Figure S17a, b, continuous trends of mean particle sizes for the cocatalysts can be obtained, that can now be readily converted to *D* for the reported loadings. When multiplied by the total cocatalyst loading (in mol) in the reaction system, it is possible to calculate the total amount of cocatalyst active sites, N_active (in mol).

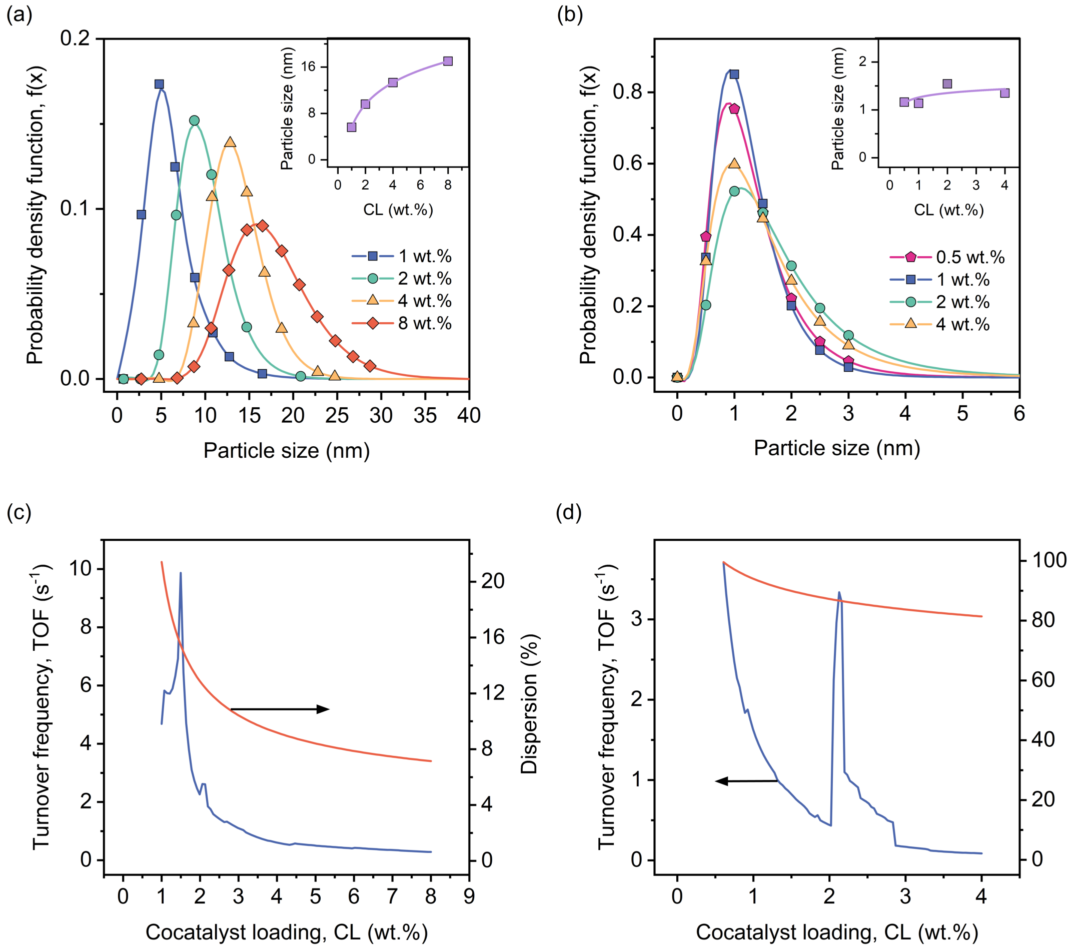


Figure S17. The cocatalyst particle size distributions of chemically-precipitated (a) Au and (b) Pt at various loadings, as reported by Idriss and coworkers.[4,5] Insets show the mean particle size based on the fittings to log-normal distributions. (c,d) The calculated dispersions of Au and Pt based on their mean sizes and subsequently the TOF based on the Rates predicted by the machine learning model under the conditions: Photocatalyst loadings: 50 mg, 10 vol.% methanol, and active photon flux 6.32×10^{16} photon s^{-1} cm^{-2}.

To calculate the TOF, the predicted Rate can be generated from the machine learning model, which we have demonstrated to exhibit unprecedented accuracies (see Figure 5 and 6 in the main manuscript). For demonstration purpose, we generated the Rate values for the case Au and Pt cocatalysts, but in both cases using the median of AcP and AC from training dataset, ATI of methanol, AcP = 6.32×10^{16} photon s^{-1} cm^{-2}. The Rates (in μmol h^{-1} of H_2) when then divided by N_{active} yield the TOF (in the unit of per time) as shown in Figure S17c, d.

References

1. T. Lookman, P. V. Balachandran, D. Xue and R. Yuan, *npj Comput. Mater.*, 2019, **5**, 21.
2. M. Bowker, C. O'Rourke and A. Mills, *Top. Curr. Chem.*, 2022, **380**, 17.
3. E. P. Melián, C. R. López, D. E. Santiago, R. Quesada-Cabrera, J. A. O. Méndez, J. M. D. Rodríguez and O. G. Díaz, *Appl. Catal., A*, 2016, **518**, 189-197.
4. M. Murdoch, G. I. N. Waterhouse, M. A. Nadeem, J. B. Metson, M. A. Keane, R. F. Howe, J. Llorca and H. Idriss, *Nat. Chem.*, 2011, **3**, 489-492.
5. Z. H. N. Al-Azri, M. AlOufi, A. Chan, G. I. N. Waterhouse and H. Idriss, *ACS Catal.*, 2019, **9**, 3946-3958.