Electronic Supplementary Information


Meirong Huang¹², Sutong Wang³, Hongwei Zhu¹*

¹ State Key Laboratory of New Ceramics and Fine Processing, School of Materials Science and Engineering, Tsinghua University, Beijing 100084, China
² State Key Laboratory of Photocatalysis on Energy and Environment, College of Chemistry, Fuzhou University, Fuzhou 350108, China
³ School of Economics and Management, Dalian University of Technology, Dalian 116023, China

*Email: hongweizhu@tsinghua.edu.cn

Methods:
• Influencing factors

Supplementary Figures:
• Figures S1–S25
Systematic study about the influencing factors:

The screening of cocatalysts for photoanodes is greatly influenced by several factors, including the physical and chemical properties of the cocatalyst (e.g., type, dopant, thickness, morphology, synthesis method), the photoanode (e.g., dopant, microscopic size, morphology, synthesis method), and the electrolyte (e.g., composition, concentration, pH). These factors can cause various cocatalysts to work with different effectiveness and mechanisms1-3.

**Cocatalyst type** can affect various properties of the cocatalyst/photoanode catalytic system, such as band alignment, light absorption, characteristics of active sites, sedimentary morphology, and surface wettability. (1) Band alignment is determined by the relative energy band positions of the cocatalyst and photoanode. Semiconductor-type cocatalysts can form heterojunction with photoanodes. Since different band alignments can change hole extraction ability and charge separation efficiency, the alignment should be appropriately regulated to achieve optimal performance4-6. (2) The light absorption ability of the catalytic system is affected by the cocatalyst’s light absorption properties. While metal-type cocatalysts may reduce light harvesting efficiency due to the shading effect, semiconductor-type cocatalysts can enhance light absorption and produce more photo-generated charge carriers6, 7. (3) The active sites of a metal-containing cocatalyst are typically the metal, and the number of active sites depends on their exposure way, which is determined by the cocatalyst type, typical examples are metal-organic frameworks8-10. The type of active sites affects the surface reaction kinetics, and more exposed active sites can improve catalytic activity by enhancing the hole injection efficiency11. (4) The sedimentary morphology of cocatalysts, such as isolated particles, conformal layer, and isolated layer, depends on the cocatalyst type, typical examples are molecular catalysts. A conformal layer usually provides better stability than isolated particles by passivating surface states and separating photoanode from electrolyte6, 12, 13. (5) The hydrophilicity of the catalytic system is mainly determined by the number of hydrophilic groups on the cocatalyst surface. Higher hydrophilicity can improve the full contact between electrolyte and cocatalyst, accelerate the release of gaseous products, and thus improve surface reaction kinetics and catalyst stability14, 15.

**Doping cocatalyst** can optimize the interfacial contact between photoanode and cocatalyst to promote charge separation16, active catalytic sites to provide higher surface OER kinetics17, expose more defects and edges to act as active sites18, as well as improve conductivity for efficient charge transfer17, 18. Nevertheless, it is important to optimize the doping concentration since high levels may introduce defects that can cause charge recombination9.
**Cocatalyst thickness** is also a trade-off parameter that controls the cocatalyst’s effectiveness. In the size range of the depletion width introduced by the semiconductor/electrolyte interface, thicker cocatalysts layers promote charge separation by extracting holes more effectively, which results in improved catalytic performance\(^9\). However, an excessively thick cocatalyst layer can cause shading effects\(^7\), create shunting pathways\(^8\), and hinder charge transport\(^4\), resulting in reduced performance. Thus, optimizing cocatalyst thickness is essential to balance these competing factors and maximize catalytic efficiency\(^21,22\).

**Cocatalyst morphology** is determined by the cocatalyst type and its preparation method, usually presented in the form of isolated particles, a conformal layer, or an isolated layer. Conformal layers have the advantage of providing superior stability over isolated particles by passivating surface states and isolating photoanode from electrolyte\(^5,12\). However, conformal layers composed of multiple nanoparticles may exhibit limited catalytic activity due to inferior charge transport within the layer caused by numerous defects and granular interfaces.

The preparation method of cocatalysts can impact their specific composition, defect state, crystallinity, morphology, size, and specific surface area, which ultimately affects their catalytic activity and stability\(^23\). These changes can make a more rational band alignment between the cocatalyst and photoanode, facilitating charge separation and injection\(^5\). In-situ synthesis techniques, such as electrodeposition, photo-deposition, and photo-assisted electrodeposition, can establish better contact between the photoanode and cocatalyst compared to post-deposition methods like dip coating, spin coating, and drop coating. The enhanced contact can promote charge transport and prolong the carrier lifetime, thereby facilitating the OER process\(^24\).

**Photoanode morphology and size** can change the effectiveness of cocatalysts. Although several studies have investigated the impact of photoanode morphology and size on its performance, the effect on cocatalysts is often overlooked. Recent research suggests that isolated layered cocatalysts can improve the charge separation efficiency and catalytic activity of denser photoanode layers more efficiently\(^25\). Denser photoanode layers have superior photon absorption and electron transport but inferior hole transport, whereas porous layers exhibit the opposite characteristics. With the help of cocatalysts that promote hole extraction, denser layers show greater performance enhancement due to more excited charge carriers, good electron transport, and improved hole transport\(^26\).

**Doping photoanode** can improve the effectiveness of cocatalysts by promoting charge separation. This improvement is likely a result of optimized interfacial bonding strength and band alignment between the photoanode and cocatalyst. Typically, doping the photoanode and
loading the cocatalyst work together in a synergistic manner\textsuperscript{27}. However, it is crucial to carefully regulate the doping concentration to prevent excessive dopants or defects from causing serious charge recombination\textsuperscript{28}.

The preparation method of photoanodes not only affects their morphology and size but also their surface states. Choosing an appropriate preparation method can improve the interaction between the photoanode and cocatalyst, resulting in reduced charge transfer resistance and more efficient charge separation.

Electrolyte composition plays a crucial role in determining the activity and stability of the catalytic system. The interaction between electrolyte anions and metal ions in cocatalysts may cause structural changes in the cocatalysts\textsuperscript{29, 30}. In addition, the addition of specific ions to the electrolyte can help to inhibit photoanode dissolution\textsuperscript{31} or regenerate the cocatalyst layer\textsuperscript{32}, leading to improved catalyst stability. Nevertheless, most anions, such as borate, phosphate, and sulfate, are electrochemically inert and do not impact the OER process\textsuperscript{33}.

Electrolyte pH usually varies with the change of electrolyte composition possessing various p\textsubscript{Ka} values. Although alkaline conditions are typically favorable for the surface OER on cocatalysts\textsuperscript{34}, extremely high pH levels may cause cocatalyst dissolution\textsuperscript{35}. The influence of electrolyte pH on cocatalyst performance is also decided by the working mechanism of cocatalysts. Therefore, cocatalysts work better in a solution with a proper pH value\textsuperscript{36}, some cocatalysts can even perform better in acidic solutions\textsuperscript{35, 37}.

Electrolyte concentration should be high enough to provide effective charge injection at the catalyst/electrolyte interface and reduce solution resistance. When anions are absent or present in low concentrations in the electrolyte, photoanodes and/or cocatalysts show poor activity. However, excessively high electrolyte concentration may lead to the detachment of cocatalysts. Therefore, optimizing the electrolyte concentration is essential for achieving optimal performance of the catalytic system\textsuperscript{33}.

References:


7. Sol A Lee, T. H. L., Changyeon Kim, Mi Gyoung Lee, Min-Ju Choi, Hoonkee Park, Seokhoon Choi, Jihan Oh, and Ho Won Jang, Tailored NiO\textsubscript{x}/Ni Cocatalysts on Silicon for Highly Efficient Water Splitting Photoanodes via Pulsed Electrodeposition. ACS Catalysis 2018, 8, 7261-7269.


16. Tsyganok, A.; Klotz, D.; Malviya, K. D.; Rothschild, A.; Grave, D. A., Different Roles of Fe\textsubscript{1-x}Ni\textsubscript{x}OOH Cocatalyst on Hematite (α-Fe\textsubscript{2}O\textsubscript{3}) Photoanodes with Different Dopants. ACS Catalysis 2018, 8 (4), 2754-2759.


Supplementary Figures

**Figure S1.** Distribution of the output variable after being categorized. The insert table is the classification basis.

<table>
<thead>
<tr>
<th>Class</th>
<th>Low</th>
<th>Medium</th>
<th>High</th>
</tr>
</thead>
<tbody>
<tr>
<td>Range ((\sqrt{y}))</td>
<td>~0.082</td>
<td>~1.391</td>
<td>~2.695</td>
</tr>
</tbody>
</table>

**Figure S2.** Data integrity of some input variables: anneal temperature (T), deposition voltage (V), deposition time (t), the morphology of photoanode, and the size of photoanode (l1 and l2 are lateral dimensions, h1 is the thickness of a single grain, h2 is the overall thickness). Here, only photoanodes prepared by the electrodeposition-anneal method were considered.
Figure S3. (a) Thickness error of the photoanode thickness filled using single imputation methods: linear regression, bagging, random forest, k-nearest neighbor (kNN), and Datawig (based on a neural network). (b) Thickness distribution of the overall thickness of photoanode.

Figure S4. Convergence of the multiple imputations by chained equations (MICE) algorithm used for filling in different input variables: (a) h1, (b) h2. Distributions of the imputed input variables: (c) h1, (d) h2. The red line is the original data, black lines represent the data from each imputation.
**Figure S5.** Data splitting for model training, validation, and testing.

**Figure S6.** (a) Five data normalization methods and related scaling principles. (b) Test accuracy of neural networks trained on the normalized datasets. The box plot uses boxes and lines to depict the distribution of statistical results, where box limits show the range of the middle 50% of the data with an orange line marking the median value.
Figure S7. (a) Some solutions to data imbalance. Predicted results of neural networks trained on the preprocessed datasets: (b) test accuracy, (c) macro-F1-score and area under the curve (AUC) of the receiver operating characteristic curve (ROC).

Figure S8. (a) Several dimensionality reduction methods, the first six methods belong to feature selection, while the last four methods belong to feature extraction. (b) Test accuracy of neural networks trained on the preprocessed datasets.
**Figure S9.** (a) A multi-layer perceptron neural network with two hidden layers. (b) Optimized hyperparameters of the neural network.

**Figure S10.** Hyperparameters and their values considered in the neural network.

**Figure S11.** The basic architecture of a random forest model.
Figure S12. Cross-validation accuracy of random forest models varying with hyperparameters: (a) n_estimators, (b) max_depth, (c) min_samples_leaf, (d) min_samples_split, (e) max_features, (f) n_estimators (second optimization)

Figure S13. Optimized hyperparameters of the random forest model. The preceding lines show the best values and results (cross-validation accuracy) of each hyperparameter individually optimized, and the last line shows the best parameter combination optimized by the grid search of all parameters.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>n_estimators</td>
<td>61</td>
<td>70</td>
</tr>
<tr>
<td>max_depth</td>
<td>7</td>
<td>67.78</td>
</tr>
<tr>
<td>min_samples_leaf</td>
<td>3</td>
<td>69</td>
</tr>
<tr>
<td>min_samples_split</td>
<td>5</td>
<td>70</td>
</tr>
<tr>
<td>max_features</td>
<td>11</td>
<td>72.11</td>
</tr>
<tr>
<td>'max_depth': 8, 'max_features': 8, 'min_samples_leaf': 2, 'min_samples_split': 5, 'n_estimators': 46</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
**Figure S14.** Optimized hyperparameters of the bagging model.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>n_estimators</td>
<td>116</td>
<td>68.89</td>
</tr>
<tr>
<td>max_samples</td>
<td>0.95</td>
<td>66.67</td>
</tr>
<tr>
<td>max_features</td>
<td>0.35</td>
<td>70</td>
</tr>
<tr>
<td>max_depth</td>
<td>11</td>
<td>74.22</td>
</tr>
<tr>
<td>min_samples_split</td>
<td>2</td>
<td>74.22</td>
</tr>
<tr>
<td>min_samples_leaf</td>
<td>1</td>
<td>74.22</td>
</tr>
</tbody>
</table>

- n_estimators=80, max_samples=0.8, max_features=0.2
- max_depth=11, min_samples_split=2, min_samples_leaf=1
- max_depth=11, min_samples_split=2, min_samples_leaf=1, max_samples=0.8, max_features=0.2, n_estimators=80

**Figure S15.** Optimized hyperparameters of the AdaBoost model.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>n_estimators</td>
<td>71</td>
<td>61.33</td>
</tr>
<tr>
<td>learning_rate</td>
<td>0.7</td>
<td>62.22</td>
</tr>
<tr>
<td>n_estimators</td>
<td>80</td>
<td>66.67</td>
</tr>
<tr>
<td>max_depth</td>
<td>2</td>
<td>66.67</td>
</tr>
<tr>
<td>min_samples_split</td>
<td>14</td>
<td>68.78</td>
</tr>
<tr>
<td>min_samples_leaf</td>
<td>3</td>
<td>68.89</td>
</tr>
</tbody>
</table>

- max_depth=5, min_samples_split=10, min_samples_leaf=2
- max_depth=5, min_samples_split=10, min_samples_leaf=2, n_estimators=80, learning_rate=0.8
**Figure S16.** Optimized hyperparameters of the gradient boosting model.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>n_estimators</td>
<td>51</td>
<td>62.44</td>
</tr>
<tr>
<td>max_depth</td>
<td>2</td>
<td>62.67</td>
</tr>
<tr>
<td>min_samples_split</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td></td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>min_samples_leaf</td>
<td>1</td>
<td>73.11</td>
</tr>
<tr>
<td>max_features</td>
<td>8</td>
<td>71.11</td>
</tr>
<tr>
<td>subsample</td>
<td>0.8</td>
<td>71.11</td>
</tr>
<tr>
<td>learning_rate</td>
<td>0.15</td>
<td>73.22</td>
</tr>
<tr>
<td>n_estimators</td>
<td>46</td>
<td></td>
</tr>
</tbody>
</table>

*learning_rate=0.15, n_estimators=46, max_depth=2, min_samples_split=3, min_samples_leaf=1, max_features=8, subsample=0.8*

**Figure S17.** Effect of different cocatalyst parameters on the performance boost: (a) preparation method of cocatalyst, (b) cocatalyst morphology, (c) dopant of cocatalyst, (d) cocatalyst thickness, (e) cocatalyst type, (f) metal type of cocatalyst.
Figure S18. Testing results of the neural network trained on the dataset normalized by the StandardScaler method: (a) confusion matrix, (b) ROC curve.

Figure S19. Training curves of the neural network trained on the dataset normalized by the StandardScaler method, showed in the form of (a) accuracy and (b) loss.
Figure S20. Training (a, b) and testing (c, d) results of the neural network trained on the dataset using the SMOTE oversampling method: (a, c) confusion matrix, (b, d) ROC curve.

Figure S21. Training curves of the neural network trained on the dataset using the SMOTE oversampling method, showed in the form of (a) accuracy and (b) loss.
Figure S22. Confusion matrixes of testing results for different ensemble methods: (a) random forest, (b) bagging, (c) AdaBoost, (d) gradient boosting

Figure S23. Testing results of the random forest used for the binary classification: (a) confusion matrix, (b) ROC curve.
Figure S24. Random forest built-in feature importance of top input variables.
Figure S25. SHAP dependence plots of several input variables: (a) CocatalystMetalCarbonate, (b) ElectrolytePotassiumBorateBuffer, (c) CocatalystHydroxide, (d) CocatalystCo, (e) PhotoanodeAllThickness, (f) PhotoanodeLength, (g) CocatalystConformalFilm, (h) CocatalystMetalOxyhydroxide, (i) Cocatalyst2DenseFilm, (j) CocatalystSilicate, (k) CocatalystNanoparticle, and (l) CocatalystPhotoelectrophoreticDeposition.