

## Supporting Information to

# Machine learning aided design of high performance of copper-based sulfide photocathode

### S1. Copper-based sulfide photocathode dataset.

Download link: [https://github.com/cyxxxx24/Performance-prediction-platform-for-copper-based-sulfide-](https://github.com/cyxxxx24/Performance-prediction-platform-for-copper-based-sulfide-photocathode.git)

photocathode.git

**Table S1.** The details for input and output variables

Input Variables	Ranges for numeric or sub-categories for categoric variables
<b>Substrate</b>	Mo, FTO, ITO, Cu
<b>HTL</b>	Au, FeOOH, NiO, (none)
<b>First layer</b>	Cu <sub>2</sub> ZnSnS <sub>4</sub> , Cu <sub>3</sub> BiS <sub>3</sub> , Cu <sub>2</sub> BaSnS <sub>4</sub> , CuInGaS <sub>2</sub> , Cu <sub>2</sub> S, CuFeS <sub>2</sub> , CuGaS <sub>2</sub> , CuInS <sub>2</sub> , CuS, CuSbS <sub>2</sub>
<b>First layer doping</b>	Ag, Bi, Cd, Fe, Ge, S, Se, Zn, (none)
<b>First layer synthesis method</b>	AAO template growth, Chemical bath deposition, Chemical vapor deposition, Colloidal method, Electrodeposition technique, Hydrothermal/Solvothermal synthesis, Physical vapor deposition, SILAR, Spin-coating deposition, Spray pyrolysis deposition, Thermal evaporation method, Wet chemical route
<b>First layer thickness (nm)</b>	0~50000
<b>First layer grain size (nm)</b>	0~5000
<b>First layer Eg (eV)</b>	0~2.8
<b>Second layer</b>	CdS, CdSe, In <sub>2</sub> S <sub>3</sub> , InCdS, MoS <sub>x</sub> , Ni-MoS <sub>x</sub> , PNDI3OT-Se1, PNDI3OT-Se2, Sb <sub>2</sub> S <sub>3</sub> , Sb <sub>2</sub> Se <sub>3</sub> , (none)
<b>Second layer synthesis method</b>	Chemical bath deposition, Electrochemical deposition, Photoelectrochemical, Physical vapor deposition, SILAR, Spin-coating deposition, Hydrothermal/Solvothermal synthesis, (none)
<b>Second layer thickness (nm)</b>	0~500
<b>Second layer grain size (nm)</b>	0~500
<b>Second layer Eg (eV)</b>	0~2.5
<b>Third layer</b>	(Ta,Mo) <sub>x</sub> (O,S) <sub>y</sub> , AZO/TiO <sub>2</sub> , HfO <sub>2</sub> , In <sub>2</sub> S <sub>3</sub> , TaO <sub>x</sub> , TiO <sub>2</sub> , TiO <sub>x</sub> , ZnO/ZnO:Al/Au, ZnS, TiMo, (none)
<b>Third layer dopant</b>	Al, (none)
<b>Third layer synthesis method</b>	Atomic layer deposition, Chemical bath deposition, RF, (none)
<b>Third layer thickness (nm)</b>	0~150
<b>Third layer Eg (eV)</b>	0~3.5
<b>Fourth layer</b>	Au, MoS <sub>x</sub> , NiO, Pt, Ru, RuO <sub>x</sub> , TaO <sub>x</sub> , (none)
<b>Fourth layer</b>	Chemical bath deposition, E-beam evaporation-sulfurization, Electrodeposition

<b>synthesis method</b>	technique, Hydrothermal/Solvothermal synthesis, Photoelectrochemical, Physical vapor deposition, (none)
<b>Electrolyte</b>	HClO <sub>4</sub> , K <sub>2</sub> HPO <sub>4</sub> , KH <sub>2</sub> PO <sub>4</sub> , KPi, Na <sub>2</sub> HPO <sub>4</sub> , Na <sub>2</sub> SO <sub>4</sub> , Na <sub>2</sub> S, KCl, K <sub>2</sub> SO <sub>4</sub> , H <sub>2</sub> SO <sub>4</sub> , Eu(NO <sub>3</sub> ) <sub>3</sub>
<b>Electrolyte Concentration (M)</b>	0~1
<b>PH</b>	0~14
<b>Bias (V vs RHE)</b>	-2~1.1
<b>Output Variables</b>	
<b>Photocurrent Density (mA/cm<sup>2</sup>)</b>	-40~1

---

## S2. Metrics for performance evaluation

(1) R-Square (R<sup>2</sup>):

$$R^2 = 1 - \frac{\sum_i (\hat{y}_i - y_i)^2}{\sum_i (y_i - \bar{y})^2}$$

According to the value of R-Squared, the quality of the model is judged, and the value range is [0,1]. The larger the R-Squared, the better the model fitting effect. In this paper, the R<sup>2</sup> value is used as the accuracy.

(2) Mean absolute error (MAE):

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i|$$

The range [0, +∞) is equal to 0 when the predicted value is completely consistent with the real value, that is, the perfect model; the greater the error, the greater the value. The smaller the value of MAE, the better the accuracy of the prediction model.

(3) Root mean square error (RMSE):

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2}$$

It represents the expected value of the square of the error. The smaller the value, the higher the prediction accuracy of the model.

(4) Mean Absolute Error (MAPE)

$$MAPE = \frac{100\%}{n} \sum_{i=1}^n \left| \frac{\hat{y}_i - y_i}{y_i} \right|$$

Range [0, +∞), The smaller the value of MAPE, the better the accuracy of the prediction model.

In these formulas,  $y_i$  is real value,  $\hat{y}_i$  is predicted value

## S3. Data normalization methods and related scaling principles.

(1) Min-Max

Min-Max standardization refers to the linear transformation of the original data, mapping the values between [ 0,1], and the data distribution is unchanged. The formula is as follows:

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)}$$

(2) Z-Score

Z-Score standardization refers to the standardization of data based on the mean and standard deviation of the original data. The formula is as follows:

$$x' = \frac{x - \mu}{\sigma}$$

(3) Mean Scaler

The standardization of decimal scaling is to map the data to the [-1,1] interval by moving the decimal digits of the data, and the moving decimal digits depend on the maximum value of the absolute value of the data. The formula is as follows:

$$x' = \frac{x}{10^j}$$

(4) Vector Scaler

Mean normalization refers to the standardization of data through the mean, maximum and minimum values in the original data. The formula is as follows:

$$x' = \frac{x - \mu}{\max(x) - \min(x)}$$

In these formulas,  $x$  is a data in the original data,  $\max(x)$  represents the maximum value in the original data,  $\min(x)$  represents the minimum value in the original data,  $\mu$  represents the mean of the original data,  $\sigma$  represents the standard deviation of the original data, and  $j$  denotes the number of decimal moving bits.

**S4. Copper-based sulfide photocathode prediction platform.**

Download link: <https://github.com/cyxxxx24/Performance-prediction-platform-for-copper-based-sulfide-photocathode.git>

After decompression, please click in order: dist-main-Forecasting platform

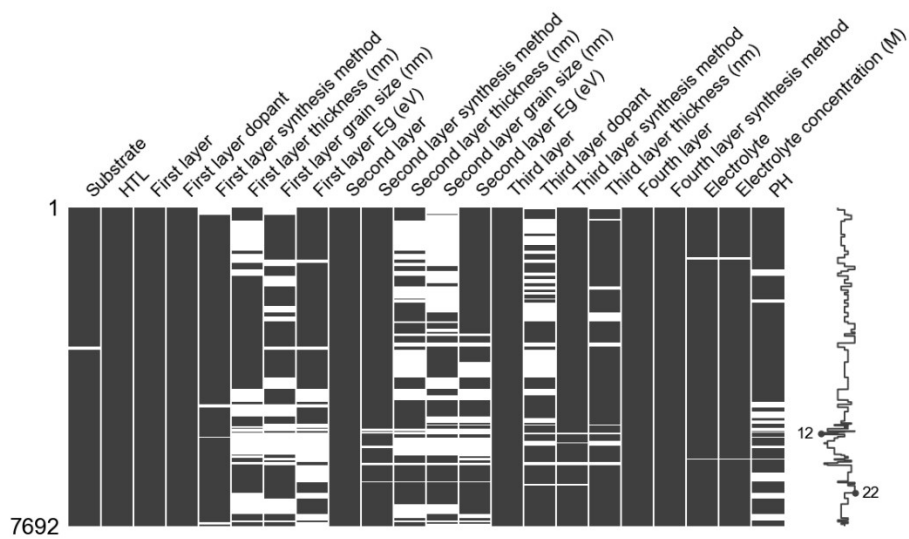


Figure S1. Data integrity of some input variables

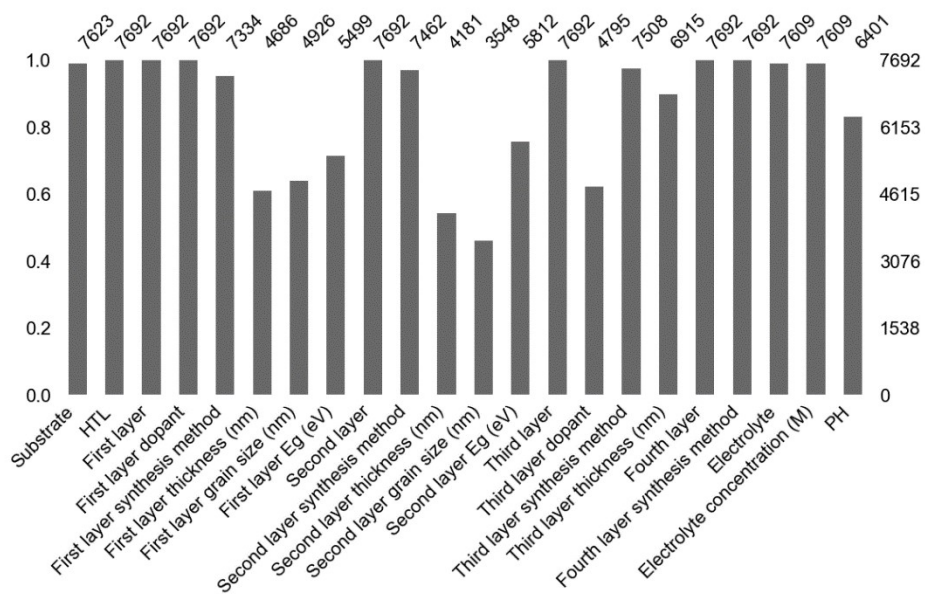
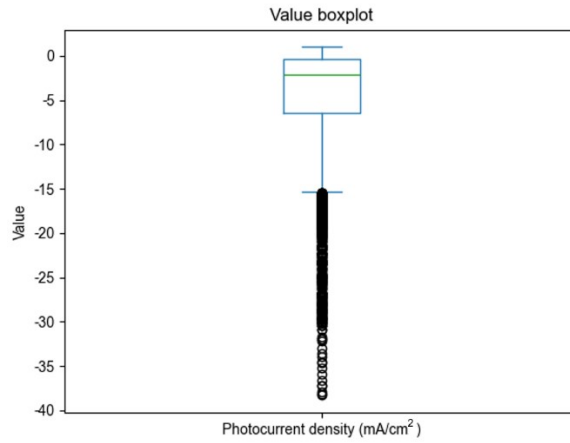
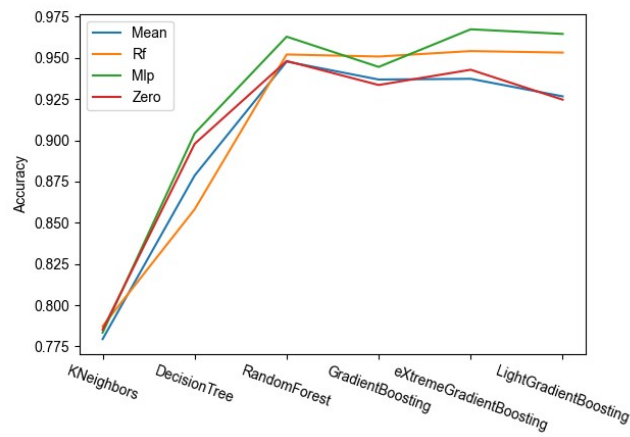


Figure S2. Data integrity of some input variables



**Figure S3.** Response variable box plot of photocurrent density



**Figure S4.** The accuracy of different filling methods (mean, random forest, neural network (mlp), zero) on different ML models

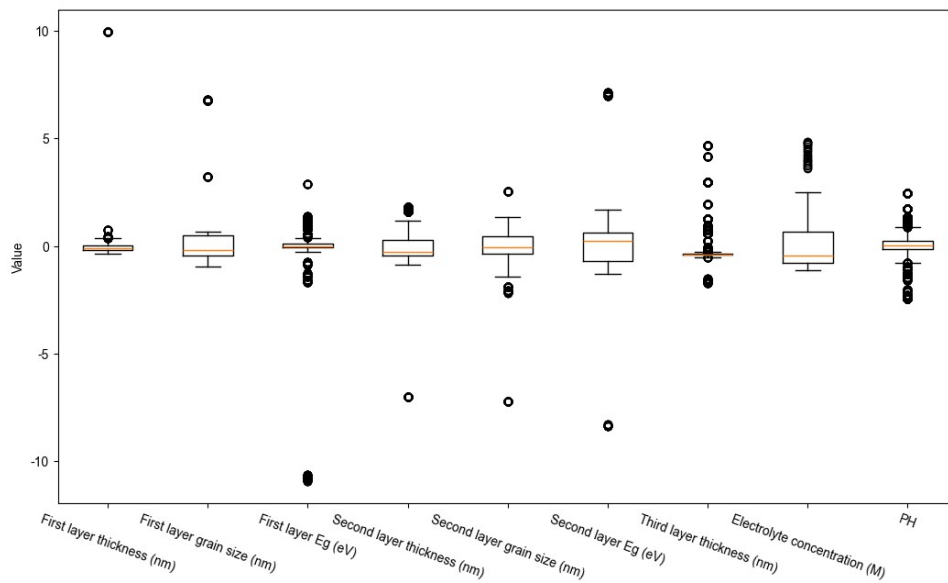


Figure S5. Input variable outliers check box plot

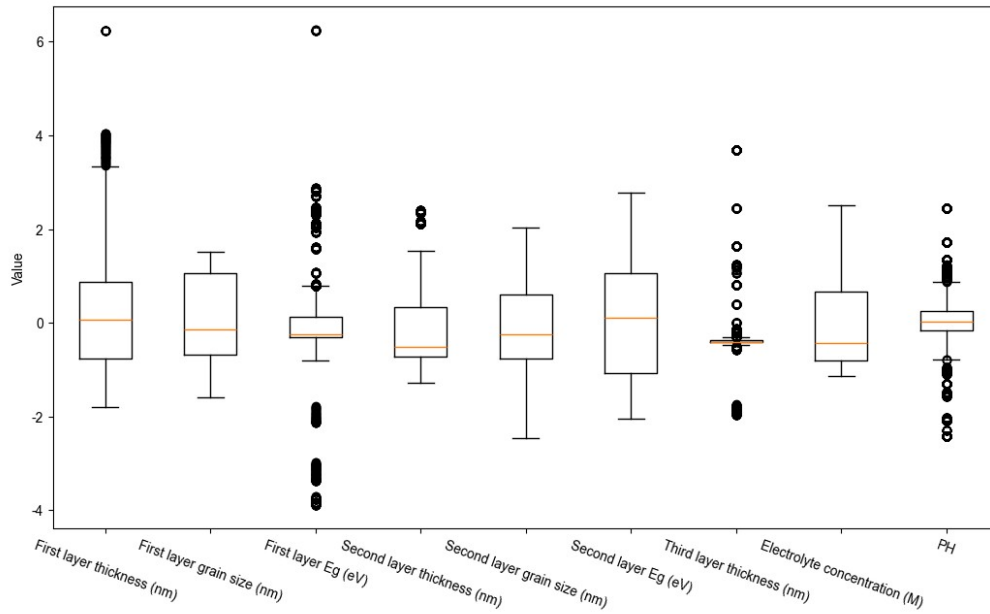


Figure S6. The processed input variable outliers check the box plot

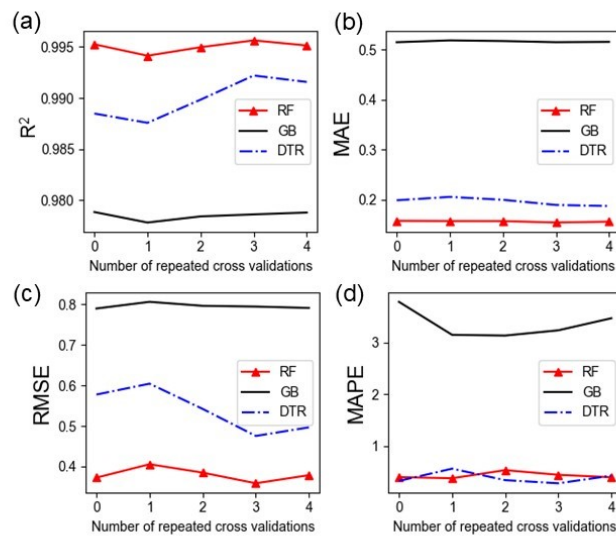
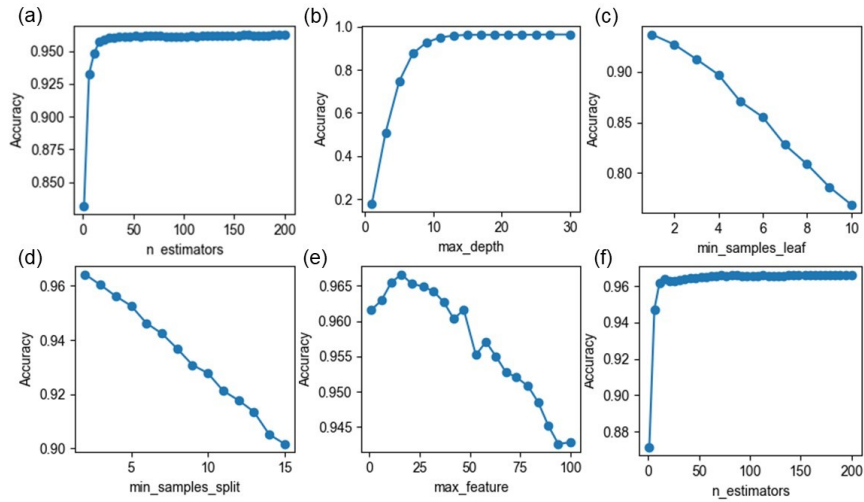
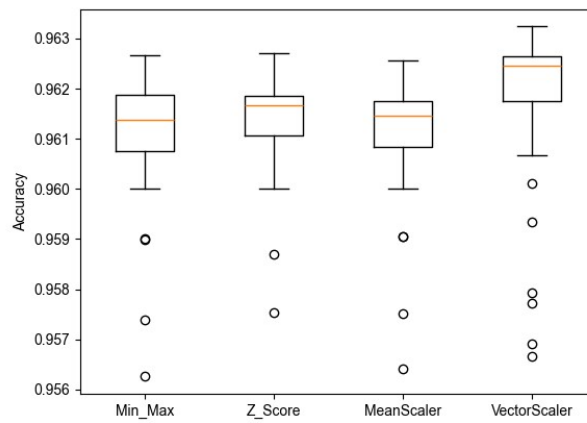


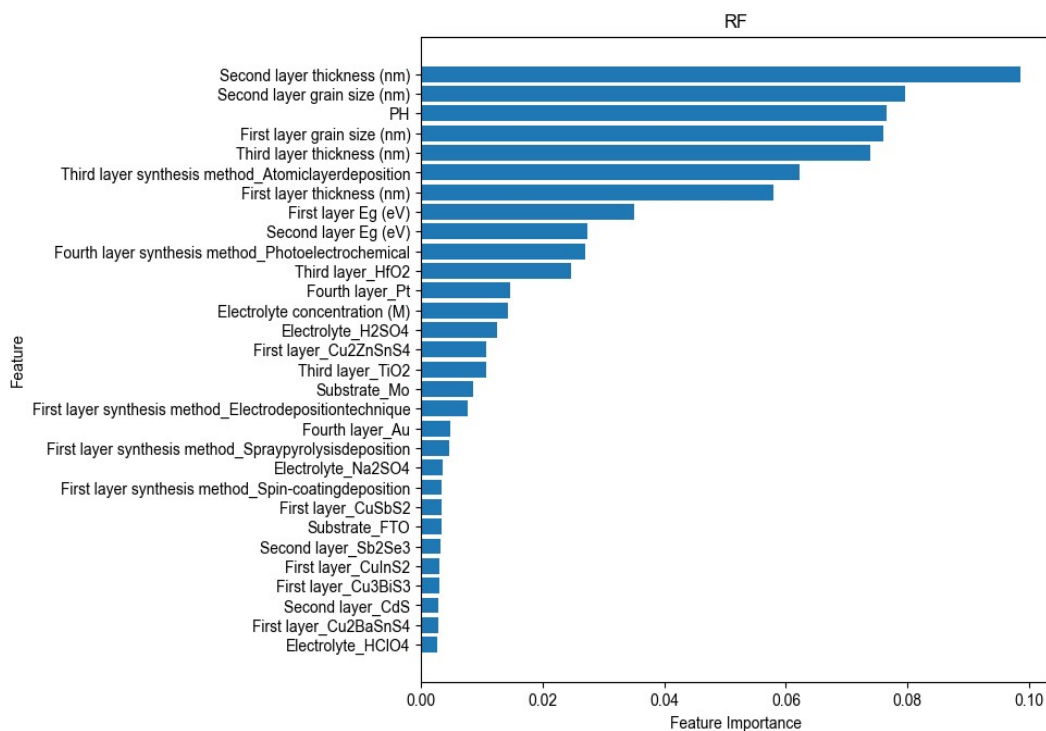
Figure S7. Comparative evaluation on different models (Random forest, Gradient boost, Decision tree) using the data set



**Figure S8.** 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 S9.** 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 S10.** Random forest built-in feature importance of top input variables



**"Performance prediction platform for copper-based sulfide photocathode"** is a prediction platform developed by Professor Feng Jiang's team at South China Normal University.

Jiang Feng's team is dedicated to the research and development of copper-based sulfide photocathodes, and has published a number of high-impact journal papers, including Nature Comm., Energy & Environ. Sci., ACS Energy Lett., ACS Catal., Appl. Catal. B: Environ., etc. Based on this, we have developed this prediction platform to better study the performance of copper-based sulfide photocathodes and to discover better copper-based sulfide photocathodes.

If you have any suggestions or questions about this platform, please contact the main person in charge of the project Cao Yuxi (Email: caoyuxi324@163.com).

Thanks.

**Figure S11.** "Home" of Performance prediction platform for copper-based sulfide photocathode



Performance prediction platform for copper-based sulfide photocathode								
Home			Prediction		Historical record		User guide	
ID	Opening voltage	Current density (OVRHE)	Substrate	HTL	First Layer	First Layer dopant	First Layer Synthesis Method	First Layer thickness(nm)
<b>Export</b>						<b>Clear All</b>		

**Figure S12.** “Historical record” of Performance prediction platform for copper-based sulfide photocathode

Performance prediction platform for copper-based sulfide photocathode				
Home		Prediction	Historical record	User guide
<p><b>Step 1:</b> inputted the parameters for the material you want to predict. If you do not know some parameters, you can refer to the parameters in the relevant paper. If there is no parameter, it defaults to none or 0.</p> <p><b>Step 2:</b> When you have inputted all the parameters. Click ‘ <b>Submit</b> ’, and the platform will automatically predict the open voltage and current density at OVRHE.</p> <p><b>Step 3:</b> If you want to Save your data, please click ‘ <b>Save</b> ’ and you can see your saved data in the history. Click ‘ <b>Refresh</b> ’, and the input and output on this page will clear to 0, and you can start a new forecast.</p> <p><b>Step 4:</b> If you want to view the saved data, please click on ‘ <b>Historical record</b> ’. Point ‘ <b>Export</b> ’ can export historical data to excel table, and point ‘ <b>Empty data</b> ’ will empty all historical data.</p>				

**Figure S13.** “User guide” of Performance prediction platform for copper-based sulfide photocathode