Supplementary Information (SI) for Journal of Materials Chemistry A.

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-photocathode.git

Table S1. The details for input and output variables

Input Variables	Ranges for numeric or sub-categories for categoric variables
Substrate	Mo, FTO, ITO, Cu
HTL	Au, FeOOH, NiO, (none)
First layer	Cu_2ZnSnS_4, Cu_3BiS_3, Cu_2BaSnS_4,$CuInGaS_2$, Cu_2S, $CuFeS_2$, $CuGaS_2$, $CuInS_2$, CuS, $CuSbS_2$
First layer doping	Ag, Bi, Cd, Fu, Ge, S, Se, Zn, (none)
First layer synthesis method	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
First layer thickness (nm)	0~50000
First layer grain size (nm)	0~5000
First layer Eg (eV)	0~2.8
Second layer	CdS, CdSe, In_2S_3, InCdS, MoS_x, Ni-MoS_x, PNDI3OT-Se1, PNDI3OT-Se2, Sb_2S_3, Sb_2Se_3, (none)
Second layer synthesis method	Chemical bath deposition, Electrochemical deposition, Photoelectrochemical, Physical vapor deposition, SILAR, Spin-coating deposition, Hydrothermal/Solvothermal synthesis, (none)
Second layer thickness (nm)	0~500
Second layer grain size (nm)	0~500
Second layer Eg (eV)	0~2.5
Third layer	(Ta,Mo)x,(O,S)y, AZO/TiO_2, HfO_2, In_2S_3, TaO_x, TiO_2, TiO_x, ZnO/ZnO:Al/Au, ZnS, TiMo, (none)
Third layer dopant	Al, (none)
Third layer synthesis method	Atomic layer deposition, Chemical bath deposition, RF, (none)
Third layer thickness (nm)	0~150
Third layer Eg (eV)	0~3.5
Fourth layer	Au, MoS_x, NiO, Pt, Ru, RuO_x, TaO_x, (none)
Fourth layer	Chemical bath deposition, E-beam evaporation–sulfurization, Electrodeposition

synthesis method	technique, Hydrothermal/Solvothermal synthesis, Photoelectrochemical, Physical vapor deposition, (none)
Electrolyte	$HClO_4$, K_2HPO_4, KH_2PO_4, KPi, Na_2HPO_4, Na_2SO_4, Na_2S, KCl, K_2SO_4, H_2SO_4, $Eu(NO_3)_3$
Electrolyte Concentration (M)	0~1
PH	0~14
Bias (V vs RHE)	-2~1.1
Output Variables	
Photocurrent Density (mA/cm^2)	-40~1

S2. Metrics for performance evaluation

(1) R-Square (R^2):

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

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^2 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 maximum value in the original data, μ represents the mean of the original data, σ 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

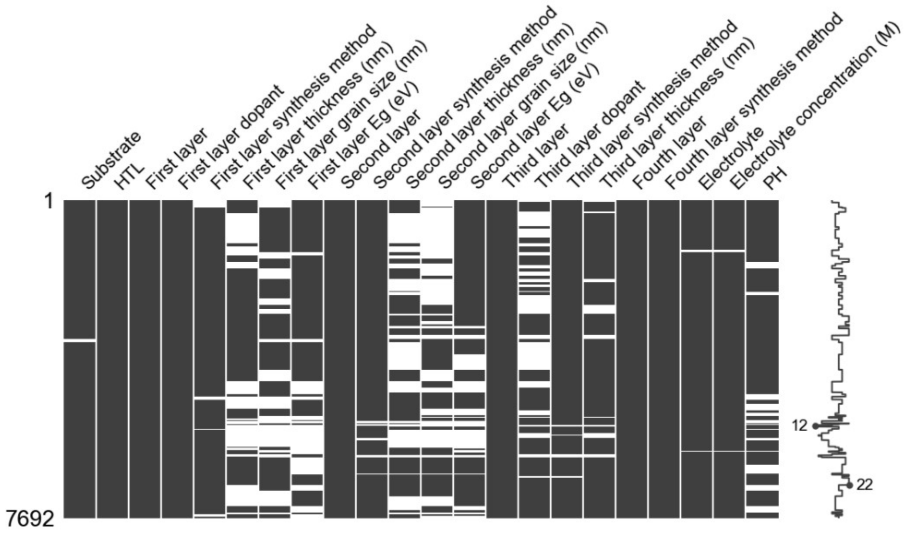

Figure S1. Data integrity of some input variables

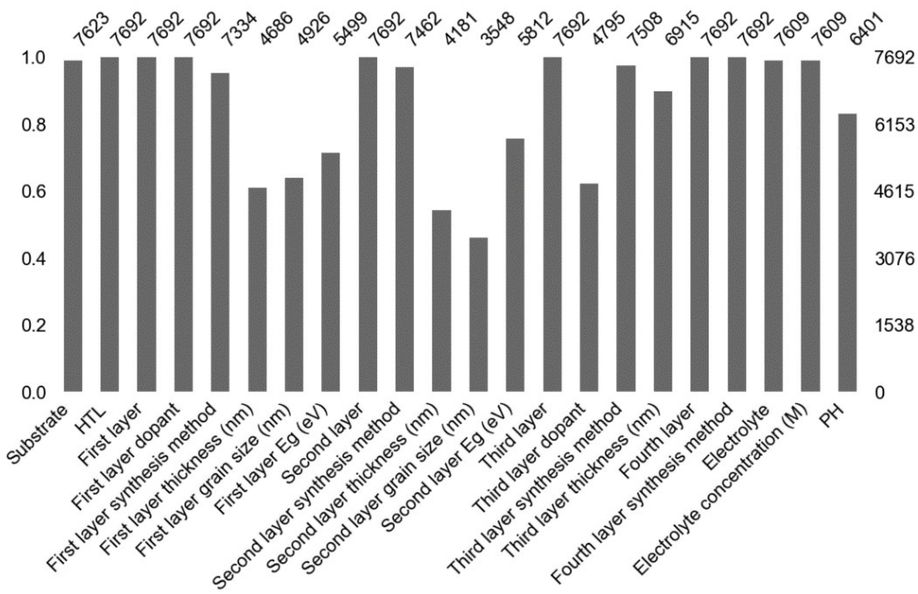

Figure S2. Data integrity of some input variables

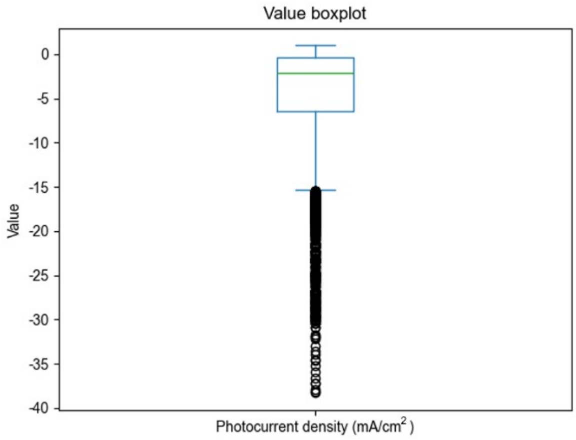

Figure S3. Response variable box plot of photocurrent density

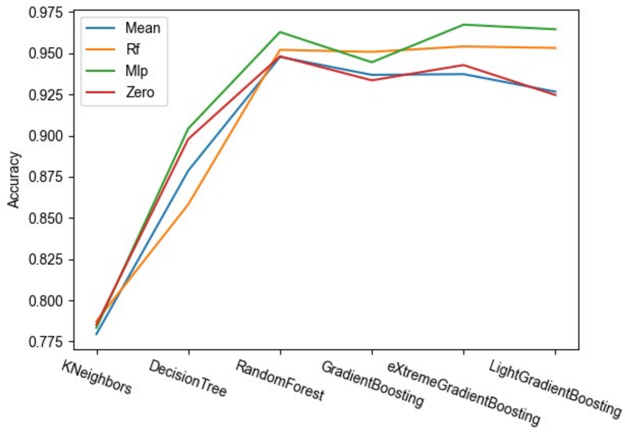

Figure S4. The accuracy of different filling methods (mean, random forest, neural network (mlp), zero) on different ML models

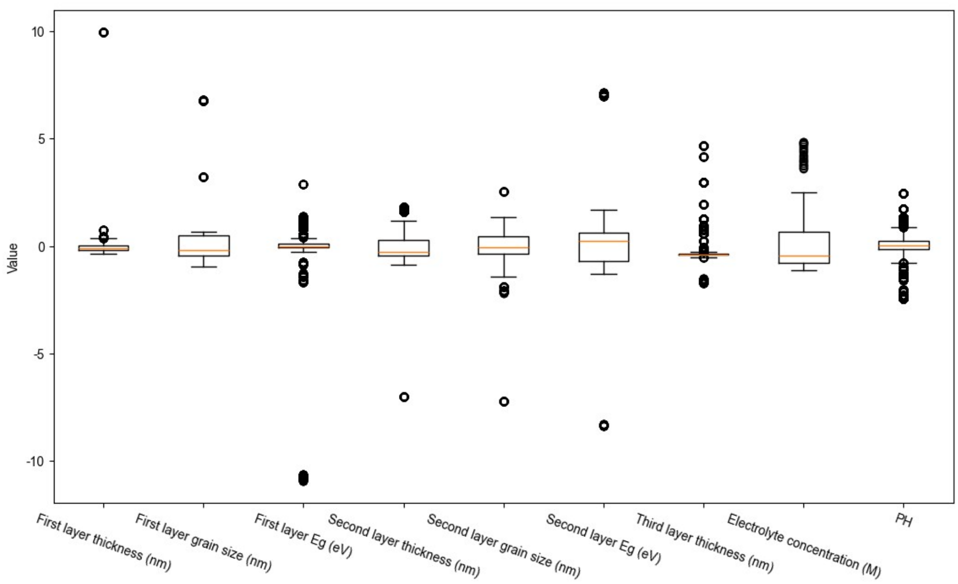

Figure S5. Input variable outliers check box plot

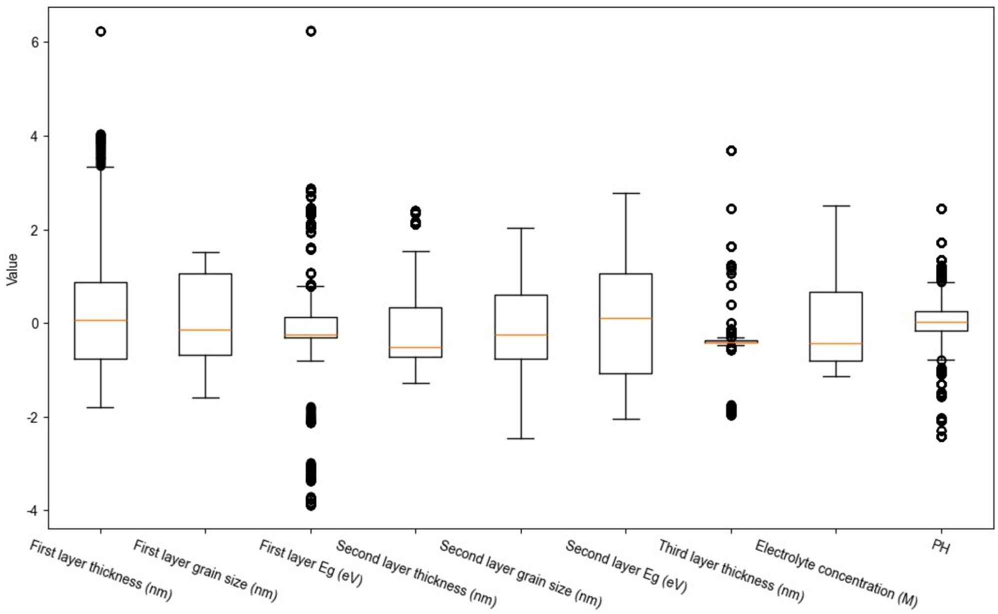

Figure S6. The processed input variable outliers check the box plot

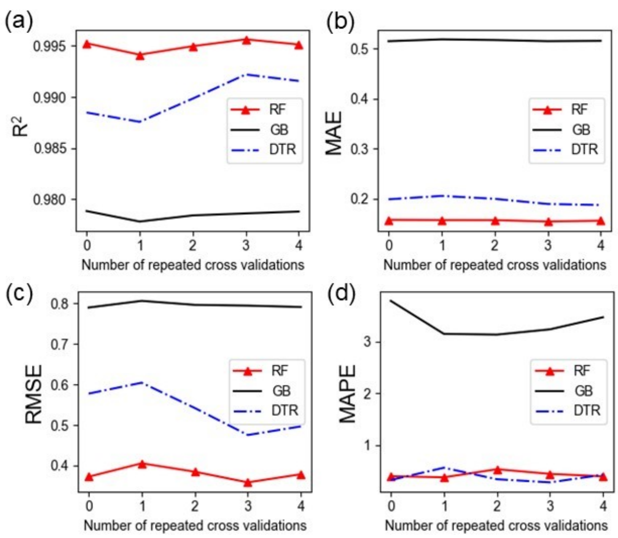

Figure S7. Comparative evaluation on different models (Random forest, Gradient boost , Decision tree) using the data set

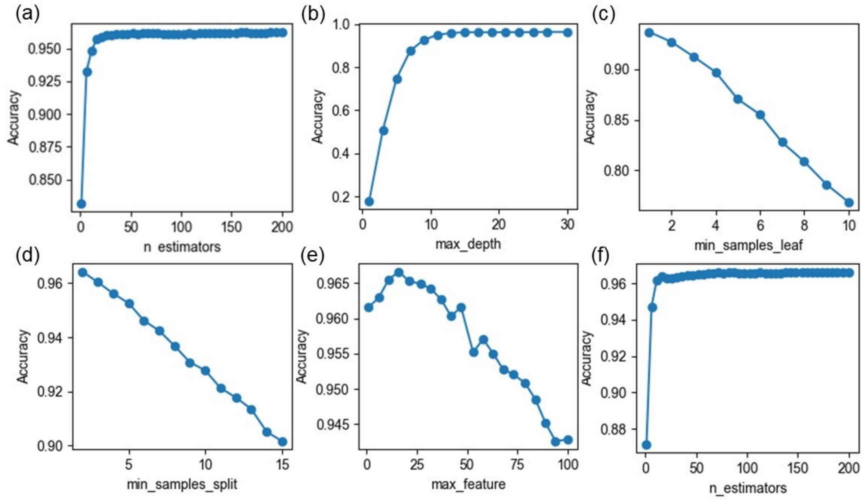

Figure S8. Cross-validation accuracy of random forest models varying with hyperparameters: (a) n_estimators, (b) max_depth, (c) min_samples_leaf, (d) min_samples_split, (e) max_features, (f) n_estimators (second optimization)

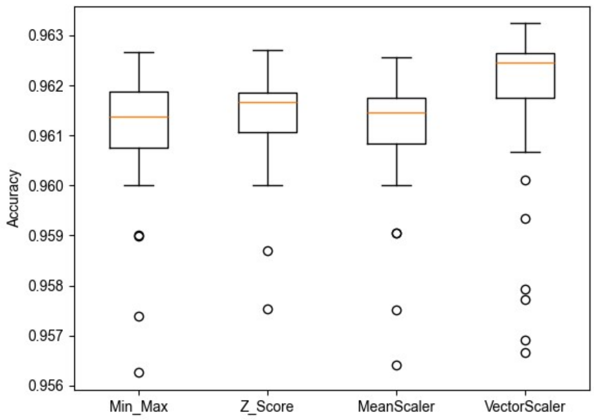

Figure S9. Test accuracy of neural networks trained on the normalized datasets. The box plot uses boxes and lines to depict the distribution of statistical results, where box limits show the range of the middle 50% of the data with an orange line marking the median value.

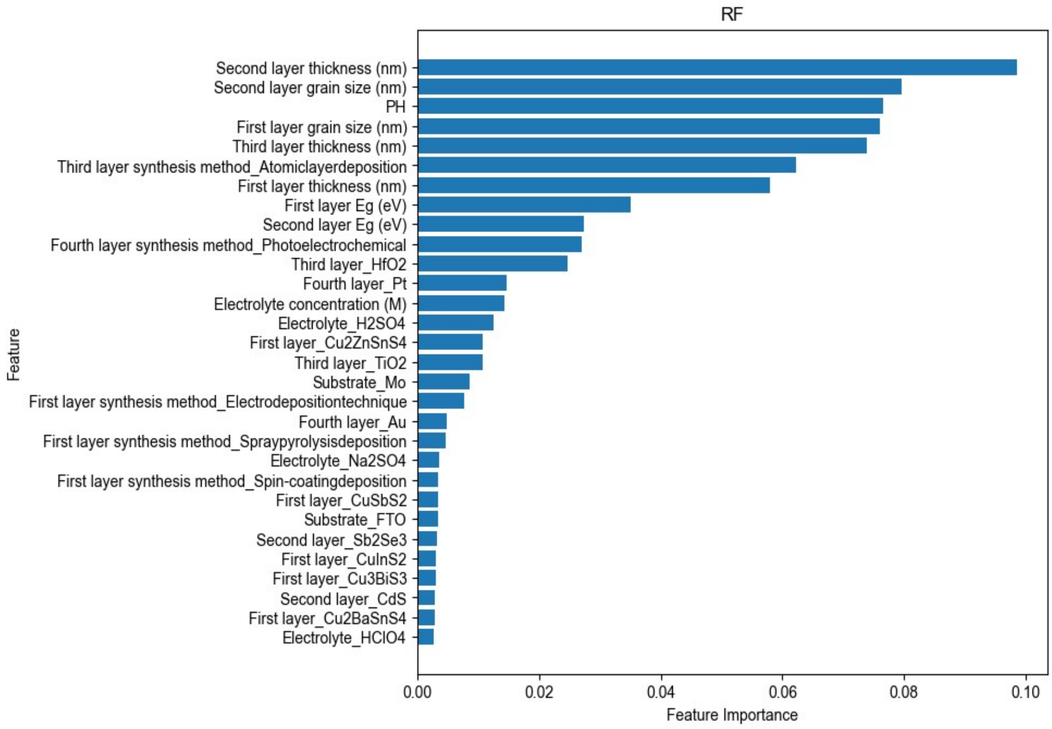

Figure S10. Random forest built-in feature importance of top input variables

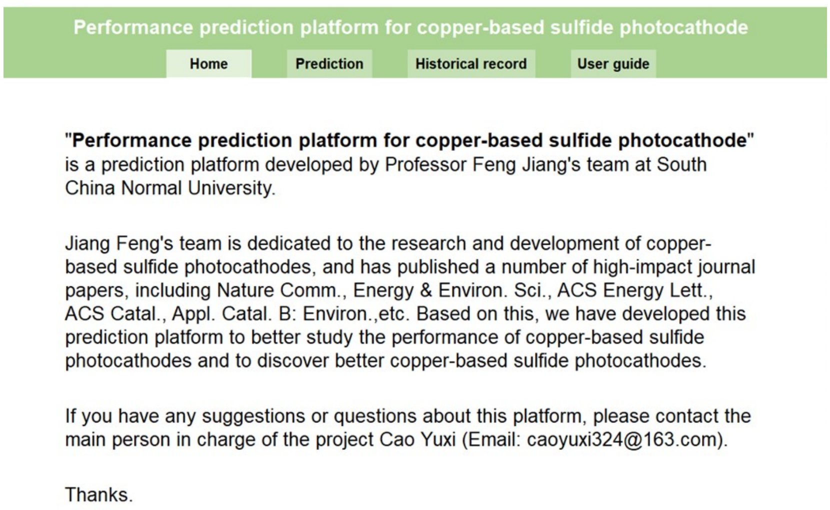

Figure S11. “Home” of Performance prediction platform for copper-based sulfide photocathode

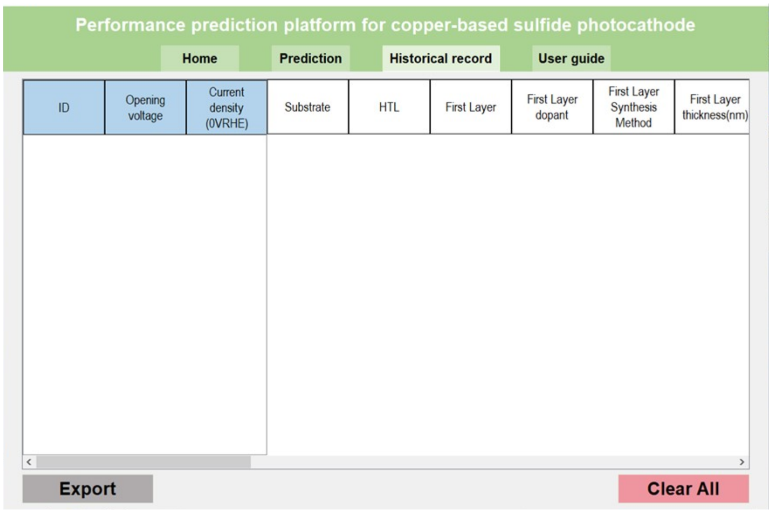

Figure S12. “Historical record” of Performance prediction platform for copper-based sulfide photocathode

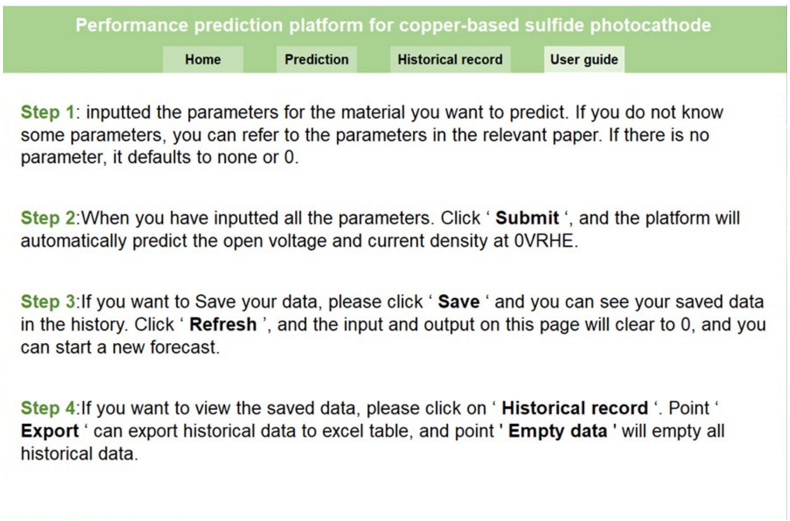

Figure S13. “User guide” of Performance prediction platform for copper-based sulfide photocathode