Supporting Information

Effectiveness and limitation of the performance prediction of perovskite solar cells by process informatics

Ryo Fukasawa,[1] Toru Asahi,[1] Takuya Taniguchi*[2]

[1] Department of Advanced Science and Engineering, Graduate School of Advanced Science and Engineering, Waseda University, 3-4-1 Okubo, Shinjuku-Ku, Tokyo, 169-8555, Japan

[2] Center for Data Science, Waseda University, 1-6-1 Nishiwaseda, Shinjuku-ku, Tokyo 169-8050, Japan

* Correspondence to takuya.taniguchi@aoni.waseda.jp

Contents

*Supporting Tables are supplied in another file (Supporting Information.xlsx)

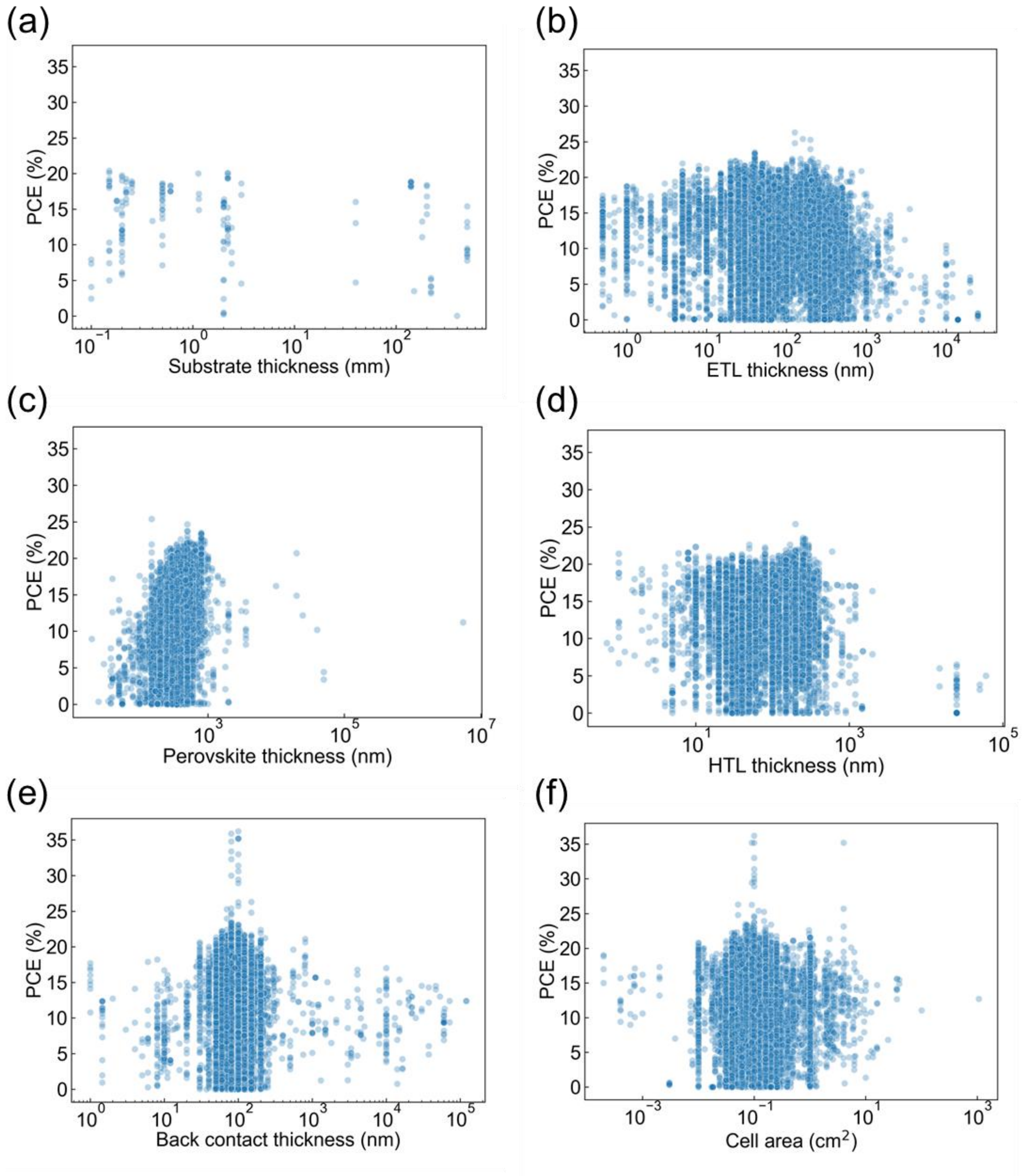


Figure S1. Scatter plots of quantitative variables and PCE. (a) Relationship between substrate thickness and PCE. (b) Relationship between ETL thickness and PCE. (c) Relationship between perovskite thickness and PCE. (d) Relationship between HTL thickness and PCE. (e) Relationship between back contact thickness and PCE. (f) Relationship between cell area and PCE.

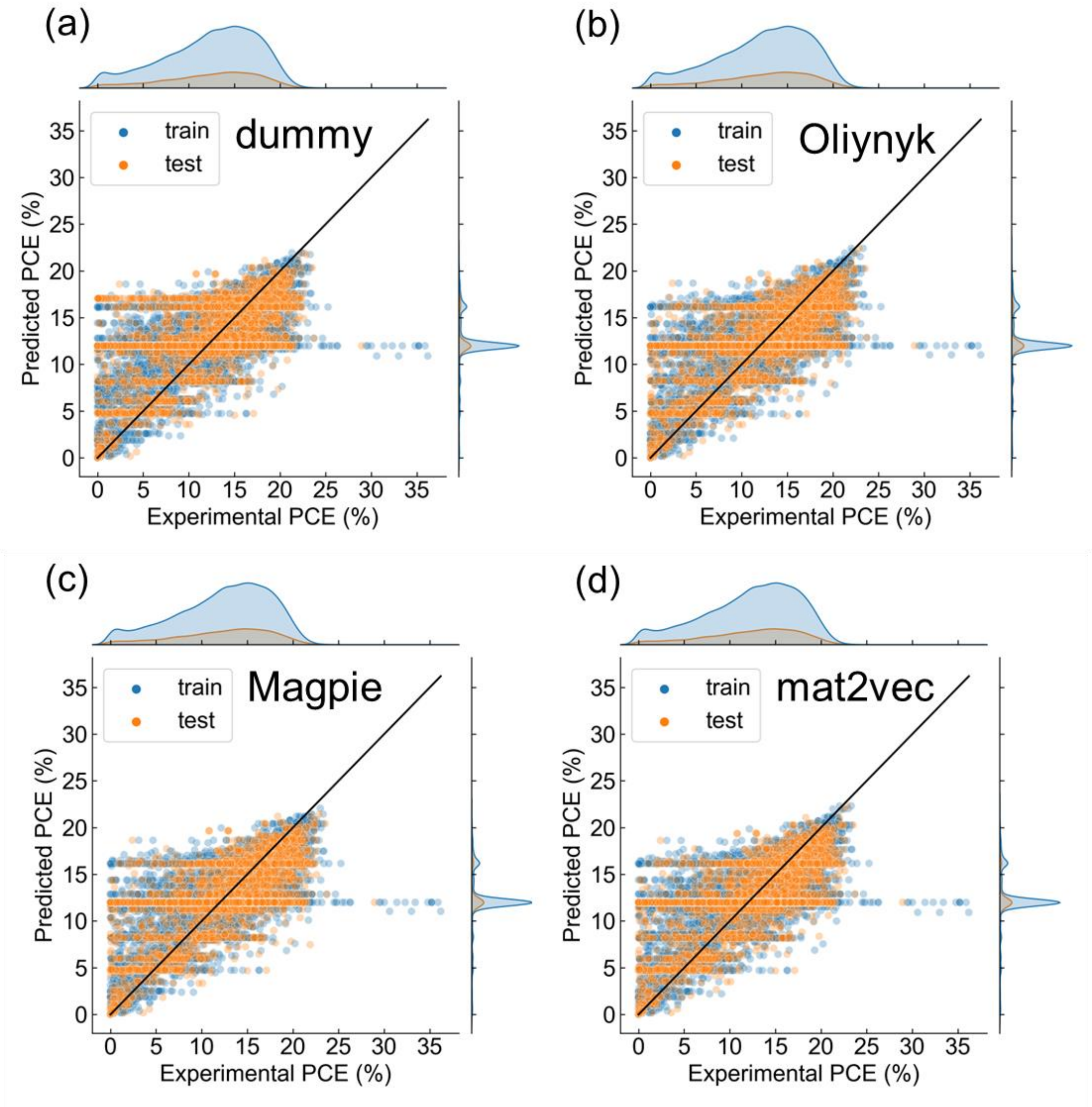


Figure S2. Comparison of joint plots of PCE regression using perovskite composition only represented by several methods. (a) Dummy variable, (b) Oliynyk, (c) Magpie, and (d) mat2vec.

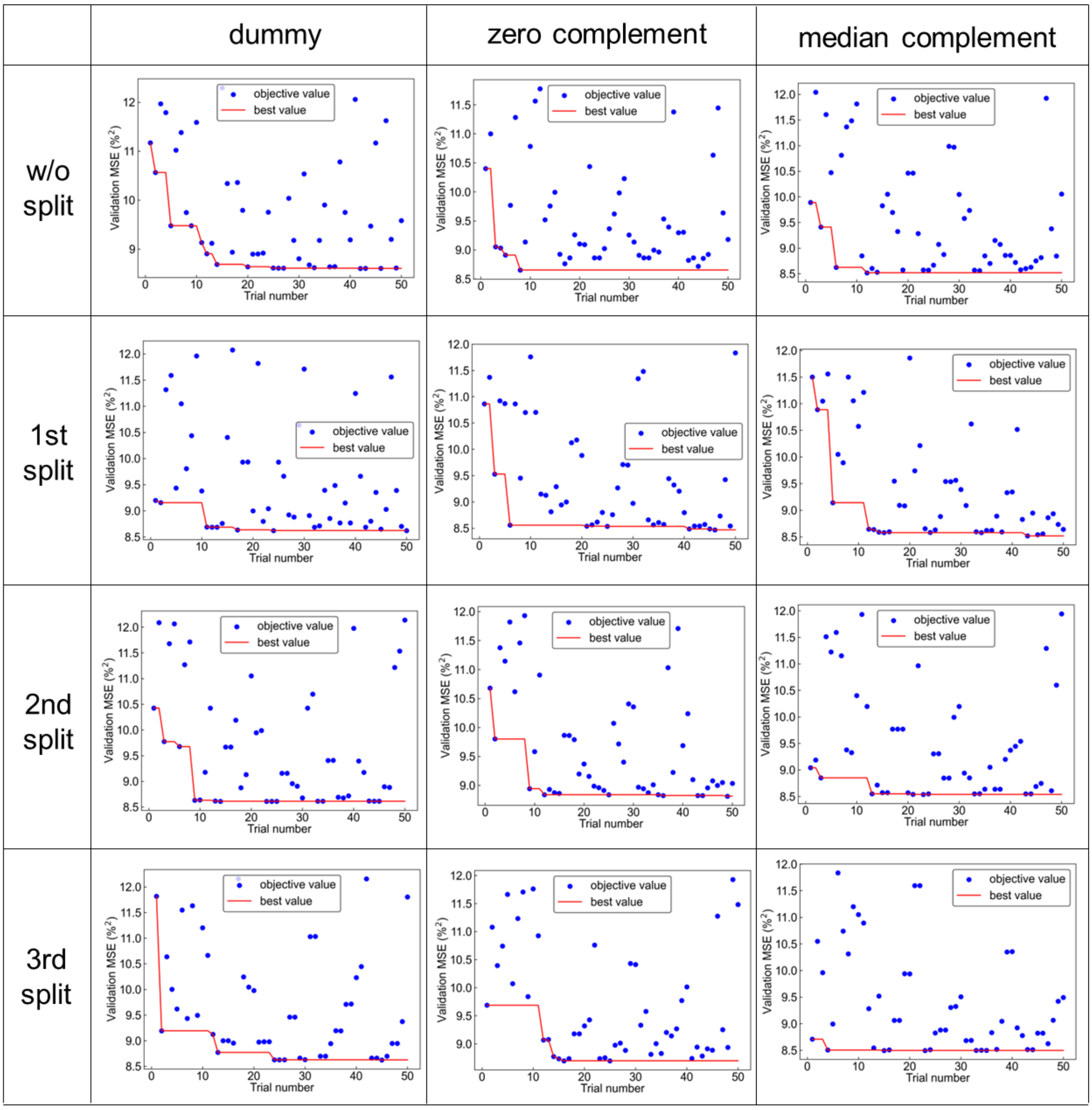


Figure S3. The hyperparameter optimization process of random forest (RF) model.

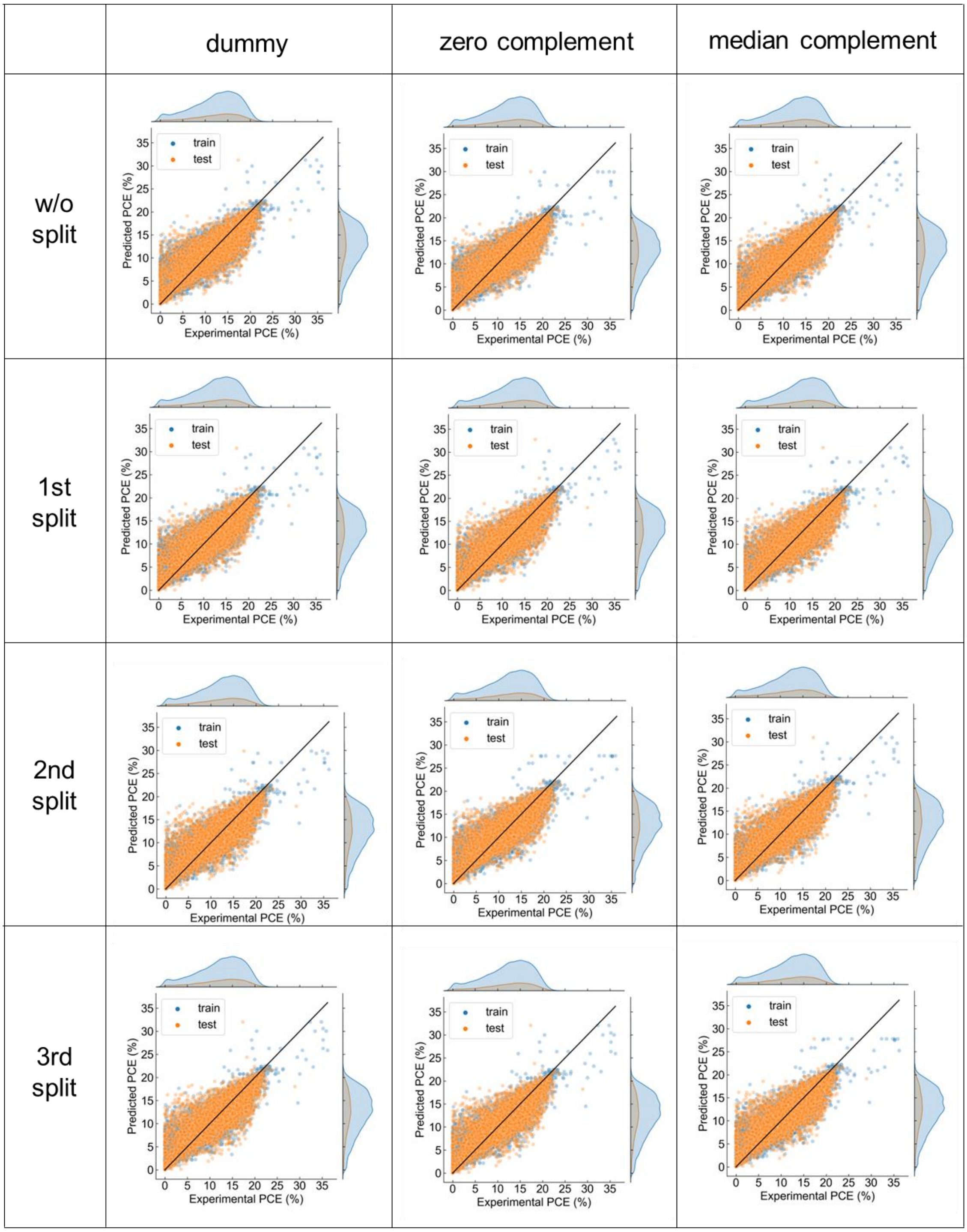


Figure S4. Comparison of joint plots of PCE regression using all columns of materials and processes.

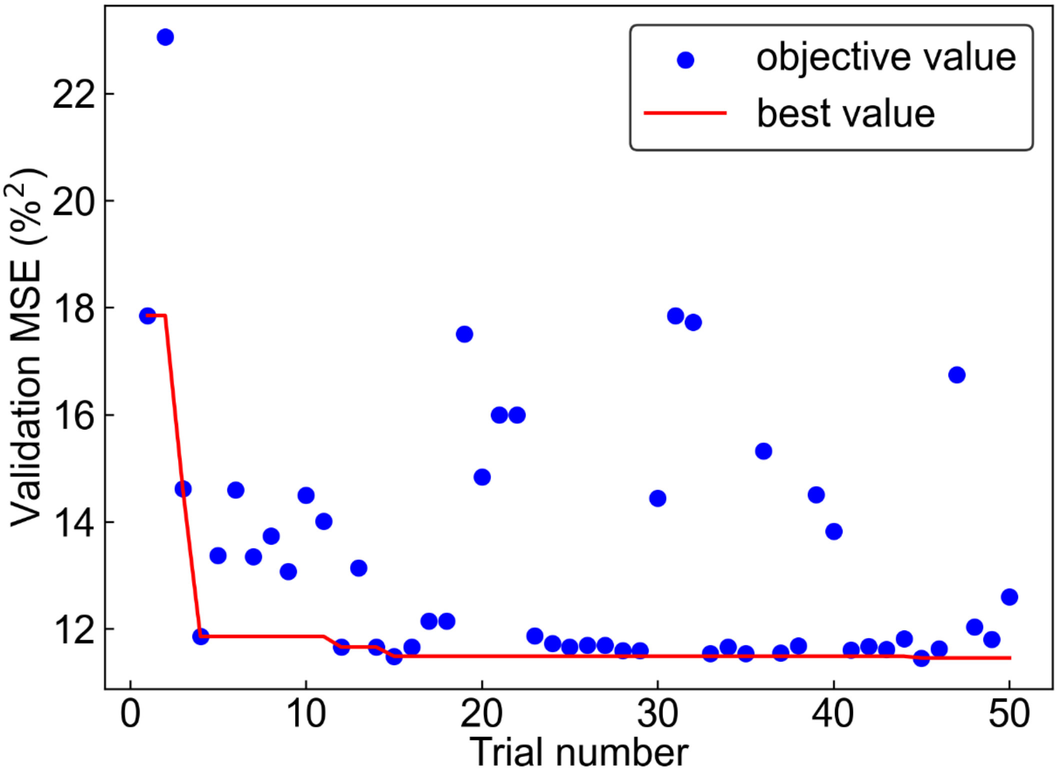


Figure S5. The hyperparameter optimization process of gradient boosting decision tree (GBDT) model when data was vectorized by 1st split and zero complements.

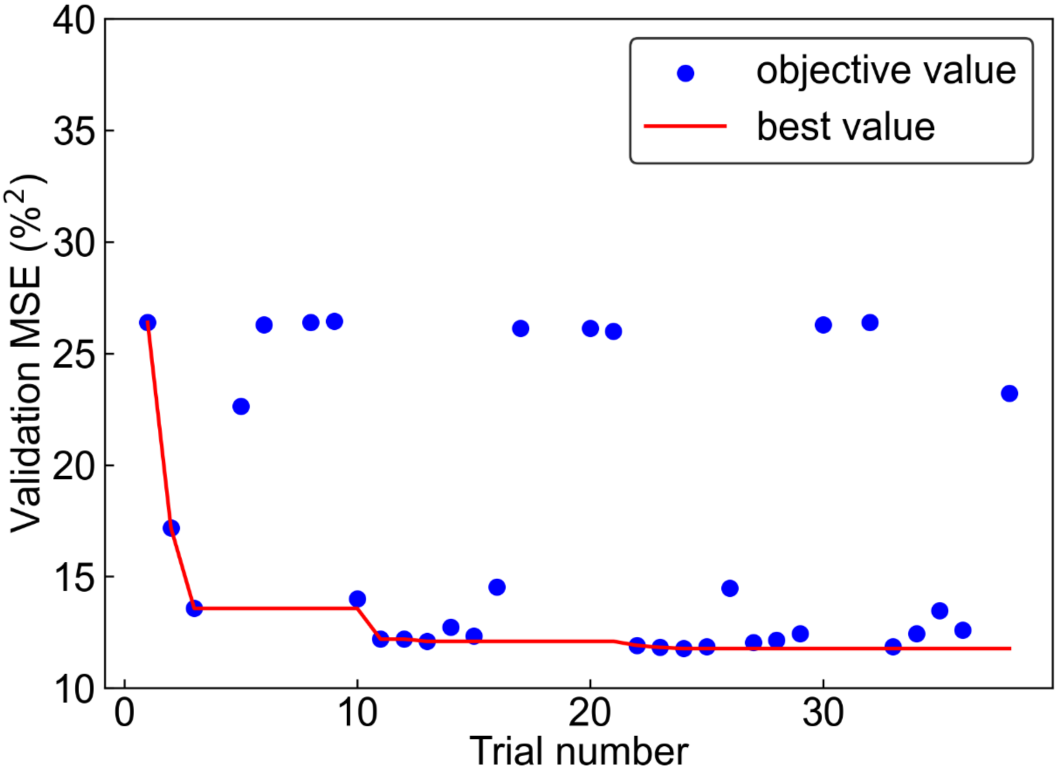


Figure S6. The hyperparameter optimization process of the neural network (NN) model when data was vectorized by 1st split and zero complements.

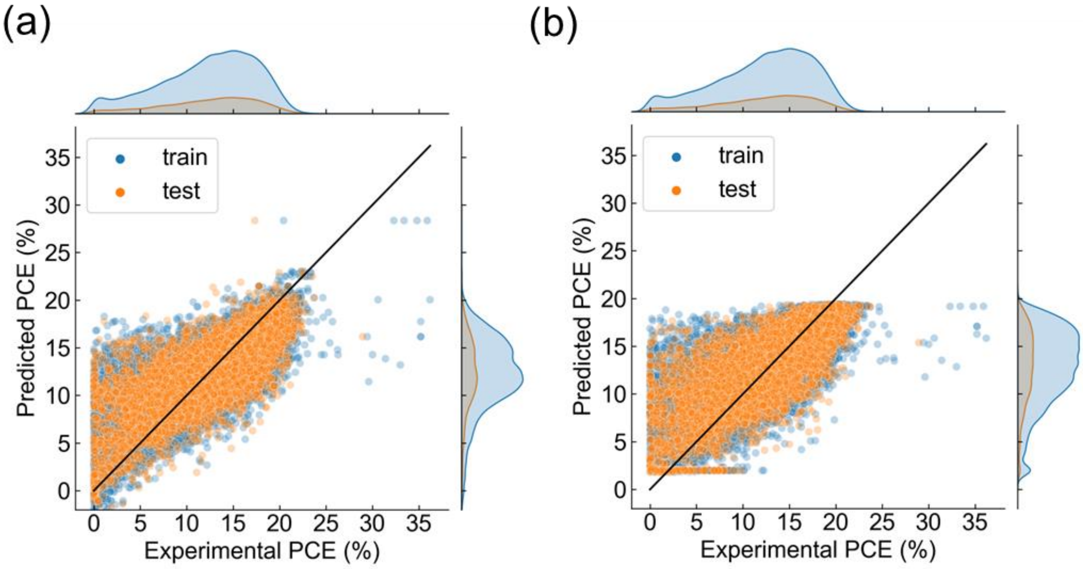


Figure S7. Joint plots of PCE regression of (a) GBDT and (b) NN considering all columns of materials and processes. The data was vectorized by 1st split and zero complements.

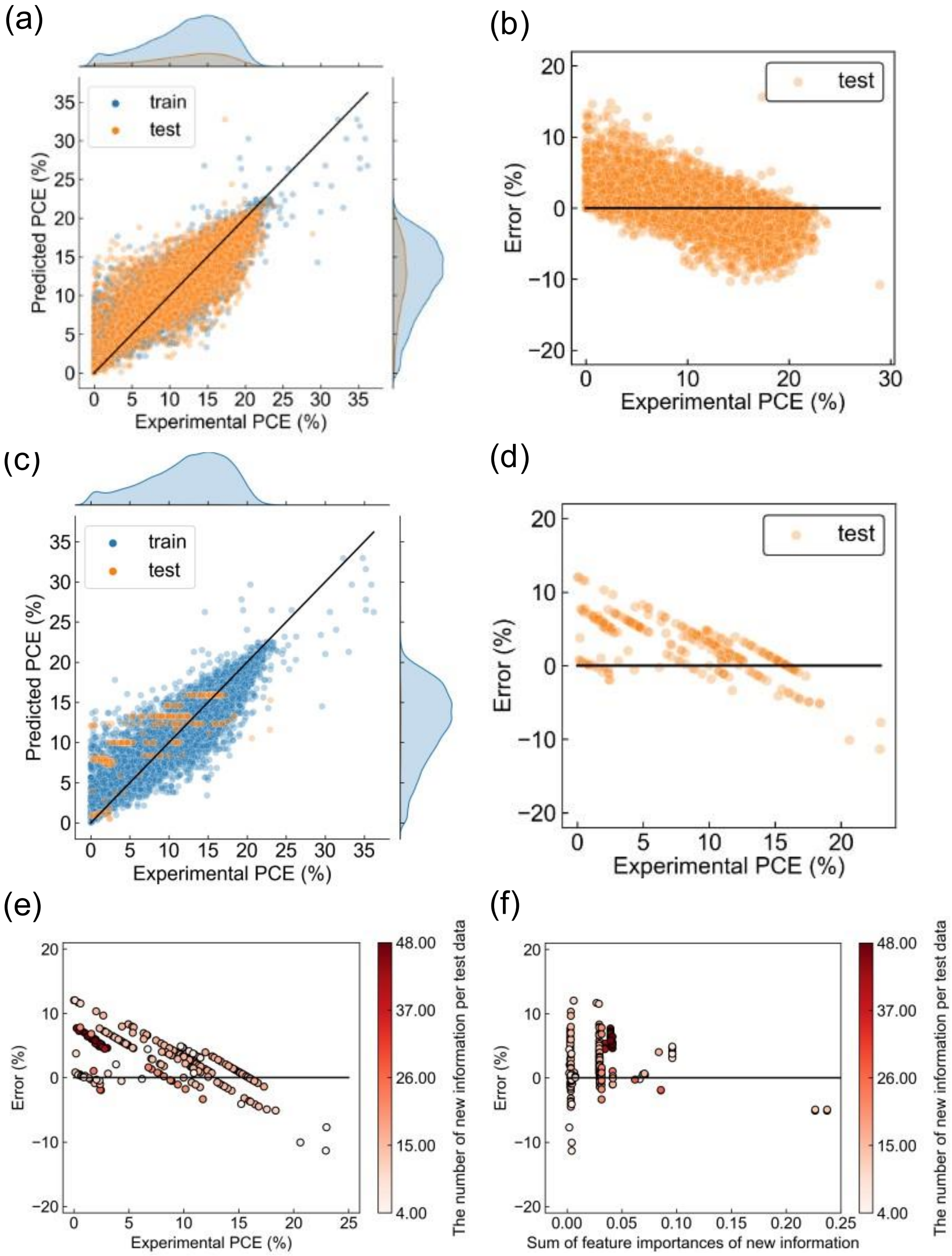


Figure S8. Comparison of test results using different test dataset. (a,b) Test result using the dataset as of 31 March 2022. They are the same as Figure 4d,f in the main text. (c,d) Test result using the newly registered data as of 24 August 2023. The trained model is the same as panel a and b. (e) Error plots color-coded by the number of new information in the vector. (f) Scatter plot showing the relationship between the prediction error and the sum of the feature importances where new information appeared.

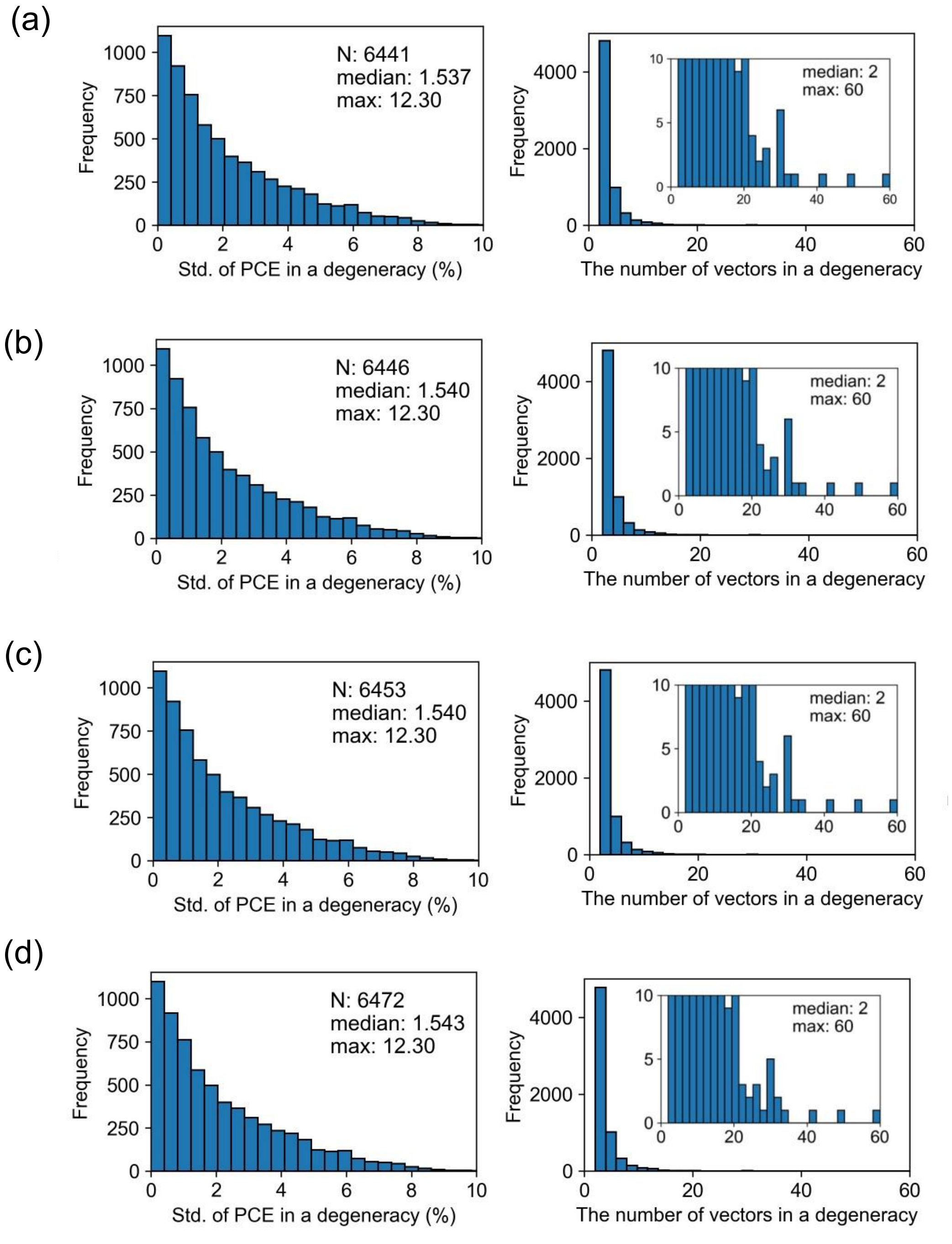
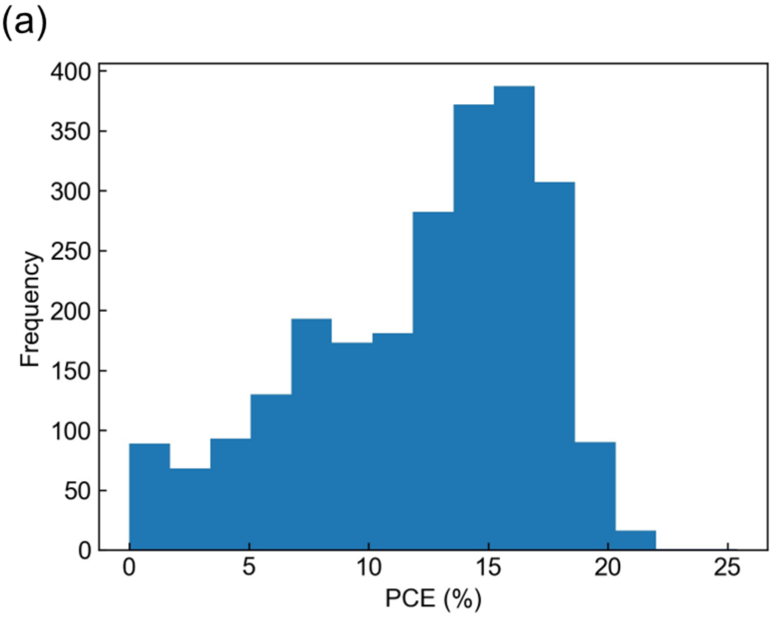


Figure S9. Comparison of distributions of standard deviation of PCE and the number of vectors in a degeneracy when the division of delimiters is changed. (a) without split, (b) 1st split, (c) 2nd split, and (d) 3rd split.



(b)

	value
Counts	2383
mean PCE (%)	12.27
min PCE (%)	0.00
median PCE (%)	13.46
max PCE (%)	25.40

Figure S10. PCE distribution of a perovskite solar cell composed of the most common combination, spincoated MAPbI3 on TiO2 with Spiro. (a) Histgram, and (b) statistics of the distribution.

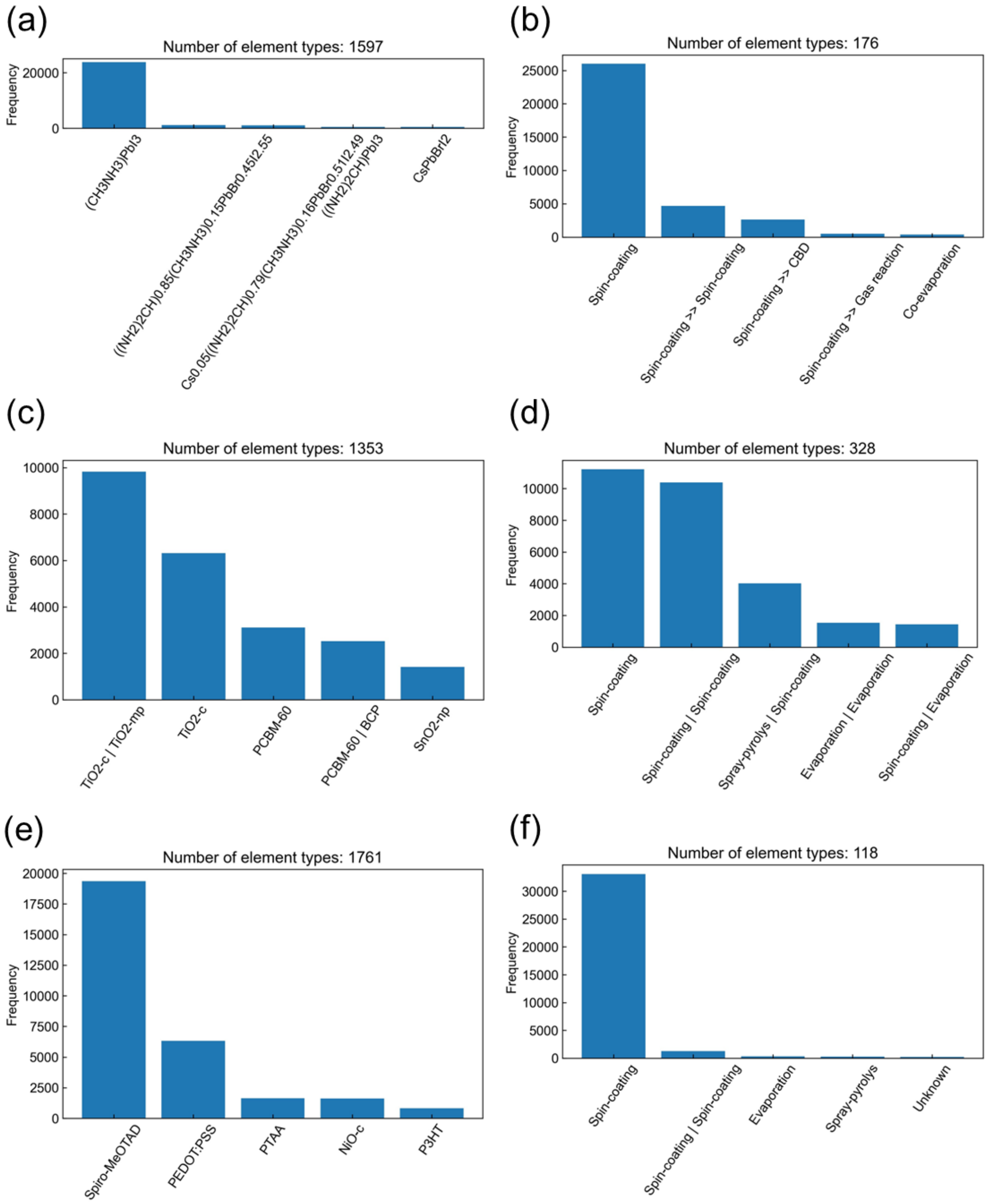
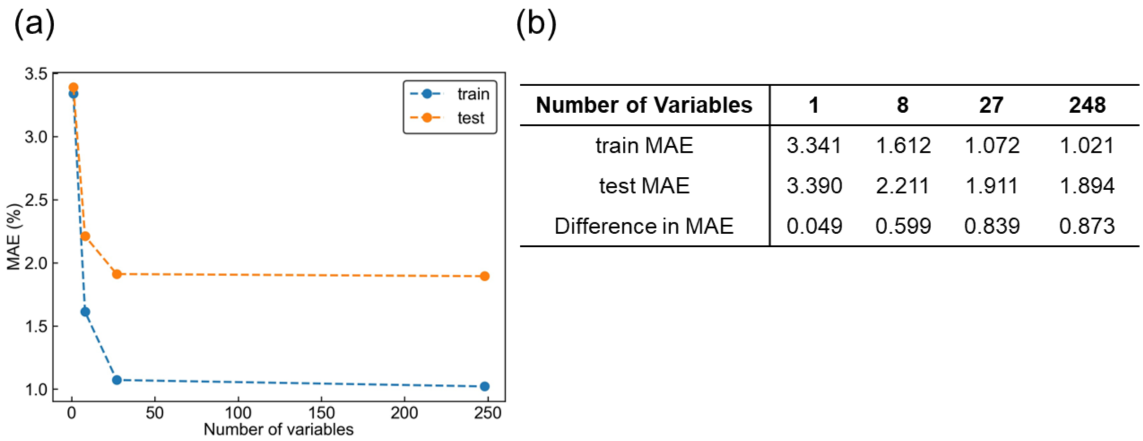


Figure S11. Frequencies of top-5 representative categorical variables in each layer. Ranking of (a) perovskite materials, (b) perovskite deposition methods, (c) ETL materials, (d) ETL deposition methods, (e) HTL materials, and (f) HTL deposition methods.



Number of Variables	1	8	27	248
train MAE	3.341	1.612	1.072	1.021
test MAE	3.390	2.211	1.911	1.894
Difference in MAE	0.049	0.599	0.839	0.873

Figure S12. Comparison of training and test errors depending on the number of selected variables. (a) Plot of MAE versus the number of variables, and (b) MAE values.