# Supplementary material

# Machine learning facilitates the development of interconnecting layers for perovskite/silicon heterojunction tandem solar cells with proof-of-concept efficiency > 38%

Xuejiao Wang[1,2,3,4,5,ξ], Guanlan Chen [1,2,3,4,5,ξ], Ying Liu[1,2,3,4,5], Guangyi Wang[1,2,3,4,5], Wei Han[1,2,3,4,5], Jin Wang[1,2,3,4,5], Pengfei Liu[1,2,3,4,5], Jilei Wang[6], Shaojuan Bao[6], Bo Yu[7], Ying Liu[7], Xinliang Chen[1,2,3,4,5], Shengzhi Xu[1,2,3,4,5], Ying Zhao [1,2,3,4,5,u] and Xiaodan Zhang [1,2,3,4,5,u]

[1]Institute of Photoelectronic Thin Film Devices and Technology, Nankai University, Tianjin 300350, P. R. China

[2]Tianjin Key Laboratory of Efficient Utilization of Solar Energy, Tianjin 300350, P. R. China

[3]State Key Laboratory of Photovoltaic Materials and Cells, Tianjin 300192, P. R. China

[4]Engineering Research Center of Thin Film Photoelectronic Technology of Ministry of Education, Tianjin 300350, P. R. China

[5]Collaborative Innovation Center of Chemical Science and Engineering (Tianjin), Tianjin 300072, P. R. China

[6]Anhui Soltrend New Energy Technology Co., Ltd, Anhui 230000, P. R. China

[7]Yingli Energy Development Co., Ltd, Baoding 071002, P. R. China

ξ These authors contributed equally to this work

u Correspondences to Xiaodan Zhang, E-mail: xdzhang@nankai.edu.cn; Ying Zhao: zhaoygds@nankai.edu.cn;

## Introduction to machine learning

Machine learning (ML) is a field of study that gives computers the ability to learn without being explicitly programmed, this is Arthur Samuel's definition of ML. With the rapid development of Artificial intelligence (AI) technology, ML as a powerful computer tool has been widely used in all walks of life. The main task of ML is to generate a certain "model" from the data through the computer with the help of an algorithm, which is called an ML algorithm. Through ML algorithms, as long as the data obtained from experience is learned, a certain model will be obtained. In the actual situation, new data will be generated constantly, and at this time, as long as the new data is substituted into the obtained model, the new data can be analyzed, predicted and judged. The two main types of ML are supervised learning and unsupervised learning. Generally speaking, supervised learning is "teaching machine learning", and unsupervised learning is "machine learning by itself". Supervised learning is mainly for regression issues and classification issues, and clustering is a typical representative of unsupervised learning. This paper mainly adopts the way of supervised learning to carry out machine learning.

## Pearson correlation coefficient (*r*)

The Pearson correlation coefficient (*r*) is a statistical measure used to measure the degree of linear correlation between two variables. It has a value between -1 and 1 and is commonly used to evaluate the correlation between two continuous variables.

$$r = \frac{\sum(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum (x_i - \overline{x})^2 \sum (y_i - \overline{y})^2}}$$

Where $x_i$ and $y_i$ are the ith observations of the variables X and Y. $\overline{x}$ and $\overline{y}$ are the mean of X and Y. The Pearson correlation coefficient *r* ranges from -1 to 1. An $r > 0$ indicates that the mean of the continuous variable is higher when the binary variable equals 1, demonstrating a positive correlation. Conversely, an $r < 0$ suggests that the mean of the continuous variable is lower when the binary variable equals 1, indicating a negative correlation. The closer *r* is to 1, the stronger the correlation; the closer it is to 0, the weaker the correlation.

## Point-biserial correlation coefficient

The Point-biserial correlation coefficient is a measure of the correlation between a continuous variable and a binary variable. It is a special case of Pearson correlation coefficient, which is suitable for the analysis of the relationship between discrete and continuous variables of binary classification (0/1).

$$r = \frac{\overline{X}_1 - \overline{X}_0}{s} \cdot \sqrt{\frac{n_1 - n_0}{n^2}}$$

Where $\overline{X}_1$ and $\overline{X}_0$ are the means of the continuous variables when the binary variables are 1 and 0, respectively, $s$ is the population standard deviation of a continuous variable, $n_1$ and $n_0$ are the number of samples of 1 and 0 in a binary categorical variable and n is the total number of samples ($n = n_1+n_0$). The point-biserial correlation coefficient $r$ ranges from -1 to 1. An $r > 0$ indicates that the mean of the continuous variable is higher when the binary variable equals 1, demonstrating a positive correlation. Conversely, an $r < 0$ suggests that the mean of the continuous variable is lower when the binary variable equals 1, indicating a negative correlation. The closer $r$ is to 1, the stronger the correlation; the closer it is to 0, the weaker the correlation.

The p-value reflects the statistical significance of the correlation coefficient:

- $p < 0.05$: The correlation is considered significant, implying it is unlikely to have occurred by chance.
- $p \geq 0.05$: The correlation may not be significant, suggesting it could be random.

## One-way ANOVA

One-way ANOVA is a statistical method used to determine whether there are significant differences between the means of three or more independent groups. It compares the variance between groups to the variance within groups to evaluate whether the grouping factor affects the dependent variable. The ratio of these variances is called the F-statistic, calculated as:

$$\mathrm{F} = \frac{\mathrm{Between-Group\ Variance}}{\mathrm{Within-Group\ Variance}}$$

The P-value represents the probability of observing the current F-value or a more

extreme result under the null hypothesis (that is, assuming that all groups have equal means). P ≤ 0.05 means that at least one group has a significantly different mean from the others, while P > 0.05 was considered to have no significant difference in the mean values of all groups.

**F-Statistic**:

- Represents the ratio of variance between groups to variance within groups.
- A higher F-value indicates greater differences between group means relative to random noise.

**p-Value**:

- If p < 0.05, reject the null hypothesis. This means at least one group mean is significantly different.
- If p ≥ 0.05, fail to reject the null hypothesis, indicating no significant difference between group means.

## Lasso regression model

Lasso regression, or Least Absolute Shrinkage and Selection Operator Regression, is a type of linear regression that performs both variable selection and regularization to improve the accuracy and interpretability of the statistical model it produces. There are three key characteristics of Lasso regression, the first one is regularization, which adds a penalty term to the loss function to prevent overfitting and handle multicollinearity. The second one is featuring selection, which shrinks coefficients of less important features to exactly zero, effectively removing them from the model and the last one is penalty term, which is based on the L1 norm of the coefficients, encouraging sparsity in the model. The objective function for Lasso regression is:

$$\text{Minimize: } \frac{1}{2n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

Where $y_i$ is actual value, $\hat{y}_i$ is predicted value, $\beta_j$ is coefficients of the features and λ is regularization strength (controls the degree shrinkage).

## Random forest model

Random Forest (RF) is a versatile ensemble learning method primarily used for classification and regression tasks. It is based on constructing a multitude of decision

trees during training and combining their outputs (majority vote for classification or average for regression) to improve predictive performance and reduce overfitting. There are also three key features of RF. The first one is ensemble of decision tress. A random forest is essentially a collection of decision trees, each tree is trained on a randomly sampled subset of the data (with replacement, i.e., bootstrapping). The second one is featuring randomness, which means during training, only a random subset of features is considered at each split in the tree, this helps reduce correlation among trees and increases diversity. The third one is Aggregation: for classification, the final prediction is made by majority voting among the trees. For regression, the output is the average of the predictions from all trees.

The RF algorithm consists of an ensemble of decision trees. During the tree-building process, each tree selects features to split the data in a way that maximizes the improvement in data purity. Feature importance is calculated based on the average contribution of each variable to the reduction in impurity (e.g., root mean squared error) across all trees. When a feature is frequently used as a splitting variable and significantly reduces prediction error, its importance score increases. In RF, the contributions of each feature across all trees are aggregated and normalized. The resulting values represent the normalized importance of each feature, typically summing to one.

## Neural network model

Neural network (NN) model is a machine learning method inspired by the structure and functioning of biological neural networks. It is designed to identify complex patterns and relationships in data by using a layered structure composed of interconnected nodes (neurons). Neural networks are widely used in tasks like regression, classification, image recognition, natural language processing, and more. Its basic structure can be divided into the following five parts, as listed in Table S1.

Table S1. Five parts of NN.

| Structure | Explanation |
|---|---|
| Input layer | Accepts the input data, with each feature of the data corresponding to one neuron in this layer. |

| | |
|---|---|
| Hidden layers | Does not perform computations—simply passes the data forward.<br>Consist of one or more layers of neurons between the input and output layers.<br>Each neuron applies a mathematical function (activation function) to the inputs and passes the results to the next layer.<br>These layers extract and learn features from the data. |
| Output layer | Provides the final predictions or classifications.<br>For regression, the output is usually a single continuous value.<br>For classification, the output layer size corresponds to the number of classes, often with a SoftMax function for probabilities. |
| Connections | Every neuron is connected to the neurons in adjacent layers through weights. Each connection is assigned a weight that determines its importance. |
| Bias | An additional parameter added to neurons to shift the activation function, improving the model's ability to learn |

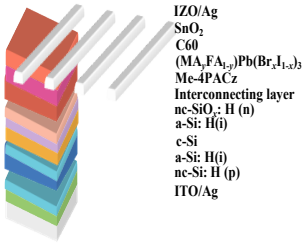


Fig. S1. The schematic diagram of the tandem solar cell structure.

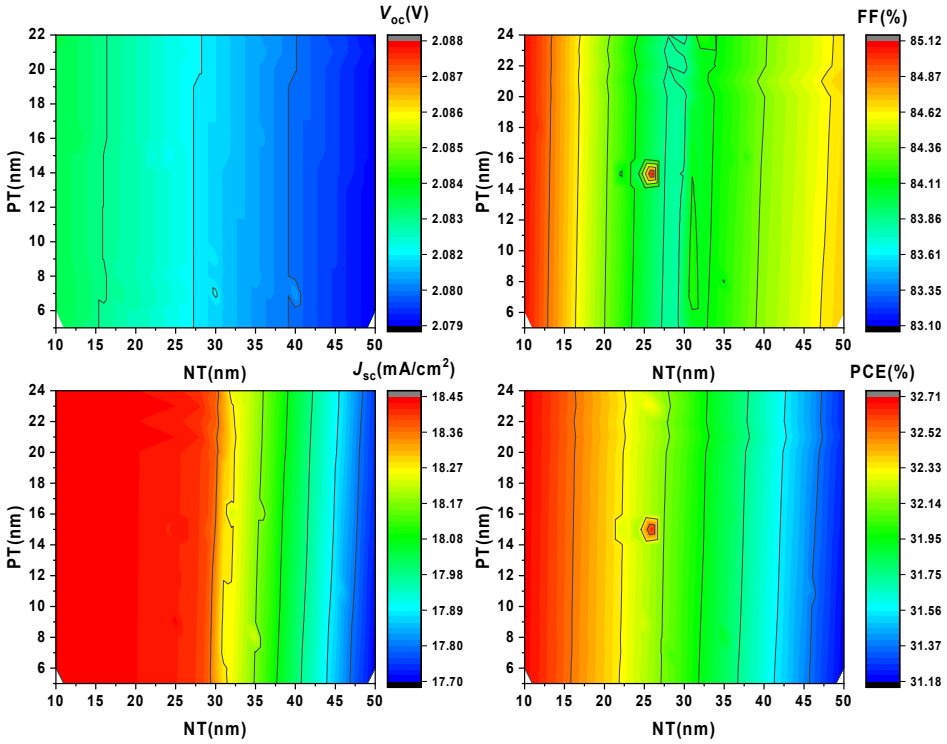


Fig. S2. Simulated device performance as a function of Si-ICL.

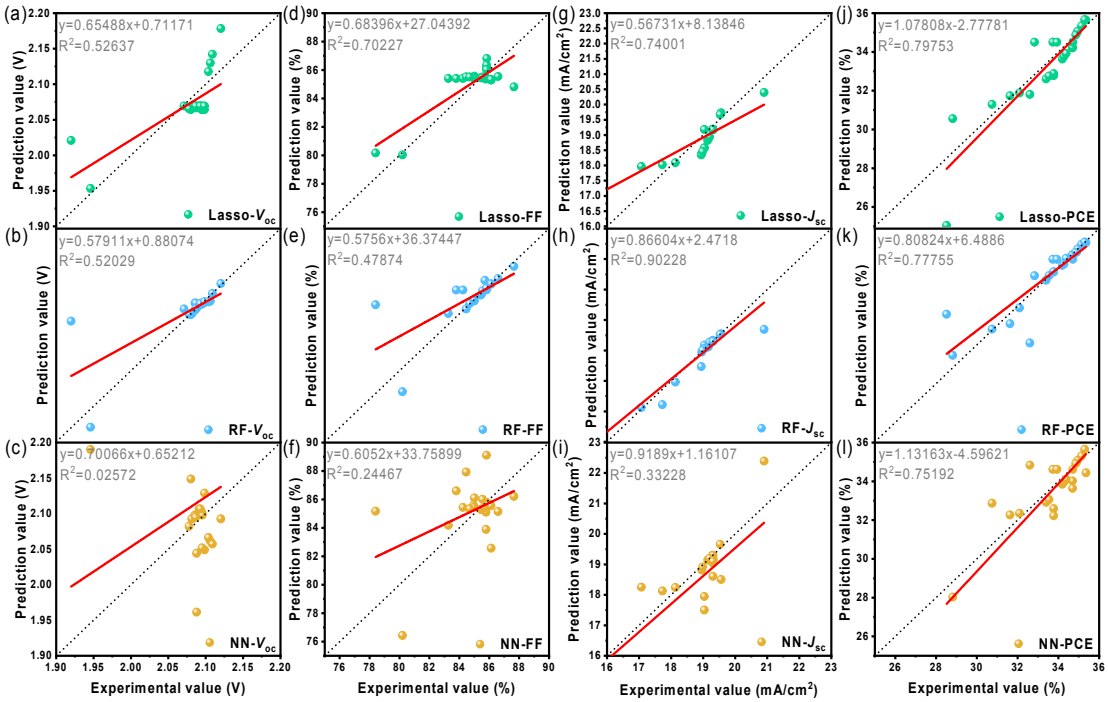


Fig. S3. Tandem solar cell performance predicted by different models based on dataset of TCO-ICL: (a-c) $V_{oc}$, (d-f) FF, (g-i) $J_{sc}$, (j-l) PCE. Models used include Lasso, RF and NN. The dashed line represents a perfect fitting, while the solid red line represents the actual fitting result.

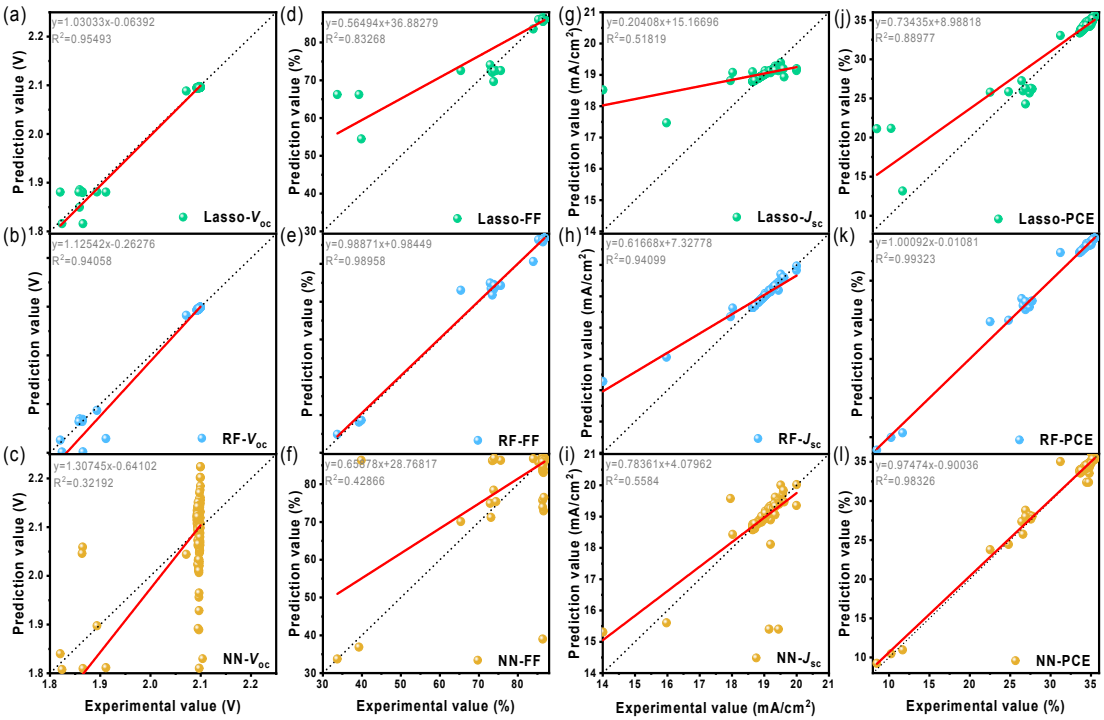


Fig. S4. Tandem solar cell performance predicted by different models based on dataset of Si+TCO-ICL: (a-c) $V_{oc}$, (d-f) FF, (g-i) $J_{sc}$, (j-l) PCE. Models used include Lasso, RF and NN. The dashed line represents a perfect fitting, while the solid red line represents the actual fitting result.

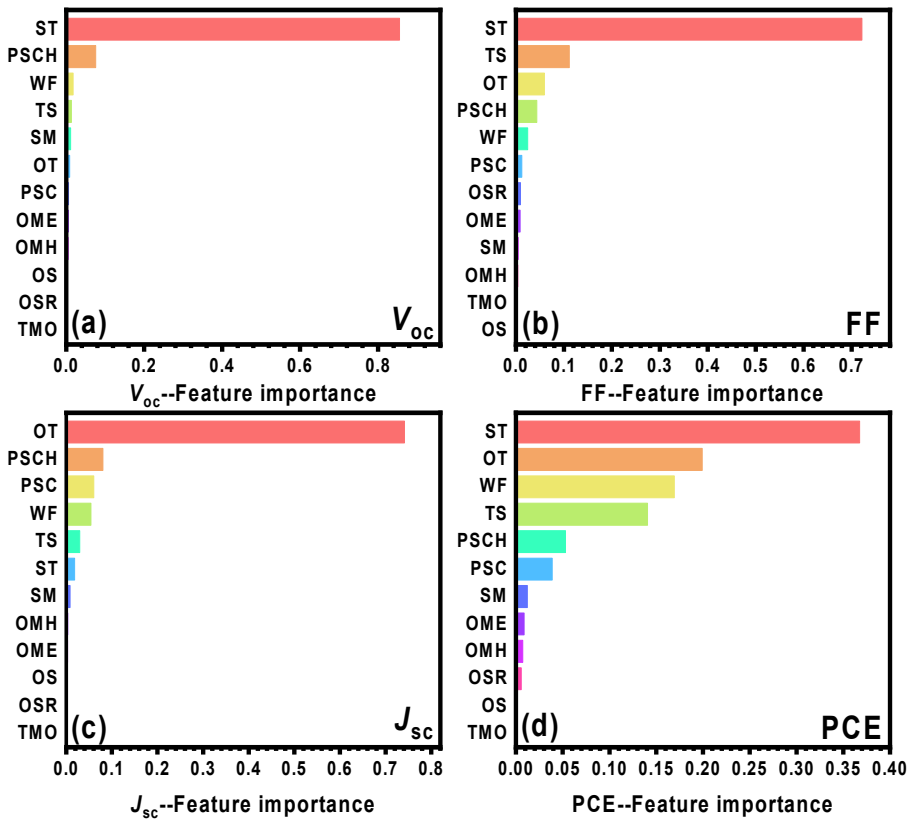


Fig. S5. Feature importance for dataset of TCO-ICL based on the RF model. (a) $V_{oc}$, (b) FF, (c) $J_{sc}$, (d) PCE.

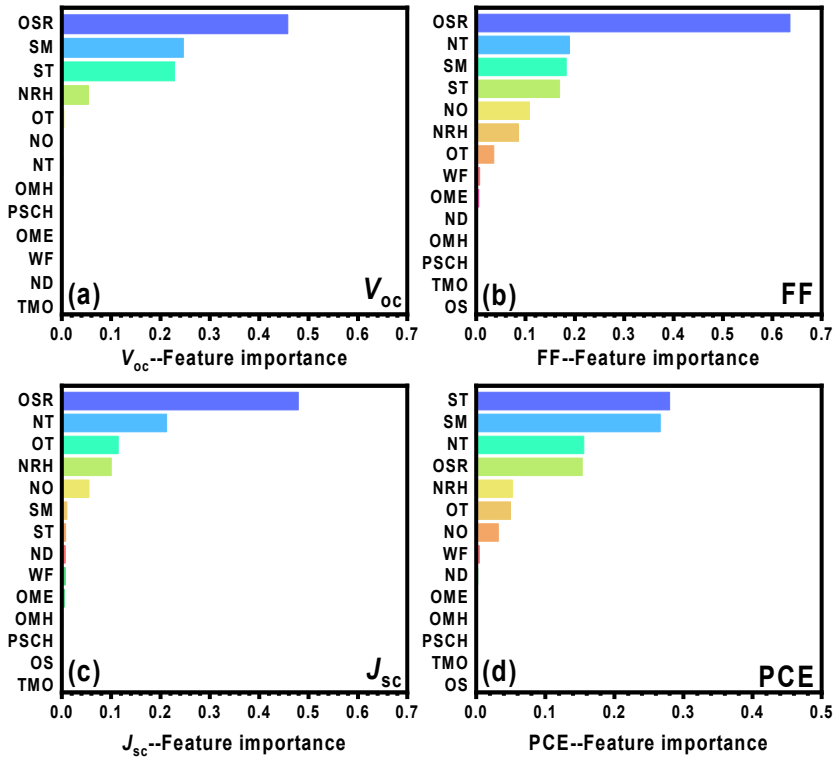


Fig. S6. Feature importance for dataset of Si+TCO-ICL based on the RF model. (a) $V_{oc}$, (b) FF, (c) $J_{sc}$, (d) PCE.

Table S2. Point-biserial correlation coefficient results of dataset of Si-ICL.

| Category Variable | Numeric Variables | Correlation coefficient *r* | P Value |
|---|---|---|---|
| SM | $V_{oc}$ | –0.9414323515397796 | $9.7483407e^{-99}$ |
| SM | $J_{sc}$ | 0.14571982356195037 | 0.0361671 |
| SM | FF | –0.9125988991043305 | $1.4347159e^{-81}$ |
| SM | PCE | –0.9073200492344362 | $4.4338796e^{-79}$ |

Table S3. Point-biserial correlation coefficient results of dataset of TCO-ICL.

| Category Variables | Numeric Variables | Correlation coefficient *r* | P Value |
|---|---|---|---|
| SM | $V_{oc}$ | –0.579666053 | $4.93e^{-10}$ |
| SM | $J_{sc}$ | 0.270775911 | 0.007305854 |
| SM | FF | –0.563437622 | $1.87e^{-9}$ |
| SM | PCE | –0.375159162 | $1.53e^{-4}$ |
| PSC | $V_{oc}$ | 0.637346795 | $2.23e^{-12}$ |
| PSC | $J_{sc}$ | 0.31318726 | $1.79e^{-3}$ |
| PSC | FF | 0.614959877 | $2.07e^{-11}$ |
| PSC | PCE | 0.64375799 | $1.14e^{-12}$ |

Table S4. Point-biserial correlation coefficient results of dataset of Si+TCO-ICL.

| Category Variable | Numeric Variables | Correlation coefficient *r* | P Value |
|---|---|---|---|
| SM | $V_{oc}$ | –0.591615133 | $8.82e^{-48}$ |
| SM | $J_{sc}$ | –0.158119715 | 0.00043071 |
| SM | FF | –0.687284807 | $4.91e^{-70}$ |
| SM | PCE | –0.766577847 | $2.59e^{-96}$ |

Table S5. One-way ANOVA results of dataset of Si+TCO-ICL.

| Categorical Variables | Numerical Variables | F Value | P Value |
|---|---|---|---|
| TMO | $V_{oc}$ | 0.055604 | 0.994236 |
| TMO | $J_{sc}$ | 0.027379 | 0.998549 |
| TMO | FF | 0.058606 | 0.993622 |
| TMO | PCE | 0.090014 | 0.985569 |
| PSCH | $V_{oc}$ | 0.077154 | 0.998243 |
| PSCH | $J_{sc}$ | 0.064096 | 0.998962 |
| PSCH | FF | 0.031515 | 0.999867 |
| PSCH | PCE | 0.059965 | 0.999143 |

Table S6. Details of relevant hyperparameters for the ML models in Si-ICL dataset.

| Models | Output Features | Alpha | Max_iter | Tol |
|---|---|---|---|---|
| Lasso | $V_{oc}$ | 0.001 | 500 | 10 |
| | $J_{sc}$ | 0.01 | 3000 | 0.1 |
| | FF | 0.1 | 2000 | 0.01 |
| | PCE | 0.001 | 500 | 10 |
| Models | Output Features | Max_depth | | N_estimators |
| RF | $V_{oc}$ | 5 | | 400 |
| | $J_{sc}$ | 11 | | 200 |
| | FF | 11 | | 550 |
| | PCE | 5 | | 550 |
| Models | Output Features | Alpha | | Hidden_layer_sizes |
| NN | $V_{oc}$ | 0.001 | | (50,50) |
| | $J_{sc}$ | 0.001 | | (100,100) |
| | FF | 0.01 | | (50,50) |
| | PCE | 0.01 | | (100,100) |

Table S7. Details of relevant hyperparameters for the ML models in TCO-ICL dataset.

| Models | Output Features | Alpha | Max_iter | Tol |
|---|---|---|---|---|
| Lasso | $V_{oc}$ | 0.01 | 8000 | 0.1 |
| | $J_{sc}$ | 0.08 | 8000 | 0.001 |
| | FF | 0.05 | 8000 | 0.001 |
| | PCE | 0.01 | 8000 | 0.01 |
| Models | Output Features | Max_depth | | N_estimators |
| RF | $V_{oc}$ | 5 | | 200 |
| | $J_{sc}$ | 8 | | 100 |
| | FF | 7 | | 400 |
| | PCE | 13 | | 200 |
| Models | Output Features | Alpha | | Hidden_layer_sizes |
| NN | $V_{oc}$ | 0.1 | | (100,100) |
| | $J_{sc}$ | 0.0001 | | (50,50) |
| | FF | 0.000001 | | (50,50) |
| | PCE | 0.00001 | | (100,100) |

Table S8. Details of relevant hyperparameters for the ML models in Si+TCO-ICL dataset.

| Models | Output Features | Alpha | Max_iter | Tol |
|---|---|---|---|---|
| Lasso | $V_{oc}$ | 0.001 | 8000 | 0.01 |
| | $J_{sc}$ | 0.1 | 8000 | 0.001 |
| | FF | 1 | 8000 | 0.001 |
| | PCE | 0.1 | 8000 | 0.01 |
| Models | Output Features | Max_depth | | N_estimators |
| RF | $V_{oc}$ | 7 | | 500 |
| | $J_{sc}$ | 14 | | 500 |
| | FF | 8 | | 500 |
| | PCE | 9 | | 500 |
| Models | Output Features | Alpha | | Hidden_layer_sizes |
| NN | $V_{oc}$ | 0.000001 | | (100,100) |
| | $J_{sc}$ | 0.1 | | (100,100) |
| | FF | 0.1 | | (50,50) |
| | PCE | 0.001 | | (50,50) |

Table S9. Performances of different ML model in the prediction of interconnecting layer of the perovskite/silicon heterojunction tandem solar cells in dataset of Si-ICL.

| Photovoltaic Parameters | ML model | RMSE |
|---|---|---|
| $V_{oc}$ | Lasso | 0.015446611151380513 |
| | RF | 0.015099898358605316 |
| | NN | 0.1602053538591303 |
| $J_{sc}$ | Lasso | 0.286992626872376 |
| | RF | 0.27951616820264574 |
| | NN | 1.120907945816788 |
| FF | Lasso | 0.9567102056018761 |
| | RF | 0.8682119604307589 |
| | NN | 3.183064186323462 |
| PCE | Lasso | 0.6176109384517436 |
| | RF | 0.5764471184651523 |
| | NN | 1.9239972952538564 |

Note: RMSE refer to root mean squared error.

Table S10. Performances of different ML model in the prediction of interconnecting layer of the perovskite/silicon heterojunction tandem solar cells in dataset of TCO-ICL.

| Photovoltaic Parameters | ML model | RMSE |
|---|---|---|
| $V_{oc}$ | Lasso | 0.042788177574249193 |
| | RF | 0.020475879062268873 |
| | NN | 0.12167004946699936 |
| $J_{sc}$ | Lasso | 0.6108993088784077 |
| | RF | 0.36550811937001726 |
| | NN | 1.2081575596461083 |
| FF | Lasso | 1.8687498275760739 |
| | RF | 1.3873067486307413 |
| | NN | 4.827784352395491 |
| PCE | Lasso | 1.2841844521563393 |
| | RF | 1.2143686975396648 |
| | NN | 1.9579579108977 |

Note: RMSE refer to root mean squared error.

Table S11. Performances of different ML model in the prediction of interconnecting layer of the perovskite/silicon heterojunction tandem solar cells in dataset of Si+TCO-ICL.

| Photovoltaic Parameters | ML model | RMSE |
|---|---|---|
| $V_{oc}$ | Lasso | 0.059321347021387714 |
| | RF | 0.0544010083819987 |
| | NN | 0.17480281827306443 |
| $J_{sc}$ | Lasso | 0.9267432287180262 |
| | RF | 0.7116399034301324 |
| | NN | 1.271500876890284 |
| FF | Lasso | 5.419571813587112 |
| | RF | 4.16788514484323 |
| | NN | 5.552792000546463 |
| PCE | Lasso | 2.2687260161718403 |
| | RF | 1.7919828243834686 |
| | NN | 2.4296643153898856 |

Note: RMSE refer to root mean squared error.