Solar Energy 228 (2021) 45–52

Contents lists available at ScienceDirect

Solar Energy

journal homepage: www.elsevier.com/locate/solener

Key factors governing the device performance of CIGS solar cells: Insights from machine learning

Chengwan Zhu [a,b,1], Wu Liu [a,b,1], Yaoyao Li [a,b], Xiaomin Huo [a,b], Haotian Li [a,b], Kai Guo [c], Bo Qiao [a,b], Suling Zhao [a,b], Zheng Xu [a,b], Honge Zhao [a,b], Dandan Song [a,b,*]

[a] Key Laboratory of Luminescence and Optical Information, Beijing Jiaotong University, Ministry of Education, Beijing 100044, China
[b] Institute of Optoelectronics Technology, Beijing Jiaotong University, Beijing 100044, China
[c] Chongqing Shenhua Thin Film Solar Technology Co., Ltd, Chongqing 400714, China

ARTICLE INFO

Keywords:
Solar cells
Chalcopyrite
$Cu(In,Ga)Se_2$ (CIGS)
Machine Learning
Alkali Post Deposition Treatment (PDT)
Ga gradient

ABSTRACT

$Cu(In_{1-x},Ga_x)Se_2$ (CIGS) solar cells are a kind of highly efficient thin film solar cells, further breakthrough in their device efficiency relies on the development of advanced methods and/or deep insights into the factors governing the device efficiency. Herein, we use the machine learning (ML) algorithms to explore the key factors governing the device performance of the CIGS solar cells and the underlying correlations. The datasets for ML are obtained from the experimental reports, which enables the results more referable for experimental optimization. Key factors governing the device performance are screened based on the correlation studies. The ML algorithms including linear regression, neural network, random forest (RF) and extreme gradient boosting are employed, among which RF performs best in predicting the efficiency of the CIGS solar cells with high accuracy (root mean square error of 0.9% and1.8%, Pearson coefficient r of 0.9 and 0.88 for validation and test sets, respectively). Furthermore, the factors and their optimal scales for high device efficiency are predicted, which provides essential guidance for experimental device optimization.

1. Introduction

Photovoltaic technology is crucial for establishing the global renewable energy system, and among the photovoltaic techniques, $Cu(In,Ga)Se_2$ (CIGS) thin film solar cells are of great potential due to their compatibility to the building integrated photovoltaics and their ability in large amount of power supply in photovoltaic power station (Ochoa et al., 2020; Muzzillo, 2017; Dhere, 2007). Maximization of the photoelectric conversion efficiency (PCE) is essential for the competitive development of photovoltaic technologies. The record PCE of CIGS solar cells has surpassed 23% (Nakamura et al., 2019), which is close to these of the lead halide perovskite solar cells and crystalline silicon solar cells.

Recent developments focus on the growth conditions and device engineering, which are key factors to improve the device performance and material quality (Ochoa et al., 2020). Specifically, the introduction of post-deposition treatments (PDT) (Salomé et al., 2015; Khatri et al., 2016; Reinhard et al., 2015), the GGI ([Ga]/([Ga] + [In])) ratio gradient (Li et al., 2014; Schleussner et al., 2011; Witte et al., 2015), and the developing of Cd-free buffer layers (Friedlmeier et al., 2015; Nakada et al., 2001; Ohtake et al., 1997; Li et al., 2019a; Bhattacharya and Ramanathan, 2004) are efficient ways, which promote the record efficiencies of the CIGS solar cells. However, the device efficiencies are still far from the Shockley-Queisser limit, ~30%. In practical terms, there is still significant room for improvement in comparison with the best single junction solar cell (GaAs with ~29%) (Steiner et al., 2013) and the best thin film solar cell (Perovskite, 25.5%) (NREL, 2020). Hence, it is necessary to explore the quantificational dependence of the device photovoltaic performance on the growth conditions and the device structures, and to figure out a whole picture for the approaches optimizing device performance.

To learn the dependence of the device photovoltaic performance on the growth conditions and the device structures, a traditional way is by doing trial and error experiments, which requires lots of time, materials, equipment, and manpower. Meanwhile, in experimental studies, the factors influencing the device performance are typically adjusted solely and independently. As a result, the interplays between different factors

* Corresponding author at: Key Laboratory of Luminescence and Optical Information, Beijing Jiaotong University, Ministry of Education, Beijing 100044, China.
E-mail address: ddsong@bjtu.edu.cn (D. Song).
[1] These authors contributed equally: Chengwan Zhu, Wu Liu.

https://doi.org/10.1016/j.solener.2021.09.031
Received 10 February 2021; Received in revised form 13 July 2021; Accepted 10 September 2021
Available online 23 September 2021

are under-evaluated, which leads to deviations in device optimization. Furthermore, experimental investigation is difficult to give quantified model to predict the device performance. Numerical device simulation method (like AMPS (AMPS, 2020) is helpful in quantificational understanding of the effects of the factors on the device performance, whereas it relies on acquiring the physical parameters of the device and it is not able to simulate the fabrication processes directly.

Nowadays, machine-learning (ML) approach is the scientific modeling that can effectively learn from past massive datasets and mechanisms with relatively small error (Choudhary et al., 2019; Li et al., 2019b; Hartono et al., 2020; Majeed et al., 2020; Sahu et al., 2018). Hence, ML is beneficial for overcoming the experimental limitations to investigate the underlying the key factors governing the device performance and their relations. In the recent past, researchers have made progress in exploring the physical properties of the materials with their structural and chemical features (Hartono et al., 2020; Li et al., 2018; Buratti et al., 2020), screening new materials for certain applications (Lu et al., 2018). In terms of solar cells, limited studies generally focus on novel types of thin film solar cells including perovskite solar cells (Li et al., 2019b) and organic solar cells (Majeed et al., 2020; Sahu et al., 2018). These studies present preliminary demonstrations for the advantages of ML in helping optimizing device performance.

Hence, in this work, we attempt to use ML approach to explore the quantificational dependence of the device photovoltaic performance on the growth conditions and the device structures of the CIGS solar cells. The dataset for ML is obtained from the experimental values in the literatures, which enable the results show sufficient low deviation from the experiments. Key factors governing the device performance and their relationships are analyzed based on the correlation studies and ML algorithms. Furthermore, the factors and their exact values responsible for high device efficiency are predicted, which provides essential guidance for experimental device optimization.

2. Methods

Input features and dataset for ML. The fabrication procedures and the typical device architecture of CIGS solar cells are shown in Fig. 1. Based on previous reports, we screened 11 important factors including [Cu]/([Ga] + [In]) ratio (CGI), [Ga]/([Ga] + [In]) ratio (GGI), thickness of CIGS layer (CT), the highest substrate temperature during fabrication (ST), fabrication method (co-evaporation or others, ES), the alkali doping enabled by the substrate or pre-deposited alkali-containing layer (AD), post deposition treatment by alkali ions (PDT), the barrier layer for Na or Fe diffusion (with or without, BL) (Bae et al., 2013), substrate type (flexible substrates or glass), n-type buffer layer (material type (BFL), thickness (BFT) and conduction band offset (CBO)). To reflect the gradient GGI concentration through the CIGS layer, we use three features for GGI, including the GGI at the back side (close to Mo contact, GGI_B), that inner the CIGS layer (GGI_M) and that at the front side (close to n-type layer interface, GGI_F). Hence, there are 15 features in total. The output performance are the photovoltaic parameters including the photoelectric conversion efficiency (PCE), the open circuit voltage (V_{OC}), the short circuit current (J_{SC}) and the fill factor (FF).

To build ML dataset, we search for the literatures reporting the experimental optimization of the CIGS solar cells. Herein, we got more than 300 data points (dataset A listed in Table S1) from more than 120 recently published papers. The maximum PCE is 23.35%, and more data information are listed in Table S2.

Machine learning settings. *R* (version 3.6.2) tool was employed, and the linear regression (LR), neural network (NN), random forest (RF) and extreme gradient boosting (XG) algorithms were used for learning based on *glm*, *neuralnet*, *randomForest* and *xgboost* functions, respectively. The NN algorithm has 2 layers, which have 6 and 4 neurons, respectively. The tree number in RF algorithm was 5000. The max depth and the rounds in xgboost function are 10 and 500, respectively. These parameters used in the algorithms were optimized in advance. The performances of the algorithms are evaluated by root mean square error (RMSE) and Pearson's coefficient (*r* value) on the test set. Here,

$$\mathrm{RMSE} = \sqrt{\sum_{i=1}^{n} \frac{(X_i - Y_i)^2}{n}}$$

$$r = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \overline{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \overline{Y})^2}}$$

X_i, Y_i, $\overline{X}$, $\overline{Y}$, and n represent for the i^{th} value of experimental dataset, the i^{th} value of predicted dataset, the mean value of the experimental

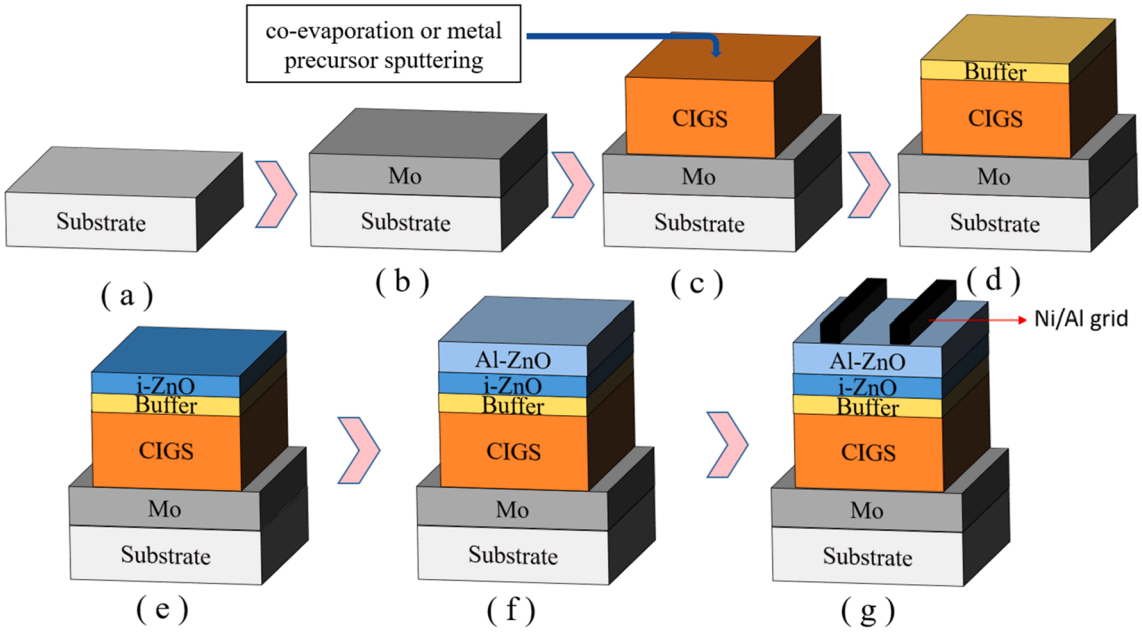

Fig. 1. Schematic fabrication processes of CIGS solar cells: (a) glass cleaning, (b) Mo sputtering, (c) CIGS thin film deposition, (d) Buffer layer (CdS, $Zn(O,S)_x$ or others) deposition, (e) i-ZnO deposition, (f) Al:ZnO deposition, and (g) Ni/Al grid deposition.

dataset, the mean value of the predicted dataset, and the number of the dataset points, respectively. The ratio of the test set is 0.3. To train the ML algorithms, we use 5-fold cross-validation to optimize the hyper-parameters, which divided the dataset into 5 parts (80% data points for training and 20% for validation) and proceeded the learning for 5 times. The model of the algorithm performing the lowest RMSE on validation set was screened for testing on the test set and further prediction.

3. Results and discussion

3.1. Data distribution and correlation analysis

We firstly analyzed the data distribution of the photovoltaic parameters and the information on the CIGS solar cells. Data summaries are listed in Table S2, and violin plots of the photovoltaic parameters and the CIGS layer thickness are shown in Fig. S1. In general, the commonly obtained open circuit voltage (V_{OC}), short circuit current (J_{SC}), fill factor (FF) and photoelectric conversion efficiency (PCE) values are in the range of 600–700 mV, 30–37 mA/cm^2, 65–78% and 8–18%, respectively. It is surprising that the reports on low PCE values are so much despite of the highest efficiency exceeding 20%. Specifically, the reported photoelectric conversion efficiency (PCE) values present large deviations. The highest PCE is 23.35% (Nakamura et al., 2019) while the lowest one is 0.69% (Chiou and Peng, 2017). The commonly used thickness value for CIGS layer is 2 μm, but a large scale of 0.4–3.2 μm were reported. As a result, large derivations and the outliers exist in the photovoltaic parameters and the CIGS thickness, which probably induce prediction inaccuracy in ML and need to be concerned. Hence, a clear and full view of the factors governing the device performance is critical, which will greatly avoid the fabrication of poorly performed devices and promote suitable choices on the fabrication/device parameters.

To get high device efficiency, it is essential to know what the limit is. Hence, we use the correlation matrix using the data listed in Table S1 to learn the correlation between the photovoltaic parameters. The results are depicted in Fig. 2. The high and positive value means the strong positive relation. PCE shows stronger correlation with V_{OC} and FF, especially with FF, than with J_{SC}. FF and V_{OC} also show strong correlation, while V_{OC} and J_{SC} show relatively weak correlation. These results imply that the improvements in PCE in the literatures mainly derive from the approaches simultaneously enhancing FF and V_{OC}. Moreover, these methods generally improve V_{OC} or J_{SC} independently. Therefore, further improvements in PCE utilizing these reported methods need to focus on the improvement mechanisms for V_{OC} and/or FF and the parameter optimization in these approaches.

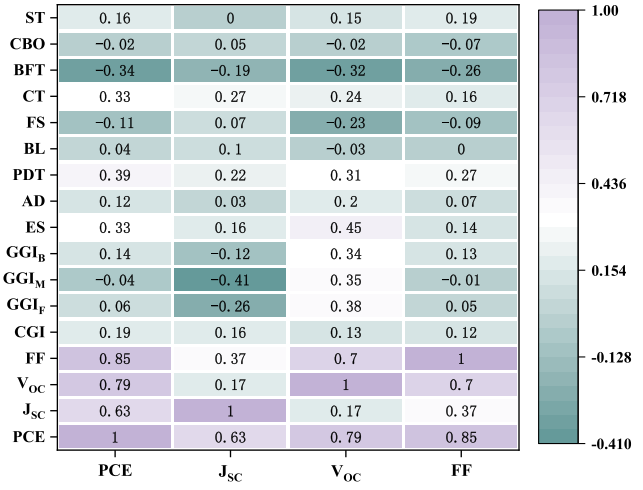

Fig. 2. Correlation matrix of the photovoltaic parameters and the factors influencing the device performance.

To learn the specific contributions of the factors to the photovoltaic parameters, we analyze their correlations using the correlation matrix shown in Fig. 2 and Table S3. Overall, to get high device performance, reasonably high CGI value, co-evaporation method (ES) and pre/post alkali doping (AD and PDT) are essential, which show positive correlations with the device performance. Also, in the reported thickness scale, thicker CIGS layer (large CT value) and thinner BFL (small BFT value) lead to better device performance. GGI value presents a trade-off in device PCE, as it positively correlates with V_{OC} while negatively correlates with J_{SC}. The factors like w/wo blocking layer (BL) has weak effect on device performance. CdS and Zn(O,S)$_x$ buffer layers show comparable record device performance, while other Cd-free buffer layers show poor device performance. The conduction band offset (CBO) at buffer layer/CIGS interface shows neglected correlation with the device performance, as most of the buffer layers used are CdS and Zn(O,S)$_x$, which have CBO smaller than 0.4 eV and has no detrimental effects on device performance (Vurgaftman et al., 2001).

3.2. Key factors governing the device performance

To clearly show the relation between the factors and the device performance, we plot the statistics of the photovoltaic parameters' values changing with the factors including co-evaporation method, alkali post deposition treatment, substrate temperature and the substrate type, as shown in Fig. 3.

(a) *Co-evaporation versus other methods (like sputtering)*. From Fig. 3a, most highly efficient devices are obtained through co-evaporation method, other methods (mainly sputtering) generally show inferior device performance. The difference in PCE induced by fabrication method mainly derives from the change of V_{OC}. It is because that in the co-evaporation method enables the control of CGI and GGI, which are critical approaches in optimizing device performance, especially in V_{OC}, as revealed by correlation results shown in Fig. 2.

(b) *Effects of alkali treatment*. Alkali treatment is also essential to get high device performance, which can be realized by sodium diffusion from soda-lime glass (SLG) or prefabricated sodium contained layer, and by post-deposition of alkali (Na and/or K) fluoride material (post deposition treatment, PDT). Fig. 3b shows the effect of alkali PDT on the device performance, which clearly reveals its importance. Alkali PDT enables alkali cations (Na^+ or K^+) diffusion into the CIGS layer, which passivates the donor defects (In_{Cu}) inside CIGS layer through replacing In and forming electrically neutral Na_{Cu} defects (Kronik et al., 1998). As a result, the hole density is increased in the CIGS layer. Meanwhile, alkali cations can also passivate the point defects at the grain boundaries and the interface, leading to the suppressed recombination. For example, by K treatment, it is revealed that the minority carrier lifetime was markedly prolonged from 22 to 58 ns, and the activation energy of recombination derived from the V_{OC}-T measurements was increased from 1.20 to 1.22 eV (Kato et al., 2017). The high hole density in CIGS layer and the suppressed recombination all benefit for high J_{SC} and V_{OC}. In addition, KF-PDT also enables a significant reduction in CdS thickness (represent by BFT in this work) (Chirilă et al., 2013), which is also revealed by their negative correlation (-0.28 between PDT and BFT shown in Table S3) and favors high device performance (BFL has negative correlations with V_{OC} and PCE). Therefore, alkali PDT is necessary for high device performance.

(c) *Effects of substrate temperature*. The substrate temperature (ST) also has a positive correlation with the photovoltaic parameters, as higher substrate temperature favors smooth surface, large grain size and single-phase CIGS films (Zhang et al., 2009), which in turn improves the device performance.

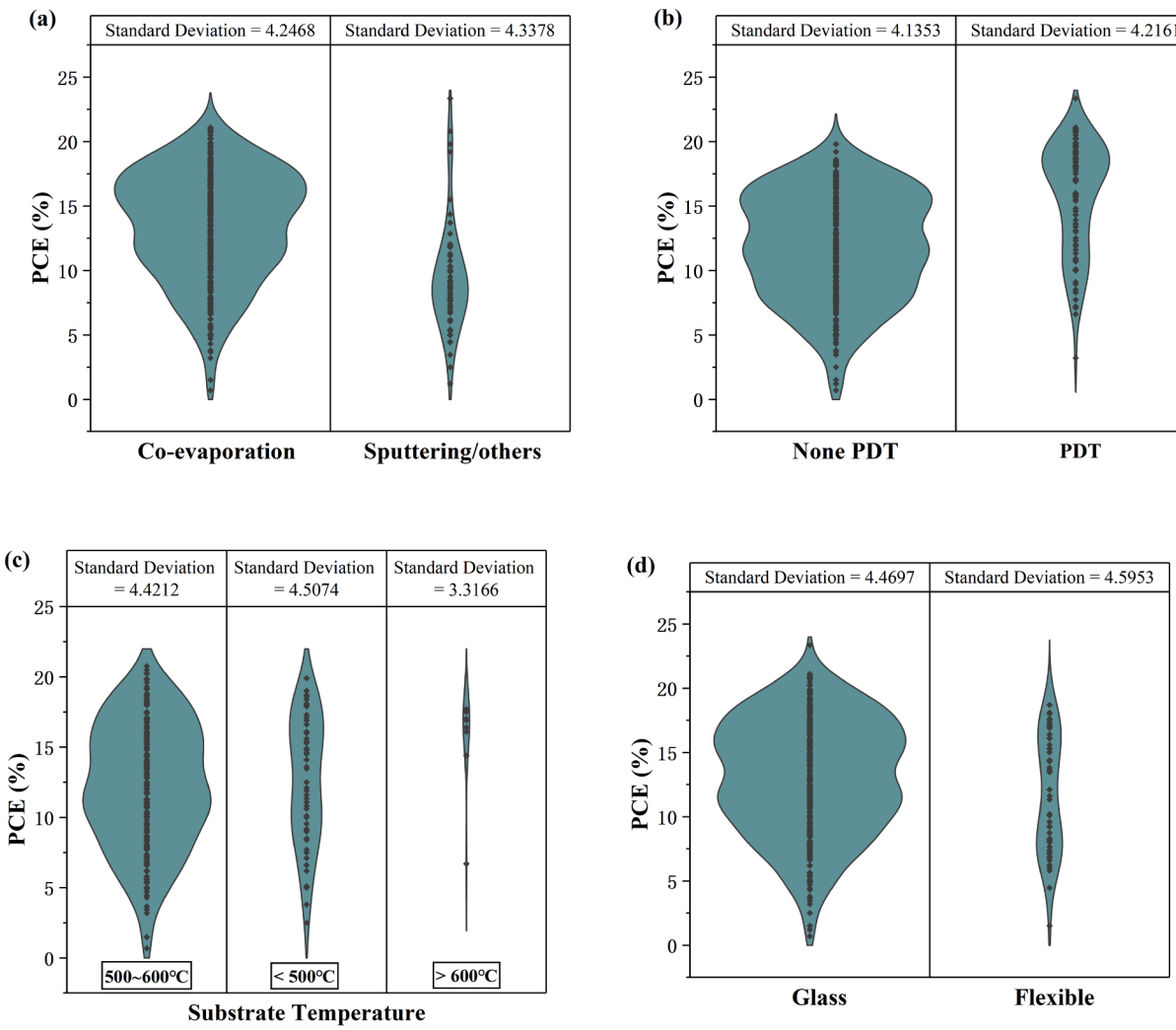

Fig. 3. Effects of different factors on device performance, (a) co-evaporation method (ES), (b) alkali post deposition treatment (PDT), and (c) substrate temperature, and (d) the substrate type.

(d) *Rigid glass substrates versus flexible substrates.* CIGS solar cells are typically fabricated on rigid glass substrates, especially rigid soda lime glass (SLG), which enable the devices generally performing better than these on flexible substrates (like polyimide, stainless steel). SLG substrates allow high processing temperatures of 600 °C or even more, have a coefficient of thermal expansion like CIGS, and provide the right amount of Na directly during film growth (AD factor) (Chirilă et al., 2013; Ma et al., 2018). These benefit for high V_{OC} and PCE. In contrast, flexible substrates do not contain alkali elements and have a large thermal mismatch with CIGS. Indeed, the lower PCE of flexible solar cells results from their negative effect on V_{OC} (shown in Fig. 2). From the correlation matrix shown in Table S3 of the input features, the flexible substrates (FS factor) show relatively high and negative correlation (-0.55) with AD factor, whereas AD promotes V_{OC}. AD is typically realized by sodium diffusion from SLG substrates, which is not applicable for flexible substrates.

Hence, it can be concluded that the key factors governing the device performance including CGI ratio, GGI ratio and gradient, CIGS thickness (CT), buffer layer type (BFL) and its thickness (BFT) and substrate temperature (ST). Moreover, to get high device performance, it is important to employ co-evaporation method (ES) and alkali post deposition treatment (PDT). In addition, substrates enabling alkali diffusion are also essential.

3.3. Prediction performance of different ML algorithms

To explore the features of the highly efficient CIGS solar cells, we use ML algorithms to learn the correlations between fabrication procedures/device structure and the device performance. To establish the datasets for these learning, we screen the data points of datasets A shown in Table S1 with the following rules to get new datasets to do the machine learning. Firstly, we screen 11 features including CGI, GGI_F, GGI_M, GGI_B, ES, AD, PDT, CT, BFL, BFT and ST to learn their correlations with the device performance. The data points with flexible substrates were removed. The features CBO and BL were eliminated, as they show weak effect on device performance. The data points using CdS and/or $Zn(O,S)_x$ BFLs were reserved, while the data points using other BFLs or non-BFLs were removed. Secondly, as many of the data points miss one or more features, then we handle the dataset with two different ways: 1) removing the data rows with missing data (dataset B); 2) supplementing the missing data with the typical values (dataset C). The details for the rules in supplementing the missing data are summarized in Table S4. Thirdly, to avoid the effect of the experimental error, we also cleaned the data points with PCE less than 10%, as this value is essential for a normal device, and we got dataset D (by removing data points with low PCE from dataset B) and dataset E (by removing data points with low PCE from dataset C).

We use *R* tool employing 4 algorithms including linear regression (LR), neural network (NN), random forest (RF) and extreme gradient

boosting (XGBoost, abbreviated as XG in Table 1 and Fig. 4). LR is simple and facile to be established manually, hence, it is used for comparison to show the performance of other complex ML models. As the input features correlate with each other, so NN is employed as it involves the correlations among the features and may performs good. In addition, as several of input features can be classified, so tree-based RF algorithm is probably suitable to process such conditions. XGBoost provides a parallel tree boosting and thus is also employed. The input features for the ML algorithms are CGI ratio, GGI ratio and gradient (GGI_F, GGI_M, GGI_B), CIGS thickness (CT), BFL thickness (BFT) and substrate temperature (ST), and the output is the photovoltaic parameters (PCE, J_{SC}, V_{OC}, FF) of the CIGS solar cells. Before establishing the models by these 4 algorithms, Lasso algorithm was employed to evaluate the necessity of the input features. We found that reducing one or more input features made the mean square error increasing rapidly, which means that these 11 input features are not able to be further removed. For ML, we use 5-fold cross-validation method to optimize the performances of the ML algorithms. The dataset was randomly divided by 7:3 for training and testing (the test set). The 5-fold cross-validation was used for optimizing the hyperparameters of the algorithms, in which the datapoints for training are further dived into 5 parts: 4 parts (80%) for training (the training set) and 1 part (20%) for validation (the validation set). The performances of the algorithms are evaluated using root mean square error (RMSE) and Pearson's coefficient (*r* value). RMSE directly evaluates the error between the predicted values and the experimental values of the datasets, which is straight and is a typical parameter to reflect the accuracy of experimental results. Pearson's coefficient (*r* value) shows the correlation between the predicted values and the experimental values of the dataset, and a larger *r* value means a stronger correlation. Therefore, an algorithm with lower RMSE and higher *r* value has higher accuracy and reliability in prediction.

Table 1 summarizes the performances of different algorithms with RMSE and *r* value on predicting the PCE values in the training set, validation set and the test set of different datasets. Fig. 4 presents the comparison of the experimental PCE and the predicted values from different algorithms including LR, NN, RF and extreme gradient boosting (XGBoost, abbreviated as XG in Table 1 and Fig. 4). The prediction performances of these algorithms on other photovoltaic parameters are listed in Table S5-S7 and Fig. S2-S4. Among the four algorithms, RF performs best on predicting PCE from both validation and test sets, which have relatively low RMSE (0.9–3%) and high *r* value (>0.83 for all datasets). The low RMSE indicates the high accuracy of RF algorithm, and the high *r* value means that the predicted values and the experimental values have strong correlation. LR is a simple model and not able to process such complex system; NN and XGBoost algorithms are overfitted (excellent performance on training set and relatively poor performance on test set) due to relatively large number of input features and limited data points. RF avoids the over-fitted situation in principle and exhibits its efficacy in complex systems like in this work. Though the device performance can be well predicted in general, the device with ultra-low or ultra-high PCE shows larger derivation between the predicted and the experimental values (shown in Fig. 4). This can be attributed to the fact that though the key features are included in training the algorithms, experimental derivations induced by the raw materials, the equipment and the tricks in fabricating devices are not able to be covered. Another reason for this is that most of the data points have moderate performance, and the algorithms learns better due to relatively larger data size and thus present higher accuracy in the devices with moderate performance. Despite of these, the approach of using ML algorithm to predict the device performance is sufficiently referable for experimental design of highly efficient devices.

Datasets: Dataset A is the full records obtained from the literatures. Dataset B is obtained by removing the missing information in dataset A. Dataset C is obtained by supplementing the missing data with the typical values in dataset A. Datasets D and E are the high efficiency records (PCE $\geq$ 10%) in dataset B and C, respectively.

Comparing the results on different datasets, it is clear that datasets D and E are much more predictable than datasets B and C. As datasets D and E only keep the data points with relatively higher PCE (PCE $\geq$ 10%), the effects of the experimental errors other than the intrinsic mechanisms in the device are greatly excluded. Moreover, datasets C and E, which have supplementary information of some data points, suffer larger derivation than datasets B and D with only reported values. This indicates that the data points with missing information may not use the typically used values, which also indicates that these factors have certain influence on the device performance. From these comparisons, it can also be seen that the smart screening of data points for ML dataset is critical.

Compared with the reported ML prediction performance on other type of solar cells, the results on CIGS solar cells obtained in this work are comparable or superior. For example, the lowest RMSE and the highest *r* value are 1.08% and 0.78 for PCE of the organic solar cells using RF algorithm (Sahu et al., 2018). In comparison, the RMSE and the *r* value based on the validation/test sets of dataset D are 0.9%/1.8% and 0.97/0.88 in this work using RF algorithm. The good performance achieved by ML algorithm in predicting the PCE of CIGS solar cells correlates with their relatively mature fabrication procedures and device architecture, and the stable properties of the inorganic materials like CIGS and other layers.

Table 1
Performances of different ML algorithms in PCE prediction of the CIGS solar cells.

Datasets	ML algorithms	Training set		Validation set		Test set	
		RMSE (%)	*r* value	RMSE (%)	*r* value	RMSE (%)	*r* value
B	*LR*	2.9	0.74	2.8	0.81	3.4	0.67
	RF	1.6	0.95	1.5	0.96	2.2	0.88
	NN	1.2	0.96	1.0	0.96	3.6	0.70
	XG	0.9	0.98	1.6	0.91	3.2	0.73
C	*LR*	3.4	0.64	3	0.65	3.7	0.59
	RF	1.7	0.94	1.5	0.93	3.0	0.83
	NN	1.9	0.89	1.7	0.91	3.8	0.70
	XG	0.8	0.98	2.7	0.77	3.4	0.66
D	*LR*	2.0	0.76	1.8	0.85	2.2	0.65
	RF	1.1	0.95	0.9	0.97	1.8	0.88
	NN	0.9	0.95	0.7	0.96	2.1	0.75
	XG	0.2	1.00	1.2	0.91	2.2	0.79
E	*LR*	2.5	0.62	2.0	0.80	2.4	0.61
	RF	1.1	0.94	1.0	0.93	1.9	0.83
	NN	1.0	0.95	0.9	0.96	3.2	0.68
	XG	0.4	0.99	1.6	0.87	2.5	0.59

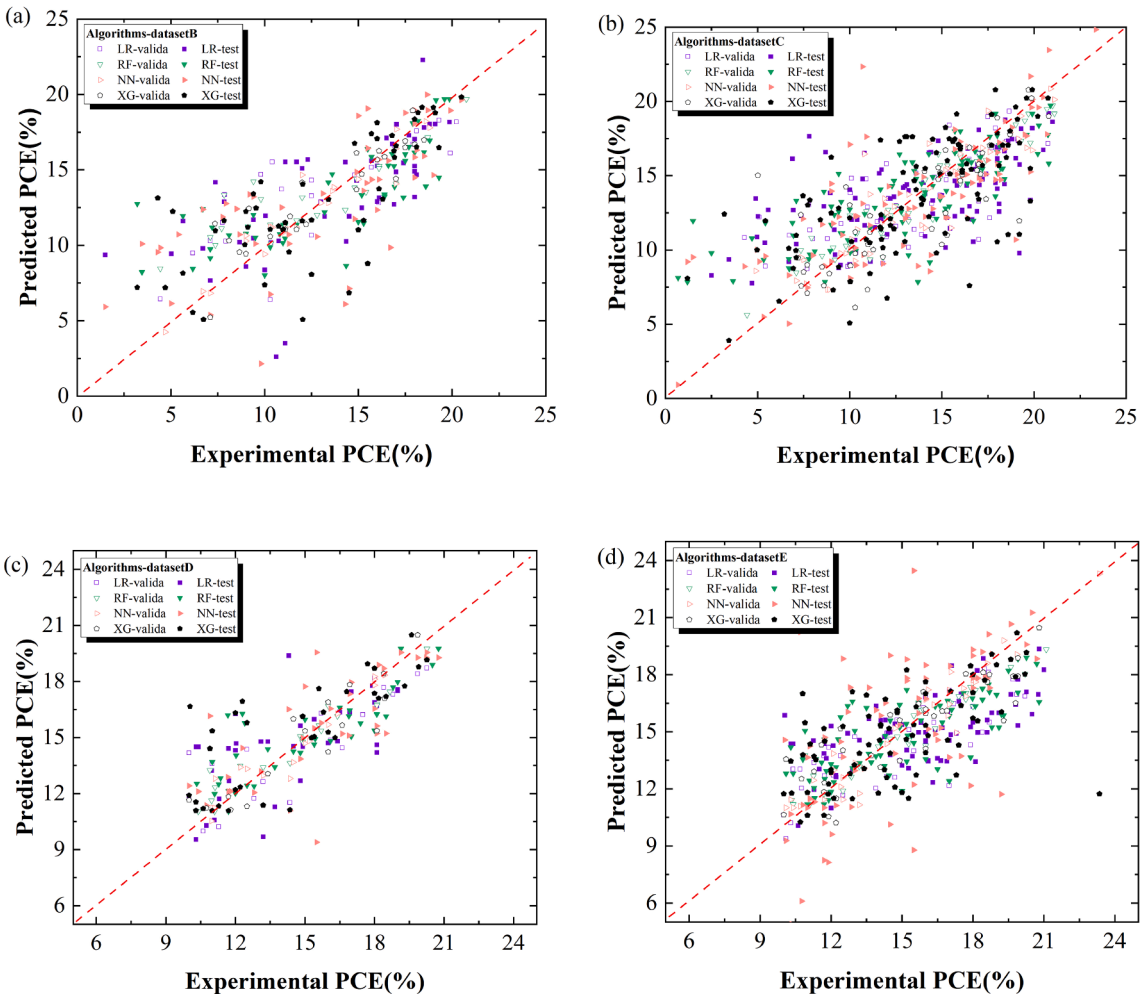

Fig. 4. Comparison of the experimental PCE values and the predicted values by different algorithms on validation (*valida*) and test sets. The red dash line presents the condition in which the predicted value equals to the experimental value. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

3.4. Approaches for highly efficient CIGS solar cells: Implications from machine learning

Comparing the performance of the five highest efficient CIGS solar cells reported up to now (Nakamura et al., 2019; Kato et al., 2019; Jackson et al., 2016, 2015; Tai et al., 2017), it can be seen from Table S8 that the photovoltaic parameters vary by a large content: V_{OC} (738.20 mV, RMSE 9.43 mV), J_{SC} (38.36 mA/cm^2, RMSE 1.13 mA/cm^2), FF (79.66%, RMSE 0.85 %) and PCE (22.574%, RMSE 0.56 %). This comparison shows that there are still room for efficiency improvement even on this high-performance level.

To further enhance the device performance, it is better to clearly understand the effects of the fabrication procedures and the device structure factors on the device performance. Adjusting the factors experimentally to learn the effects is not so efficient in such complex systems with so many correlated factors, as this requires a huge number of experiments. Hence, we use the RF algorithm trained by the experimental results (dataset D) to predict the device performance of massive assuming CIGS solar cells.

The importance of the input features can be evaluated by RF model (shown in Fig. S5c from dataset D), which reveals that the most important features include alkali post deposition treatment (PDT), CIGS thickness (CT), buffer layer thickness (BFT), CGI and GGI gradient. Considering the importance of the input features, and the correlations among the factors and the device performance, we set a series ranges for the fabrication procedures and the device structure factors (detailed settings are illustrated in Table S9) which make the generation of more than 600 thousand CIGS solar cells. Combining the prediction results and the experimental findings in literature, the effects and the underlying mechanisms of the key factors on the device performance are discussed below.

GGI ratio and GGI gradient. GGI ratio and GGI gradient are critical factors in governing the device performance and their modifications are key approaches for high device performance. GGI ratio affects the device performance in two aspects: the energy band-gap energy (Eg) of CIGS and the defects in CIGS film. In principle, $Eg = 1.01(1 - x) + 1.65x - 0.15x(1 - x)$, with x being the respective GGI value (Jackson et al., 2011). Bulk defects are minimal in CIGS when the Ga atomic ratio is around $x = 0.3$. With a higher Ga ($x > 0.3$) ratio, the number of defects in the film increases with Ga content (Hanna et al., 2001; Ramanujam and Singh, 2017).

The change of predicted PCE with the GGI ratio and gradient is shown in Fig. 5. For CIGS solar cells employing CdS as the buffer layer and co-evaporation method for CIGS, optimal GGI_F and GGI_M ratios are 0.37–0.40 and optimal GGI_B is > 0.4. These scales are a bit higher than the typically used GGI values (0.30–0.33), but they are also demonstrated to be able to get highly efficient devices. For example, with a GGI ratio of 0.4 and CGI ratio of 0.95, the device PCE of 21.18% is achieved (Koida et al., 2018). For better comparison, the device performance, and the corresponding CGI and GGI ratio values from experimental reports

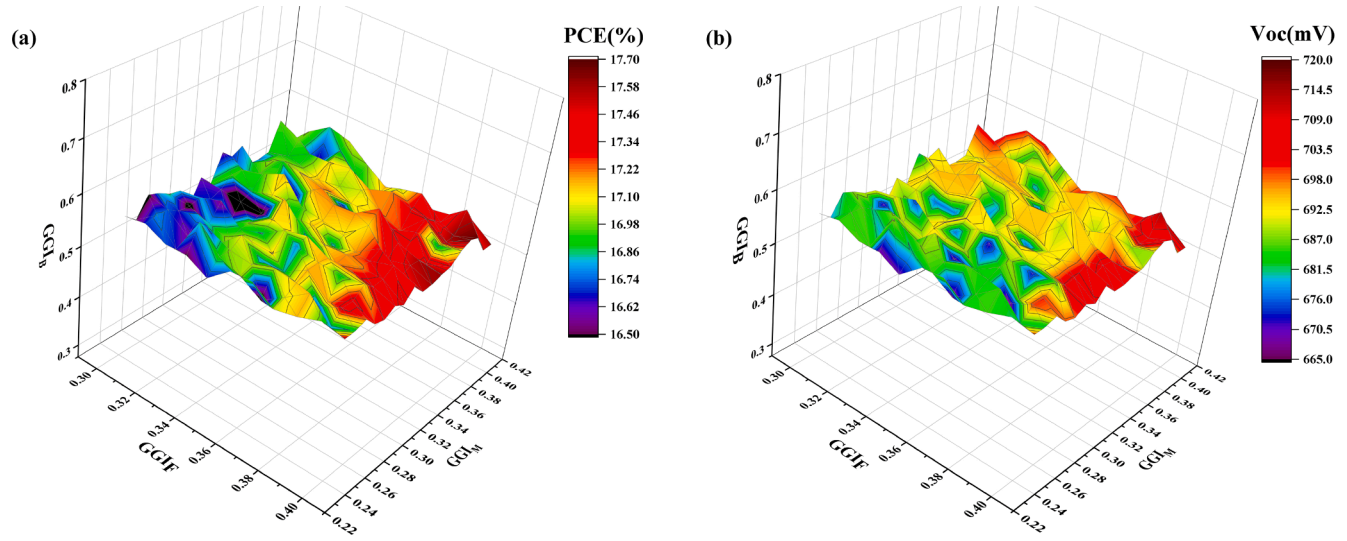

Fig. 5. Effects of GGI values on the device performance, (a) PCE and (b) V_{OC}. The results are from the prediction of RF algorithm trained by datasets D. Here, CGI ratio is 0.89.

with PCE > 20% are summarized in Table S8. Further increasing the GGI ratio to 0.4 is probably a way to further improve the device performance, as high GGI ratio promotes V_{OC}. However, for GGI ratio > 0.3, though high Ga content favors for high built-in potential, increased band-gap energy and increased bulk defects reduce the light absorption and deteriorate device performance, leading to the optimized PCE in these optimal scales.

Therefore, improvements in PCE is limited by solely increasing or decreasing Ga content. Notch type gradient, i.e., with smallest GGI_M ratio, is beneficial for boosting V_{OC} and J_{SC}. Front grading improves V_{OC} due the large Eg, small GGI_M ratio promotes the light absorption and the resultant J_{SC}, and the back grading decreases bulk/surface recombination at the back-contact interface due to the back-surface field (BSF) created by Ga grading (Jackson et al., 2011; Chantana et al., 2015). Here, notch type (with smallest GGI_M ratio) is also proved to be efficient in improving device performance by the prediction results. Moreover, high GGI ratio at the back-contact is found to highly benefit for high device efficiency. The optimal GGI_B ratio can be as high as 0.80, which is obviously higher than these of optimal GGI_F and GGI_M ratios.

Hence, to get highly efficient CIGS solar cells, the results from ML suggest that relatively higher GGI ratio (0.37–0.40) and notch type gradient with high GGI ratio at back side (up to 0.8) are effective ways.

CGI effect. Increasing CGI ratio increases the electrical performance of the CIGS film, but also increases the interface recombination. Hence, CGI ratio of highly efficient CIGS solar cells generally falls into the range of 0.87–0.95 (as listed in Table S8). Here, CGI ratio is adjusted from 0.87 to 0.95, which shows minor effects on device performance, as shown in Fig. 6.

Others. For CIGS solar cells employing CdS as the buffer layer and co-evaporation method for CIGS, optimal CIGS thickness is 2–2.3 μm and optimal substrate temperature is 540–600 °C, which cover the typically used parameters in experimental reports for highly efficient CIGS solar cells. The thickness of CdS in the range of 10–50 nm has minor effect on device performance.

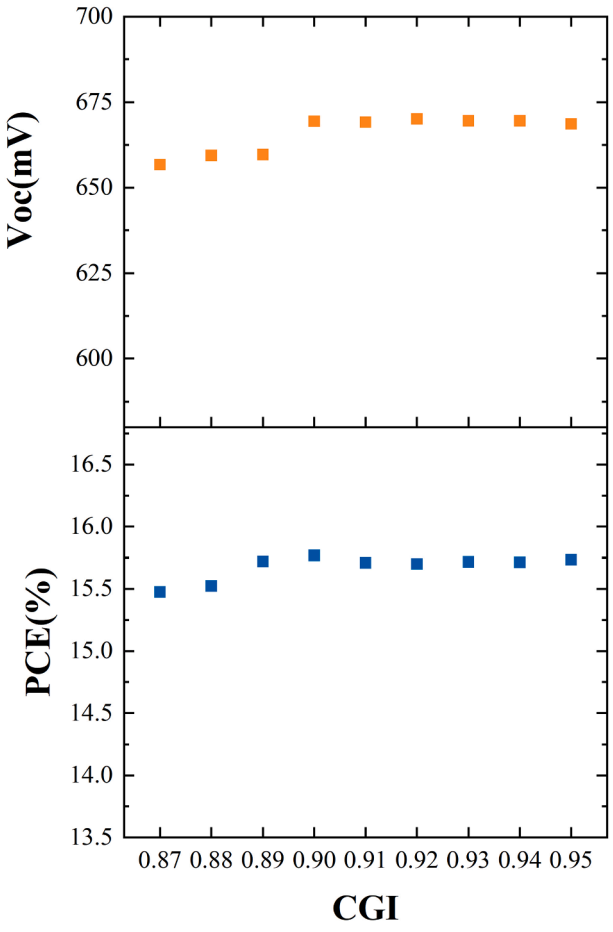

Fig. 6. Effects of CGI values on the device performance. The results are from the prediction of RF algorithm trained by dataset D. Here, GGI ratio is 0.3.

4. Summary

Factors governing the device efficiency of CIGS solar cells are analyzed, and the correlation between the preparation parameters and the configuration of the devices with their photovoltaic performance are modeled by machine learning (ML) algorithms. Among the used four algorithms, random forest performs best in predicting the efficiency of the CIGS solar cells with high accuracy (RMSE of 0.9%/1.8% and Pearson coefficient *r* of 0.9/0.88 on validation/test sets). Based on the optimized model of RF algorithm, the factors and their exact values responsible for high device efficiency are predicted. These results prove the great potential and the power of machine learning tool in studying the mechanisms and optimizing the device performance of CIGS solar

cells.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work was supported by the National Key R&D Program of China Grant Nos. 2018YFB1500200; Beijing Natural Science Foundation Nos. 2192045; the National Natural Science Foundation of China under Grant Nos. 61775013 and 62075006.

Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.solener.2021.09.031.

References

AMPS, 2020. Analysis of Microelectronic and Photonic Structures, http://www.ampsmodeling.org/materials/ncds.htm (accessed 15 December 2020).
Bae, D., Kwon, S., Oh, J., Kim, W.K., Park, H., 2013. Investigation of Al2O3 diffusion barrier layer fabricated by atomic layer deposition for flexible Cu(In, Ga)Se-2 solar cells. Renewable Energy 55, 62–68.
Bhattacharya, R.N., Ramanathan, K., 2004. Cu(In, Ga)Se-2 thin film solar cells with buffer layer alternative to CdS. Sol. Energy 77, 679–683.
Buratti, Y., Le Gia, Q.T., Dick, J., Zhu, Y., Hameiri, Z., 2020. Extracting bulk defect parameters in silicon wafers using machine learning models. npj Comput. Mater. 6, 142.
Chantana, J., Hironiwa, D., Watanabe, T., Teraji, S., Kawamura, K., Minemoto, T., 2015. Estimation of open-circuit voltage of Cu(In, Ga)Se-2 solar cells before cell fabrication. Renew. Energy 76, 575–581.
Chiou, C.S., Peng, H.C., 2017. Influence of process parameters on the gallium composition of a CuIn1-xGaxSe2 solar cell on the efficiency of non-vacuum blade coating stacking. Sol. Energy 146, 436–442.
Chirilă, A., Reinhard, P., Pianezzi, F., Bloesch, P., Uhl, A.R., Fella, C., Kranz, L., Keller, D., Gretener, C., Hagendorfer, H., Jaeger, D., Erni, R., Nishiwaki, S., Buecheler, S., Tiwari, A.N., 2013. Potassium-induced surface modification of Cu(In, Ga)Se-2 thin films for high-efficiency solar cells. Nat. Mater. 12 (12), 1107–1111.
Choudhary, K., Bercx, M., Jiang, J., Pachter, R., Lamoen, D., Tayazza, F., 2019. Accelerated discovery of efficient solar cell materials using quantum and machine-learning methods. Chem. Mater. 31, 5900–5908.
Dhere, N.G., 2007. Toward GW/year of CIGS production within the next decade. Sol. Energy Mater. Sol. Cells 91 (15-16), 1376–1382.
Friedlmeier, T.M., Jackson, P., Bauer, A., Hariskos, D., Kiowski, O., Wuerz, R., Powalla, M., 2015. Improved Photocurrent in Cu(In, Ga)Se-2 Solar Cells: From 20.8% to 21.7% Efficiency with CdS Buffer and 21.0% Cd-Free. IEEE J. Photovolt. 5, 1487–1491.
Hanna, G., Jasenek, A., Rau, U., Schock, H.W., 2001. Influence of the Ga-content on the bulk defect densities of Cu(In, Ga)Se-2. Thin Solid Films 387, 71–73.
Hartono, N.T.P., Thapa, J., Tiihonen, A., Oviedo, F., Batali, C., Yoo, J.J., Liu, Z., Li, R., Fuertes Marron, D., Bawendi, M.G., Buonassisi, T., Sun, S., 2020. How machine learning can help select capping layers to suppress perovskite degradation. Nat. Commun. 11, 4172.
Jackson, P., Hariskos, D., Lotter, E., Paetel, S., Wuerz, R., Menner, R., Wischmann, W., Powalla, M., 2011. New world record efficiency for Cu(In, Ga)Se-2 thin-film solar cells beyond 20%. Progr. Photovolt. 19, 894–897.
Jackson, P., Hariskos, D., Wuerz, R., Kiowski, O., Bauer, A., Friedlmeier, T.M., Powalla, M., 2015. Properties of Cu(In, Ga)Se-2 solar cells with new record efficiencies up to 21.7%. Physica Status Solidi-Rapid Res. Lett. 9 (1), 28–31.
Jackson, P., Wuerz, R., Hariskos, D., Lotter, E., Witte, W., Powalla, M., 2016. Effects of heavy alkali elementsin Cu(In, Ga)Se-2 solar cells with efficiencies up to 22.6%. Physica Status Solidi-Rapid Res. Lett. 10 (8), 583–586.
Kato, T., Handa, A., Yagioka, T., Matsuura, T., Yamamoto, K., Higashi, S., Wu, J.-L., Tai, K.F., Hiroi, H., Yoshiyama, T., Sakai, T., Sugimoto, H., 2017. Enhanced Efficiency of Cd-Free Cu(In, Ga)(Se, S)(2) Minimodule Via (Zn, Mg)O Second Buffer Layer and Alkali Metal Post-Treatment. IEEE J. Photovolt. 7 (6), 1773–1780.
Kato, T., Wu, J.-L., Hirai, Y., Sugimoto, H., Bermudez, V., 2019. Record Efficiency for Thin-Film Polycrystalline Solar Cells Up to 22.9% Achieved by Cs-Treated Cu(In, Ga)(Se, S)(2). IEEE J. Photovolt. 9 (1), 325–330.
Khatri, I., Fukai, H., Yamaguchi, H., Sugiyama, M., Nakada, T., 2016. Effect of potassium fluoride post-deposition treatment on Cu(In, Ga)Se-2 thin films and solar cells fabricated onto sodalime glass substrates. Sol. Energy Mater. Sol. Cells 155, 280–287.
Koida, T., Nishinaga, J., Ueno, Y., Higuchi, H., Takahashi, H., Iioka, M., Shibata, H., Niki, S., 2018. Impact of front contact layers on performance of Cu(In, Ga)Se-2 solar cells in relaxed and metastable states. Progr. Photovolt. 26, 789–799.
Kronik, L., Cahen, D., Schock, H.W., 1998. Effects of sodium on polycrystalline Cu(In, Ga)Se-2 and its solar cell performance. Adv. Mater. 10 (1), 31–36.
Li, J.V., Grover, S., Contreras, M.A., Ramanathan, K., Kuciauskas, D., Noufi, R., 2014. A recombination analysis of Cu(In, Ga)Se-2 solar cells with low and high Ga compositions. Sol. Energy Mater. Sol. Cells 124, 143–149.
Li, J., Ma, Y., Chen, G., Gong, J., Wang, X., Kong, Y., Ma, X., Wang, K., Li, W., Yang, C., Xiao, X., 2019a. Effects of Ammonia-Induced Surface Modification of Cu(In, Ga)Se-2 on High-Efficiency Zn(O, S)-Based Cu(In, Ga)Se-2 Solar Cells. Solar Rrl 3, 1800254.
Li, J., Pradhan, B., Gaur, S., Thomas, J., 2019b. Predictions and strategies learned from machine learning to develop high-performing perovskite solar cells. Adv. Energy Mater. 9, 1901891.
Li, Z., Omidvar, N., Chin, W.S., Robb, E., Morris, A., Achenie, L., Xin, H., 2018. Machine-learning energy gaps of porphyrins with molecular graph representations. J. Phys. Chem. A 122, 4571–4578.
Lu, S., Zhou, Q., Ouyang, Y., Guo, Y., Li, Q., Wang, J., 2018. Accelerated discovery of stable lead-free hybrid organic-inorganic perovskites via machine learning. Nat. Commun. 9, 3405.
Ma, X., Ma, Y., Yang, S., Yang, C., Lin, T., Wang, K., Xiao, X., 2018. Pre-incorporation of Na into flexible Cu(In, Ga)Se-2 thin film solar cells. Sol. Energy 173, 1080–1086.
Majeed, N., Saladina, M., Krompiec, M., Greedy, S., Deibel, C., MacKenzie, R.C.I., 2020. Using deep machine learning to understand the physical performance bottlenecks in novel thin-film solar cells. Adv. Funct. Mater. 30, 1907259.
Muzzillo, C.P., 2017. Review of grain interior, grain boundary, and interface effects of K in CIGS solar cells: Mechanisms for performance enhancement. Sol. Energy Mater. Sol. Cells 172, 18–24.
Nakada, T., Mizutani, M., Hagiwara, Y., Kunioka, A., 2001. High-efficiency Cu(In, Ga)Se-2 thin-film solar cells with a CBD-ZnS buffer layer. Sol. Energy Mater. Sol. Cells 67, 255–260.
Nakamura, M., Yamaguchi, K., Kimoto, Y., Yasaki, Y., Kato, T., Sugimoto, H., 2019. Cd-Free Cu(In, Ga)(Se, S)(2) Thin-Film Solar Cell With Record Efficiency of 23.35%. IEEE J. Photovoltaics 9 (6), 1863–1867.
NREL, 2020. National Renewable Energy Laboratory, https://www.nrel.gov/pv/cell-efficiency/.html (accessed 15 December 2020).
Ochoa, M., Buecheler, S., Tiwari, A.N., Carron, R., 2020. Challenges and opportunities for an efficiency boost of next generation Cu(In, Ga)Se(2)solar cells: prospects for a paradigm shift. Energy Environ. Sci. 13 (7), 2047–2055.
Ohtake, Y., Okamoto, T., Yamada, A., Konagai, M., Saito, K., 1997. Improved performance of Cu(InGa)Se2 thin-film solar cells using evaporated Cd-free buffer layers. Sol. Energy Mater. Sol. Cells 49, 269–275.
Ramanujam, J., Singh, U.P., 2017. Copper indium gallium selenide based solar cells - a review. Energy Environ. Sci. 10, 1306–1319.
Reinhard, P., Bissig, B., Pianezzi, F., Avancini, E., Hagendorfer, H., Keller, D., Fuchs, P., Doebeli, M., Vigo, C., Crivelli, P., Nishiwaki, S., Buecheler, S., Tiwari, A.N., 2015. Features of KF and NaF Postdeposition Treatments of Cu(In, Ga)Se-2 Absorbers for High Efficiency Thin Film Solar Cells. Chem. Mater. 27, 5755–5764.
Sahu, H., Rao, W., Troisi, A., Ma, H., 2018. Toward predicting efficiency of organic solar cells via machine learning and improved descriptors. Adv. Energy Mater. 8, 1801032.
Salomé, P.M.P., Rodriguez-Alvarez, H., Sadewasser, S., 2015. Incorporation of alkali metals in chalcogenide solar cells. Sol. Energy Mater. Sol. Cells 143, 9–20.
Schleussner, S., Zimmermann, U., Watjen, T., Leifer, K., Edoff, M., 2011. Effect of gallium grading in Cu(In, Ga)Se-2 solar-cell absorbers produced by multi-stage coevaporation. Sol. Energy Mater. Sol. Cells 95, 721–726.
Steiner, M.A., Geisz, J.F., Garcia, I., Friedman, D.J., Duda, A., Kurtz, S.R., 2013. Optical enhancement of the open-circuit voltage in high quality GaAs solar cells. J. Appl. Phys. 113, 123109.
Tai, K.F., Kamada, R., Yagioka, T., Kato, T., Sugimoto, H., 2017. From 20.9 to 22.3% Cu(In, Ga)(S, Se)(2) solar cell: Reduced recombination rate at the heterojunction and the depletion region due to K-treatment. Jpn. J. Appl. Phys. 56 (8S2), 08MC03. https://doi.org/10.7567/JJAP.56.08MC03.
Vurgaftman, I., Meyer, J.R., Ram-Mohan, L.R., 2001. Band parameters for III-V compound semiconductors and their alloys. J. Appl. Phys. 89 (11), 5815–5875.
Witte, W., Abou-Ras, D., Albe, K., Bauer, G.H., Bertram, F., Boit, C., Brueggemann, R., Christen, J., Dietrich, J., Eicke, A., Hariskos, D., Maiberg, M., Mainz, R., Meessen, M., Mueller, M., Neumann, O., Orgis, T., Paetel, S., Pohl, J., Rodriguez-Alvarez, H., Scheer, R., Schock, H.-W., Unold, T., Weber, A., Powalla, M., 2015. Gallium gradients in Cu(In, Ga)Se-2 thin-film solar cells. Progr. Photovolt. 23, 717–733.
Zhang, L.i., He, Q., Jiang, W.-L., Liu, F.-F., Li, C.-J., Sun, Y., 2009. Effects of substrate temperature on the structural and electrical properties of Cu(In, Ga)Se-2 thin films. Sol. Energy Mater. Sol. Cells 93 (1), 114–118.