Received: 13 October 2022 | Revised: 1 February 2023 | Accepted: 3 February 2023

DOI: 10.1002/pip.3683

RESEARCH ARTICLE

PROGRESS IN PHOTOVOLTAICS WILEY

Advanced analysis of internal quantum efficiency measurements using machine learning

Zubair Abdullah-Vetter[1] | Priya Dwivedi[1] | Yoann Buratti[1] | Arcot Sowmya[2] | Thorsten Trupke[1] | Ziv Hameiri[1]

[1]School of Photovoltaics and Renewable Energy Engineering (SPREE), University of New South Wales (UNSW), Sydney, New South Wales, Australia

[2]School of Computer Science and Engineering (CSE), University of New South Wales (UNSW), Sydney, New South Wales, Australia

Correspondence
Zubair Abdullah-Vetter, School of Photovoltaics and Renewable Energy Engineering (SPREE), University of New South Wales (UNSW), Sydney, NSW 2052, Australia
Email: z.abdullahvetter@unsw.edu.au

Funding information
Australian Renewable Energy Agency, Grant/Award Number: 2020/RND016

Abstract

The internal quantum efficiency (IQE) is given as the ratio between the externally collected electron current and the photon current absorbed by the device. Spectral analysis of IQE measurements is a powerful method to identify performance-limiting mechanisms in solar cells. It also enables the extraction of key electrical and optical parameters. However, the potential of IQE measurements is only rarely fully utilized, presumably due to the significant complexity associated with the fitting process and its sensitivity to noise. In this study, machine learning is proposed as an efficient method to extract quantitative information from IQE measurements. The extraction method is automated and easy to use, providing an array of specific device parameters. By simplifying the analytical process, the developed machine learning algorithms also extract the parasitic absorption of the antireflection coating, a key parameter that is difficult to obtain by traditional methods. Although the method has been developed for and tested on silicon solar cells, it can be adapted and applied to other types of solar cells.

KEYWORDS
antireflection coating, chain regression, photovoltaics, random forest, spectral response

1 | INTRODUCTION

The spectral response (SR) of a photovoltaic (PV) device is determined by illuminating the device with variable monochromatic light and measuring the generated short circuit current (I_{sc}) in Ampere and the incident light intensity (I_{light}) in Watts at each wavelength (λ).[1] By also measuring the wavelength-dependent reflection [R(λ)] of the device, SR(λ) can be converted into the internal quantum efficiency (IQE)[2]:

$$\mathrm{IQE}(\lambda) = \frac{1}{1-R(\lambda)} \cdot \frac{hc}{q\lambda} \cdot \mathrm{SR}(\lambda), \text{where } \mathrm{SR}(\lambda) = \frac{I_{sc}(\lambda)}{I_{light}(\lambda)}. \quad (1)$$

Here the constants *h*, c, and *q* are Planck's constant, the speed of light, and the elementary charge constant, respectively. The wavelength-dependent IQE defines the ratio between the photogenerated electron current collected by the device and the non-reflected photon current at each wavelength.[1] Hence, it describes the probability that a photon, assuming that it is not reflected at the front surface, contributes to I_{sc}. IQE measurements are often used to conduct loss analysis of solar cells[1,2] and to quantify parameters representing performance-limiting mechanisms such as surface recombination velocity and the diffusion length of carriers.[3]

Traditional analysis methods of IQE measurements of silicon (Si) solar cells require fitting different mathematical expressions to

specific sections of the spectral IQE.[4] An example is the extraction of the effective minority carrier diffusion length (L_{eff}) in the bulk of the cell.[4] It is extracted by plotting the inverse IQE against the wavelength-dependent absorption depth of Si. The inverse slope within the near-infrared wavelength range is then used to extract L_{eff}. The resulting L_{eff} is strongly impacted by measurement noise and the selected wavelength range for the fit.[4,5] This approach is applied for other parameters, which are similarly extracted by fitting IQE curves or sections of the curve using known mathematical models that have been derived and simplified in earlier studies.[2,4–7] However, this process requires testing different combinations of input parameters, which can be time-consuming and complex. Software packages such as LASSIE[8] have been developed to automate the fitting process. However, these packages are not easily available. Another approach is based on simulating solar cells and obtaining their IQE curves through the use of software programs such as PC1D,[9] SunSolve,[10] Griddler,[11] and others.[12,13] Nevertheless, these simulations often require knowledge or assumptions about a large number of input parameters, resulting in the need to also input many different combinations of those parameters.

Parasitic absorption within the antireflection coating (ARC) can significantly reduce the photo-generated current.[14,15] It is an important parameter in the context of IQE analysis that is difficult to assess.[14] Absorption within ARC materials can be studied with transmission measurements on glass substrates.[16] However, they do not always represent the ARC properties when deposited on Si wafers, particularly after undergoing thermal processing steps.[14–16] SunSolve[10] obtains the parasitic absorption by using complex ray tracing algorithms, which is computationally expensive, while LASSIE[8] references a generalized database of ARC absorption profiles. Therefore, it is difficult to decouple the impact of the ARC parasitic absorption when analyzing IQE measurements.

In this study, we propose and demonstrate the use of machine learning (ML) regression models to automatically analyze IQE measurements. The trained models use the measured IQE curve as input and determine key electrical and optical parameters of the solar cell. There is no requirement for the user to manipulate the values or wavelength ranges. It is shown that the ML models are also capable of obtaining the parasitic absorption within the ARC layer from the measured IQE curves, addressing the above-mentioned limitations associated with current approaches. This simplified approach allows the users to automatically analyze solar cells by providing important information extracted from IQE measurements.

2 | METHODOLOGY

2.1 | Datasets

In this study, large training and testing datasets of IQE curves were generated using the formulas adapted from Fischer[7,8]:

$$\mathrm{IQE} = X_{\mathrm{EMI}} \cdot \int \eta(z) G(z) \cdot \mathrm{d}z, \tag{2}$$

where X_{EMI} is the carrier collector (emitter) term, while the bulk term is the integral of the bulk collection efficiency, $\eta(z)$, multiplied by the normalized generation profile, $\boldsymbol{G}(z)$, over the cell depth, z. The key independent parameters of X_{EMI} are the emitter's internal collection efficiency (IQE_0) and the width of the emitter (w_e), while $\eta(z)$ is determined by the effective diffusion length (L_{eff}) and rear surface recombination velocity (SRV_B). The value of $G(z)$ is impacted by the internal reflectance of the rear surface (R_B) and by the parameter, D_B, which describes how diffused the light is after the first internal reflection and ranges between zero (no diffusion) and one (complete diffusion). The complete set of the formulas is provided in Appendix A with definitions of the variables in Table A1.

Randomized combinations of values of these six key parameters (IQE_0, w_e, L_{eff}, SRV_B, R_B, and D_B) were used to generate a simulated dataset of 100 000 labeled IQE curves in the wavelength range of 280 to 1200 nm at 10 nm steps. Table 1 lists the range of values for each parameter.

The separate effect of each parameter on the shape of the simulated IQE curves is displayed in Figure 1. The X_{EMI} terms (IQE_0 and w_e) affect the short wavelength region. The value of IQE_0 directly shifts the y-intercept of the curve (Figure 1A) and thicker w_e reduce the IQE in the 300 nm to 600 nm range of the curve (Figure 1B). The values of $\eta(z)$ and $G(z)$ affect the long wavelength region. Higher values for L_{eff} increase the IQE (Figure 1C), while larger SRV_B reduces the IQE (Figure 1D) in the 800 nm to 1000 nm spectral range of the curve. Higher values of both R_B and D_B increase the IQE at the near-bandgap wavelengths, as shown in Figure 1E,F, respectively.

As discussed, the current of solar cells can be significantly reduced by parasitic absorption in the ARC layer, causing a reduction of the IQE in the short wavelength range. To incorporate the parasitic absorption, an additional dataset was created, integrating silicon nitride (SiN_x) as the ARC layers. Randomized SiN_x layer thickness and extinction coefficients (taken from PV Lighthouse[15,17]) were used to

TABLE 1 List of parameters and their range of values used to simulate the IQE curves and parasitic ARC absorption.

Parameter	Range or value
IQE_0	0.3–0.99
w_e (nm)	10–1000
L_{eff} (cm)	4×10^{-3} – 10
SRV_B (cm s^{-1})	1–400
R_B	0–1
D_B	0–1
SiN_x thickness (nm)	60–90
SiN_x extinction coefficients	PV lighthouse[15,17]

Abbreviations: ARC, antireflection coating; IQE, internal quantum efficiency; PV, photovoltaic.

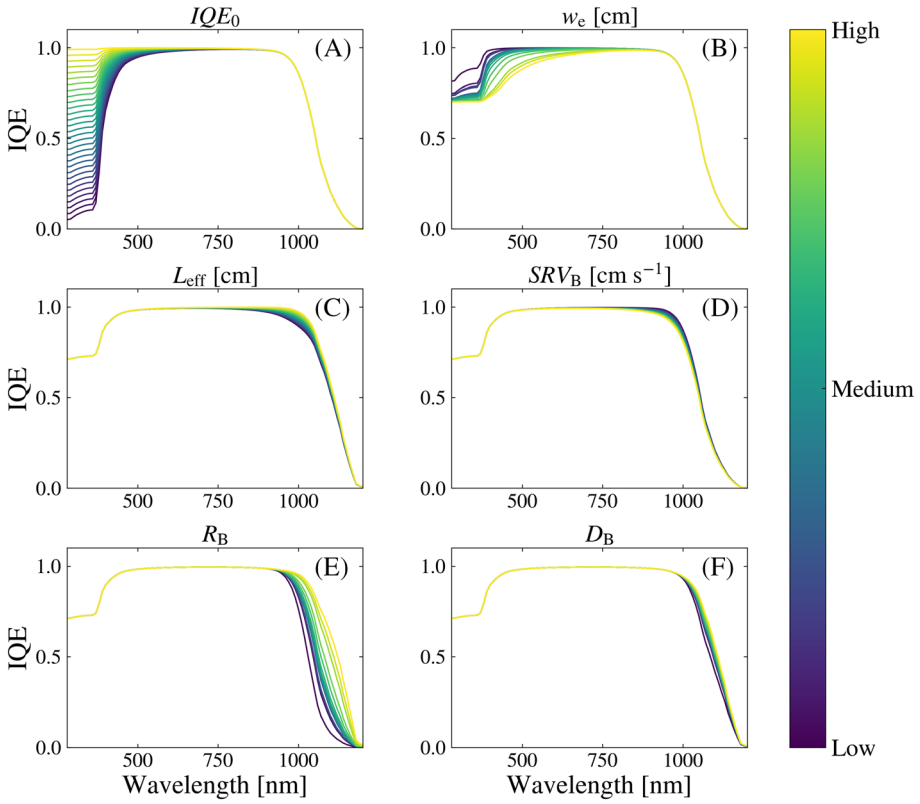

FIGURE 1 Representation of the effect of individual parameters on the simulated internal quantum efficiency (IQE) data, (A) IQE_0, (B) w_e, (C) L_{eff}, (D) SRV_B, (E) R_B, and (F) D_B.

calculate a large range of different parasitic absorption curves (A_{SiN_x}). The thickness range and extinction coefficients are listed in Table 1. These generated curves were then randomly applied to the existing dataset to produce 100 000 IQE curves of solar cells with the wide range of ARCs combined with the variation in electrical and optical parameters described above.

For experimental validation, IQE measurements of 12 industrial passivated emitter and rear contact (PERC) cells fabricated using mono-, multi-, and cast-monocrystalline Si wafers were collected. The measurements were fitted using the LASSIE software[8] to extract the six key solar cell parameters discussed above. The extracted parameters were compared with the parameters obtained by the proposed machine learning method.

2.2 | Machine learning models

To train the ML regressor models, the dataset was randomly split into training and testing datasets with an 80:20 ratio. The 80% training set was further split into six folds, and six different regressor models were trained, one on each fold, to predict each of the six parameters. To improve the performance of the model, chain regression was applied. Chain regression involves training an array of ML models in a specified order of the given targets.[18] After the first model is trained, that model makes predictions on the subsequent fold of data, and these predictions are added to the feature vector that is used to train the next model. Hence, the subsequent models use previous parameter predictions in their training, forming a chain link between the ML models.[18] The training methodology used in training the chain regressor is summarized in Figure 2. The predictions for IQE_0 and w_e were generally very good using single ML regressor instances so were kept in the first and second places of the chain order. In contrast, the order of the remaining regressors (L_{eff}, R_B, D_B, and SRV_B) has slight effects on the resulting predictions. Different permutations of the order were tested, and the results of this procedure are provided in Appendix B. Hereafter, the optimized order is used for the discussion.

Additional features were also added before training to further improve the prediction results. These additional features were formed by calculating the discrete differences in IQE values between each 10 nm step. As a result, the feature vector input to the chain model contains both the IQE curve and IQE difference curve. This is known as feature engineering,[19] and it was found to improve the prediction results by increasing the information provided to the models. A comparison between using and not using feature engineering is provided in Appendix C and their performance metrics are listed in Table C1.

Different regressor models, available in the scikit-learn package,[18] were compared in this study. They included random forest (RF),[20] support vector machine (SVM),[21] and stochastic gradient descent linear regression (SGD).[22] These models were optimized using hyperparameter tuning (values that control the learning process), where repeated random selections of their hyperparameters are trained via cross-validation (CV).[18] The optimum results of each random search are supplied in Appendix B with the performance metrics for each regressor type included in Tables B1-B3. The optimized chain regression model with the best results was used for later steps in the methodology.

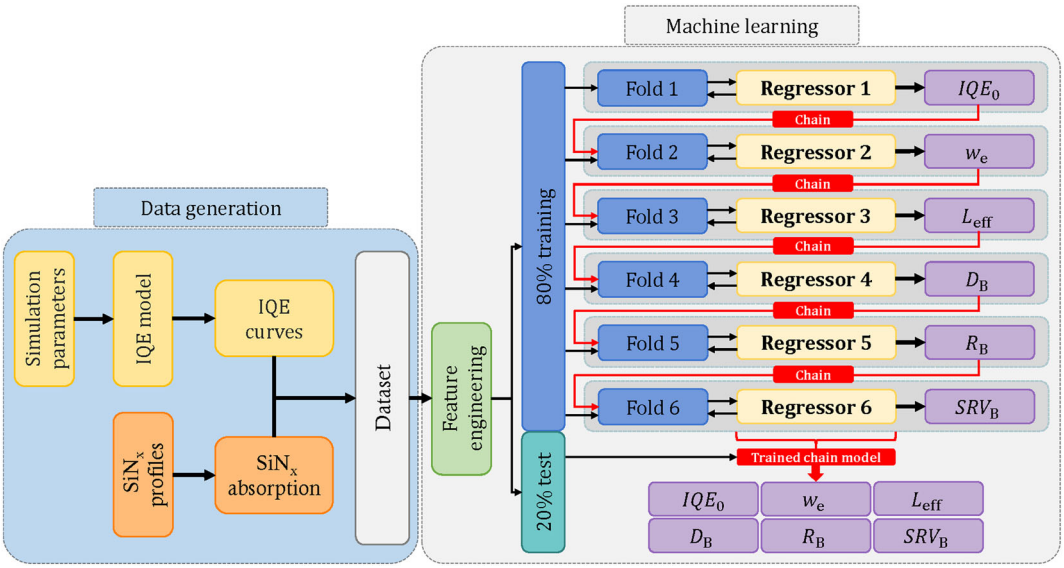

FIGURE 2 Flowchart of the data generation and machine learning steps. The generated internal quantum efficiency (IQE) curves and SiN_x absorption profiles were combined to form the complete dataset. Feature engineering was then used before splitting the data into 'training' and 'test' sets, while the training set was further split into six folds. A dedicated model was trained to predict each of the single parameters while the chain regression used the predicted parameters as inputs for the subsequent models. The unseen test set was then used to evaluate the trained chain of regressors.

The ML models were trained to predict the six parameters from IQE curves of solar cells with ARCs. As the short wavelength region of the IQE curve is dominated by X_{EMI}, the predicted IQE_0 and w_e values can be used to generate an IQE curve without the impact of A_{SiN_x}. By subtracting the short wavelength region of the two IQE curves (with and without the impact of A_{SiN_x}), A_{SiN_x} is determined. A representative example of this procedure is shown in Figure 3. The current density losses (J_{SiN_x}) due to A_{SiN_x} can then be calculated by integrating the product of A_{SiN_x} and the AM1.5G spectrum[23] over the wavelength range 280–600 nm.

To evaluate the trained models, the predicted and true parameters of the 20% test set were compared using the root mean square error (RMSE)[18]:

$$\mathrm{RMSE}(y,\hat{y}) = \frac{1}{N}\sqrt{\sum_{i=1}^{N}(y_i-\hat{y}_i)^2}, \tag{3}$$

where N is the number of samples, y_i is the true value of the i^{th} sample, and $\hat{y}_i$ is the corresponding predicted value. The performance of the ARC decoupling was evaluated by comparing the simulated and predicted J_{SiN_x} from the test dataset using the RMSE.

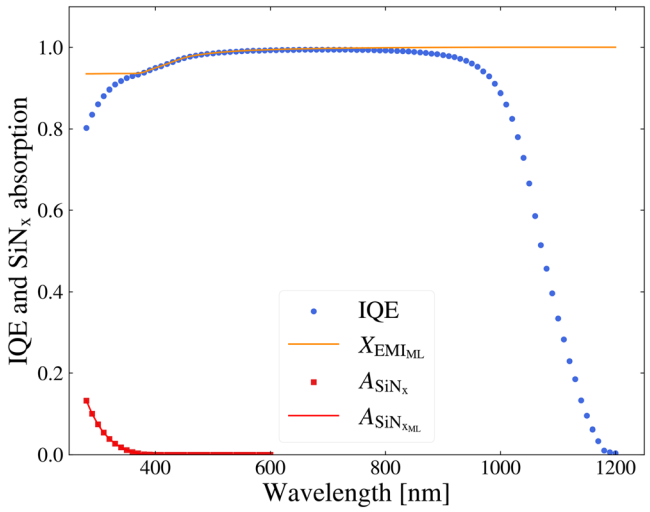

FIGURE 3 Simulated/measured internal quantum efficiency (IQE) with A_{SiN_x} (blue dots) and the IQE curve generated considering only IQE_0 and w_e, $X_{EMI_{ML}}$ (orange line). The subtraction of the IQE from $X_{EMI_{ML}}$ provides the predicted parasitic absorption curve, $A_{SiN_{x_{ML}}}$ (red line), which is compared with the true parasitic absorption curve, A_{SiN_x} (red squares).

3 | RESULTS

The results of the three trained models (RF, SVM, and SGD) on the simulated dataset are compared in Table 2. The optimized chain order for each model structure is also included. The RF chain regressor obtains the best RMSE across the six parameters. This is due to a large number of decision trees (1000 in this study) that constitutes the entirety of the model structure. These decision trees are also uncorrelated, so each makes a prediction on its own, allowing for the ensemble of trees to obtain the best possible prediction. The best RF model and its optimized chain order (IQE_0, w_e, L_{eff}, D_B, R_B, and SRV_B) are used in the following sections.

TABLE 2 The RMSE values obtained by the ML models with their optimized chain order.

Parameter/model	RF	SVM	SGD
IQE_0	0.00481	0.0334	0.0413
w_e (nm)	12.1	77.0	285
L_{eff} (cm)	0.192	0.833	0.673
D_B	0.0383	0.158	0.246
R_B	0.00192	0.045	0.0347
SRV_B (cm s^{-1})	12.2	23.0	29.0
Optimized chain order	IQE_0, w_e, L_{eff}, D_B, R_B, SRV_B	IQE_0, w_e, SRV_B, D_B, L_{eff}, R_B	IQE_0, w_e, L_{eff}, R_B, D_B, SRV_B

Abbreviations: ML, machine learning; RMSE, root mean square error; RF, random forest; SGD, stochastic gradient descent; SVM, support vector machine.

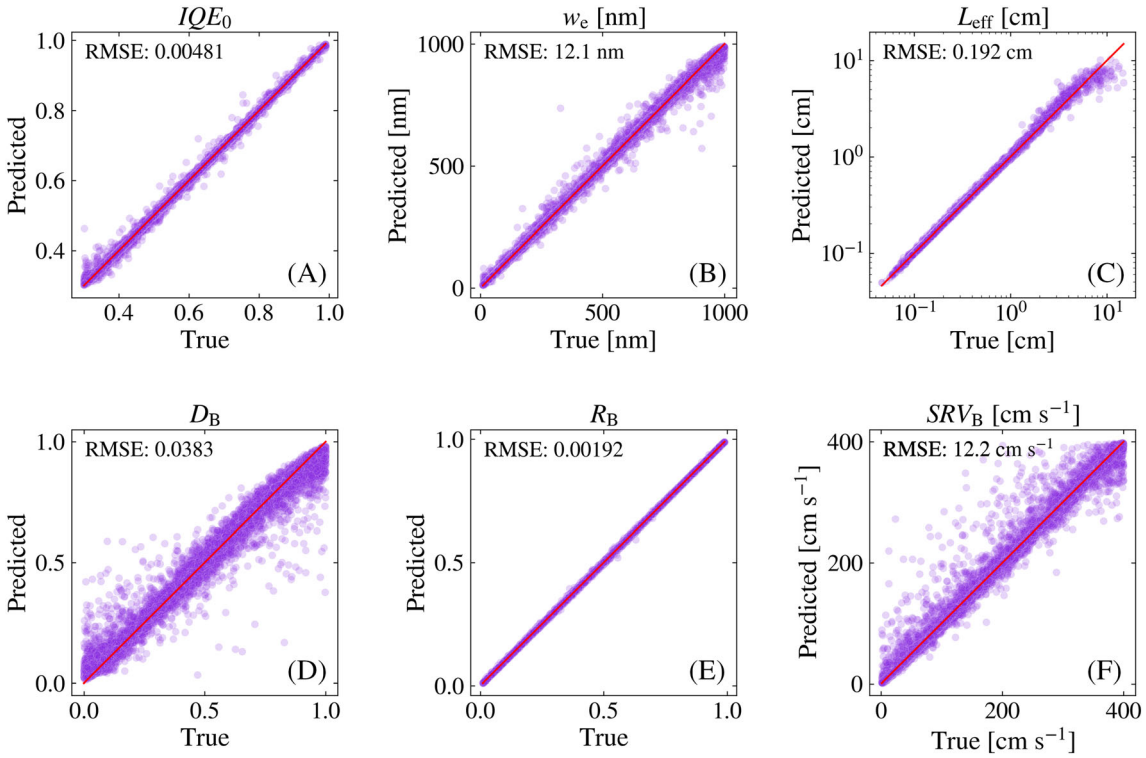

FIGURE 4 The chain regression results displayed as predicted value versus true value plots (A) IQE_0, (B) w_e, (C) L_{eff}, (D) D_B, (E) R_B, and (F) SRV_B.

The results of the optimized RF chain model applied to the test set are shown in Figure 4. The model predicts IQE_0, w_e, L_{eff}, and R_B with very low RMSE values (included in the figure). However, the deviations in Figure 4D, F indicate a larger uncertainty for D_B (RMSE of 0.0383) and SRV_B (RMSE of 12.2 cm s^{-1}). When investigating the errors in the D_B predictions, it was found that at low R_B values, D_B has a negligible impact on the IQE curve. Therefore, when R_B is low, the trained model naturally struggles to predict the correct value of D_B. This effect is described in Figure 5 below. However, modern solar cells are designed to have a high rear internal reflection to increase the capture of long wavelength light.[24] In these cases, the RF model predicts D_B with a high level of accuracy. For example, after removing samples with $R_B < 0.5$ from the test set, the RMSE of D_B predictions reduces from 0.0383 to 0.0171.

Errors in the SRV_B predictions were also investigated. As shown in Figure 4F, the larger deviations from the $y = x$ line are due to the overestimation of SRV_B. Representative samples were further investigated. It was found that the values suggested by the ML model still produced good fits to the IQE, compensating for the higher SRV_B with higher than the true L_{eff} values. An example of this occurrence is displayed in Figure 6. To overcome these few cases, a larger training dataset with more cases of very similar IQEs (in the 800 to 1050 nm range) needs to be generated to fine tune the regressors. Alternatively, fixing known parameters may improve the prediction in these cases.

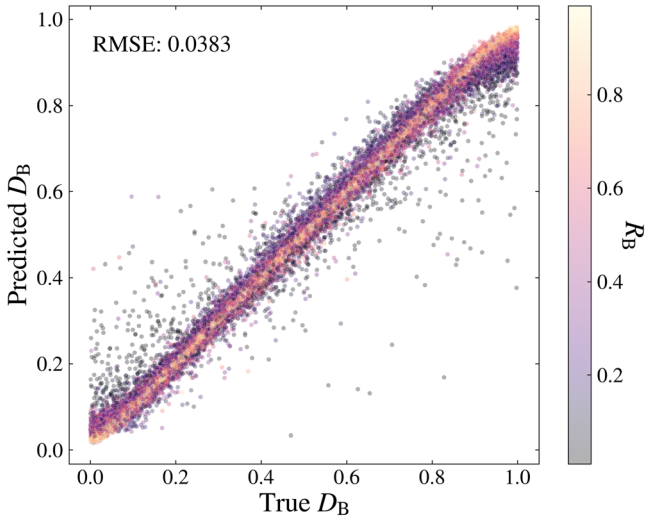

FIGURE 5 The predicted versus true plot of D_B with a color scale showing the corresponding R_B. As R_B increases, the prediction performance of D_B improves.

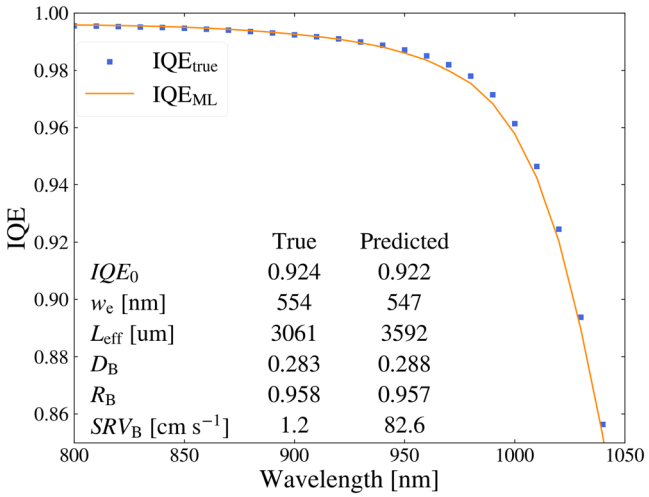

FIGURE 6 An example where the internal quantum efficiency (IQE) generated from the wrongly predicted machine learning (ML) values (IQE_{ML}) is similar to the true IQE (IQE_{true}).

3.1 | Decoupling of the ARC absorption

The predicted versus true values of J_{SiN_x} are shown in Figure 7. The discrete nature of the plot is due to the discrete combinations of ARC layer thickness and optical properties used in the simulation (as listed in Table 1). The model successfully extracts the absorption within the ARC layer with a very low RMSE (8×10^{-5} mA cm^{-2}). The results demonstrate the ability of the developed method to automatically decouple the parasitic absorption of the ARC layer from IQE measurements. This is an otherwise difficult problem to overcome when using traditional fitting methods in the analysis of IQE measurements.

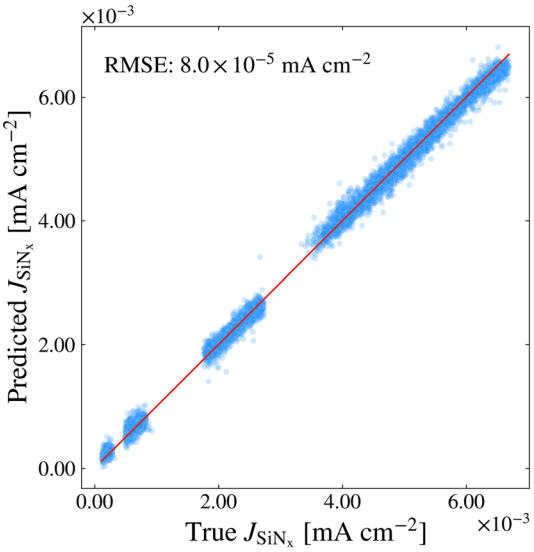

FIGURE 7 Predicted versus true J_{SiN_x} for the test set. The gaps in the plot are due to the discrete combinations of antireflection coating (ARC) layer thickness and extinction coefficients[15,17] used in the simulation of A_{SiN_x}.

TABLE 3 Comparison of the RMSE values of the ML models trained on a dataset with and without noise.

Parameter	RMSE without noise	RMSE with noise
IQE_0	0.00481	0.0054
w_e (nm)	12.1	23.6
L_{eff} (cm)	0.192	0.280
D_B	0.0383	0.0966
R_B	0.00192	0.00518
SRV_B (cm s^{-1})	12.2	13.9
J_{SiN_x} (mA cm^{-2})	8×10^{-5}	1.25×10^{-4}

3.2 | Noise sensitivity

The proposed ML approach should be capable of predicting the six solar cell parameters in the presence of noise, inevitably expected in experimental IQE data. Another RF chain model was trained on a dataset with randomly generated Gaussian noise added to the simulated dataset. The standard deviation of the added noise at each wavelength was determined based on the measured noise in our measurement system. The six RMSE prediction values without noise (as shown in Figure 4) and with noise are compared in Table 3. As expected, the errors increase when compared with the dataset without noise; however, the RMSE values are still low. For D_B, the larger errors were still found to be in samples where the simulated R_B values were small. The ARC decoupling also achieves an RMSE of 1.25×10^{-4} mA cm^{-2}, showing that the developed approach is still capable of decoupling the parasitic absorption from noisy IQE measurements. Overall, it seems that the ML-based method is resilient to

noise. Supporting information on this dataset is included in Appendix D with Figures D1-D3 displaying the results of the ML model trained with noise added to the dataset.

To further demonstrate this resiliency, 20 randomly chosen IQE curves simulated with noise were also fit using LASSIE.[8] The extracted parameters by the two methods (ML and LASSIE) were then com-

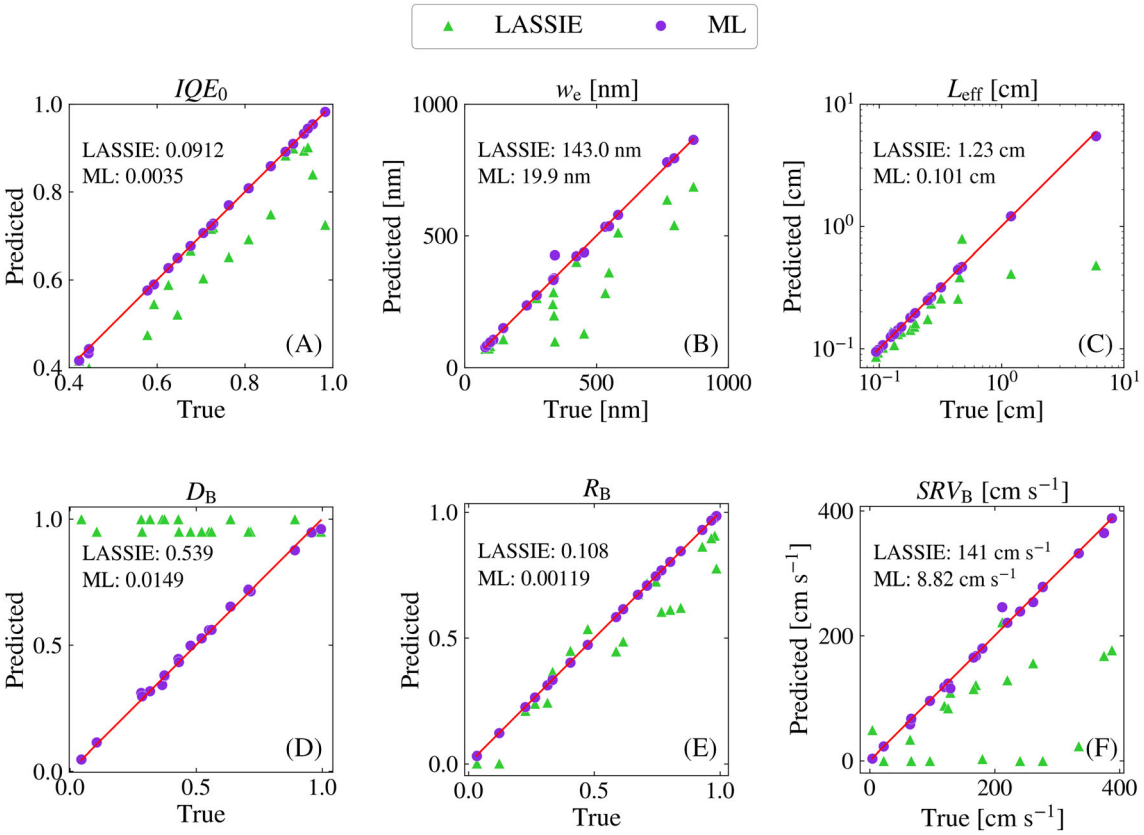

FIGURE 8 The predicted values versus true value plots of (A) IQE_0, (B) w_e, (C) L_{eff}, (D) D_B, (E) R_B, and (F) SRV_B fit in LASSIE (green triangle) and predicted by machine learning (ML) (purple dots) of 20 different simulated curves with noise. The root mean square error (RMSE) for each method is printed in the axis of each parameter.

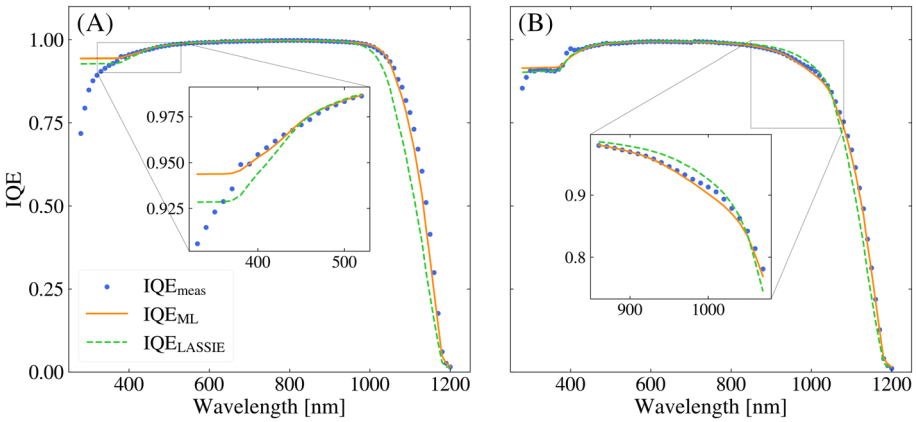

FIGURE 9 Two examples comparing generated internal quantum efficiency (IQE) data with experimental IQE measurements (blue dots) of (A) passivated emitter and rear contact (PERC) monocrystalline and (B) PERC cast-monocrystalline solar cells (IQE_{meas}). Using Equation (2), the machine learning (ML)-based and the LASSIE fit parameters generate the curves IQE_{ML} (orange) and IQE_{LASSIE} (dashed green), respectively. Insets are used to highlight the differences in the fit quality between the two methods.

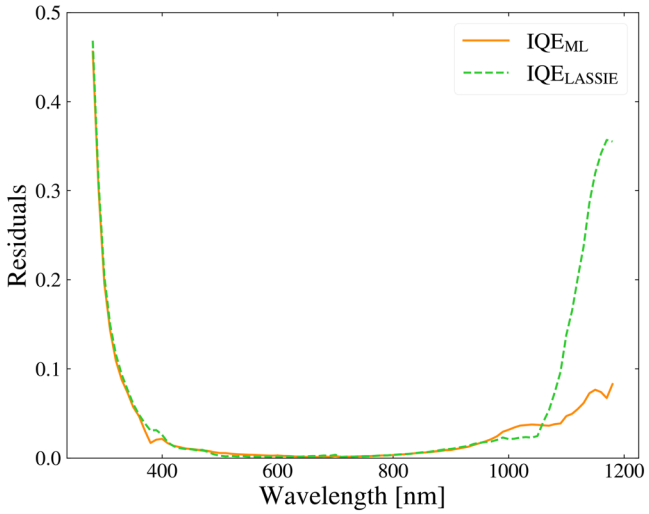

FIGURE 10 Average residuals at each wavelength for the generated internal quantum efficiency (IQE) curves from machine learning (ML)-based (IQE_{ML} [orange]) and LASSIE (IQE_{LASSIE} [dashed green]) parameters when compared with the validation dataset of 12 different solar cells.

pared. The predicted and true parameters are presented in Figure 8. The ML predictions obtain much lower RMSE for all six parameters. These results show that while noise in IQE measurements increases the error in the ML predictions, the proposed approach is more noise resilient than the traditional fitting methods.

3.3 | Validation

Experimental IQE data were used to validate the trained RF model. The measured cells consisted of 12 PERC solar cells fabricated using mono-, multi-, and cast-monocrystalline wafers. For comparison, the experimental IQE data were also fit using LASSIE,[8] which provides the same six parameters. Using Equation (2), the extracted parameters (by the ML-based approach and by LASSIE) were used to generate IQE curves. The quality of the fit was then evaluated by comparing the measured and generated IQE. However, as ARC decoupling is obtained by the subtraction of the generated IQE from the measured IQE, any errors in the short wavelength region will be compensated by this decoupling process. Therefore, ARC decoupling was not included in the comparisons, resulting in larger residuals in the short wavelength range. Two representative examples of the validation set are provided in Figure 9. The ML-generated IQE curve (IQE_{ML}) achieves a closer fit than the LASSIE-generated curve (IQE_{LASSIE}) in the short (see the inset of Figure 8A) and long (see the inset of Figure 8B) wavelength ranges. Above 1050 nm, it is clear that the IQE_{ML} achieves a better fit than IQE_{LASSIE} in both examples.

A comparison of the whole validation set is provided in Figure 10. The differences between the measured (IQE_{meas}) and generated IQE curves (IQE_{ML} and IQE_{LASSIE}) were calculated for all the cells. These residual curves were then averaged, and the resulting averaged residuals are presented in Figure 10. As expected, due to not applying ARC decoupling, the residuals are higher in the range below 600 nm. While the two methods provide similar fits in the range between 600 nm and 1000 nm, IQE_{LASSIE} achieves a slightly lower error in the 1000 to 1050 nm range. However, the ML approach vastly outperforms the LASSIE-based fit above 1050 nm. This section is predominantly affected by the electrical and optical parameters (L_{eff}, SRV_B, R_B, and D_B). Calculating the area under the curves (AUC) for the two residual curves shown in Figure 10, the ML method has an AUC of 13.02, whereas the fitting method (LASSIE) has an AUC of 21.90.

4 | CONCLUSION

In this study, ML is proposed as a powerful method to extract key performance metrics from IQE measurements of Si solar cells. A chain of RF models was trained on a large dataset of simulated IQE curves to predict six key parameters relating to the cells' electrical and optical performance. The trained model achieved low RMSE scores showing that this method is capable of learning the essential features of an IQE measurement and is thus, capable of replacing traditional fitting methods. The method was also shown to be reasonably resilient to noise in IQE measurements. Furthermore, using the ML approach, it was shown that the effect of parasitic absorption of the ARC layer can be decoupled from the IQE measurement, which is otherwise difficult to extract using traditional fitting methods. The trained model was validated on real IQE measurements of PERC solar cells, indicating that the developed ML model outperforms the fitting method. The method can be easily adapted to analyze IQE measurements of other solar cell types.

AUTHOR CONTRIBUTIONS

All authors contributed to the development of the methodology. Z.A. V developed the models, did the analysis, and wrote the initial version of the manuscript. Z. H. initiated and supervised the work.

ACKNOWLEDGEMENTS

The authors would like to thank A. Krzywicki (UniofAdelaide) for his helpful discussions at the beginning of this project and A. Samadi (UNSW) for sharing his assortment of IQE measurements used for the validation. This work was supported by the Australian Government through the Australian Renewable Energy Agency (ARENA, Grant 2020/RND016). The views expressed herein are not necessarily the views of the Australian Government, and the Australian Government does not accept responsibility for any information or advice contained. Open access publishing facilitated by University of New South Wales, as part of the Wiley - University of New South Wales agreement via the Council of Australian University Librarians.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available from the corresponding author upon reasonable request.

ORCID

Zubair Abdullah-Vetter https://orcid.org/0000-0002-2743-0102

Yoann Buratti https://orcid.org/0000-0002-3621-9712

REFERENCES

1. Hartman JS, Lind MA. Spectral response measurements for solar cells. *Solar Cells*. 1982;7(1):147-157. doi:10.1016/0379-6787(82)90099-0
2. Yang WJ, Ma ZQ, Tang X, Feng CB, Zhao WG, Shi PP. Internal quantum efficiency for solar cells. *Solar Energy*. 2008;82(2):106-110. doi:10.1016/j.solener.2007.07.010
3. Brendel R, Hirsch M, Plieninger R, Werner JH. Quantum efficiency analysis of thin-layer silicon solar cells with back surface fields and optical confinement. *IEEE Trans Electron Dev*. 1996;43(7):1104-1113. doi:10.1109/16.502422
4. Basore PA, "Extended spectral analysis of internal quantum efficiency," in 23rd IEEE Photovoltaic Specialists Conference, 1993, pp. 147–152.
5. Lan D, Green MA. Extended spectral response analysis of conventional and front surface field solar cells. *Sol Energy Mater sol Cells*. 2015;134:346-350. doi:10.1016/j.solmat.2014.12.018
6. Basore PA. Numerical modelling of textured silicon solar cells using PC-1D. *IEEE Trans Electron Dev*. 1990;37(2):337-343. doi:10.1109/16.46362
7. Altermatt PP, Müller J, Fischer B, "A simple emitter model for quantum efficiency curves and extracting the emitter saturation current," in 28th European Photovoltaic Solar Energy Conference and Exhibition, 2013, pp. 840–845.
8. Fischer B. *Loss Analysis of Crystalline Silicon Solar Cells Using Photoconductance and Quantum Efficiency Measurements*," PhD Thesis, Konstanz University; 2003.
9. Clugston DA, Basore PA, "PC1D version 5: 32-bit solar cell modelling on personal computers," in 26th IEEE Photovoltaic Specialists Conference, 1997, pp. 207–210.
10. "SunSolve™." https://www.pvlighthouse.com.au/sunsolve (accessed Mar. 08, 2022).
11. Wong J, "Griddler: intelligent computer aided design of complex solar cell metallization patterns," in 39th IEEE Photovoltaic Specialists Conference, 2013, pp. 933–938.
12. Fell A. A free and fast three-dimensional/two-dimensional solar cell simulator featuring conductive boundary and quasi-neutrality approximations. *IEEE Trans Electron Dev*. 2013;60(2):733-738. doi:10.1109/TED.2012.2231415
13. Froitzheim A, Stangl R, Elstner L, Kriegel M, Fuhs W, "AFORS-HET: a computer program for the simulation of heterojunction solar cells to be distributed for public use," in 3rd World Conference on Photovoltaic Energy Conversion, 2003, pp. 279–282.
14. Doshi P, Jellison GE, Rohatgi A. Characterization and optimization of absorbing plasma-enhanced chemical vapor deposited antireflection coatings for silicon photovoltaics. *Appl Optics*. 1997;36(30):7826-7837. doi:10.1364/AO.36.007826
15. Duttagupta S, Ma F, Hoex B, Mueller T, Aberle AG. Optimised antireflection coatings using silicon nitride on textured silicon surfaces based on measurements and multidimensional modelling. *Energy Procedia*. 2012;15:78-83. doi:10.1016/j.egypro.2012.02.009
16. Nagel H, Aberle AG, Hezel R. Optimised antireflection coatings for planar silicon solar cells using remote PECVD silicon nitride and porous silicon dioxide. *Progress in Photovoltaics: Research and Applications*. 1999;7(4):245-260. doi:10.1002/(SICI)1099-159X(199907/08)7:4%3C245::AID-PIP255%3E3.0.CO;2-3
17. Vogt MR. *Development of Physical Models for the Simulation of Optical Properties of Solar Cell Modules*," PhD Thesis, Leibniz Universität Hannover; 2015.
18. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: machine learning in python. *J Mach Learn Res*. 2011;12:2825-2830.
19. Raschka S, Mirjalili V. *Python Machine Learning*. 3rded. Birmingham - UK: Packt Publishing; 2019.
20. Breiman L. Random forests. *Mach Learn*. 2001;45(1):5-32. doi:10.1023/A:1010933404324
21. Hastie T, Tibshirani R, Friedman J. Support vector machines and flexible discriminants. In: Hastie T, Tibshirani R, Friedman J, eds. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. New York, NY: Springer; 2009:417-458. doi:10.1007/978-0-387-84858-7_12
22. Kiefer J, Wolfowitz J. Stochastic estimation of the maximum of a regression function. *Ann Math Stat*. 1952;23(3):462-466. doi:10.1214/aoms/1177729392
23. Hulstrom R, Bird R, Riordan C. Spectral solar irradiance data sets for selected terrestrial conditions. *Solar Cells*. 1985;15(4):365-391. doi:10.1016/0379-6787(85)90052-3
24. Sabater AA, Fell A, Brand AA, Müller M, Greulich JM. Parameterization of the back-surface reflection for PERC solar cells including the variation of back-contact coverage. *IEEE J Photovolt*. 2021;11(5):1136-1140. doi:10.1109/JPHOTOV.2021.3082402

How to cite this article: Abdullah-Vetter Z, Dwivedi P, Buratti Y, Sowmya A, Trupke T, Hameiri Z. Advanced analysis of internal quantum efficiency measurements using machine learning. *Prog Photovolt Res Appl*. 2023;31(8):790-802. doi:10.1002/pip.3683

APPENDIX A

Complete formulae for the IQE generation used in Equation (2):[7,8]

$$\mathrm{IQE} = X_{\mathrm{EMI}} \cdot \int \eta(z) G(z) \cdot \mathrm{d}z, \tag{A1}$$

$$X_{\mathrm{EMI}} = IQE_0 + \left(1 - e^{-\frac{2w_{\mathrm{d}}}{(1-IQE_0)L_\alpha}}\right)(1-IQE_0)^2 \cdot \frac{L_\alpha}{2w_{\mathrm{d}}}, \tag{A2}$$

where $w_{\mathrm{d}} = \frac{1-IQE_0}{2} \cdot w_{\mathrm{e}}$

$$\eta(z) = \cosh\left(\frac{z}{L}\right) - \frac{L}{L_{\mathrm{eff}}} \cdot \sinh\left(\frac{z}{L}\right), \tag{A3}$$

where $L_{\mathrm{eff}} = \mathrm{L} \cdot \frac{1+SRV_{\mathrm{B}} \cdot L_{\mathrm{D}} \cdot \tanh\left(\frac{W}{L}\right)}{SRV_{\mathrm{B}} \cdot L_{\mathrm{D}} + \tanh\left(\frac{W}{L}\right)}$

$$G(z) = \frac{1}{L_\alpha \cos(v_1)} \cdot e^{-\frac{z}{L_\alpha \cos(v_1)}} + T_1 R_{\mathrm{B}} \cdot \frac{1}{L_\alpha \cos(v_2)} \cdot e^{-\frac{W-z}{L_\alpha \cos(v_2)}} + \frac{T_1 T_2 R_{\mathrm{B}} R_{f1}}{L_\alpha \cos(v_n)\left(1 - T_n^2 R_{\mathrm{B}} R_{fn}\right)} \cdot \left(e^{-\frac{z}{L_\alpha \cos(v_1)}} + T_n R_{\mathrm{B}} \cdot e^{-\frac{W-z}{L_\alpha \cos(v_2)}}\right) \tag{A4}$$

where

$$T_1 = e^{-\frac{W}{L_\alpha \cos(v_1)}}, T_n = e^{-\frac{W}{L_\alpha \cos(v_n)}}, T_2 = D_{\mathrm{B}} T_n + (1 - D_{\mathrm{B}}) T_1$$

$$\cos(v_2) = \frac{-W}{L_\alpha \cdot \ln T_2}, R_{f1} = \frac{D_{\mathrm{B}} T_n R_{fn} + (1 - D_{\mathrm{B}}) T_1 R_{fs}}{D_{\mathrm{B}} T_n + (1 - D_{\mathrm{B}}) T_1}.$$

TABLE A1 Variables and constant values used in the simulation of the IQE curves (Equations (A1)–(A4)).

Variables or constants	Definition	Units or value
L_α	Absorption depth of silicon (wavelength-dependent)	cm
IQE_0	Emitter internal collection efficiency	N/A (0-1)
w_e	Width of the emitter	cm
w_d	Width of the 'dead layer' model[7]	cm
$\eta(z)$	Bulk collection efficiency	N/A (0-1)
z	Depth within the bulk of the cell	cm
L	Diffusion length of minority carriers in the bulk	cm
SRV_B	Rear surface recombination velocity	cm/s
L_{eff}	Effective diffusion length of minority carriers in the bulk	cm
$G(z)$	Normalized generation profile	N/A (0-1)
R_B	Rear internal reflectance	N/A (0-1)
D_B	How diffused light is after the first rear internal reflection	N/A (0-1)
W	The thickness of the sample, constant value used	0.018 cm (180 μm)
T_1, T_2, T_n	Light transmission of the first pass, second pass after rear internal reflection, and remaining passes respectively	N/A (0-1)
$cos(v_1)$	Angle of light penetration due to front texturing	0.75
$cos(v_n)$	Angle of the diffused light after internal reflection	0.5
$cos(v_2)$	Effective angle of specular light after rear internal reflection	N/A (0-1)
R_{f1}	Effective front internal reflectance	N/A (0-1)
R_{fn}, R_{fs}	Internal reflectance of the diffused and specular light, respectively	0.62 and 0.92

Abbreviation: IQE, internal quantum efficiency.

APPENDIX B

B.1 | Hyperparameter and chain order optimization

Different permutations of the R_B, L_{eff}, D_B, and SRV_B were tested to identify the best chain order for each of the three tested ML models. For each permutation, a cross-validated random search was used to optimize the hyperparameters for the model and its chain order of parameter targets. The total dataset size was reduced to 20 000 curves to speed up this optimization process. The average RMSE[18] was used to evaluate the best models, which can be affected by the scale of the parameters. To ensure that the different scales of the parameters do not affect the averaged RMSE, the parameters were all scaled to the range [0 – 1][19] before training. This scaling was not necessary for the RF model used in the main paper. The results for each type of model are in the tables below.

TABLE B1 Ascending order of RMSE results for the optimized hyperparameters for each chain order for random forest.

First	Second	Third	Fourth	Number of estimators	Minimum samples before split	Minimum samples per leaf	Maximum features	Bootstrap	Average RMSE
L_{eff}	D_B	R_B	SRV_B	1000	2	1	auto	True	0.18697
L_{eff}	R_B	SRV_B	D_B	1000	2	1	auto	True	0.18732
R_B	L_{eff}	D_B	SRV_B	700	2	2	auto	True	0.18735
L_{eff}	R_B	D_B	SRV_B	1300	5	2	0.75	True	0.18889
L_{eff}	D_B	SRV_B	R_B	300	2	1	0.75	True	0.18974
R_B	L_{eff}	SRV_B	D_B	500	10	1	0.75	False	0.1918
R_B	SRV_B	L_{eff}	D_B	300	2	2	0.75	False	0.19191
SRV_B	L_{eff}	D_B	R_B	500	2	1	0.75	True	0.19236
L_{eff}	SRV	R_B	D_B	1300	10	2	0.75	True	0.19245
SRV_B	L_{eff}	R_B	D_B	700	2	6	0.75	False	0.19281
SRV_B	R_B	L_{eff}	D_B	300	5	2	auto	True	0.19305

(Continues)

TABLE B1 (Continued)

First	Second	Third	Fourth	Number of estimators	Minimum samples before split	Minimum samples per leaf	Maximum features	Bootstrap	Average RMSE
R_B	D_B	L_{eff}	SRV_B	700	5	1	0.75	True	0.19313
SRV_B	R_B	D_B	L_{eff}	700	5	2	0.25	False	0.1937
L_{eff}	SRV_B	D_B	R_B	700	10	2	auto	True	0.19398
D_B	L_{eff}	R_B	SRV_B	1300	5	1	auto	True	0.19499
R_B	SRV_B	D_B	L_{eff}	1300	10	1	0.75	False	0.19505
SRV_B	D_B	R_B	L_{eff}	500	10	2	0.75	True	0.19591
D_B	SRV_B	R_B	L_{eff}	500	2	2	auto	True	0.196645
D_B	L_{eff}	SRV_B	R_B	300	5	1	auto	True	0.196655
D_B	SRV_B	L_{eff}	R_B	700	2	1	auto	True	0.197253
SRV_B	D_B	L_{eff}	R_B	700	5	2	0.75	False	0.19726
D_B	R_B	SRV_B	L_{eff}	700	2	2	auto	True	0.197864
D_B	R_B	L_{eff}	SRV_B	1000	2	2	0.25	False	0.199142
R_B	D_B	SRV_B	L_{eff}	1000	10	4	0.25	True	0.20937

Abbreviation: RMSE, root mean square error.

TABLE B2 Ascending order of RMSE results for the optimized hyperparameters of each chain order for support vector machine regression. Poly and rbf are polynomial and radial basis function kernel types, respectively.

First	Second	Third	Fourth	Kernel type	Kernel coefficient	Regularization	Average RMSE
SRV_B	D_B	L_{eff}	R_B	rbf	1	0.5	0.241643
D_B	SRV_B	L_{eff}	R_B	poly	1	0.5	0.256816
D_B	L_{eff}	SRV_B	R_B	poly	1	1	0.259002
L_{eff}	SRV_B	D_B	R_B	rbf	0.1	100	0.265067
D_B	SRV_B	R_B	L_{eff}	poly	0.1	10	0.270219
SRV_B	D_B	R_B	L_{eff}	poly	Scale	100	0.271332
L_{eff}	D_B	SRV_B	R_B	poly	Scale	10	0.284915
SRV_B	L_{eff}	D_B	R_B	poly	Scale	10	0.286102
D_B	R_B	L_{eff}	SRV_B	rbf	0.1	10	0.298974
D_B	L_{eff}	R_B	SRV_B	poly	0.1	0.5	0.305471
D_B	R_B	SRV_B	L_{eff}	poly	0.01	100	0.308156
L_{eff}	D_B	R_B	SRV_B	poly	0.1	10	0.322739
SRV_B	R_B	D_B	L_{eff}	poly	Auto	100	0.342365
R_B	SRV_B	D_B	L_{eff}	rbf	1	0.1	0.343206
SRV_B	R_B	L_{eff}	D_B	rbf	1	0.1	0.343656
SRV_B	L_{eff}	R_B	D_B	rbf	0.1	0.5	0.344259
R_B	D_B	SRV_B	L_{eff}	poly	0.01	100	0.345552
R_B	L_{eff}	SRV_B	D_B	poly	0.01	100	0.347602
R_B	SRV_B	L_{eff}	D_B	rbf	0.1	1	0.347649
R_B	L_{eff}	D_B	SRV_B	poly	0.1	0.1	0.348882
L_{eff}	SRV_B	R_B	D_B	rbf	0.001	100	0.350953
L_{eff}	R_B	SRV_B	D_B	poly	0.1	0.1	0.351682
R_B	D_B	L_{eff}	SRV_B	rbf	1	0.5	0.353524
L_{eff}	R_B	D_B	SRV_B	rbf	1	0.5	0.353614

Abbreviation: RMSE, root mean square error.

TABLE B3 Ascending order of RMSE results for the optimized hyperparameters for each chain order for stochastic gradient descent linear regression.

First	Second	T hird	Fourth	Penalty type	Learning rate change	Initial learning rate	Regularization	Average RMSE
L_{eff}	R_B	D_B	SRV_B	Elastic net	Adaptive	0.01	0.0001	0.35422
R_B	L_{eff}	D_B	SRV_B	Elastic net	Adaptive	0.01	0.0001	0.35472
SRV_B	D_B	R_B	L_{eff}	l2	Constant	0.01	0.0001	0.36289
R_B	L_{eff}	D_B	SRV_B	l2	Adaptive	0.01	0.001	0.36812
L_{eff}	SRV_B	R_B	D_B	l2	Adaptive	0.01	0.001	0.37132
R_B	SRV_B	D_B	L_{eff}	Elastic net	Adaptive	0.01	0.001	0.37194
R_B	D_B	SRV_B	L_{eff}	Elastic net	Adaptive	0.01	0.001	0.37398
SRV_B	L_{eff}	D_B	R_B	elastic Net	Constant	0.01	0.0001	0.37566
L_{eff}	SRV_B	D_B	R_B	l2	Constant	0.01	0.001	0.37664
L_{eff}	D_B	SRV_B	R_B	Elastic net	Constant	0.01	0.001	0.38584
D_B	R_B	SRV_B	L_{eff}	l2	Constant	0.01	0.0001	0.38754
R_B	L_{eff}	SRV_B	D_B	l1	Adaptive	0.001	0.0001	0.40583
L_{eff}	D_B	R_B	SRV_B	l2	Adaptive	0.001	0.001	0.40674
D_B	R_B	L_{eff}	SRV_B	l2	Constant	0.001	0.0001	0.40993
SRV_B	R_B	D_B	L_{eff}	l2	Adaptive	0.001	0.001	0.41065
SRV_B	L_{eff}	R_B	D_B	Elastic net	Adaptive	0.001	0.001	0.41079
D_B	SRV_B	L_{eff}	R_B	l2	Adaptive	0.001	0.001	0.41111
D_B	L_{eff}	SRV_B	R_B	l2	Inverse scaling	0.01	0.0001	0.41287
D_B	SRV_B	R_B	L_{eff}	Elastic net	Constant	0.01	0.01	0.41366
SRV_B	D_B	L_{eff}	R_B	l2	Inverse scaling	0.01	0.0001	0.41612
D_B	L_{eff}	R_B	SRV_B	Elastic net	Adaptive	10	0.01	0.41906
SRV_B	R_B	L_{eff}	D_B	l1	Adaptive	0.01	0.01	0.42533
R_B	SRV_B	L_{eff}	D_B	l1	Constant	0.01	0.01	0.4259
L_{eff}	R_B	SRV_B	D_B	Elastic net	Adaptive	0.001	0.1	0.44873

Abbreviation: IQE, internal quantum efficiency.

APPENDIX C

C.1 | Effectiveness of using feature engineering

Feature engineering was used by calculating the discrete differences in IQE values between each 10 nm step and added as extra features for the input feature vector for training the models. This step was used to improve the prediction results, in particular, the optical parameters D_B and R_B. Below is a comparison of the RMSE test set results for when feature engineering is used and when it is not.

TABLE C1 Comparison of the RMSE values of the ML models trained on the dataset with and without feature engineering.

Parameter predicted	RMSE with feature engineering	RMSE without feature engineering
IQE_0	0.00481	0.00786
w_e (nm)	12.1	16.0
L_{eff} (cm)	0.192	0.195
D_B	0.0383	0.0877
R_B	0.00192	0.00521
SRV_B (cm s^{-1})	12.2	13.6

APPENDIX D

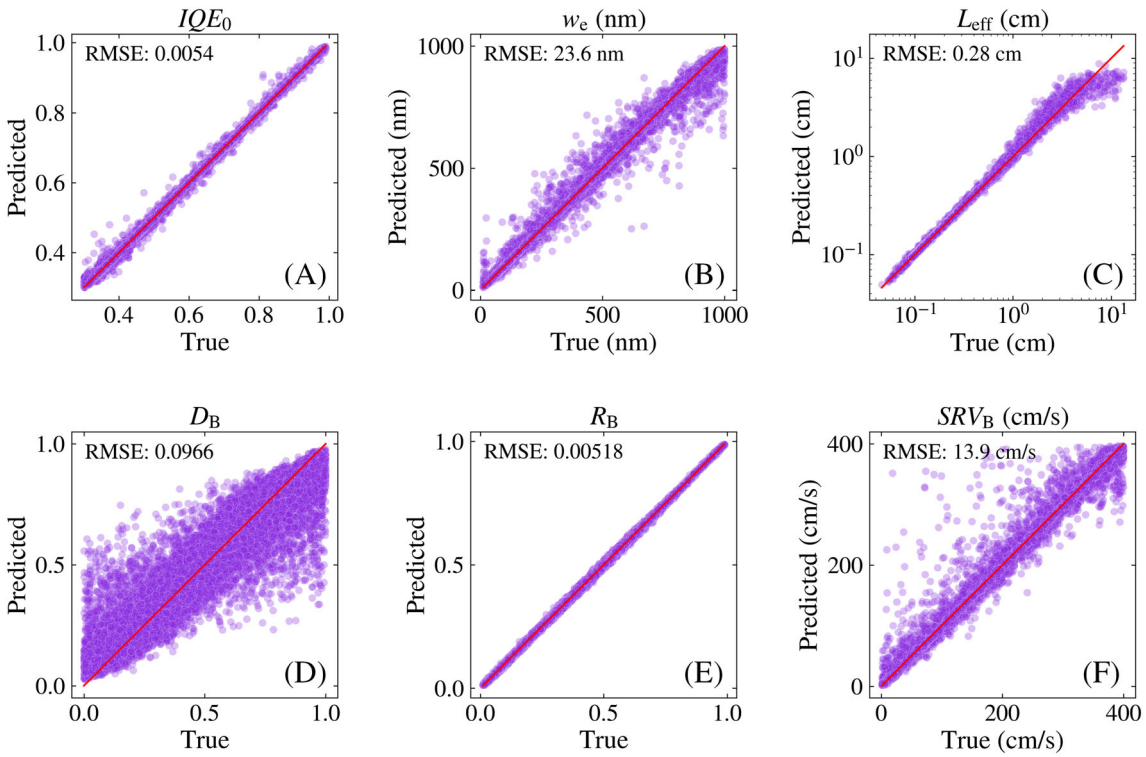

FIGURE D1 The chain regression results obtained from data with noise simulation displayed as predicted value vs true value plots (A) IQE_0, (B) w_e, (C) L_{eff}, (D) D_B, (E) R_B, and (F) SRV_B.

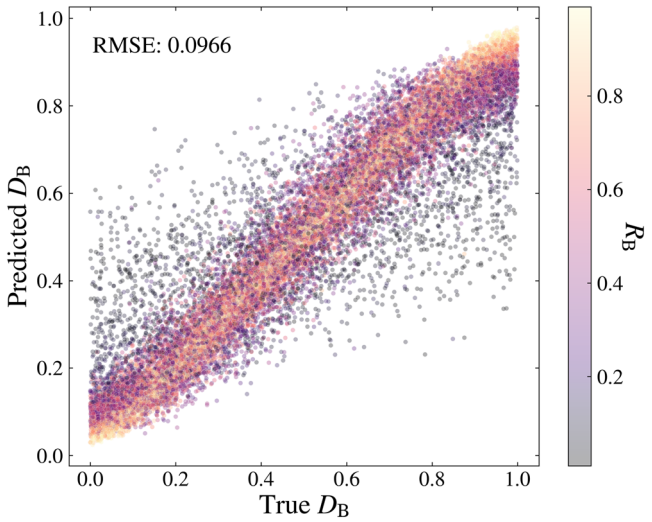

FIGURE D2 The predicted versus true plot of D_B predictions (from the noise simulated test set) with a color scale showing the corresponding R_B. As R_B values increase, the prediction performance of D_B improves. Removing R_B values below 0.5, the RMSE improves to 0.0495.

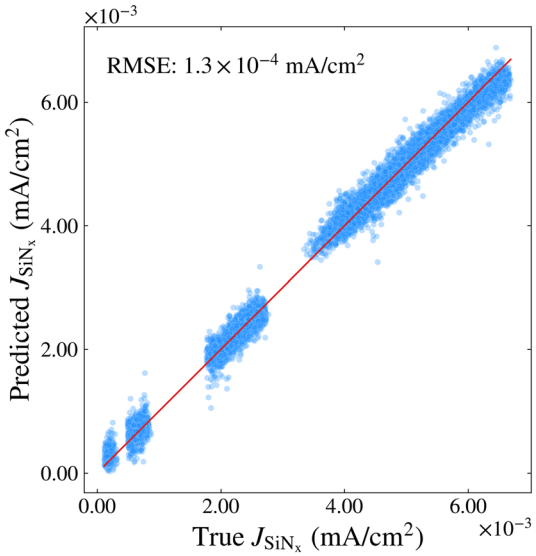

FIGURE D3 The true versus predicted J_{SiN_x} for the test set on data with noise simulation. The gaps in the plot are due to the discrete combinations of ARC layer thickness and extinction coefficients[15,17] used in the simulation of A_{SiN_x}.