

# Energy & Environmental Science

Accepted Manuscript

This article can be cited before page numbers have been issued, to do this please use: N. G. An, L. Ng Wei Tat, Y. Liu, S. Song, M. Gao, Y. Zhou, C. Ma, Z. Wei, J. Y. Kim, U. Bach and D. Vak, *Energy Environ. Sci.*, 2025, DOI: 10.1039/D5EE02815A.



This is an Accepted Manuscript, which has been through the Royal Society of Chemistry peer review process and has been accepted for publication.

Accepted Manuscripts are published online shortly after acceptance, before technical editing, formatting and proof reading. Using this free service, authors can make their results available to the community, in citable form, before we publish the edited article. We will replace this Accepted Manuscript with the edited and formatted Advance Article as soon as it is available.

You can find more information about Accepted Manuscripts in the [Information for Authors](#).

Please note that technical editing may introduce minor changes to the text and/or graphics, which may alter content. The journal's standard [Terms & Conditions](#) and the [Ethical guidelines](#) still apply. In no event shall the Royal Society of Chemistry be held responsible for any errors or omissions in this Accepted Manuscript or any consequences arising from the use of any information it contains.

## Broader context

Combining high-throughput experimentation with machine learning (ML) accelerates material discovery and optimizes fabrication processes. While prior works in photovoltaics have explored ML-driven research using roll-to-roll (R2R) technology, no systematic study has guided selection of ML models to precisely predict device behavior. The next critical step is to forecast new material device properties using insights from existing ones, requiring expandable data structure. We present a universal and expandable data structure designed to improve ML models and develop photovoltaic parameter and  $J-V$  prediction models. Using MicroFactory platform, we fabricated over 26,000 unique organic photovoltaics (OPVs), creating datasets used to train these models. Random forest proved the most effective model, and its predicted formulation led to PCE of 11.8% for fully R2R printed OPVs. We showed that models trained on binary systems can predict ternary device performance, demonstrating how accumulated, compatible datasets enable the development of generalized ML models with broader applicability to new material systems.



# Mass-customization of organic photovoltaics and data production for machine learning models precisely predicting device behavior

View Article Online  
DOI: 10.1039/C5EE02815A

Na Gyeong An<sup>1, 2, 3\*</sup>, Leonard Ng Wei Tat<sup>4</sup>, Yang Liu<sup>5</sup>, Seyeong Song<sup>3</sup>, Mei Gao<sup>1</sup>, Yinhua Zhou<sup>5</sup>, Chang-Qi Ma<sup>6</sup>, Zhixiang Wei<sup>7</sup>, Jin Young Kim<sup>3\*</sup>, Udo Bach<sup>2\*</sup>, and Doojin Vak<sup>1\*</sup>

## Author affiliations

- 1 Commonwealth Scientific and Industrial Research Organisation (CSIRO) Manufacturing, Clayton, Victoria, 3168, Australia
- 2 Department of Chemical and Biological Engineering, Monash University, Clayton, Victoria, 3800, Australia
- 3 Graduate School of Carbon Neutrality, School of Energy and Chemical Engineering, Ulsan National Institute of Science and Technology (UNIST), Ulsan, 44919, South Korea
- 4 School of Materials Science and Engineering (MSE), Nanyang Technological University (NTU), 50 Nanyang Avenue, Singapore 639798, Singapore
- 5 Wuhan National Laboratory for Optoelectronics, Huazhong University of Science and Technology, Wuhan, China
- 6 i-Lab & Printable Electronics Research Center, Suzhou Institute of Nano-Tech and Nano-Bionics, Chinese Academy of Sciences (CAS), Suzhou, 215123, China
- 7 Chinese Academy of Sciences (CAS) Key Laboratory of Nanosystem and Hierarchical Fabrication, National Center for Nanoscience and Technology, Beijing, 100190, China

## Broader context

Combining high-throughput experimentation with machine learning (ML) accelerates material discovery and optimizes fabrication processes. While prior works in photovoltaics have explored ML-driven research using roll-to-roll (R2R) technology, no systematic study has guided selection of ML models to precisely predict device behavior. The next critical step is to forecast new material device properties using insights from existing ones, requiring expandable data structure. We present a universal and expandable data structure designed to improve ML models and develop photovoltaic parameter and  $J-V$  prediction models. Using MicroFactory platform, we fabricated over 26,000 unique organic photovoltaics (OPVs), creating datasets used to train these models. Random forest proved the most effective model, and its predicted formulation led to PCE of 11.8% for fully R2R printed OPVs. We showed that models trained on binary systems can predict ternary device performance, demonstrating how accumulated, compatible datasets enable the development of generalized ML models with broader applicability to new material systems.



## Abstract

View Article Online  
DOI: 10.1039/D5EE02815A

High-throughput experimentation (HTE) combined with machine learning (ML) has emerged as a powerful tool to accelerate material discovery and optimize fabrication processes. However, in photovoltaics field, only a few studies have successfully applied this approach using industrially relevant techniques, roll-to-roll (R2R) process. We developed universal and extendable data structure for ML training that accommodates upcoming materials while retaining compatibility with existing dataset. Using MicroFactory platform, which enables mass-customization of organic photovoltaics (OPVs), we fabricated and characterized over 26,000 unique cells within four days. To guide selection of ML model for precisely predicting device behavior, photovoltaic parameter and  $J-V$  prediction models to forecast device parameters and  $J-V$  curves were developed, respectively. Random forest model proved most effective, achieving a PCE of 11.8% (0.025 cm<sup>2</sup>)—the highest record for fully R2R-fabricated OPVs. By integrating accumulated datasets with smaller new-component datasets, we enhanced model performance for PM6:Y6:IT-4F and PM6:D18:L8-BO systems, showing that models trained on binary systems can predict ternary performance and enabling the development of generalized ML models for future high-performance materials.

## Introduction

Organic photovoltaics (OPVs) hold immense promise as a clean and renewable energy source due to their myriad advantages, including being lightweight and flexible<sup>1</sup>, non-toxic<sup>2, 3</sup> and colour-tunable<sup>4-6</sup>. Recent advancements largely driven by materials discovery have pushed the power conversion efficiencies (PCEs) of OPVs beyond 20%<sup>7-11</sup>. However, a gigantic parameter space including an astronomical number of materials and processing parameter combinations remain unexplored, leaving much of OPVs' potential untapped.

High-throughput experimentation (HTE) has emerged as a promising avenue to address the limitations of conventional labor-intensive experimentations. It allows for the rapid screening of large combinational parametric libraries with fast characterization tools, thereby reducing labor, time and resource requirements. By combining HTE with machine learning (ML), vast datasets can be efficiently processed, accelerating material discovery or optimizing fabrication processes through data-driven decision-making.<sup>12</sup> However, many ML-driven studies in photovoltaics have relied on fabrication techniques incompatible with large-scale and mass production, such as spin coating<sup>13-18</sup>. It is difficult to utilize ML models trained with data from industry-incompatible fabrication techniques for a lab-to-fab transition. This is because such fabrication techniques require redevelopment of processing conditions specific to scalable manufacturing methods.<sup>19</sup> Furthermore, high-efficiency materials optimized under small-scale laboratory conditions often fail to perform reliably in large-scale processing, emphasizing the need to develop materials tailored for scalable manufacturing rather than modifying existing processes to accommodate lab-optimized materials.<sup>20, 21</sup> While some pioneering works successfully demonstrated ML-driven research with roll-to-roll (R2R) technology, industrial relevant method<sup>22-24</sup>, no systematic study has been conducted to give a guidance to determine ML models for accurately forecasting device behavior. Furthermore, the next critical step is to predict a device behavior of new materials leveraging insights from existing materials. This requires an expandable data structure for ML training to be able to accommodate new high-performance materials to be developed in the future, the critical gap yet to be addressed.

Here, we present a universal and expandable data structure designed to train ML models that continuously improve as more data are accumulated, regardless of the OPV material components. Through feature engineering, the training features are crafted to accommodate new materials while maintaining compatibility with existing dataset, enabling ML models to develop a comprehensive knowledge of OPV materials. Using the MicroFactory platform that mimics industrial processes<sup>24, 25</sup>,





we applied R2R slot die coating in a desktop machine to develop manufacturing process with minimum material usage and automated *in situ* formulation. This approach allowed for mass-customization of OPVs by continuously altering fabrication parameters and formulations while a R2R tester facilitated testing of the completed devices, leading to a production of over 26,000 unique OPV cells.

The fabrication parameters and device outcomes were digitalized and used to train ML models. Using this unprecedented amount of OPV data, we developed not only photovoltaic parameter (PP) prediction models forecasting device parameters but also  $J-V$  prediction models creating  $J-V$  curves of unexperimented devices. These models guided the selection of ML models through systematic evaluation by screening various algorithms, followed by hyperparameter optimizations for precisely predicting device behavior. The random forest (RF) model proved the most effective. Using the RF-predicted formulation, we fabricated a device achieving PCE of 11.8%, a record-breaking efficiency for fully R2R-fabricated OPVs. Furthermore, we demonstrated that combining large, accumulated datasets with smaller datasets from different material systems improves ML model performance, showing that models trained on binary systems can predict ternary performance and supporting the creation of generalized models capable of forecasting device behavior for new materials.

## Results and discussion

### High-throughput device fabrication/characterization using the MicroFactory platform

In this study, we selected PM6 as a donor and L8-BO as an acceptor material, and optimized R2R device configuration of PET (polyethylene terephthalate)/TCE (transparent conducting electrode)/PM6:L8-BO/BM-HTL-1/S315 as detailed in Note S1 and S2 (ESI). The material structure, energy diagram, device configuration, and resulting device performance are presented in Fig. 1a–c, Fig. S1–S12 and Table S1–S8 (ESI). Additionally, the effects of various additives including 1,8-diiodooctane (DIO), 1-chloronaphthalene (CN), 2-hydroxy-4-methoxybenzophenone (2-HM), and 1,4-dichlorobenzene (DCB) were investigated, but no significant impact on device performance was observed (Fig. S11 and Table S7, ESI). To generate a large dataset efficiently, we leveraged the MicroFactory platform (Fig. 1d, Fig. S13 and S14, ESI) developed with resources from the 3D printing community. In our previous work, we introduced a 3D printer-based slot die coater as a lab-to-fab translation tool<sup>26</sup>, which only required minimal modifications to an existing 3D printer. Building on this concept, it involves the construction of a fully customized machine derived from a 3D printing framework, with operations governed by the open-source G-code protocol. Additional technical details are provided in Note S3 (ESI).

With this optimized device configuration and the advanced capabilities of the MicroFactory platform, we fabricated over 26,000 OPV cells across two experimental batches. These experiments systematically explored fabrication parameters, with each batch divided into multiple sub-batches to alter donor:acceptor (D:A) ratio or film thickness. We previously reported a dual-feed deposition method to formulate solutions *in situ* and a way to digitalize active layer composition by introducing deposition density (DD) and total deposition density (TDD)<sup>23</sup>. DD is an absolute quantity of a component per unit area (in  $\mu\text{g cm}^{-2}$ ), and TDD is the sum of DDs of all materials and has a strong correlation with film thickness. Due to the lossless nature of slot die coating, DD can be derived from deposition parameters and solution compositions. This lossless deposition significantly reduces material consumption compared to traditional spin-coating methods. In our experiments, we used only 10.3 mg of PM6 and 12.2 mg of L8-BO per 1,000 devices, corresponding to  $\sim 964$  devices per mL of solution for both materials. In contrast, spin coating typically yields only 50–58 devices per mL. This high material efficiency highlights the practicality of our approach for screening new materials, even when available sample quantities are limited.

While our prior study required several months to collect data for about 2,000 devices<sup>23</sup>, the MicroFactory platform allowed for the high-throughput fabrication of over ten thousand devices in a



single day by one researcher, with testing completed the following day. Fabricating and characterizing over 26,000 OPV cells took only four days with the coater and tester operating in parallel. All data were saved online and could be simply combined using Python scripts. Although we paused fabrication for data analysis, this unprecedented fabrication capability demonstrates significant potential for creating big data on OPV manufacturing parameters.

Fig. 2a presents individual PPs, DD of each material, and L8-BO/PM6 ratios for 26,422 devices based on the device position on a 150 m-long substrate (Fig. 2b). Notably, failed devices were intentionally fabricated as markers between sub-batches and used to verify the combined dataset of fabrication and testing parameters. Since it is impossible to display  $J-V$  curves for all devices, we show  $J-V$  curves for the four most characteristic devices with the highest PPs in Fig. 2c and Table 1. The best device achieved a power conversion efficiency (PCE) of 11.6% at a D:A ratio of 1:1.13 (Fig. S15, ESI), similar to an optimal formulation found in a previous study<sup>27</sup>. Although this PCE is already significantly higher than that of the best fully R2R-fabricated OPVs in the literature (Fig. S16 and Table S9, ESI), further improvement might be possible through the ML-assisted device optimization. However, the achievable PCE would remain below 14.1%, a boundary determined by the best fill factor (FF), current density ( $J_{sc}$ ) and open circuit voltage ( $V_{oc}$ ) due to the inherent trade-offs among those photovoltaic parameters. However, identifying PP trends within a multi-dimensional fabrication parameter space from such large data sets is impractical for humans, necessitating the use of ML.

### Developing ML models

Utilizing ML for data analysis involves several steps, as illustrated in Fig. 3a. With the data already collected, the next important step is designing features as input parameters for ML training. Extracting meaningful features from available data is common practice. As we previously reported, ML models using simplified features, such as the DD of three materials, were able to visualize their effects<sup>23</sup>. However, we found this approach is not expandable for additional materials or fabrication parameters. Hence, we designed generic and material specific features (DD of each material) that can accommodate upcoming materials. Although some features may look like a duplication of another feature, it ensures the data can be used universally. Further details can be found in Note S4 (ESI), with descriptions of the data structures and features available in Tables S10–S12 (ESI), and the corresponding numerical values stored in Data S1 and S2.

After preparing and cleaning the data, ML models using various algorithms were trained and compared for a given multi-dimensional regression problem. Seven bagging and boosting algorithms were initially selected for evaluation. Bagging involves training multiple weak learners independently in parallel and averaging their predictions, which reduces variance and improves stability. Boosting, in contrast, builds weak learners sequentially, with each new model correcting the errors made by the previous ones.<sup>28</sup> These ensemble approaches are generally less prone to overfitting than a model learns the training data too closely including noise or fluctuations, resulting in poor generalization to new data. To identify the best-performing models, we optimized hyperparameters for each algorithm using GridSearchCV with 5-fold cross-validation. This approach systematically explores all possible combinations of predefined hyperparameter values and evaluates performance to find the optimal set. Due to the training time required for large numbers of hyperparameter combinations up to two million, we used a small fraction (2.5%) of data for training (Note S5 for detail, ESI). Each hyperparameter combination was evaluated based on performance metrics: coefficient of determination ( $R^2$ ), mean absolute percentage error (MAPE), root mean squared error (RMSE)<sup>23</sup> and overfitting index (OI, defined by  $R^2$  of test or validation dataset/ $R^2$  of train dataset where lower values indicate stronger overfitting whereas an OI close to 1 implies that the model generalizes well). This evaluation utilized 80% of the trained dataset for model training and the remaining 20% for model validation (Fig. S17, ESI). While RMSE is inherently tied to the scale of target variables, making it a scale-dependent metric<sup>29–31</sup>. MAPE can



become undefined or problematic when actual values are zero or near zero, resulting in infinite or extreme values<sup>30–32</sup>. In contrast,  $R^2$  offers a scale-independent evaluation<sup>33</sup>, quantifying how well the model accounts for the variance in the data relative to the total variance. Given these considerations, we employed  $R^2$  as a universal metric to determine the optimal hyperparameter combination of models generated by each algorithm and to enable cross-scale comparison between different prediction models. The performance metrics for the best model from each algorithm are shown in Fig. S18–S24 and Table S13–S19 (ESI) with detailed discussions in Note S6 (ESI). We found no outstanding model while Ada Boost algorithm produced model with poorer performance than others.

Based on this evaluation, we selected one widely used algorithm from each category for subsequent analysis: Random Forest (RF) for bagging and Extreme Gradient Boosting (XGBoost) for boosting. While optimizing most of the available hyperparameters (Note S7, ESI), we discovered that only a few hyperparameters (max leaf nodes, max features and number of estimators for RF and learning rate, number of estimators and max depth for XGBoost) made meaningful difference across all PPs. Therefore, we picked the top three most effective hyperparameters and selectively showcased their impact on model performance for predicting  $J_{sc}$  and PCE (Fig. S25–S28, ESI). This information will serve as a comprehensive guide for future hyperparameter optimization.

The models were then re-optimized with different dataset sizes (2.5%, 20 %, and 80%) using only the three most influencing hyperparameters. We observed general agreement in the optimal hyperparameters across various data fractions, indicating that hyperparameters optimized with a small dataset do not require re-optimization for larger datasets (Fig. S29–S32 and Table S20 and S21, ESI). The final optimized hyperparameters are shown in Note S8 (ESI).

### Performance of ML models for photovoltaics

While PCE is a key parameter of interest, we developed prediction models for all PPs including series resistance ( $R_s$ ) and shunt resistance ( $R_{sh}$ ) for a comprehensive understanding of device behavior based on fabrication parameters. Additionally, we trained models to create  $J$ – $V$  curves, referred to as  $J$ – $V$  prediction models, by introducing reading voltage values in  $J$ – $V$  curves as features in the train dataset. This approach resulted in about 1.8 million data points from 71 data points per  $J$ – $V$  curve, with the model making 71 predictions using various reading voltages. The selection of the best  $J$ – $V$  prediction models and their optimized hyperparameter combinations is described in Note S9 (ESI). Fig. 3 shows the predicted PPs and  $J$ – $V$  curves alongside the original measured  $J$ – $V$  curves of three randomly selected devices from the test dataset.

Therefore, our models offer three ways for predicting PCE: (i) direct prediction from the PCE prediction model, (ii) PCE calculated from the predicted FF,  $V_{oc}$  and  $J_{sc}$ , and (iii) PCE calculated from the created  $J$ – $V$  curves. We found general agreement in PCE from all three approaches, with similar and exceptionally high  $R^2$  values (Fig. 3 and S33 and Table 2 and S22, ESI). This high model performance is attributed to a consistently produced train dataset rather than collection from multiple sources and optimized hyperparameters. Although any of these three approaches can predict PCE, the  $J$ – $V$ -based approach provides comprehensive device insights while demanding more computing resources. However, once hyperparameters are optimized, training each model with 1.8 million datasets of 10 input parameters takes only 15 minutes for RF and 7 seconds for XGBoost on a desktop computer, suggesting that computing resources will not be an issue soon.

Despite XGBoost's advantage over RF in training time due to parallelized computing, it performed poorly in the  $J$ – $V$  prediction model, as observed in 3 and 15 additional randomly selected devices (Fig. 3 and S34, ESI). Although it still achieves comparable model evaluation metrics to the RF model (Fig. S35 and Table S23, ESI), these metrics fail to capture localized fluctuations in the predicted  $J$ – $V$  curves, which become evident only when the curves are plotted and compared against experimental data. To



further understand model behavior, we calculated feature importance for both RF and XGBoost  $J-V$  prediction model, where higher values indicate greater influence on the model's predictions. Among all features,  $V$  showed the highest importance for both RF and XGBoost model. However, since  $J$  and  $V$  are directly correlated in  $J-V$  prediction, we excluded  $V$  from subsequent analyses to better highlight the relative importance of the remaining features (Fig. S36, ESI). Our analysis revealed that 'Acceptor Frac' was the most influential feature for both models, while TDD showed much lower contributions. The lower importance of 'Donor Frac' is expected because it is linearly dependent on 'Acceptor Frac' (Donor Frac = 1 – Acceptor Frac), meaning the model primarily relies on one of the two to avoid redundancy. Importantly, this does not imply that donor content is less significant for device behavior but rather that its contribution is already captured through 'Acceptor Frac'. These results suggest that the D:A ratio plays a more critical role than active layer thickness in predicting  $J-V$  characteristics. Taking together, these results indicate that while RF and XGBoost identified similar trends in feature importance, RF provided more stable and consistent predictions across datasets. Therefore, RF-based  $J-V$  prediction models were chosen for photovoltaic optimization.

We observed poor performances in resistance prediction models. For the  $R_s$  model, this was attributed to a small number of extreme outliers, possibly from measurement error (poor contact during the measurement). After excluding about 3% of outliers, the model showed good performance with an  $R^2$  score of  $\sim 0.88$ . The  $R_{sh}$  model underperformed because the slope of the  $J-V$  curves at 0 V for good devices is near zero, making the inverse value highly sensitive to measurement noise. Despite this, we found the model proved useful in predicting defective devices ( $R_{sh} < 336 \Omega \text{ cm}^2$ , bottom 30% in this dataset) with  $R^2$  score of 0.80.

### Photovoltaic optimization using ML models

The  $J-V$  prediction ML model provides a whole new way of device optimization. We could browse experimental parameters with an interactive ML tool instantly showing  $J-V$  curves of virtually planned experiments (Video S1). The tool provides insights into parameter-dependent device performance to researchers, while full optimization was carried out programmatically by generating all possible parameter combinations and sorting by predicted PCE. However, this approach will provide a global optimum point, all top formulations would be in the same zone and miss serendipitous discoveries of high-performance devices that may not be designed by human knowledge.

To address this, we split all possible parameter combinations with a coarse data resolution ( $2 \mu\text{l min}^{-1}$  or  $\mu\text{g cm}^{-2}$ ) into clusters and examined the optimum condition for each cluster. Using K-means clustering, an unsupervised ML algorithm, we created ten clusters (Fig. S37 and Table S24, ESI). While the clusters of evenly spaced datasets differ from conventional meaning of clusters, it provides a simple solution for grouping multidimensional data. Subsequently, we generated all possible combinations with finer data resolution ( $0.5 \mu\text{l min}^{-1}$  or  $\mu\text{g cm}^{-2}$ , 6,561 parameter combinations for each cluster) around each cluster's top formulation to refine them further. Finally, we compared the Euclidean distance (ED) (Eq. 1) of each cluster's top formulation from the global best formulation in cluster 0 (Table S25 and Note S10 for detail, ESI).

Based on ED and predicted PCE, we fabricated devices using the predicted formulation of the global best and that of the cluster farthest from cluster 0. We found other clusters further than the selected cluster, however, parameters of those clusters were challenging to fabricate due to too high flow rates causing overflow issues and slightly lower PCEs. The device results are shown in Fig. 4 and summarized in Table 3. The global best formulation achieved up to 11.8% PCE, significantly higher than the best PCE for fully R2R fabricated OPVs in the literature. Previously, the highest reported PCE was 5.6% as summarized in Fig. S16 and Table S9 (ESI). Only recently, we reported a 9.35% using this high-throughput setup<sup>24</sup> and this work marks another leap in the record PCE for R2R fabricated OPVs.





It clearly demonstrates how laboratory innovation and digital technologies such as ML can accelerate OPV progress and potentially other printed electronics.

View Article Online  
DOI: 10.1039/D5EE02815A

With a view towards the eco-friendly manufacturing of OPV, we tested the parameters using a non-halogenated solvent, o-xylene (Xyl), and found Xyl-based devices performed comparably (Fig. 4 and Table 3). These results show great potential for a commercial application of fabrication parameters found in this work. External quantum efficiency (EQE) spectra were obtained after printing the silver electrode and encapsulating the cells for manual testing (Fig. S37, ESI). All four devices display good agreement between  $J_{SC}$  measured from the tester and calculated from EQE spectra (Fig. 4e and f).

### The Future of data-driven and ML-assisted photovoltaic research

While the MicroFactory platform has a unique strength in thoroughly screening parameters of given materials, it is limited in the number of materials that can be automatically processed in a batch. Conversely, robotic-arm-based HTE systems can screen a variety of materials, albeit with a relatively lower throughput compared to the MicroFactory platform<sup>12</sup>. It would be ideal to combine two complementary HTE platforms to accelerate the research cycle from material screening to the development of manufacturing techniques. A key question is whether large dataset can be reused to optimize newly developed materials. To address this, we aimed to create a data structure that can accommodate upcoming high-performance materials, enabling the continuous improvement of ML models with accumulated data. To achieve this, we created generic features as well as material specific features. For example, “Donor Frac” in Table S11 (ESI) may seem redundant since it can be easily derived from “PM6 DD” and “L8-BO DD”. However, this feature becomes crucial when introducing a co-donor or different donor material. Without such features, ML models would not understand the relationship between a new donor and PM6. By leveraging the established knowledge of known materials, ML models could be trained for new materials without re-creating extensive datasets.

To prove this hypothesis, we combined the large dataset with smaller datasets including new components, with one new material-specific feature per component (Fig. S2 for detailed chemical structure, ESI), trained RF-based models, and compared their performance depending on the size of additional datasets. We first used a dataset created from our previous study on PM6:Y6:IT-4F ternary blend<sup>23</sup>. In that case, we simplified the system and minimized the number of features to visualize all feature-dependent performance in 3D space, manually re-creating features based on experimental parameters. New dataset of a PM6:D18:L8-BO ternary and a PM6:PYF-T-o all polymer systems were also created by running relatively small batch. For training, we used batches of data rather than individual cells, specifically 294, 467, 1240 and 1990 for PM6:Y6:IT-4F, 257, 517, 1289 and 2053 for PM6:D18:L8-BO ternary system and 258, 784, 1538, and 2323 for PM6:PYF-T-o all polymer system, respectively. The remaining datasets were reserved for testing (Note S11 for detail, ESI). The complete datasets and model performance results are available in Data S3 and S4, respectively.

Fig. 5 illustrates performances of RF-based models trained without and with accumulated PM6:L8-BO dataset. Despite the PM6:Y6:IT-4F system having only one common material, the models trained with the accumulated data consistently outperformed those without it, especially more improvement with smaller train datasets. The same trend was observed in the PM6:D18:L8-BO system, which shares two common materials with the accumulated dataset, leading to even more significant improvements. Notably, the model trained with only 10% of the new dataset performed well in predicting the remaining 90%. Interestingly, while integrating the accumulated dataset substantially enhanced predictive accuracy for ternary systems, the improvement was not consistent for PM6:PYF-T-o all polymer system (Fig. S39, ESI). Specifically, in cases where the training dataset sizes were 784 or 1,538, the test  $R^2$  values showed a slight decrease after adding the accumulated data. This behavior may arise from intrinsic differences between all polymer systems and small molecule NFA-based systems. These



findings suggest that polymer acceptors may require a separate model category or an additional independent feature to better capture their distinct characteristics.

View Article Online  
DOI: 10.1039/D5EE02815A

Nonetheless, these results highlight the value of digitalizing research parameters and demonstrate how existing knowledge can be applied to new material systems in OPV research. By showing that models trained on binary systems can predict ternary performance, this work illustrates how accumulated, compatible datasets enable the development of ML models with a comprehensive understanding of materials and fabrication parameters. Such models could significantly accelerate OPV commercialization.

## Conclusion

We developed new data-driven research approach that will significantly accelerate the progress of printed OPV towards commercialization. It includes generating reusable and expandable training data capable of predicting the properties of new materials and developing PP and  $J-V$  prediction ML models. The data was firstly generated using the MicroFactory platform, which facilitated the mass-customization and characterization of over 26,000 OPV cells in just four days through fully digitalized processes. These data were then used to train the PP and  $J-V$  prediction models that can guide researchers to rapidly optimize OPV manufacturing parameters. We found the  $J-V$  prediction models are useful not only for predicting efficiencies, but also for understanding device behavior based on the composition and fabrication parameters of untested devices. The RF- $J-V$  prediction models accurately predicted  $J-V$  curves with 0.969 of exceptional  $R^2$  score. The PCE calculated from these curves were also as precise as PCE prediction models, both showing approximately 0.96 of  $R^2$  score. This high model performance is attributed to the systematically collected high-quality and large-quantity data, as well as finding suitable algorithms followed by thorough hyperparameters optimization. In addition, OPV device using the RF model-predicted formulation recorded 11.8% PCE, best efficiency for fully printed OPVs to date. Further data accumulation will enable the development of more generalized ML models that leverage accumulated and compatible datasets to predict device behavior for new materials and fabrication conditions. By demonstrating that models trained on binary systems can effectively predict ternary performance, this work highlights how data reuse accelerates optimization and drives OPV development toward commercialization. This study paves the way to a digital revolution in printed photovoltaic research and, potentially, solution-processed electronics.

## Experimental section

### Material preparation

PM6 (PBDB-T-2F), L8-BO (L8-BO-2F), and BM-HTL-1 were procured from Brilliant Materials. Qx-1 was synthesized following the procedures introduced in reference<sup>34</sup> and also purchased from Hyper Chem. D18 (PCE18) was obtained from 1-Material. S315 (Orgacon S315) was purchased from Agfa-Gevaert N.V. PEDOT:F was synthesized following the procedures outlined in reference<sup>35</sup> A commercial solution of ZnO NPs (5.6% w/v in isopropanol (IPA)) was sourced from Infinity PV. PEIE (polyethylenimine, 80% ethoxylated solution, 37 wt. % in H<sub>2</sub>O), CB (chlorobenzene, anhydrous, 99.8%), ortho-xylene (o-xylene, anhydrous, 97.0%), 2-methoxyethanol (anhydrous, 99.0%), and IPA (anhydrous, 99.0%) were acquired from Sigma-Aldrich. The TCE (OPV8, 8  $\Omega$  sq<sup>-1</sup>) was obtained pre-patterned on PET from Mekoprint A/S with a proprietary deposited metal composition. Ag paste (PV416) was purchased from DUPONT. A 3M<sup>TM</sup> ultra barrier solar film 512 was purchased from 3M.

### R2R device fabrication and optimization





R2R devices were fabricated on pre-patterned TCE-coated PET substrates. A customized R2R coater equipped with programmable syringe pumps facilitated the experiment. A solution of ZnO NPs was prepared by diluting it with IPA in a 1:1 ratio (v/v), resulting in a 2.8% solution (w/v). A solution of PEIE (0.4 wt% dissolved in 2-methoxyethanol) and the ZnO NP solution were concurrently deposited using two distinct slot die heads (head 1: PEIE, head 2: ZnO NP). The substrate passed through head 1 followed by head 2, enabling the deposition of the ZnO NPs layer over the PEIE layer. Both solutions were coated at a speed of 20 cm min<sup>-1</sup> with varying solution feed rates (PEIE: 15 µl min<sup>-1</sup> and ZnO NP: 20 µl min<sup>-1</sup>), resulting in wet film thicknesses (WFT) of 5.76 µm for PEIE and 7.69 µm for ZnO NP based on a coating area of 26 cm<sup>2</sup> min<sup>-1</sup> (1.3 cm width × 20 cm length). The layers were then thermally annealed at different temperatures (PEIE: 130 °C and ZnO NP: 120 °C) on curved hot plates located beneath each slot die head for ~45 seconds immediately after deposition. Subsequent layers were fabricated at the same coating speed (30 cm min<sup>-1</sup>). Two different active solutions, PM6:Qx-1 and PM6:L8-BO (D:A = 1:1.5, w/w), were dissolved in Xyl (25 mg/ml total concentration) and stirred at 80 °C overnight. During deposition, the solution feed rates were fixed at 60 µl min<sup>-1</sup> (WFT: 1.54 µm based on a coating area of 39 cm<sup>2</sup> min<sup>-1</sup>, 1.3 cm width × 30 cm length). Optimization of the active layer involved two technologies. Firstly, films were fabricated using a hot deposition technique<sup>36,37</sup> at various deposition temperatures (head/bed: RT/RT, 60/60, 60/100, 90/60 and 90/100 °C). Heat was applied directly to the slot-die head using a custom-built heater probe into a metal slot die, while a second heat source in a coating stage was applied to the substrate during slot-die solution deposition. Films were then thermally annealed at 130 °C on the curved hot plate for ~30 seconds. Next, drying kinetics were controlled by either thermally annealing the film at 130 °C on a coating stage, blowing air onto the wet film during deposition, or allowing the film to naturally dry without treatment for comparison. To fabricate a thin film HTL, a PEDOT:F solution was diluted with IPA in a 1:1 ratio (v/v). The PEDOT:F and S315 solutions were deposited from head 1 and 2, respectively, to create the PEDOT:F layer underneath the S315 layer. Solution feed rates were 30 and 168 µl min<sup>-1</sup> for the PEDOT:F and S315, respectively, corresponding to WFTs of 10 µm and 8 µm based on coating areas of 30 cm<sup>2</sup> min<sup>-1</sup> (1 cm width × 30 cm length) and 21 cm<sup>2</sup> min<sup>-1</sup> (0.7 cm width × 30 cm length) at the same coating speed. The layers were then thermally annealed at different temperatures (PEDOT:F: 100 °C and S315: 130 °C) on curved hot plates for ~30 seconds. All R2R device fabrication was conducted under ambient laboratory conditions, where the relative humidity and temperature were maintained at 35.8 ± 3.68 % and 22.4 ± 0.58 °C (mean ± standard deviation), respectively.

### High-throughput fabrication of R2R device

For high-throughput device fabrication, we selected the PM6:L8-BO system as an example. All active layers were formulated using CB solvents due to the limited solubility of L8-BO in the acceptor-rich solution when using Xyl solvents. A pre-mixed PM6:L8-BO solution (1:1.2, w/w, PM6 concentration: 10 mg ml<sup>-1</sup>) was first prepared, along with an identical solution containing various additives: DIO, CN, 2-HM and DCB. The additive concentrations were set to 2 and 3 vol% for DIO and CN, and weight ratios relative to the donor solid of 150 wt% for 2-HM and the total donor and acceptor solids of 150 wt% for DCB, respectively. Relative pump flow rates were adjusted to control the solid content of additives and to determine the optimal additive condition. Based on these evaluations, the best-performing solution (without additives) was selected for large-scale fabrication. Using this optimized formulation, a total of 26,422 OPV devices were produced across two independent batches, each yielding more than 13,000 devices. The device fabrication process mirrored those introduced in the 'R2R Device Fabrication and Optimization' section, with variations in the deposition of the active layer and HTL. These variations are detailed below for each batch.

Batch 1: Active layer solutions of PM6:L8-BO were prepared in CB at 80 °C overnight, with two distinct D:A ratios of 3:1 and 1:3 (denoted as donor- and acceptor-rich solutions), maintaining a fixed donor



concentration of 10 mg ml<sup>-1</sup>. These solutions were deposited using two programmable syringe pumps combined in a Y-connector, and the *in-situ* mixed solution fed through tubing with an inner diameter of 0.5 mm. 13 sub-batches were conducted throughout the batch. Total solution flows remained constant at 72, 48, 36, or 24  $\mu\text{l min}^{-1}$  (resulting in WFT: 1.85, 1.23, 0.923, or 0.616  $\mu\text{m}$ ) while adjusting the relative flow rates of the two pumps for first 4 sub-batches. This manipulation varied the D:A ratios of the total solution flow, leading to linear increases in TDD during deposition. In contrast, the volumetric ratios of the donor-and acceptor-rich solutions were held constant to achieve fixed D:A ratios of 1:3, 1:2, 1:1.5, 1:1.2, 1:1, 1.2:1, 1.5:1, 2:1, or 3:1 in the total solution flow. The total solution flows varied within the range of 24–96  $\mu\text{l min}^{-1}$  (resulting in WFT: 0.616–2.46  $\mu\text{m}$ ), leading to linear decreases in TDD during deposition for the remaining 9 sub-batches.

Batch 2: Active layer solutions of PM6:L8-BO prepared at 80 °C overnight consisted of a donor-rich solution (D:A ratio of 3:1 and a donor concentration of 15 mg ml<sup>-1</sup>) and an acceptor-rich solution (D:A ratio of 1:3 and a donor concentration of 5 mg ml<sup>-1</sup>) were prepared. The deposition process mirrored that of Batch 1. Total solution flows were maintained at a constant rate of 120, 90, 60, or 30  $\mu\text{l min}^{-1}$  (resulting in WFT: 3.08, 2.31, 1.54, or 0.769  $\mu\text{m}$ ) while altering the relative flow rates of the two pumps for the first 4 sub-batches. The D:A ratios of the total solution flow were varied while TDD remained constant, corresponding to 63.5, 47.7, 31.8, or 15.9  $\mu\text{g cm}^{-2}$ , respectively. The volumetric ratios of the donor-and acceptor-rich solution were held constant to achieve fixed D:A ratios of 1:3, 1:2, 1:1.5, 1:1.2, 1:1, 1.2:1, 1.5:1, 2:1, or 3:1 in the total solution flow. The total solution flows varied within the range of 24–120  $\mu\text{l min}^{-1}$  (resulting in WFT: 0.616–3.08  $\mu\text{m}$ ), leading to linear decreases in TDD during deposition for the remaining 9 sub-batches.

All active layers were fabricated at RT for both head and bed temperatures, and the films were exposed to air blowing process during deposition. The dead volume of the tubing from the Y-connector and the slot die was calculated to be ~33.7 and ~39.6  $\mu\text{l}$  for Batch 1 and Batch 2, respectively. This implies that the *in situ* mixed solution was deposited around 28 and 20 seconds after mixing at a flow rate of 72 and 120  $\mu\text{l min}^{-1}$ . The composition of the film at each position was calculated based on the relative flow rate and the deposition offset originating from the dead volume of the deposition system. BM-HTL-1, diluted with IPA in a 1:1 ratio (v/v), was used as the HTL material, and concurrently deposited with the S315 layer in the same manner as depicted in the previous section.

### Device characterization

All devices had a defined cell area of 2.5 mm<sup>2</sup>, determined by the TCE pattern and a commercially available circular aperture (Thorlab, 1.8 mm diameter). *J-V* characteristics of all OPV devices were characterized using an AAA solar simulator (Enlitech, SS-F5-3A), which employed a Xenon lamp light source (USHIO, UXL-300D-0) coupled with an AM 1.5G solar spectrum filter. This setup was integrated with a R2R tester connected to a source measurement unit (a Keithley 2400-LV) under standard test conditions (1 sun, 100 mW cm<sup>-2</sup>, AM 1.5G) and ambient surroundings. The light intensity was calibrated and monitored using an Enlitech certified secondary reference cell (Si-FigRC-5021), which had been pre-calibrated by a certified KG-2 filtered Si reference cell. Ag paste was screen printed on top of the S315 layer using a SP-006 screen printer (ESSEMTEC AG) to establish electrical contact with a test clip during EQE measurement. Subsequently, the devices were encapsulated with a barrier film. EQE spectra were acquired in air using a PEC-S20 instrument (Pecel Technologies Inc.).

### Film characterization

To analyze the properties of the thin films, several techniques were employed. Work function measurements were performed using photoelectron spectroscopy (PESA) in air. UV-vis absorption



spectra of thin films were obtained by using a SPM-C1 spectrometer (StellarNet Inc.) with a Xenon lamp serving as the light source. AFM height and phase images (5  $\mu\text{m} \times 5 \mu\text{m}$ ) were captured using a Bruker Dimension Icon (Bruker Corporation, USA). Scans were recorded using tapping mode and Bruker ScanAsyst-Air probes (cantilever frequency 70 kHz, spring constant 0.4 N m<sup>-1</sup>, nominal tip radius 2 nm) under ambient conditions. A scan rate of 0.25 Hz was maintained for all scans. Nanoscope Analysis 1.9 (Bruker Corporation, USA) was used for image processing and production. All samples were R2R slot die coated onto PET substrates.

### ML predicted formulation-based device fabrication

The fabrication process for all devices followed the steps outlined in the ‘R2R device fabrication and optimization’ section, excluding the deposition of active layer. PM6:L8-BO active layer solutions in CB and Xyl were prepared at 80 °C overnight based on the ML-predicted formulations generated by the best PCE and  $J-V$  prediction models. These formulations were divided into two groups with different donor concentrations: Cluster 0 (10.2 mg ml<sup>-1</sup>, D:A = 1:1.14, w/w) and Cluster 1 (9.29 mg ml<sup>-1</sup>, D:A = 1:1.22, w/w). The active layers were deposited using a single syringe pump at solution flow rates of 72.5 and 117.5  $\mu\text{l min}^{-1}$ , corresponding to WFT of 1.86 and 3.01  $\mu\text{m}$  based on 39 cm<sup>2</sup> min<sup>-1</sup> coating are. This resulted in TDD of 40.5 and 62  $\mu\text{g cm}^{-2}$  for Cluster 0 and Cluster 1, respectively. All films were fabricated at RT for both head and bed deposition temperatures and dried under airflow conditions, consistent with the process used for the high-throughput produced devices.

### Similarity measurement metrics

Euclidean Distance (ED) metric was utilized for quantifying similarities between two data points in a multidimensional space using following formula:

$$\text{Euclidean Distance (ED)} = \sqrt{\sum_{k=1}^M \left( \frac{F_{k,x} - F_{k,y}}{F_{k,max}} \right)^2} \quad (\text{Eq. S1})$$

$M$  represents the total number of features.  $F_{k,x}$  and  $F_{k,y}$  denote the  $k^{\text{th}}$  features of two selected devices, such as TDDs of device 1 and device 2.  $F_{k,max}$  denotes maximum value of the  $k^{\text{th}}$  features.

### Ternary or all polymer system-based device fabrication and data preparation

**PM6:Y6:IT-4F System:** The data from previously fabricated and characterized devices were utilized, as detailed in a prior publication<sup>23</sup>. The final dataset included 16 different batches, totaling 2422 data, with 10 features and 1 label for PCE prediction. Notably, the introduction of Y6, IT-4F and *o*-dichlorobenzene (DCB) is represented in the dataset as columns as ‘DCB FR’, ‘Y6 DD’ and ‘IT-4F DD’.

**PM6:D18:L8-BO System:** The device fabrication followed procedures outlined in the prior section, with the exception of the active layer deposition. The experiment consisted of two independent batches, each with 5 sub-batches. In Batch 1, donor-only solution (PM6:D18 = 4:1, total concentration: 20 mg ml<sup>-1</sup>) and acceptor-only solution (L8-BO concentration: 20 mg ml<sup>-1</sup>) were prepared. In Batch 2, pre-mixed solutions of PM6:D18:L8-BO (0.5:0.5:1.2, PM6 or D18 concentration: 5 mg ml<sup>-1</sup>) and PM6:L8-BO (1:1.2, PM6 concentration: 10 mg ml<sup>-1</sup>) dissolved in CB at 80°C overnight. During deposition, total solution flows were maintained at constant rates of 120, 96, 72, 48, or 24  $\mu\text{l min}^{-1}$  (resulting in WFT: 3.08, 2.46, 1.85, 1.23 or 0.615  $\mu\text{m}$ ), with alterations in the relative flow rates of the two pumps. The



D:A ratios (Batch 1) or PM6:D18 ratios (Batch 2) of the total solution flow were varied, while TDD were either changed (Batch 1) or kept constant (Batch 2), resulting in TDD of 68.7, 54.9, 41.2, 27.5, or 13.7  $\mu\text{g cm}^{-2}$ , respectively. The deposition occurred under identical environment and the composition of the film at each position were calculated using the same methods as in the prior experiments. Dataset preparation involved consistent feature design and data cleaning procedures as in the previous section. The final dataset comprises 2570 data, with 10 features and 1 label for PCE prediction, incorporating D18 as the third material, represented by the column 'D18 DD' in the dataset. Any inconsistent data were removed prior to analysis.

**PM6:PYF-T-o System:** Device fabrication followed the procedures described previously, except for the active layer deposition. Three independent batches were prepared, each consisting of 5, 6 sub-batches and 1 sub-batch. In Batch 1, donor-only (PM6 concentration, 14  $\text{mg ml}^{-1}$ ) and acceptor-only (PYF-T-o concentration, 14  $\text{mg ml}^{-1}$ ) solutions were prepared. In Batch 2, pre-mixed solutions of PM6:PYF-T-o at ratios of 2.5:1 and 1:2.5 (PM6 concentration: 7  $\text{mg ml}^{-1}$ ) were prepared. In Batch 3, a pre-mixed PM6:PYF-T-o solution (1:1.2, w/w) (PM6 concentration: 7  $\text{mg ml}^{-1}$ ) was prepared along with an identical solution containing PN as a solid additive (100 wt% relative to the total donor and acceptor solids). All solutions were dissolved in CB and stirred overnight at 70 °C. During deposition, the total solution flow rates were set to 120, 96, 72, 48, or 24  $\mu\text{l min}^{-1}$  in Batches 1 and 2, resulting in WFTs of 3.08, 2.46, 1.85, 1.23, and 0.615  $\mu\text{m}$ , respectively, while Batch 3 was fixed at 72  $\mu\text{l min}^{-1}$ . Relative pump flow rates were varied to control D:A ratios, film thickness, or the solid content of 2-PN. The TDD was kept constant in Batch 1 (43.1, 34.5, 25.8, 17.2, and 8.62  $\mu\text{g cm}^{-2}$ ) but varied in Batches 2 and 3. Deposition conditions were consistent with earlier experiments, and film composition at each position was determined using the same calculation methods. Dataset preparation followed the same feature design and data cleaning protocols described previously. The final dataset comprised 2,840 data points, including 10 features and 1 target label for PCE prediction. Two additional feature columns, 'PYF-T-o DD' and 'PN DD', were introduced to represent PYF-T-o as the acceptor material and 2-PN as the additive, respectively. Any inconsistent data were removed before analysis.

### Author contributions

N.G.A. and D.V. conceived the idea and designed the experiments. N.G.A. fabricated all OPV devices and performed photovoltaic characterizations. D.V. developed automatic roll-to-roll fabrication/testing machines and the software to run the machines and to collect data. N.G.A. designed all illustrations in this paper. N.G.A. wrote machine learning codes with assistance of D.V.. Y. L. synthesized PEDOT:F, and C.-Q. M. and Z.W. provided Qx-1 material. T.R. conducted AFM measurement. L. N. W. T., S.Song., M.G., Y. Z., J.Y.K., U.B. and D.V. participated in the discussion of the experiment results. N.G.A. wrote the manuscript and N.G.A., J.Y.K., U.B., and D.V. revised the manuscript.

### Data availability

All data generated or analyzed during this study are included in the published article and its ESI, supplementary data files and supplementary video. Detailed descriptions of the supplementary data files and video are as follows:

Supplementary data files

Data S1. Self-produced dataset for PP prediction composed of 25,249 data.

Data S2. Self-produced dataset for  $J-V$  prediction composed of 1,792,697 data.



Data S3. Dataset for PCE prediction in ternary and all polymer material systems. It includes 16 batches totaling 2,422 data points for the PM6:Y6:IT-4F system, and 10 sub-batches totaling 2,570 data points for the PM6:D18:L8-BO system. Additionally, the all-polymer dataset comprises 11 sub-batches with 2,840 data points for the PM6:PYF-T-o system.

Data S4. Model performance summary of ternary and all polymer material system without or with the accumulated dataset.

### Supplementary video

Video S1. The  $J-V$  curves predicted by the  $J-V$  prediction model were displayed for selected features. Predicted PPs from the PP prediction model were also listed in the top-left legend of the chart. As each feature slider is adjusted, the predicted  $J-V$  curves and PPs update interactively according to the selected feature values.

### Code availability

In this study all ensemble-based models were built using scikit-learn and xgboost, which are open-source libraries. All code, relevant datasets and supplementary files are shared at: <https://figshare.com/s/ca0b8a65d1acdadb9f28>.

### Conflicts of interest

The authors declare no competing interests.

### Acknowledgements

This work is supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (2021R1A6A3A14039635), the Australian Centre for Advanced Photovoltaics (ACAP) program funded by the Australian Government through Australian Renewable Energy Agency (ARENA), CSIRO's Labs of the Future fund, CAS-CSIRO international collaboration fund (121E32KYSB20190021), and the Australian Research Council (ARC) Centre of Excellence in Exciton Science (ACEX: CE170100026), the NRF grant funded by the Korea government (MSIT) (RS-2024-00429694). We would like to express our gratitude to Tom Raeber for his assistance with AFM measurements.

### References

1. M. Kaltenbrunner, M. S. White, E. D. Glowacki, T. Sekitani, T. Someya, N. S. Sariciftci and S. Bauer, *Nat. Commun.*, 2012, **3**, 770.
2. L. Hong, H. Yao, Z. Wu, Y. Cui, T. Zhang, Y. Xu, R. Yu, Q. Liao, B. Gao, K. Xian, H. Y. Woo, Z. Ge and J. Hou, *Adv. Mater.*, 2019, **31**, 1903441.
3. L. Xu, Y. Xiong, S. Li, W. Zhao, J. Zhang, C. Miao, Y. Zhang, T. Zhang, J. Wu, S. Zhang, Q. Peng, Z. Wang, L. Ye, J. Hou and J. Wang, *Adv. Funct. Mater.*, 2024, **34**, 2314178.
4. J. Kong, M. Mohadjer Beromi, M. Mariano, T. Goh, F. Antonio, N. Hazari and A. D. Taylor, *Nano Energy*, 2017, **38**, 36–42.
5. Y. Cui, C. Yang, H. Yao, J. Zhu, Y. Wang, G. Jia, F. Gao and J. Hou, *Adv. Mater.*, 2017, **29**, 1703080.





6. N. G. An, T. Lee, J. Heo, J. W. Kim, S. Song, W. Lee, B. Walker, E. Lim and J. Y. Kim, *Sol. RRL*, 2021, **5**, 2000742. View Article Online  
DOI: 10.1039/D3EE02815A
7. Y. Sun, L. Wang, C. Guo, J. Xiao, C. Liu, C. Chen, W. Xia, Z. Gan, J. Cheng, J. Zhou, Z. Chen, J. Zhou, D. Liu, T. Wang and W. Li, *J. Am. Chem. Soc.*, 2024, **146**, 12011–12019.
8. Y. Jiang, S. Sun, R. Xu, F. Liu, X. Miao, G. Ran, K. Liu, Y. Yi, W. Zhang and X. Zhu, *Nat. Energy*, 2024, **9**, 975–986.
9. C. Chen, L. Wang, W. Xia, K. Qiu, C. Guo, Z. Gan, J. Zhou, Y. Sun, D. Liu, W. Li and T. Wang, *Nat. Commun.*, 2024, **15**, 6865.
10. L. Zhu, M. Zhang, G. Zhou, Z. Wang, W. Zhong, J. Zhuang, Z. Zhou, X. Gao, L. Kan, B. Hao, F. Han, R. Zeng, X. Xue, S. Xu, H. Jing, B. Xiao, H. Zhu, Y. Zhang and F. Liu, *Joule*, DOI: 10.1016/j.joule.2024.08.001.
11. N. Wei, J. Chen, Y. Cheng, Z. Bian, W. Liu, H. Song, Y. Guo, W. Zhang, Y. Liu, H. Lu, J. Zhou and Z. Bo, *Adv. Mater.*, 2024, **36**, 2408934.
12. X. Rodríguez-Martínez, E. Pascual-San-José, Z. Fei, M. Heeney, R. Guimerà and M. Campoy-Quiles, *Energy Environ. Sci.*, 2021, **14**, 986–994.
13. N. Meftahi, M. A. Surmiak, S. O. Fürer, K. J. Rietwyk, J. Lu, S. R. Raga, C. Evans, M. Michalska, H. Deng, D. P. McMeekin, T. Alan, D. Vak, A. S. R. Chesman, A. J. Christofferson, D. A. Winkler, U. Bach and S. P. Russo, *Adv. Energy Mater.*, 2023, **13**, 2203859.
14. X. Du, L. Lüer, T. Heumueller, J. Wagner, C. Berger, T. Osterrieder, J. Wortmann, S. Langner, U. Vongsaysy, M. Bertrand, N. Li, T. Stubhan, J. Hauch and C. J. Brabec, *Joule*, 2021, **5**, 495–506.
15. C. Liu, L. Lüer, V. M. L. Corre, K. Forberich, P. Weitz, T. Heumueller, X. Du, J. Wortmann, J. Zhang, J. Wagner, L. Ying, J. Hauch, N. Li and C. J. Brabec, *Adv. Mater.*, 2024, **36**, 2300259.
16. J. Zhang, B. Liu, Z. Liu, J. Wu, S. Arnold, H. Shi, T. Osterrieder, J. A. Hauch, Z. Wu, J. Luo, J. Wagner, C. G. Berger, T. Stubhan, F. Schmitt, K. Zhang, M. Sytnyk, T. Heumueller, C. M. Sutter-Fella, I. M. Peters, Y. Zhao and C. J. Brabec, *Adv. Energy Mater.*, 2023, **13**, 2302594.
17. J. Liu, J. Zhang, G. Rao, Z. Peng, Y. Huang, S. Arnold, B. Liu, C. Deng, C. Li, H. Li, H. Zhi, Z. Zhang, W. Zhou, J. Hauch, C. Yan, C. J. Brabec and Y. Zhao, *ACS Energy Lett.*, 2024, **9**, ACS Energy Letters.
18. N. Majeed, M. Saladina, M. Krompiec, S. Greedy, C. Deibel and R. C. I. MacKenzie, *Adv. Funct. Mater.*, 2020, **30**, 1907259.
19. R. Søndergaard, M. Hösel, D. Angmo, T. T. Larsen-Olsen and F. C. Krebs, *Mater. Today*, 2012, **15**, 36–49.
20. F. C. Krebs, N. Espinosa, M. Hösel, R. R. Søndergaard and M. Jørgensen, *Adv. Mater.*, 2014, **26**, 29–39.
21. B. Roth, R. R. Søndergaard and F. C. Krebs, in *Handbook of Flexible Organic Electronics*, ed. S. Logothetidis, Woodhead Publishing, Oxford, 2015, DOI: <https://doi.org/10.1016/B978-1-78242-035-4.00007-5>, pp. 171–197.
22. M. Wagner, A. Distler, V. M. Le Corre, S. Zapf, B. Baydar, H.-D. Schmidt, M. Heyder, K. Forberich, L. Lüer, C. J. Brabec and H. J. Egelhaaf, *Energy Environ. Sci.*, 2023, **16**, 5454–5463.
23. N. G. An, J. Y. Kim and D. Vak, *Energy Environ. Sci.*, 2021, **14**, 3438–3446.
24. L. W. T. Ng, N. G. An, L. Yang, Y. Zhou, D. W. Chang, J.-E. Kim, L. J. Sutherland, T. Hasan, M. Gao and D. Vak, *Cell Rep. Phys. Sci.*, 2024, **5**.
25. H. C. Weerasinghe, N. Macadam, J.-E. Kim, L. J. Sutherland, D. Angmo, L. W. T. Ng, A. D. Scully, F. Glenn, R. Chantler, N. L. Chang, M. Dehghanimadvar, L. Shi, A. W. Y. Ho-Baillie, R. Egan, A. S. R. Chesman, M. Gao, J. J. Jasieniak, T. Hasan and D. Vak, *Nat. Commun.*, 2024, **15**, 1656.
26. D. Vak, K. Hwang, A. Faulks, Y.-S. Jung, N. Clark, D.-Y. Kim, G. J. Wilson and S. E. Watkins, *Adv. Energy Mater.*, 2015, **5**, 1401539.
27. C. Li, J. Zhou, J. Song, J. Xu, H. Zhang, X. Zhang, J. Guo, L. Zhu, D. Wei, G. Han, J. Min, Y. Zhang, Z. Xie, Y. Yi, H. Yan, F. Gao, F. Liu and Y. Sun, *Nat. Energy*, 2021, **6**, 605–613.
28. R. Polikar, in *Ensemble Machine Learning: Methods and Applications*, eds. C. Zhang and Y. Ma, Springer New York, New York, NY, 2012, DOI: 10.1007/978-1-4419-9326-7\_1, pp. 1–34.
29. J. W. Cort and M. Kenji, *Clim. Res.*, 2005, **30**, 79–82.

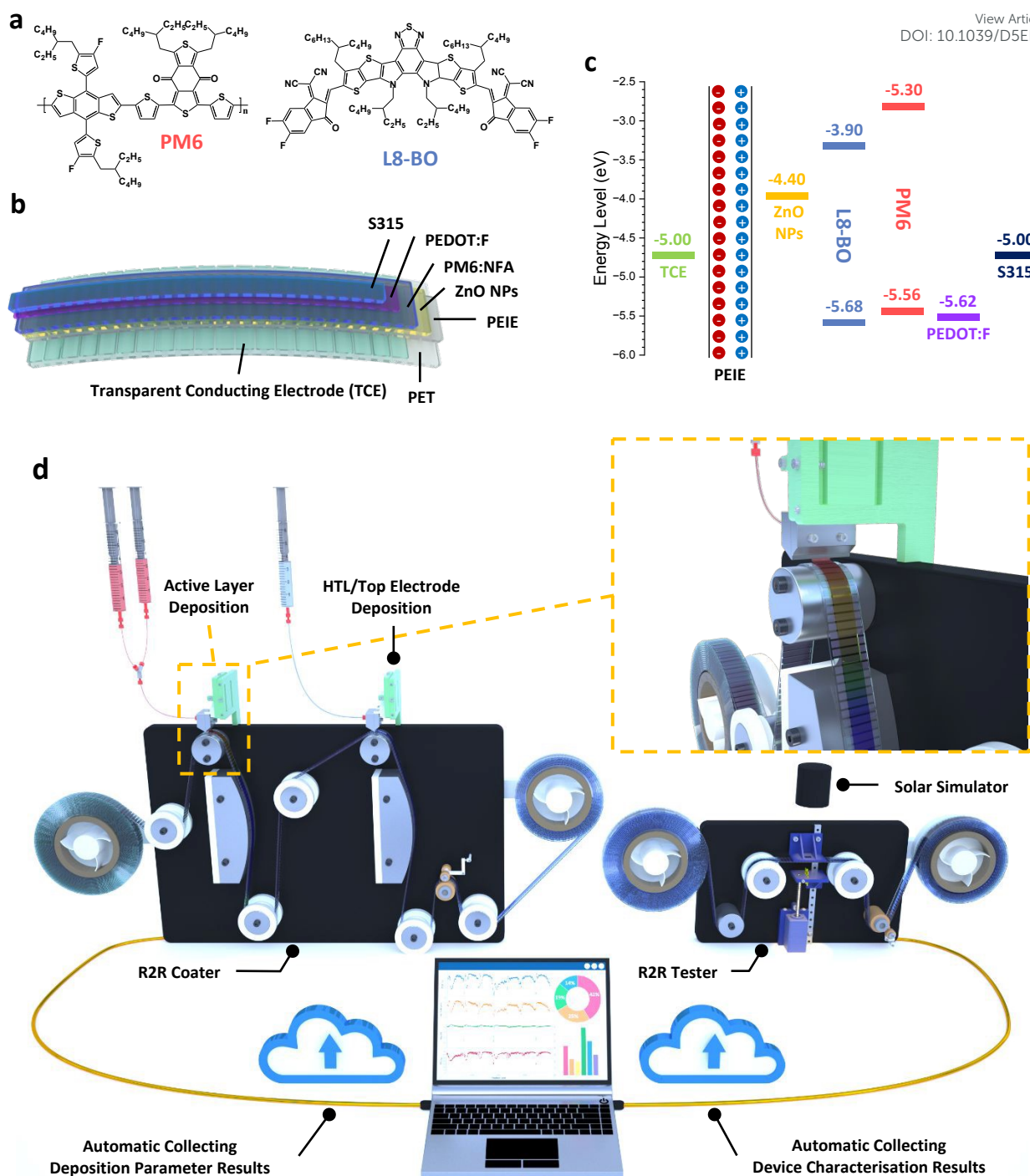




30. A. Botchkarev, *Interdiscip. J. Inf. Knowl. Manag.*, 2019, **14**, 45–79.
31. A. Jierula, S. Wang, T.-M. OH and P. Wang, *Appl. Sci.*, 2021, **11**, 2314.
32. R. J. Hyndman and A. B. Koehler, *Int. J. Forecast.*, 2006, **22**, 679–688.
33. M. J. W. Davide Chicco, Giuseppe Jurman, *PeerJ Comput. Sci.*, 2021, **7**:e623.
34. Y. Shi, Y. Chang, K. Lu, Z. Chen, J. Zhang, Y. Yan, D. Qiu, Y. Liu, M. A. Adil, W. Ma, X. Hao, L. Zhu and Z. Wei, *Nat. Commun.*, 2022, **13**, 3256.
35. Y. Jiang, X. Dong, L. Sun, T. Liu, F. Qin, C. Xie, P. Jiang, L. Hu, X. Lu, X. Zhou, W. Meng, N. Li, C. J. Brabec and Y. Zhou, *Nat. Energy*, 2022, **7**, 352–359.
36. S. Song, K. T. Lee, C. W. Koh, H. Shin, M. Gao, H. Y. Woo, D. Vak and J. Y. Kim, *Energy Environ. Sci.*, 2018, **11**, 3248–3255.
37. S.-I. Na, Y.-H. Seo, Y.-C. Nah, S.-S. Kim, H. Heo, J.-E. Kim, N. Rolston, R. H. Dauskardt, M. Gao, Y. Lee and D. Vak, *Adv. Funct. Mater.*, 2019, **29**, 1805825.

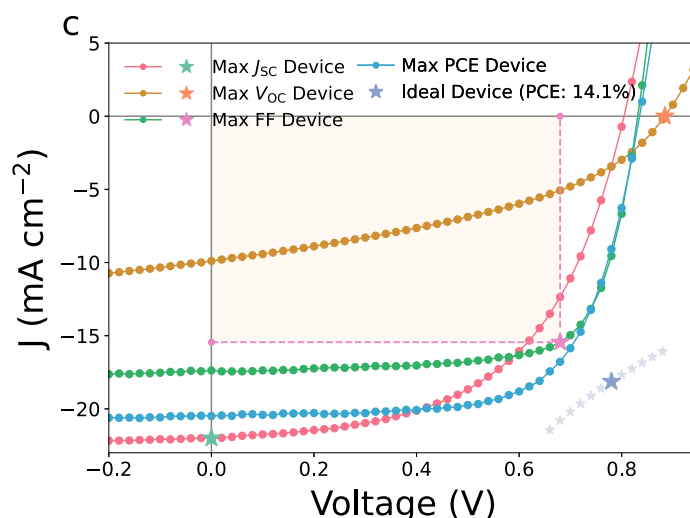
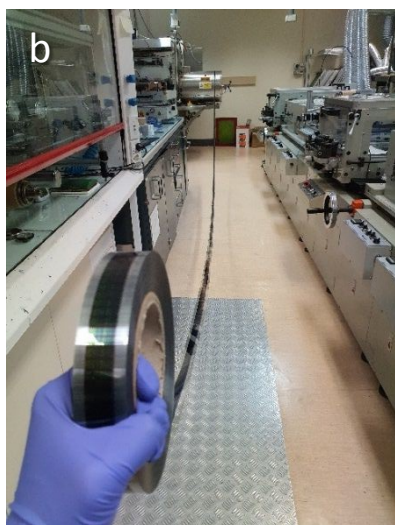
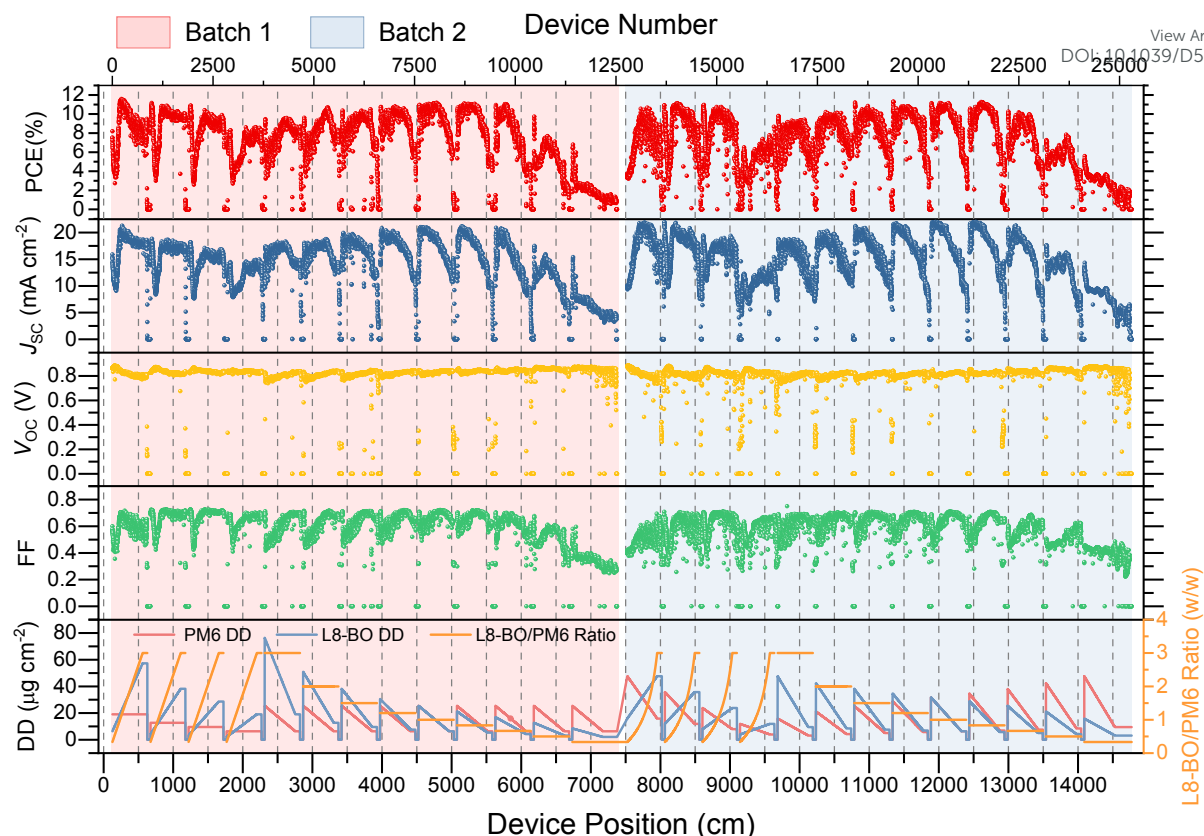
View Article Online  
DOI: 10.1039/D5EE02815A





**Fig. 1.** Materials, energy diagrams and structure of R2R OPV device and MicroFactory platform illustration (a) Chemical structures of materials used in this study. (b) Detailed device structure illustration. (c) Energy diagram of fully-R2R fabricated OPVs. (d) A 3D schematic of the MicroFactory platform featuring custom-built automatic R2R coater and tester, along with a database that stores all collected data. The inset image in Fig. 1d illustrates the fabrication of an active layer with a gradient composition, represented by a rainbow-colored film, in a single deposition process. By employing programmable syringe pumps, two solution flows were controlled linearly, and the solutions were mixed *in situ* and deposited onto a continuously moving substrate. 9 deposition parameters were collected and stored in a database during deposition. Subsequently, the devices were automatically characterized using the R2R tester until a specified number of measurements had been made. The data collected from three coating runs and one test run were integrated into one consolidated dataset based on the position of the roll (device position).





**Fig. 2.** PP distribution for a total 26,422 devices, photograph of an actual device and  $J$ - $V$  characteristics of four selected devices. (a) PP data of individual devices, DD of each material, and L8-BO/PM6 ratios (w/w) across 26,422 PM6:L8-BO devices, organized based on their positions in the device. These devices were part of 13 sub-batches in each batch, wherein various D:A ratios or film thicknesses were explored. The devices with 0% PCE were intentionally fabricated as checkpoints at the end of each sub-batch. (b) Photograph of actual devices corresponding to  $\sim 74.8$  m for one batch. (c)  $J$ - $V$  characteristics of four selected devices, each showcasing the highest  $J_{sc}$ ,  $V_{oc}$ , FF, and PCE values among 26,422 devices. The positions of each value were indicated as star shapes on each  $J$ - $V$  curve, except for the maximum PCE value. The maximum PCE, calculated to be 14.1%, was based on these values. Possible MPPs of ideal devices with the PCE were depicted as purple star shapes for visualization using an equation,  $J_{MP} = -PCE/V_{MP}$  where  $J_{MP}$  is current density and  $V_{MP}$  is voltage of the MPPs within voltage range of  $0-V_{oc}$  V ( $0-0.884$  V in this case). The larger and darker star shape indicated the proximity of MPP to the  $J$ - $V$  characteristics of devices demonstrating the highest PCEs.

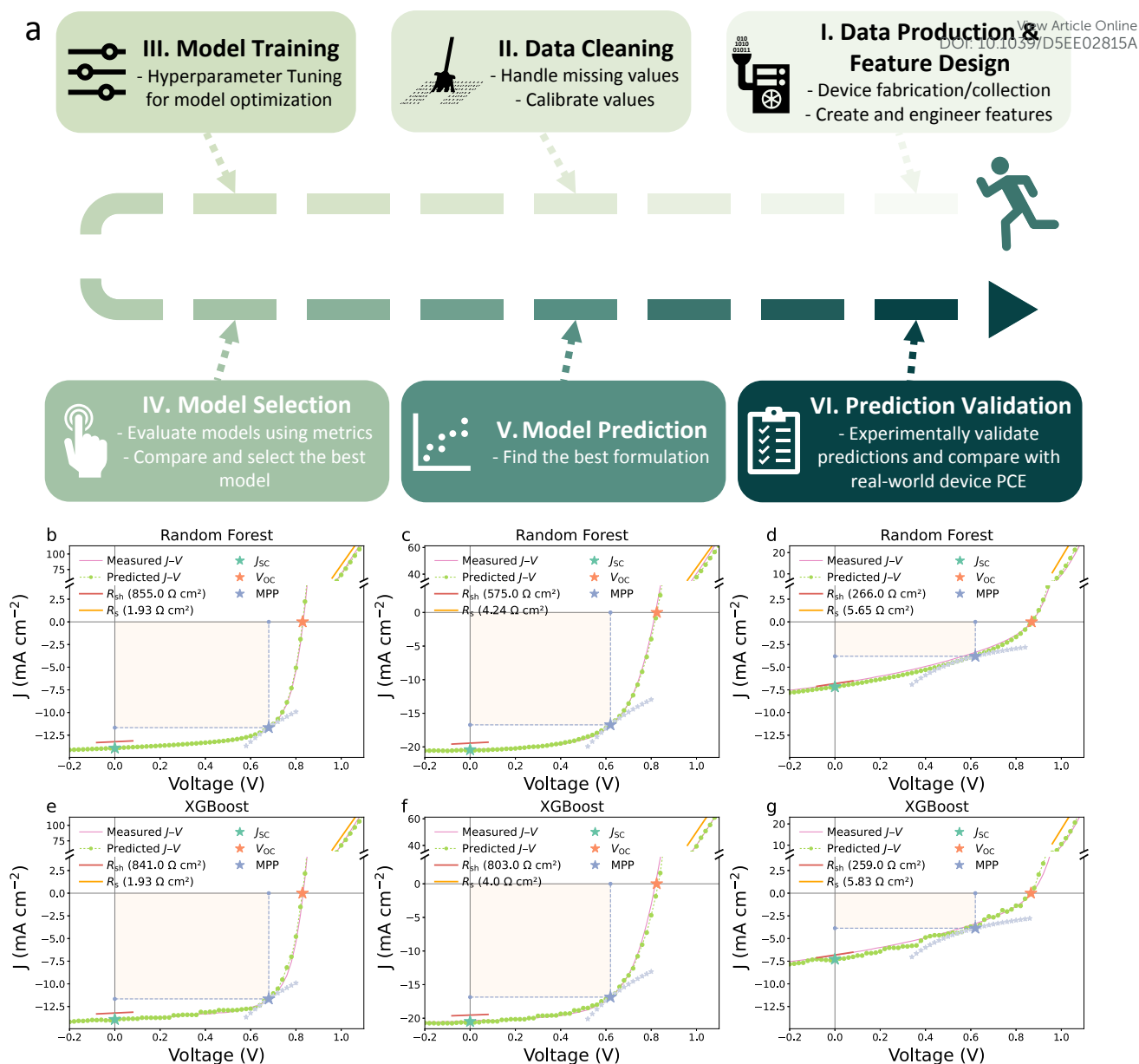


**Table 1.** Summary of PPs of four selected devices exhibiting the highest  $J_{SC}$ ,  $V_{OC}$ , FF and PCE, among 26,422 devices. View Article Online  
DOI: 10.1039/D5EE02815A

Device	$J_{SC}$ (mA cm <sup>-2</sup> )	$V_{OC}$ (V)	FF	PCE (%)	$R_{sh}$ (Ω cm <sup>2</sup> )	$R_s$ (Ω cm <sup>2</sup> )
Max $J_{SC}$	22.0	0.805	0.548	9.70	1130	3.01
Max $V_{OC}$	9.89	0.884	0.410	3.59	210	11.6
Max FF	17.4	0.831	0.727	10.5	996	2.00
Max PCE	20.5	0.835	0.677	11.6	5004	2.58

\* 200 devices were fabricated at each formulation.





**Fig. 3.** Workflow overview in this study,  $J$ - $V$  measured and predicted characteristics and predicted PPs of 3 randomly selected devices. (a) Illustration of the comprehensive workflow in this study, comprising automated device fabrication and characterization, data storage of deposition and photovoltaic data in a database, and ML application. (b)–(g)  $J$ - $V$  actual and predicted characteristics of 3 randomly selected devices. Predicted PPs, including  $J_{sc}$ ,  $V_{oc}$ ,  $R_{sh}$ , and  $R_s$ , were denoted by star-shaped markers and thick solid lines, derived from each best PP prediction model. Measured  $J$  and predicted  $J$  values were obtained from actual characterization and  $J$ - $V$  prediction model, respectively. The MPP candidates were identified using the same method described in Fig. 2b, and were represented by star-shaped markers, with larger ones indicating the MPP point closest to the actual  $J$ - $V$  curve (MPP<sub>close</sub>). Additionally, the predicted  $J_{sc}$ ,  $V_{oc}$ ,  $R_{sh}$  and  $R_s$  values were represented as star-shape markers or solid lines, derived from each best PP prediction model. The predicted FFs were depicted as ivory-colored filled rectangles, with  $J$  and  $V$  values determined from the MPP<sub>close</sub>. Predictions were generated using RF models for panels (b)–(d), and XGBoost models for panels (e)–(g).



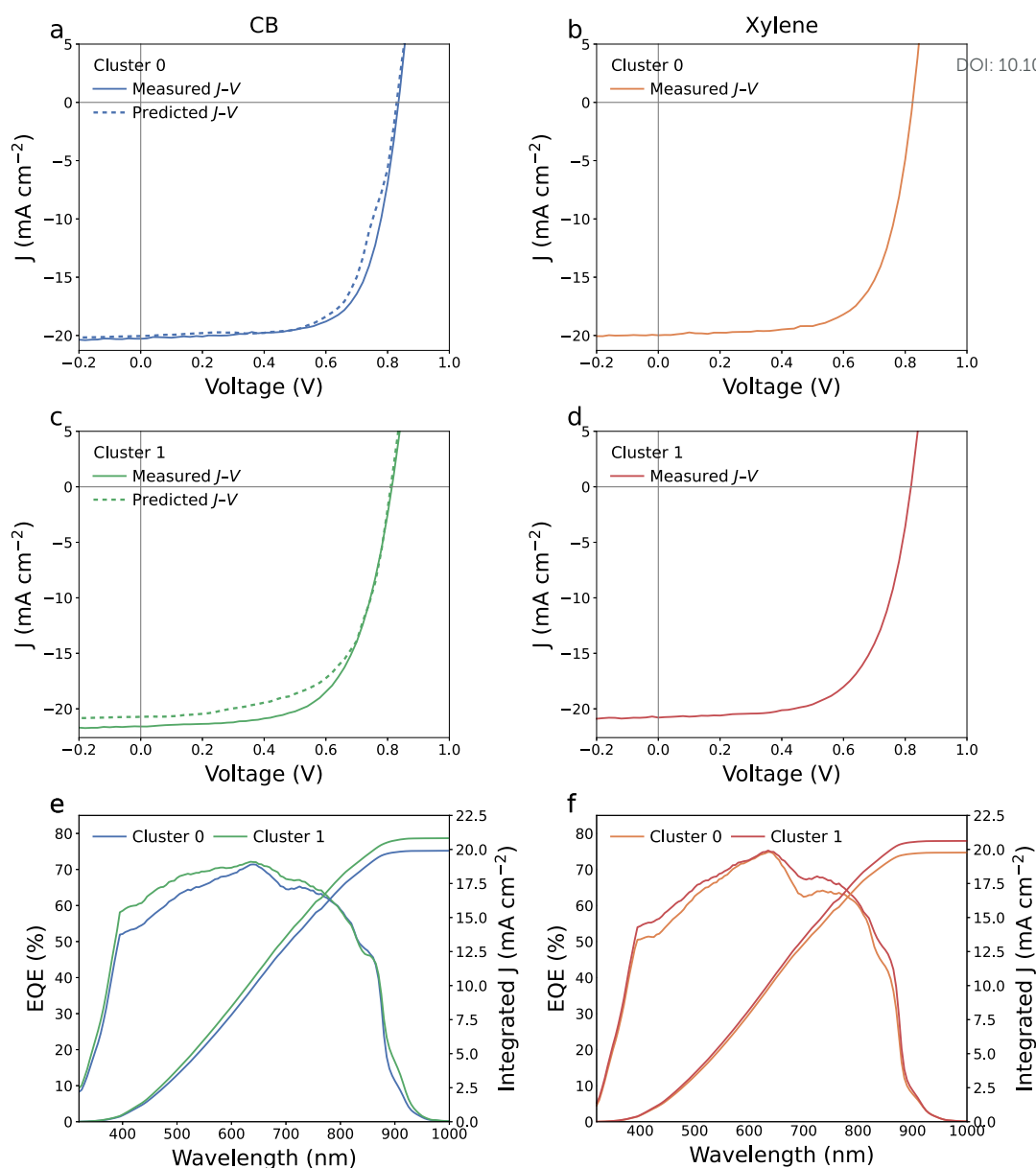
**Table 2.**  $R^2$  score, MAPE and RMSE of each PP ( $R^2_{JV}$ ,  $MAPE_{JV}$  and  $RMSE_{JV}$ ) calculated from predicted  $J-V$  curves.  $R^2$  score, MAPE and RMSE of PP prediction model ( $R^2_{PP}$ ,  $MAPE_{PP}$  and  $RMSE_{PP}$ ) were provided for comparison.

Model	Dataset	PP	$R^2_{JV}$	$MAPE_{JV}$	$RMSE_{JV}$	$R^2_{PP}$	$MAPE_{PP}$	$RMSE_{PP}$
RF	Test	$J_{sc}$	0.980	13.6	0.720	0.982	21.0	0.683
XGBoost			0.979	248	0.737	0.984	21.1	0.650
RF		$V_{oc}$	0.953	1.46	0.037	0.964	1.38	0.032
XGBoost			0.893	2.04	0.056	0.965	1.46	0.032
RF		FF	0.925	3.89	4.37	0.957	3.24	3.30
XGBoost			0.866	6.48	5.84	0.961	3.20	3.17
RF		PCE	0.949	6.82	0.681	0.958	6.83	0.613
XGBoost			0.932	9.28	0.784	0.962	6.77	0.589
RF		$R_{sh}$	-7.26	98.7	13364	-0.225 (0.839) <sup>a</sup>	75.5 (15.6) <sup>a</sup>	5146 (41.5) <sup>a</sup>
XGBoost			-9.11	528	14786	-2.65 (0.800) <sup>a</sup>	98.5 (16.4) <sup>a</sup>	8879 (46.3) <sup>a</sup>
RF		$R_s$	-0.001	10.4	17552	0.028 (0.886) <sup>b</sup>	3816 (6.80) <sup>b</sup>	17296 (0.773) <sup>b</sup>
XGBoost			-0.001	15.4	17552	0.028 (0.885) <sup>b</sup>	3751 (7.04) <sup>b</sup>	17296 (0.775) <sup>b</sup>

a) Metrics within the parentheses are derived from  $R_{sh}$  values belonging to bottom 30% ( $< 366 \Omega \text{ cm}^2$ )  
b) Metrics within the parentheses are derived after excluding outliers (about 3% of entire dataset).







**Fig. 4.** Comparison of  $J-V$  characteristics and EQE spectra for R2R OPVs with top formulations in varied solvent condition.  $J-V$  characteristics of R2R OPVs with the top formulation in cluster 0 and 1 using processing solvents (a) and (c), CB and (b) and (d), xylene, along with (e) and (f), corresponding EQE spectrum. Predicted  $J-V$  characteristics obtained from the  $J-V$  prediction model are also included for comparison.



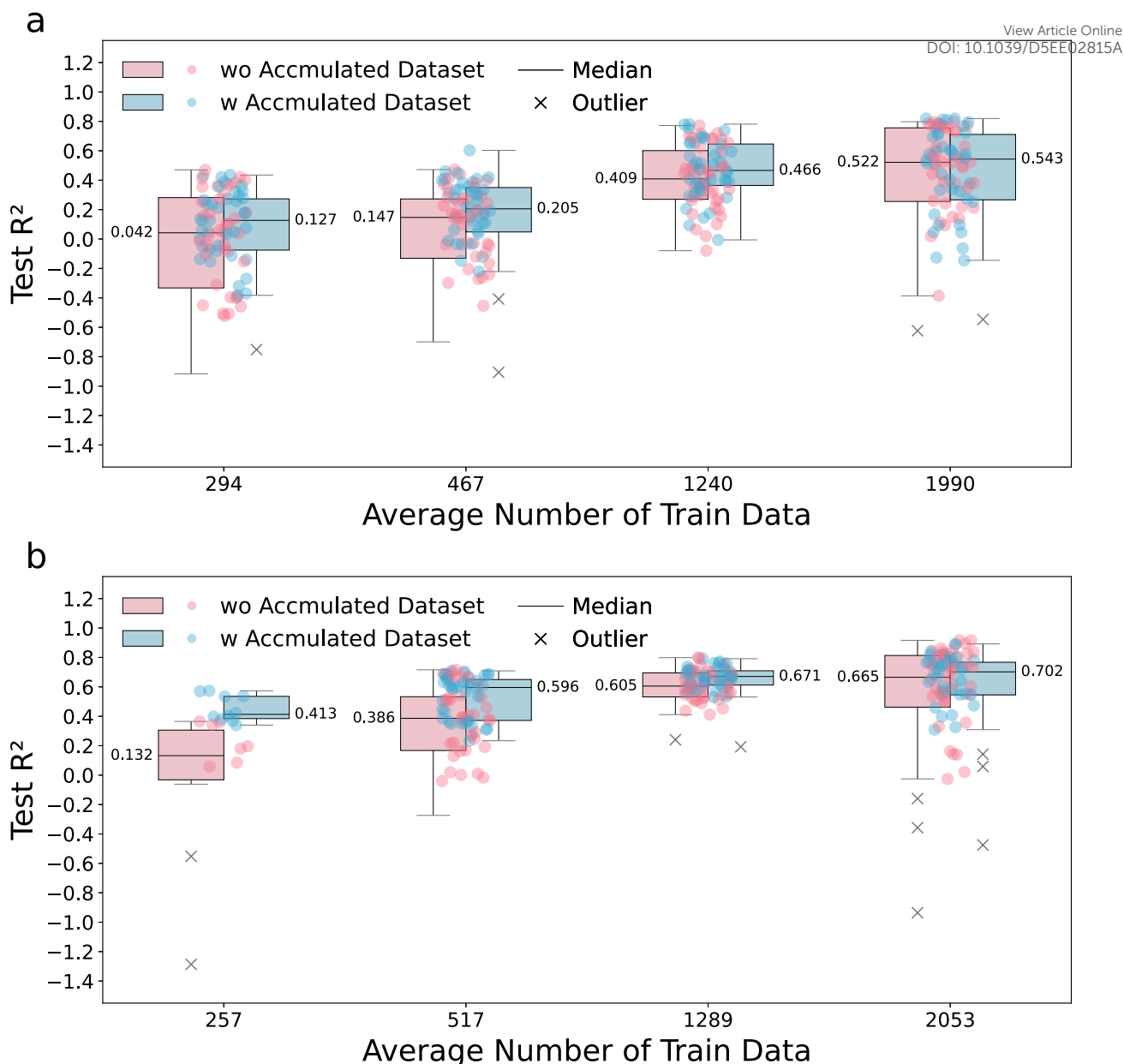
**Table 3.** Summary of PPs for R2R OPVs with the top formulation in Cluster 0 and 1 using CB and xylene processing solvents. PPs obtained from PCE and  $J-V$  prediction model were also listed for comparison.

View Article Online  
DOI: 10.1039/D5EE02815A

Cluster Number	Prediction	Solvent	$J_{sc}$ (mA cm <sup>-2</sup> )	Cal. $J_{sc}$ (mA cm <sup>-2</sup> )	$V_{oc}$ (V)	FF	PCE (%)	$R_{sh}$ ( $\Omega$ cm <sup>2</sup> )	$R_s$ ( $\Omega$ cm <sup>2</sup> )
0	PCE	CB	19.7	-	0.834	0.632	11.3	913	2.72
	$J-V$		20	-	0.829	0.679	11.3	1319	2.67
	-		20.3 (19.9 ± 0.821)	19.91	0.835 (0.831 ± 0.010)	0.694 (0.688 ± 0.035)	11.8 (11.3 ± 0.550)	874 (2065 ± 6883)	2.18 (2.14 ± 0.279)
	-	Xylene	20.0 (19.8 ± 0.674)	19.77	0.824 (0.820 ± 0.004)	0.678 (0.672 ± 0.025)	11.2 (10.9 ± 0.195)	6993 (1885 ± 5139)	1.77 (1.99 ± 0.184)
	PCE		20.5	-	0.810	0.625	10.3	534	2.09
	$J-V$		20.7	-	0.810	0.619	10.4	1730	2.01
1	-	CB	21.6 (21.3 ± 0.573)	20.83	0.814 (0.804 ± 0.008)	0.632 (0.564 ± 0.056)	11.1 (9.67 ± 1.01)	3894 (813 ± 700)	2.90 (3.80 ± 0.808)
	-		20.8 (19.7 ± 1.51)	20.63	0.819 (0.816 ± 0.010)	0.638 (0.590 ± 0.063)	10.9 (9.51 ± 1.35)	664 (1210 ± 3153)	1.72 (1.96 ± 0.447)
	PCE	Xylene	20.5	-	0.810	0.625	10.3	534	2.09
	$J-V$		20.7	-	0.810	0.619	10.4	1730	2.01

\* 200 devices were fabricated at each formulation.  
\* All statistical data were calculated from 200 devices.





**Fig. 5.** Test  $R^2$  performance across train dataset sizes without or with accumulated dataset for two ternary systems. Box plots illustrating test  $R^2$  for two material systems, (a) PM6:Y6:IT-4F and (b) PM6:D18:L8-BO, in relation to the average number of train data, corresponding to about 10, 20, 50 and 80% without and with accumulated dataset (PM6:L8-BO material system composed of 20,199 data points). Each data point for the  $R^2$ s was represented as circle-shaped markers. Pink and blue boxes represented the interquartile range (IQR) of the test  $R^2$  without and with the accumulated data, respectively, extending from the first quartile (Q1) to the third quartile (Q3). Each median was indicated by a solid line within the box, accompanied by a numerical value. The whiskers' boundaries followed the 1.5 IQR rule to identify outliers, where the lower whisker extends to the lowest data point above  $Q1 - 1.5 \times (Q3 - Q1)$ , and the upper whisker to the highest data point below  $Q3 + 1.5 \times (Q3 - Q1)$ . Outliers were denoted by a marker 'x'.



## Data availability

All data generated or analyzed during this study are included in the published article and its ESI, supplementary data files and supplementary video. Detailed descriptions of the supplementary data files and video are as follows:

### Supplementary data files

Data S1. Self-produced dataset for PP prediction composed of 25,249 data.

Data S2. Self-produced dataset for  $J$ - $V$  prediction composed of 1,792,697 data.

Data S3. Dataset for PCE prediction in ternary and all polymer material systems. It includes 16 batches totaling 2,422 data points for the PM6:Y6:IT-4F system, and 10 sub-batches totaling 2,570 data points for the PM6:D18:L8-BO system. Additionally, the all-polymer dataset comprises 11 sub-batches with 2,840 data points for the PM6:PYF-T-o system.

Data S4. Model performance summary of ternary and all polymer material system without or with the accumulated dataset.

### Supplementary video

Video S1. The  $J$ - $V$  curves predicted by the  $J$ - $V$  prediction model were displayed for selected features. Predicted PPs from the PP prediction model were also listed in the top-left legend of the chart. As each feature slider is adjusted, the predicted  $J$ - $V$  curves and PPs update interactively according to the selected feature values.

## Code availability

In this study all ensemble-based models were built using scikit-learn and xgboost, which are open-source libraries. All code, relevant datasets and supplementary files are shared at:

<https://figshare.com/s/ca0b8a65d1acdadb9f28>.

