ARTICLE OPEN

Machine learning for accelerating the discovery of high-performance donor/acceptor pairs in non-fullerene organic solar cells

Yao Wu^{1,3}, Jie Guo^{1,3}, Rui Sun¹ and Jie Min^{1,2}✉

Integrating artificial intelligence (AI) and computer science together with current approaches in material synthesis and optimization will act as an effective approach for speeding up the discovery of high-performance photoactive materials in organic solar cells (OSCs). Yet, like model selection in statistics, the choice of appropriate machine learning (ML) algorithms plays a vital role in the process of new material discovery in databases. In this study, we constructed five common algorithms, and introduced 565 donor/acceptor (D/A) combinations as training data sets to evaluate the practicalities of these ML algorithms and their application potential when guiding material design and D/A pairs screening. Thus, the best predictive capabilities are provided by using the random forest (RF) and boosted regression trees (BRT) approaches beyond other ML algorithms in the data set. Furthermore, >32 million D/A pairs were screened and calculated by RF and BRT models, respectively. Among them, six photovoltaic D/A pairs are selected and synthesized to compare their predicted and experimental power conversion efficiencies. The outcome of ML and experiment verification demonstrates that the RF approach can be effectively applied to high-throughput virtual screening for opening new perspectives to design of materials and D/A pairs, thereby accelerating the development of OSCs.

npj Computational Materials (2020)6:120; <https://doi.org/10.1038/s41524-020-00388-2>

INTRODUCTION

In the latest decade, the discovery of novel photoactive donor (D) and acceptor (A) materials has greatly promoted the development of bulk heterojunction (BHJ) organic solar cells (OSCs)^{1–7}. In this process, tens of thousands of photovoltaic materials have been designed and synthesized by chemists and materials scientists. More than 50,000 publications focusing on OSCs have been tracked by Web of Science (see Supplementary Fig. 1) in the latest decade. More than 60% of these publications are based on material science, mainly involved in material design for photoactive layers in OSCs. Apart from materials reported in these publications, there are imponderable materials, which have been synthesized and characterized while they have not been reported ever owing to low performance in OSC devices. It mainly results from failed design strategies of D/A materials and/or unsuitable synergistic effects of D/A combinations. Clearly, the consumption on trial and error experiments is huge. As shown in Fig. 1, power conversion efficiency (PCE) can only be recognized after manual designing, material engineering, and devices engineering. In other words, development of photovoltaic materials is an experience-based screening process, whose essence is to include a lifecycle of “Discovery-Synthesis and Research-Optimization-Evaluation”. In addition to material design and synthesis, which involves new route, structure characterization, material purification, etc., the process of material optimization and evaluation is also uncertain. On one hand, the efficient device performance is achieved under the synergistic effects of donor/acceptor (D/A) combinations, including complementary absorption⁸, matched energy levels⁹, favorable blend morphology with nanoscale phase separation¹⁰, high and balanced charge transport property^{11–13}, etc. On the other hand, adjusting processing parameters such as D/A ratio, processing solvent and solvent additive, pre- and post-treatments,

selection of film forming technology and so on^{14–18}, and constructing device architectures via interface engineering and device engineering^{19–21}, are also tedious and laborious work. Obviously, the input–output ratio as discussed above generally is unequal in the field of photovoltaic material research. To reduce the consumption of trial and error experiments and accelerate the discovery of high-performance materials, developing strategies based on shortening the lifecycle of material development will become extremely important in the future research of OSCs.

Thanks to the contributions of many computer scientists in terms of the combination of artificial intelligence (AI) and big data²², machine learning (ML) technique, which is a computational method, spread quickly into many research fields and commercial projects, such as drug discovery²³, genome sequencing²⁴, metal-organic frameworks materials²⁵, and organic light-emitting diodes²⁶, etc. In material science, ML is an ideal tool that can effectively learn from past massive data sets and mechanisms, automatically generate structures, assess their electronic features and other properties, determine the underlying rules among these data sets and build scientific models to make predictions with reasonable accuracy^{27–32}. As a result, compared with hundreds of hours or more it takes to evaluate a compound by experiment, the models based on appropriate algorithms can predict properties in a few minutes. It is foreseeable that applying ML methods to the design of photovoltaic materials or systems will greatly accelerate the discovery of high-efficient materials, reduce the research lifecycle, and promote the development of OSCs.

Some ML approaches have been introduced into the OSCs for the screening of complex molecules and their PCE prediction of relevant photovoltaic systems^{33–40}. For instance, Sun et al.³³ developed a deep learning model based on a convolutional neural network (CNN) that enables recognition of chemical structures

¹The Institute for Advanced Studies, Wuhan University, Wuhan 430072, China. ²Beijing National Laboratory for Molecular Sciences, Beijing 100190, China. ³These authors contributed equally: Yao Wu, Jie Guo. ✉email: min.jie@whu.edu.cn

and automatic classification for predicting the PCE of D materials. Shinji et al. reported a screening of conjugated molecules for polymer-fullerene OSC applications with artificial neural network (ANN) and random forest (RF) algorithms by importing parameters such as PCE, molecular weight, energy levels, and electronic properties with digitized chemical structures³⁶. However, these examples of the application of ML algorithms were mostly used in the fullerene-based OSCs, whereas the reports about the application of ML algorithms in the non-fullerene-based OSCs are limited^{41–43}. As non-fullerene acceptors (NFAs) draw researchers' attention and have become research hotspots^{44–46}, and most state of art OSCs with efficiency up to 13–17% were achieved by NFA OSCs in these few years^{4,47,48}, we should pay more attention to the applications of ML approaches that would tackle broader and more complex non-fullerene OSCs. Notably, the main challenges associated with screening the ideal photovoltaic materials from diverse non-fullerene systems are not only the selection of optimal algorithms but also the construction of training data sets. D/A pairs rather than single D or A materials as parameters generally have been considered to identify the device

Fig. 1 Process of materials evaluation by experiment method. This process usually involves manual design and material engineering, device engineering and performance evaluation.

performance, and give the effective instructions on materials designing, morphology optimization and device engineering. Although ML methods have great potential in discovering materials for non-fullerene OSCs, they have not been fully explored, compared and verified systematically.

In this work, we performed ML methods including linear regression (LR), multinomial logistic regression (MLR), RF, ANN, and boosted regression trees (BRT) to investigate polymer-NFA-based OSCs devices based on the D-A copolymer Ds and ladder fused non-fullerene As with A-D-A configuration. Detailed description of these algorithms was provided in Machine Computation section of Supplementary information and illustrated at Supplementary Fig. 2. Our models are constructed on the basis of 565 D/A combinations as training data sets—relevant chemical structures of NFA systems as well as their photovoltaic parameters obtained from the literature. The correlation between D/A pairs and the PCE predication for the polymer-NFA OSC devices is validated via the above-mentioned ML algorithms. After ML, RF and BRT algorithms showed the higher maximum Pearson's correlation coefficient (r) values as compared with the other algorithms. Thus, >32 million D/A combinations were further calculated by these two models for predicting the potential PCEs of these D/A combinations. Among them, six D/A pairs as photoactive systems were selected and synthesized to compare their predictive and experimental PCEs. It was found that the developed RF model can be applied effectively to high-throughput virtual screening for guiding the compatible design and characterization of new D/A combinations in non-fullerene OSCs.

RESULTS

Research Workflow

The ML-assisted materials screening and experimental validation of new D/A pairs are depicted in Fig. 2. Frequently reported new D/A pairs in recent years serve as reliable suppliers of training data. These photovoltaic materials have definite structures while

Fig. 2 Workflow of building, application, and evaluations of machine learning methods in this work. **a** Scheme of collecting experimental data and converting chemical structure to digitized data. **b** Scheme of machine training, predicting, and methods evaluating.

Fig. 3 Diagram for translating chemical structures of D/A pairs into digitized data. The example for translating structure of polymer D is provided on the left side and that of the small molecule NFA is provided at right side. Besides, R_1 , R_2 , and R_3 represent alkyl chains with different length.

they are easy to be divided into several parts and transformed into digitized chemical structures or machine languages. The analytic workflow toward discovery of D/A pairs consists of the following steps: (1) collecting published D/A combinations and experimental data as training data sets. (2) Converting the structures of investigated photovoltaic materials into binary numbers. (3) applying the five ML models to predict PCEs of D/A pairs via the digitized keys together with relevant photovoltaic parameters. Here, data of 477 D/A pairs were inputted for training to generate predictive models. Then, data of 88 D/A pairs were served as testing set to evaluate the predictive capability of models based on the five algorithms, which are introduced from Microsoft Azure ML studio for predicting PCEs. (4) The models computed ~32,076,000 potential D/A pairs, from the results of which, six new D/A pairs are selected by experiments to further confirm the performance of ML methods in screening counterpart for active layer materials. (5) The selected D/A pairs will be synthesized and characterized to evaluate the performance of ML models measured using root mean squared error (RMSE) metrics.

Data

In all, 565 D/A pairs were collected from the literature (274 papers). The details of these D/A pairs together with their photovoltaic parameters and corresponding references are provided in the Supplementary Information (see Supplementary Table 11). Transferring chemical structure into input data critically to mitigate personal impact on the accuracy of predictive models has a vital role in determining the accuracy of models. In this study, the chemical structure was divided into several fragments with a series of binary type numbers by merging highly similar structures and numbering them, and further transformed into machine language. D-A copolymer Ds usually have donor units (e.g., benzodithiophene (BDT)) and As (e.g., thiazolo[5,4-d]thiazole (TTz)) flanked with or without π bridges (such as thiophene, thiazole), whereas non-fullerene As have structure with donor core (D), side chains, electron-deficient end group (A) and, in some cases, π bridges. For example, as shown in Fig. 3, polymer D J71 was divided into four fragments as part 1 (BDT D unit), part 2 (thiophene π bridge on the left side), part 3 (TTz A unit), and part 4 (thiophene π bridge on the right side). On the total, there are 31 part 1 fragments, 14 part 2 fragments, 27 part 3 fragments, and 14 part four fragments summarized for the collected D materials. Of note that random copolymers as well as regioregularity of common polymer backbone were not collected. As for non-fullerene As, single molecular structure was divided into five parts.

Normally, during the synthesis of D backbone of A-D-A NFAs, part 1 unit was coupling linked with two part three units. Accomplished with the formation of part 2, the D backbone is obtained with a ring-cycling reaction. The π bridges and end groups were defined as part 4 and part 5, respectively. To simplify computation, non-centrosymmetric were not included. On the total, there were 30 part one fragments, 18 part two fragments, 6 part three fragments, 22 part four fragments, and 35 part five fragments summarized for A materials. The chemical structures of these fragments are listed in Supplementary Fig. 3.

In addition to structure reasons, these materials were divided into several parts for their different roles in determining properties such as energy levels and aggregation behaviors, and further determining PCE, which means different parts have different weightings in determining predictive results. For example, the end groups (part 5 of As) play a vital role in determining lowest unoccupied molecular orbit levels, whereas highest occupied molecular orbit (HOMO) levels depend more on D core (part 1, part 2, and part 3 of As). Moreover, lamellar stacking is usually adjusted by side chains (part 2 of As). For some polymer Ds and NFAs, they are not constituted by all kinds of parts. For example, many NFAs do not have π bridges (part 4 of As). In this case, missing fragment is also numbered with individual binary number. As a result, the D/A pairs based on two series of structural fragments can be numbered and transformed successfully to digitized chemical structures.

Predicted results

As shown in Fig. 4a, 84.4% of data (477 in 565 D/A pairs) were randomly inputted for training and models generation, and the rest of data were randomly selected as validation set and run with five models to test their accuracies in predicting PCE. Figure 4b–f show the training results of the five algorithms, where the horizontal and vertical axes represent the experimental PCE and predicted PCE, respectively, and r is the correlation coefficient. After training and testing the datasets, both RF and BRT models exhibit satisfactory performance in predicting PCE with higher r values of 0.70 for RF and 0.71 for BRT, respectively, as summarized in Supplementary Table 1. The experimental PCE and predicted PCE obtained from LR, MLR, and ANN methods have relatively lower correlation, with r values of 0.54, 0.59, and 0.60, respectively, which are far below the perfect positive correlation and insufficient for practical use. The RF and BRT models repeatedly fit many decision trees to improve the accuracy of the model. They are known to be robust even in the presence of several

Fig. 4 Prediction workflow and correction with laboratory results. **a** Workflow of PCE prediction from ML methods; **b–f** The correlation between the true and predicted efficiency in **b** LR, **c** MLR, **d** RF, **e** BRT, and **f** ANN methods. The diagonal line indicates the perfect positive correlation ($r = 1$).

explanatory variables, which are similar to the conditions in this work. Thus, the RF and BRT gave the improved r values between predicted PCEs and experimental PCEs. In addition, 10-fold cross validation experiments (see Supplementary Fig. 4) were conducted with 5 ML methods as displayed in Supplementary Table 2, RF method has the lowest mean absolute error value of 0.832 compared with other methods (2.116 for LR, 2.599 for MLR, 1.653 for BRT, and 2.159 for ANN, respectively).

Results of classification

PCEs in these data were divided into three groups, high level (A, $\text{PCE} > 11\%$), moderate level (B, $7\% \leq \text{PCE} < 11\%$) and low level (C, $0 \leq \text{PCE} < 7\%$). The row and column represent the experimental and predicted classifications of the PCE, respectively, such that the diagonal line corresponds to the correct answer. As shown in the columns of Fig. 4d, RF has a high accuracy in predicting PCE in A (15 out of 23, 65.2%) and B (34 out of 48, 70.8%) levels, whereas BRT performs better in B (31 out of 48, 64.6%) and C (13 out of 17, 76.5%) levels. Note that all methods have reasonable accuracy of predicting PCE in B level, which are 81.25, 52.6, 70.8, 64.6, and 72.9% for LR, MLR, RF, BRT, and ANN, respectively. It might be ascribed to higher proportion of training data in B level (see Supplementary Fig. 5). It implies that if more data were involved, the accuracy of our models can be improved³³. Unlike the evaluation of B level, the C level for predicting PCE shows the huge uncertainty, demonstrated by their accuracies of predicting PCE, which are 5.9%, 20.0%, 23.5%, 76.5%, and 58.8% for LR, MLR,

RF, BRT, and ANN, respectively. From this point of view, we should also pay close attention to the value of undesired experimental results, most of which are not published. Typically, relevant D/A pairs and their photovoltaic performance also work as important references not only for manual designing of materials but also for ML. As RF and BRT exhibit favorable performance in predicting PCE and give correct answers with great accuracy for the D/A pairs, these two methods are applied to predict the PCEs of automatically generated 32,076,000 D/A combinations, and further evaluate the practicability of these two ML algorithms.

The data set consisting of structural fragments of D/A pairs can be used as a new database that automatically generates new D/A combinations for potential PCE prediction. Based on fragments listed in Supplementary Fig. 3, resulting from 565 D/A pairs, there are over 32,076,000 new D/A combinations generated. The distributions of predictive PCE based on RF and BRT methods are shown in Fig. 5a, b, respectively. According to results outputted by RF model (in Fig. 5a), 12.27% of D/A combinations can possibly achieved PCE exceeding 11%, and it is 14.15% for BRT model (in Fig. 5b). As we known, Poly[(2,6-(4,8-bis(5-(2-ethylhexyl-3-fluoro)thiophen-2-yl)-benzo[1,2-b:4,5-b']dithiophene))-alt-(5,5-(1',3'-di-2-thienyl-5',7'-bis(2-ethylhexyl)-benzo[1',2'-c:4',5'-c']dithiophene-4,8-dione)] (PM6) is a well-developed D material, which has achieved high PCE in many D/A combinations². Here, we further extracted results of PM6 containing D/A combinations to investigate the statistical characteristics of high-performance active layer materials. As shown in Supplementary Fig. 6, there are 17,820 results for PM6 in each model, of which 13.47% is in A

Fig. 5 Distribution of predicted PCE resulting from all automatically generated D/A pairs. Results from all D/A pairs with **a** RF method and **b** BRT method.

level for RF models, and 33.63% in A level for BRT model. As with our expectation, the PCE distributions of PM6 tend to move toward district of higher PCE compared with the overall distribution of 32,076,000 results, which is more obvious in BRT model.

Experimental validation

In these D/A combinations with PM6 as D material, three NFAs including Y-ThCN, Y-ThCH₃, and Y-PhCl were selected, synthesized, and characterized to demonstrate how ML methods help to screen counterpart for active layer materials. These D/A combinations were selected for the following reasons: (i) they were predicted to have high performance with PCE over 10%; (ii) the new materials are easy to synthesized by one-step reaction with all raw materials commercially available. It is worth mentioning that not all materials in the predicted D/A combinations can be synthesized in laboratory. Some materials exist only in theory. Besides, to enrich the experiment sample size, the D/A combinations consisting of PBDB-T and Y-ThCN, Y-ThCH₃, and Y-PhCl were also characterized. These combinations were further applied to verify the PCE predication, and evaluate the predictive ability of RF and BRT models. The chemical structures and the performance of selected D/A combinations was listed in Fig. 6 with corresponding photovoltaic parameters and predictive PCEs summarized in Table 1. The details of the synthesis of NFAs are described in Experimental section of Supplementary Information and their ¹H nuclear magnetic resonance (¹HNMR) spectra were provided at Supplementary Figs. 7–9.

As the predicted results from BRT method, PM6:Y-ThCN, PM6:Y-ThCH₃, and PM6:Y-PhCl binary systems can achieve PCEs of 11.56, 11.14, and 13.30%, respectively. In addition, if the RF method is used as the ML model, these three systems can exhibit the predicted PCEs of 10.52% for PM6:Y-ThCN, 10.41% for PM6:Y-ThCH₃, and 13.33% for PM6:Y-PhCl, respectively. After the systematic optimization of photoactive layers, as presented in Supplementary Table 3–10, devices based on PM6:Y-ThCN and PM6:Y-PhCl achieved high PCEs of 13.18% and 15.71% (Fig. 6b), respectively, which are not very far off the predicted PCEs of both BRT model and RF models (Table 1). However, the predicted and experimental PCEs of PM6:Y-ThCH₃ are not at the same level. For PBDB-T based system, PBDB-T:Y-ThCN, PBDB-T:Y-ThCH₃, and PBDB-T:Y-PhCl are predicted to achieve PCEs of 12.73, 12.49, and 12.55% by BRT method, and 11.49%, 11.64%, and 11.32% by RF method. Their experimental PCEs show a high agreement with their predicted results with 11.02%, 11.08%, and 11.19% (Fig. 6c). Owing to higher HOMO levels of PBDB-T, the PBDB-T system had lower open circuit voltage (*V*_{oc})⁴⁹. All systems except PM6:Y-ThCH₃ have high current density (*J*_{sc}) and fill factor (FF) as shown in Table 1. External quantum efficiency (EQE) of the representative

devices is given in Supplementary Fig. 10, which are well correlated with *J*_{sc}. Generally, the experimental PCEs of most system agree well with the predicted PCE, indicating that both methods are qualified to provide instruction on counterpart materials screening in active layers.

To further screen the optimal methods for PCE prediction, RMSE values between predicted PCE and experimental PCE were introduced from this equation,

$$RMSE = \sqrt{\frac{\sum_{i=1}^n (y'_i - y_i)^2}{n}} \quad (1)$$

where *y_i* and *y'_i* represent predicted and experimental PCE value, respectively, and *n* is the number of D/A combinations characterized in this experiment. The unit of percentage is ignored to make the RMSE value more apparent. For BRT method, the RMSE value is 2.42, whereas it is 1.17 for RF method. Obviously, RF method shows a narrower deviation in predicting PCE of D/A combinations, indicating that it is more useful to accelerate the discovery of high-performance D/A pairs in OSCs.

DISCUSSION

The deviation between prediction and experiment seems inevitable, because experimental data are not always “reliable” enough⁵⁰. As we know, PCEs of photovoltaic materials are sensitive to the materials purity, experimental operation, external environment, etc. One D/A combination might have different experimental PCEs in different laboratories. Usually, the experiment results probably lost repeatability owing to difference in trace solvent vapor in atmosphere, humidity, temperature and sometimes critical operating details of researchers in different labs and even in same labs. Another well-known batch-to-batch vibration exists in polymer materials, which tend to have different molecule weight and molecule weight distribution. For example, during the period of our research, Hou’s group also published PM6:Y-PhCl combination. In their work, the maximum PCE was 16.5%, which is higher than that of our work⁵¹. It probably results from the different batches of PM6 materials and experimental conditions. It should be noted that, although the laboratory efficiency fluctuates below the theoretical highest efficiency that one D/A combination could achieve, and the difference is difficult to sweep out, the laboratory efficiency can still be used to roughly evaluate the real performance of the materials in OSC devices because almost every researchers spared no effort to narrow the gap between laboratory efficiency and theoretical efficiency of their materials, as higher PCE is one of the most important pursuit of material scientist in OPV filed. Therefore, the gap between the laboratory efficiency and theoretical efficiency causes deviation between prediction and experiment, yet it is not so huge to make

Fig. 6 Chemical structures of investigated D/A pairs and corresponding photovoltaic parameters. **a** Chemical structures of the investigated D/A pairs, including polymer donors (PM6 and PBDB-T) and NFAs (Y-ThCN, Y-ThCH₃, Y-PhCl). **b** *J*-*V* curves of devices based on PM6:Y-ThCN, PM6:Y-ThCH₃, and PM6:Y-PhCl. **c** *J*-*V* curves of devices based on PBDB-T:Y-ThCN, PBDB-T:Y-ThCH₃, and PBDB-T:Y-PhCl.

Table 1. The device parameters of OSCs with different D/A combinations under AM 1.5 G illumination (100 mW cm⁻²).

D/A combinations	<i>V</i> _{oc} (V)	<i>J</i> _{sc} (mA cm ⁻²)	Calculated <i>J</i> _{sc} (mA cm ⁻²)	FF (%)	PCE (%)	Predicted PCE (%) ^b	Predicted PCE (%) ^c
PM6:Y-ThCN	0.906	21.91	20.20	66.44	13.18 ^a (12.98 ± 0.23)	11.56	10.52
PM6:Y-ThCH ₃	0.952	12.89	14.81	54.30	6.67 ^a (6.18 ± 0.49)	11.14	10.41
PM6:Y-PhCl	0.845	26.82	24.68	60.70	15.71 ^a (15.60 ± 0.18)	13.30	13.33
PBDB-T:Y-ThCN	0.854	21.72	20.97	59.43	11.02 ^a (10.81 ± 0.36)	12.73	11.49
PBDB-T:Y-ThCH ₃	0.856	22.78	22.15	56.85	11.08 ^a (10.90 ± 0.18)	12.49	11.64
PBDB-T:Y-PhCl	0.763	23.14	22.62	63.37	11.19 ^a (11.02 ± 0.23)	12.55	11.32

^aMaximum PCE; ^bobtained from BRT method; ^cobtained from RF method.

the models invalid. On the other side, although we cannot guarantee that the experimental PCE is exactly the highest PCE it can achieve, because there are so many conditions we have not tested⁵², big data processing shows its superiority that it can weaken the influence of experimental errors to some extent, so results from ML methods still have valuable instructions on experiment. From this respect, the RF and BRT models can be upgraded by adding the processing conditions and the specific physical and chemical properties of materials as well as the new performance evaluation parameters in the future work. For instance, as we previously reported, an industry figure of merit, which integrates device efficiency, photostability, and synthetic complexity to evaluate materials can be introduced into these models to better described the comprehensive performance in practical application⁷. Although there is deviation between results of experiments and simulations, it doesn't hinder us from exploring its application in OSCs.

In summary, our work offered a paradigm for involving ML in solving problems in OSCs, which used to rely on vast laboratory

work. Five models were built on LR, MLR, BRT, RF, and ANN algorithms. Although 565 D/A pairs as training data sets were collected to evaluate the predictive ability of these ML algorithms. Contrary to the low correlation coefficient in LR, MLR, and ANN, both BRT and RF models exhibited improved accuracy in predicting PCE with high *r* values of 0.71 for BRT and 0.70 for RF, respectively. In addition, the BRT and RF methods were evaluated in practical application by screening 32,076,000 D/A combinations. By analyzing results outputted by these two methods, six new D/A combinations, including PM6:Y-ThCN, PM6:Y-ThCH₃, PM6:Y-PhCl, PBDB-T:Y-ThCN, PBDB-T:Y-ThCH₃, and PBDB-T:Y-PhCl combinations, were characterized for further evaluating the practicability of these two algorithms applied into OSCs. Furthermore, by integrating and comparing the predicted data and experimental results optimized, it can be concluded that the RMSE value matches nicely with the RF model in the predicting PCE, indicating that it is more suitable for screening new promising materials from predictive statistical data generated. Our results demonstrate that ML methods, especially for RF

model, are alternative strategies to accelerate the discovery of high-performance D/A combinations and can provide the decision making of molecular design in OSCs or other organic electronics.

METHODS

Data collection of non-fullerene-based OSCs

The data set was constructed by collecting 565 D/A pairs published between 2015 and 2019. And it was restricted to binary OSCs based on non-fullerene small molecule A and polymer D.

Model building

The ML model was built based on LR, MLR, RF, BRT, and ANN algorithms with 84.4% of collecting data for training set and 15.6% for testing set. The work of computation was achieved on this website: <https://studio.azureml.net/>.

Model performance metrics

The models built in this work were evaluated by mean absolute error, RMSE, and Pearson's correlation coefficient (see Supplementary Table 1). In addition, they were further validated by experiment.

DATA AVAILABILITY

All data generated or analyzed during this study are available from the corresponding author upon reasonable request.

Received: 10 November 2019; Accepted: 22 July 2020;

Published online: 13 August 2020

REFERENCES

- Bredas, J. L., Norton, J. E., Cornil, J. & Coropceanu, V. Molecular understanding of organic solar cells: the challenges. *Acc. Chem. Res.* **42**, 1691–1699 (2009).
- Yuan, J. et al. Single-junction organic solar cell with over 15% efficiency using fused-ring acceptor with electron-deficient core. *Joule* **3**, 1140–1151 (2019).
- Meng, L. et al. Organic and solution-processed tandem solar cells with 17.3% efficiency. *Science* **361**, 1094 (2018).
- Ameri, T., Khoram, P., Min, J. & Brabec, C. J. Organic ternary solar cells: a review. *Adv. Mater.* **25**, 4245–4266 (2013).
- Guo, J. & Min, J. A cost analysis of fully solution-processed ITO-free organic solar modules. *Adv. Energy Mater.* **9**, 1802521 (2019).
- Cui, Y. et al. Achieving over 15% efficiency in organic photovoltaic cells via copolymer design. *Adv. Mater.* **31**, 1808356 (2019).
- Min, J. et al. Processability: evaluation of electron donor materials for solution-processed organic solar cells via a novel figure of merit. *Adv. Energy Mater.* **7**, 1700465 (2017).
- Bakulin, A. A. et al. The role of driving energy and delocalized states for charge separation in organic semiconductors. *Science* **335**, 1340 (2012).
- Bin, H. et al. 9.73% efficiency nonfullerene all organic small molecule solar cells with absorption-complementary donor and acceptor. *J. Am. Chem. Soc.* **139**, 5085–5094 (2017).
- Zhou, C. et al. Toward high efficiency polymer solar cells: influence of local chemical environment and morphology. *Adv. Energy Mater.* **7**, 1601081 (2017).
- Ye, L. et al. Enhanced efficiency in fullerene-free polymer solar cell by incorporating fine-designed donor and acceptor materials. *ACS Appl. Mater. Interfaces* **7**, 9274–9280 (2015).
- Bin, H. et al. 11.4% Efficiency non-fullerene polymer solar cells with trialkylsilyl substituted 2D-conjugated polymer as donor. *Nat. Commun.* **7**, 13651 (2016).
- Lin, H. et al. High-performance non-fullerene polymer solar cells based on a pair of donor-acceptor materials with complementary absorption properties. *Adv. Mater.* **27**, 7299–7304 (2015).
- Zhao, F., Wang, C. & Zhan, X. Morphology control in organic solar cells. *Adv. Energy Mater.* **8**, 1703147 (2018).
- Ye, L. et al. Miscibility-function relations in organic solar cells: significance of optimal miscibility in relation to percolation. *Adv. Energy Mater.* **8**, 1703058 (2018).
- Lee, H., Park, C., Sin, D. H., Park, J. H. & Cho, K. Recent advances in morphology optimization for organic photovoltaics. *Adv. Mater.* **30**, 1800453 (2018).
- Min, J. et al. Time-dependent morphology evolution of solution-processed small molecule solar cells during solvent vapor annealing. *Adv. Energy Mater.* **6**, 1502579 (2016).
- Benanti, T. L. & Venkataraman, D. Organic solar cells: an overview focusing on active layer morphology. *Photosynth. Res.* **87**, 73–81 (2006).
- Li, Y. et al. Perylene diimide-based cathode interfacial materials: adjustable molecular structures and conformation, optimized film morphology, and much improved performance of non-fullerene polymer solar cells. *Mater. Chem. Front.* **3**, 1840–1848 (2019).
- Wang, J. et al. Regulating bulk-heterojunction molecular orientations through surface free energy control of hole-transporting layers for high-performance organic solar cells. *Adv. Mater.* **31**, 1806921 (2019).
- Kang, Q. et al. A printable organic cathode interlayer enables over 13% efficiency for 1-cm² organic solar cells. *Joule* **3**, 227–239 (2019).
- Agrawal, A. & Choudhary, A. Perspective: materials informatics and big data: Realization of the “fourth paradigm” of science in materials science. *APL Mater.* **4**, 053208 (2016).
- Alexander, T. Best practices for QSAR model development, validation, and exploitation. *Mol. Inf.* **29**, 476–488 (2010).
- Libbrecht, M. W. & Noble, W. S. Machine learning applications in genetics and genomics. *Nat. Rev. Genet.* **16**, 321–332 (2015).
- Wilmer, C. E. et al. Large-scale screening of hypothetical metal–organic frameworks. *Nat. Chem.* **4**, 83 (2011).
- Gomez-Bombarelli, R. et al. Design of efficient molecular organic light-emitting diodes by a high-throughput virtual screening and experimental approach. *Nat. Mater.* **15**, 1120–1127 (2016).
- Butler, K. T., Davies, D. W., Cartwright, H., Isayev, O. & Walsh, A. Machine learning for molecular and materials science. *Nature* **559**, 547–555 (2018).
- Xia, R. & Kais, S. Quantum machine learning for electronic structure calculations. *Nat. Commun.* **9**, 4195 (2018).
- Pereira, F. et al. Machine learning methods to predict density functional theory B3LYP energies of HOMO and LUMO orbitals. *J. Chem. Inf. Model.* **57**, 11–21 (2017).
- Friedrich, P., Konrad, M., Strunk, T. & Wenzel, W. Machine learning of correlated dihedral potentials for atomistic molecular force fields. *Sci. Rep.* **8**, 2559 (2018).
- Wu, K. et al. Prediction of polymer properties using infinite chain descriptors (ICD) and machine learning: toward optimized dielectric polymeric materials. *J. Polym. Sci. Part B* **54**, 2082–2091 (2016).
- Raccuglia, P. et al. Machine-learning-assisted materials discovery using failed experiments. *Nature* **533**, 73–76 (2016).
- Sun, W. et al. The use of deep learning to fast evaluate organic photovoltaic materials. *Adv. Theory Simul.* **2**, 1800116 (2019).
- Padula, D., Simpson, J. D. & Troisi, A. Combining electronic and structural features in machine learning models to predict organic solar cells properties. *Mater. Horiz.* **6**, 343–349 (2019).
- Lee, M. H. Insign. *Adv. Energy Mater.* **9**, 1900891 (2019).
- Nagasawa, S., Al-Naamani, E. & Saeki, A. Computer-aided screening of conjugated polymers for organic solar cell: classification by random forest. *J. Phys. Chem. Lett.* **9**, 2639–2646 (2018).
- Kaya, M. & Hajimirza, S. Application of artificial neural network for accelerated optimization of ultra thin organic solar cells. *Sol. Energy* **165**, 159–166 (2018).
- Jorgensen, P. B. et al. Machine learning-based screening of complex molecules for polymer solar cells. *J. Chem. Phys.* **148**, 241735 (2018).
- Lopez, S. A., Sanchez-Lengeling, B., de Goes Soares, J. & Aspuru-Guzik, A. Design principles and top non-fullerene acceptor candidates for organic photovoltaics. *Joule* **1**, 857–870 (2017).
- Sahu, H., Rao, W., Troisi, A. & Ma, H. Toward predicting efficiency of organic solar cells via machine learning and improved descriptors. *Adv. Energy Mater.* **8**, 1801032 (2018).
- Lee, M.-H. Insights from machine learning techniques for predicting the efficiency of fullerene derivatives-based ternary organic solar cells at ternary blend design. *Adv. Energy Mater.* **9**, 1900891 (2019).
- Lee, M.-H. Robust random forest based non-fullerene organic solar cells efficiency prediction. *Org. Electron.* **76**, 105465 (2020).
- Lin, Y.-C. et al. Enhancing photovoltaic performance by tuning the domain sizes of a small-molecule acceptor by side-chain-engineered polymer donors. *J. Mater. Chem. A* **7**, 3072–3082 (2019).
- Wang, T. et al. A wide-bandgap D–A copolymer donor based on a chlorine substituted acceptor unit for high performance polymer solar cells. *J. Mater. Chem. A* **7**, 14070–14078 (2019).
- Zhang, G. et al. Nonfullerene acceptor molecules for bulk heterojunction organic solar cells. *Chem. Rev.* **118**, 3447–3507 (2018).
- Yang, J. et al. Aromatic-diimide-based n-type conjugated polymers for all-polymer solar cell applications. *Adv. Mater.* **0**, 1804699 (2018).
- Xu, X. et al. Single-junction polymer solar cells with 16.35% efficiency enabled by a platinum(ii) complexation strategy. *Adv. Mater.* **31**, 1901872 (2019).

48. Wang, T. et al. Solution-processed polymer solar cells with over 17% efficiency enabled by an iridium complexation approach. *Adv. Energy Mater.* **10**, 2000590 (2020).
49. Fan, Q. et al. Overcoming the energy loss in asymmetrical non-fullerene acceptor-based polymer solar cells by halogenation of polymer donors. *J. Mater. Chem. A* **7**, 15404–15410 (2019).
50. Sutherland, B. R. Beyond photovoltaic lab efficiency. *Joule* **2**, 1032–1034 (2018).
51. Cui, Y. et al. Over 16% efficiency organic photovoltaic cells enabled by a chlorinated acceptor with increased open-circuit voltages. *Nat. Commun.* **10**, 2515 (2019).
52. Tabor, D. P. et al. Accelerating the discovery of materials for clean energy in the era of smart automation. *Nat. Rev. Mater.* **3**, 5–20 (2018).

ACKNOWLEDGEMENTS

This work was financially supported by the National Natural Science Foundation of China (NSFC) (grant no. 21702154 and 51773157) and the Fundamental Research Funds for the Central Universities. We also thank the support of the opening project of Key Laboratory of Materials Processing and Mold and Beijing National Laboratory for Molecular Sciences (BNLMS201905). We thank Suiyang Dai and Lezhi Yi for assisting Rui Sun in devices characterization.

AUTHOR CONTRIBUTIONS

Y.W. and J.G. contributed to the work equally. J.M. put forward the idea. J.M., J.G., and Y.W. designed the work plan for this research. J.G. and Y.W. collected the training data and performed mechanical calculations for the training set, testing set, final prediction, and data processing. Y.W. synthesized new materials used in this manuscript. Devices characterization and EQE measurement were conducted by R.S. Y.W. wrote the original draft, and all authors, including J.M., J.G., and R.S. contributed to analysis and revision. J.M. mentored the other authors throughout the investigation.

COMPETING INTERESTS

The authors declare no competing interests.

ADDITIONAL INFORMATION

Supplementary information is available for this paper at <https://doi.org/10.1038/s41524-020-00388-2>.

Correspondence and requests for materials should be addressed to J.M.

Reprints and permission information is available at <http://www.nature.com/reprints>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2020