Journal of Applied Physics ARTICLE pubs.aip.org/aip/jap

Machine learning study on organic solar cells and virtual screening of designed non-fullerene acceptors

Cite as: J. Appl. Phys. **134**, 153104 (2023); doi: 10.1063/5.0169284
Submitted: 24 July 2023 · Accepted: 2 October 2023 ·
Published Online: 20 October 2023

Cai-Rong Zhang,[1,a)] Ming Li,[1] Miao Zhao,[1] Ji-Jun Gong,[1] Xiao-Meng Liu,[1] Yu-Hong Chen,[1] Zi-Jiang Liu,[2] You-Zhi Wu,[3] and Hong-Shan Chen[4]

AFFILIATIONS

[1]Department of Applied Physics, Lanzhou University of Technology, Lanzhou, Gansu 730050, China
[2]School of Mathematics and Physics, Lanzhou Jiaotong University, Lanzhou 730070, China
[3]School of Materials Science and Engineering, Lanzhou University of Technology, Lanzhou, Gansu 730050, China
[4]College of Physics and Electronic Engineering, Northwest Normal University, Lanzhou, Gansu 730070, China

[a)]**Author to whom correspondence should be addressed:** zhcrxy@lut.edu.cn

ABSTRACT

Machine learning (ML) is effective to establish the complicated trilateral relationship among structures, properties, and photovoltaic performance, which is fundamental issue in developing novel materials for improving power conversion efficiency (PCE) of organic solar cells (OSCs). Herein, we constructed the database of 397 donor–acceptor pairs of OSCs with photovoltaic parameters and descriptor sets, which include donor–acceptor weight ratio within the active layer of the OSCs, root mean square of roughness, and 1024-bit Morgan molecular fingerprint for donor (Fp-D) and acceptor (Fp-A). The ML models random forest (RF), adaptive boosting (AdaBoost), extra trees regression, and gradient boosting regression trees were trained based on the descriptor set. The metrics determination coefficient (R^2), Pearson correlation coefficient (r), root mean square error, and mean absolute error were selected to evaluate ML model performances. The results showed that the RF model exhibits the highest accuracy and stability for PCE prediction among these four ML models. Moreover, based on the decomposition of non-fullerene acceptors L8-BO, BTP-ec9, AQx-2, and IEICO, 20 acceptor molecules with symmetric A–D–A and A–π–D–π–A architectures were designed. The photovoltaic parameters of the designed acceptors were predicted using the trained RF model, and the virtual screening of designed acceptors was conducted based on the predicted PCE. The results indicate that six designed acceptors can reach the predicted PCE higher than 12% when P3HT was adopted as a donor. While PM6 was applied as a donor, five designed acceptors can achieve the predicted PCE higher than 16%.

I. INTRODUCTION

Organic solar cells (OSCs) are important renewable energy technology due to their sustainability, flexibility, and market potential.[1,2] OSCs convert solar energy into electricity, reducing reliance on finite resources, and lowering carbon emissions and environmental pollution. Their flexibility and adaptability make them suitable for various shapes and curved surfaces. Additionally, with lower manufacturing costs and sustainable production and recycling methods, OSCs have the potential for widespread application in the renewable energy market. However, further research and technological improvements are still required to enhance efficiency and stability for commercialization.

In OSCs, the active layer is a critical component that plays a crucial role in photovoltaic processes. The active layer typically consists of bulk heterojunction,[3] which is composed of multiple materials, usually referred to as electron donor and acceptor materials.[4,5] High-performance OSCs were reported one after another,[6] and the power conversion efficiency (PCE) of OSCs has been significantly improved. In 2015, Hou *et al.* developed polymeric donor PM6 and prepared OSCs with PCE up to 9.2% using PM6/$PC_{71}BM$ as an active layer.[7] In 2019, Wang *et al.* added the small

molecule donor BIT-4F-T as the third component to the PTB7-Th: IEICO-4F blend to obtain ternary OSCs with a PCE of 14%.[8] In 2020, Ding *et al.* reported a more efficient polymer donor D18 and prepared OSCs with PCE up to 18.22% using D18/Y6 as the active layer.[9] The PCE of OSCs reported so far can reach up to 20% or more.[10] Therefore, the development of novel donor and acceptor materials is a significant route to improve PCE.

OSCs have been studied for decades, and many excellent performance donor and acceptor materials have been reported. However, the research and development of new materials are still conducted using the expensive and time-consuming traditional experimental trial-and-error method. Therefore, it is an important issue to find the effective method that can accelerate the development of photovoltaic donor and acceptor materials based on the complex relationship between molecular structures and device performances.

With the continuous advancement of computational science, applying data-driven machine learning (ML) approaches can accelerate the research and development of OSCs.[11] In recent years, ML has been widely used for the prediction of material activity,[12,13] drug development,[14,15] PCE of OSCs,[16] and inverse design of high-performance materials.[17–20] ML algorithms can find intrinsic connections between features and target parameters, established by traversing all features in the database. The trained ML model can be used for the prediction of performance parameters and molecular design for OSCs. Hence, ML provides a new choice and an alternative pathway for the development of OSCs. In 2018, Saeki *et al.* employed extended connectivity fingerprints and molecular access system fingerprints to train the ML model.[21] In 2019, Troisi *et al.* computed geometric and electronic property descriptors for a database of 249 donor–acceptor pairs and used them to train ML models. It was observed that either electronic properties or structural parameters yielded similar results.[20] In 2020, Min *et al.* utilized ML analysis to identify the optimal donor–acceptor pairs for OSCs. They trained ML models on a data set consisting of 565 samples, and the enhanced regression tree model and the random forest (RF) model performance achieved Pearson correlation coefficient (r) of 0.71 and 0.70, respectively.[22] In recent years, Wang *et al.* have reiterated the importance of ML and discussed its applications in material design.[23] By incorporating open-circuit voltage loss and dielectric constants into the descriptor sets, we presented the ML model for predicting OSC performances are prevailing to that based on previous descriptor sets,[24] and the importance of dielectric constants for developing OSC materials was approved by the experimental study.[25] Ma *et al.* trained a gradient boosting regression model using morphology descriptors to predict the charge transfer characteristics ($r = 0.96$).[26] Yi *et al.* trained ML models using three highly correlated molecular property descriptors, namely, singlet–triplet energy gap, optical gap, and driving force, achieved good predictive accuracy ($r = 0.81$).[27]

In this work, we established a database comprising molecular structures, heterojunction parameters, and OSC parameters for 397 donor–acceptor pairs from 193 published articles. The descriptor sets include the 1024-bit Morgan fingerprints of the donor (Fp-D) and acceptor molecules (Fp-A), the donor–acceptor weight ratio (D/A ratio) within the active layer of the OSCs, and root mean square of roughness (RMSR). In comparison to the research conducted by Nagasawa *et al.*[21] and Troisi *et al.*,[20] the predictive accuracy of the model has been enhanced by utilizing descriptors of molecular property and structural parameters simultaneously. Four kinds of ML algorithms were conducted to explore the prediction of photovoltaic performance. The RF model demonstrated the best predictive performance and stability, with *r* values of 0.83 and 0.81 on the test set and training set, respectively. Based on the selected optimal model, the photovoltaic parameters of 20 newly designed non-fullerene acceptors were predicted and virtually screened using the RF model. The acceptor–donor pairs PM6:A1 and PM6:A9 can achieve the top two predicted PCEs of 16.23% and 16.12%, respectively.

II. METHODS

A. Database

Because of the importance of non-fullerene acceptors in the development of OSCs and their obvious advantages over fullerenes, most of the reported high-performance OSCs use non-fullerene acceptors. In this work, the OSC data, including donor and acceptor molecular structures, D/A ratio, RMSR, PCE, open-circuit voltage (V_{OC}), short-circuit current density (J_{SC}), and fill factor (*FF*), were collected from 193 experimental articles (from 2016 to 2022) that reported 411 sets of donor and non-fullerene acceptor pairs. Based on this, we have constructed a completely new database. In order to prevent the impact of duplicating donor and non-fullerene acceptor pairs on the predictive performance of the ML model, only the data with relatively higher PCE for the same donor–acceptor pairs are retained in the database. This ensures that the ML model can identify the best features of data throughout the training process. The final whole database includes 397 donor and non-fullerene acceptor pairs, which are used for the subsequent ML work.

ML methods can discover intrinsic connections among quantities by scanning databases and thereby construct reasonable logical relationships. The training set is used to complete the learning and training of the selected ML model, and then the test set is used to test the performance of the trained target model. The data between the test set and the training set should be completely independent of each other to ensure that the ML model testing cannot be influenced by the exposed data set (training set) and also to ensure the reliability of the trained model. Therefore, the database was partitioned into a mutually independent training set (317 data points, 80%) and a test set (80 data points, 20%) using the stratified sampling method.[28] The stratified sampling method divides samples into different strata based on their characteristics and then samples randomly within each stratum. This method is used to maintain a uniform sample distribution, cover important features adequately, and reduce bias and variance issues, thereby improving the reliability of test results.

B. ML algorithms

Four decision tree type ML algorithms were selected in this work based on supervised learning, including RF,[29] adaptive boosting (AdaBoost),[30] extra trees regression (ETR),[31] and gradient boosting regression trees (GBRTs).[32] The advantages of the four

ML models are that (1) the RF algorithm is one of the Bagging models, which can repeatedly generate a new training set of k randomly selected (with put-back) samples from the original training data set M to train the decision tree. The training results are then arithmetically averaged to obtain the final RF model output. Since the RF training process can be parallelized, the computation is fast, and due to applying random sampling during training processes, the model gained good generalization ability. Though tree-based models may have limitations when predicting values beyond the range of the training data, it does not mean that they are incapable of handling such situations entirely. With proper ML model adjustments and parameter optimization, tree-based models can exhibit some degree of extrapolation ability.[31,33] They can make predictions based on the patterns and relationships observed in the training data. Zhu *et al.* explained the importance and correlation between each functional group and the predicted properties through the RF model, and the design and prediction of new molecules were achieved.[34] (2) The AdaBoost model is a typical boosting algorithm, which achieves model training by continuously adjusting the weights of multiple weak classifiers and obtains the final strong classifier after several iterations. Chan *et al.* evaluated the utility of two tree-based models, AdaBoost and RF, using hyperspectral data for detailed classification, and the results showed that AdaBoost showed the highest overall accuracy.[35] (3) The ETR algorithm is very similar to the RF that is composed of many decision trees, but, ETR does not use random sampling, each decision tree is obtained by all training samples, i.e., each decision tree is trained by the same samples. Although the prediction accuracy of a single decision tree is low, multiple decision trees can be combined to get an accurate model, so the generalization of ETR is better than that of RF in some cases. Agrawal *et al.* used the ETR model to predict the HOMO values of donor compounds, and the results indicated that the ETR model outperformed other ML algorithms.[36] (4) GBRT is an iterative decision tree algorithm, consisting of multiple regression trees. Each tree needs to learn the results of all previous trees and the residuals, and the results of all the regression trees are summed up to get the final result. This algorithm can process various types of data and is very robust to abnormal data. Zou *et al.* constructed two different GBRT models based on different numbers of descriptors, showing excellent predictive accuracy, to explain the complex relationship between Y6's OSCs PCE and donor characteristic descriptors.[37]

When predicting the performance of OSCs by ML, four parameters, including the coefficient of determination (R^2), Pearson correlation coefficient (r), root mean square error (RMSE), and mean absolute error (MAE), were chosen to evaluate the prediction accuracy of the ML model. In addition, to avoid the chance arising from simple data alignment, which has an impact on the final prediction ability of the model, the tenfold cross-validation method was applied during model training to improve the model generalization ability.

C. Descriptor sets

Apart from dividing the training and test sets, the most important impact on the prediction performance of the ML model is the selection of descriptors, i.e., input features, in the database. Therefore, when selecting a descriptor, it is necessary to consider whether the selected descriptor can accurately and comprehensively contain the key information of OSCs. Ma *et al.* predicted device performance parameters by filtering out well behaved descriptors as input through ML models and also designed new molecules for active layer materials with high PCE.[18,38,39] Therefore, the selection of descriptors is crucial in the training process of ML. We have chosen Fp-D, Fp-A, D/A ratio, and RMSR as descriptors for training the model. The descriptors for the ML study on OSCs can be categorized as molecular structure-like descriptors[18,36,40,41] and molecular property-like descriptors.[38,42] Based on the constructed database, the D/A ratio, RMSR, Fp-D, and Fp-A are defined as the descriptor set for characterization. This combination of descriptors has not been reported in the previous work to our knowledge.

In OSCs, the D/A ratio refers to the relative content of the donor and acceptor materials in the active layer.[43] The D/A ratio has a significant impact on the performance of OSC. An appropriate D/A ratio can regulate the processes of light absorption, charge separation, and transport, leading to efficient energy conversion. Additionally, the right D/A ratio can also influence the long-term stability and durability of OSCs. Therefore, choosing an appropriate D/A ratio is crucial in the design of OSCs.

Moderate RMSR can indeed increase the effective surface area of OSC, thereby enhancing light absorption.[44,45] Additionally, a suitable RMSR can help to optimize OSCs. The optimal RMSR can improve the contact efficiency between the electrode and active layer, facilitating efficient electron transfer and collection. Therefore, RMSR can serve as important descriptors for OSCs.

Molecular fingerprinting was first proposed in 1987 to enable the description of molecular structures.[39] Zhao *et al.* completed the generation and prediction of convolutional neural network models using strings as input and designed non-fullerene acceptor

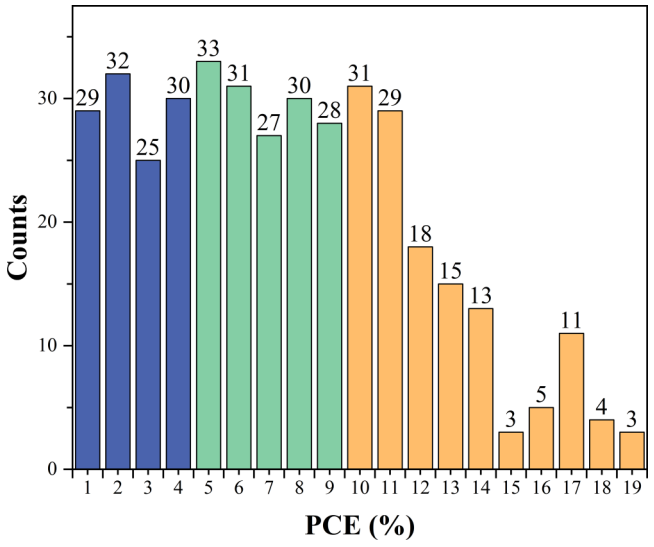


FIG. 1. Experimental PCE distribution of non-fullerene based organic solar cells (OSCs) in the data set.

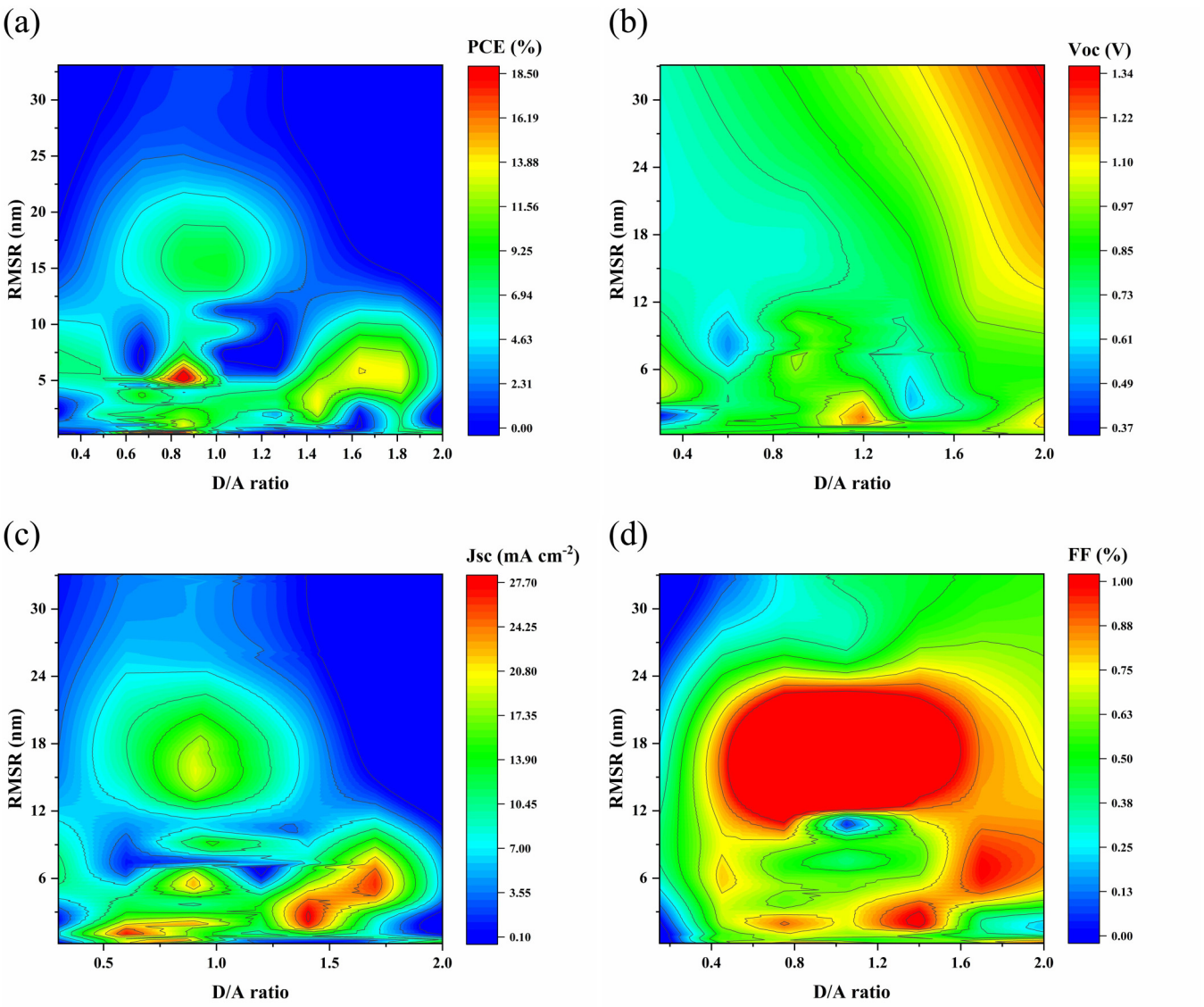


FIG. 2. Contour color fill plot of the device parameters: (a) PCE (%), (b) V_{OC} (V), (c) J_{SC} (mA/cm^2), and (d) FF with a donor–acceptor weight ratio (D/A ratio) and root mean square (RMS) roughness for the 397 donor–acceptor pairs.

materials, which were finally further validated by theoretical calculations.[46] Troisi *et al.* found a moderate correlation between Daylight fingerprints and Morgan fingerprints with the OSC parameters.[20,47] Sun *et al.* found that molecular fingerprints of different lengths predicted PCE differently, while the results showed that the prediction accuracy of molecular fingerprints around 1000 bits was the best.[19] In 2020, Troisi *et al.* described the molecular structure through Morgan fingerprints and found that it contains information that some new property descriptors can represent.[17] Since molecular fingerprints indicate the presence or absence of specific molecular substructures using binary values of "0" and "1," they provide more accurate and comprehensive information about a molecule's description. Therefore, Morgan fingerprints are considered important descriptors as they effectively capture and describe molecular structural information. In recent studies, there have been many successful attempts in using descriptors such as OSCs molecular fingerprints to train ML models for prediction purposes. Sun *et al.* developed an automated design framework using La FREMD Fingerprint and ML algorithms and successfully predicted candidate materials with a PCE exceeding 15%.[48] Sharma *et al.* trained GBRT and other ML

TABLE I. The corresponding determination coefficients (R^2), Pearson's correlation coefficients (r), root means square errors (RMSE), and means absolute errors (MAE) of the predicted device parameters PCE (%), V_{OC} (V), J_{SC} (mA/cm^2), and FF (%) using four ML algorithms on the train set (317 data points) and the test set (80 data points). The tenfold cross validation was performed on the train set. The data outside and inside parentheses are the results predicted using the test set and the training set, respectively.

Model	Device parameters	R^2	R	RMSE	MAE
RF	PCE	0.69 (0.66)	0.83 (0.81)	2.51 (2.45)	2.10 (1.96)
	V_{OC}	0.56 (0.40)	0.75 (0.64)	0.09 (0.09)	0.06 (0.07)
	J_{SC}	0.61 (0.66)	0.79 (0.82)	4.16 (3.61)	3.25 (2.82)
	FF	0.50 (0.46)	0.73 (0.69)	9.21 (9.10)	7.54 (7.17)
AdaBoost	PCE	0.61 (0.65)	0.78 (0.81)	2.80 (2.62)	2.16 (1.89)
	V_{OC}	0.66 (0.42)	0.82 (0.68)	0.07 (0.09)	0.05 (0.06)
	J_{SC}	0.62 (0.65)	0.79 (0.81)	4.10 (3.61)	3.24 (2.75)
	FF	0.47 (0.47)	0.71 (0.69)	9.50 (9.07)	7.34 (6.83)
ETR	PCE	0.68 (0.62)	0.83 (0.80)	2.53 (2.59)	2.11 (1.99)
	V_{OC}	0.56 (0.35)	0.77 (0.62)	0.08 (0.10)	0.06 (0.07)
	J_{SC}	0.62 (0.63)	0.79 (0.80)	4.11 (3.80)	3.18 (2.87)
	FF	0.50 (0.41)	0.73 (0.66)	9.27 (9.36)	7.52 (7.27)
GBRT	PCE	0.63 (0.64)	0.79 (0.80)	2.74 (2.57)	2.23 (2.01)
	V_{OC}	0.62 (0.40)	0.82 (0.67)	0.08 (0.09)	0.06 (0.06)
	J_{SC}	0.64 (0.66)	0.80 (0.82)	4.03 (3.67)	3.07 (2.79)
	FF	0.48 (0.48)	0.71 (0.70)	9.47 (8.90)	7.64 (6.88)

models using frontier molecular orbitals, optical bandgap, and MACCS molecular fingerprints, achieving a Pearson correlation coefficient of 0.85 for prediction accuracy.[49] In this work, we utilized the D/A ratio, RMSR, and 1024-bit Morgan molecular fingerprints to train RF, GBRT, ETR, and AdaBoost models. Additionally, the Morgan molecular fingerprint for polymer in this study is represented as that of one repeating unit because the property difference among polymers is mainly determined by their repeating units. The dangling bonds are saturated using hydrogen atoms.

The data collection for the D/A ratio and RMSR was conducted from published articles of OSCs. The acquisitions of Fp-D and Fp-A were done in the Python environment. The molecular structure files were imported into Open Babel[50] software and converted into SMILES strings. Then, the SMILES strings were saved to the data set and used in the conversion from SMILES string to 1024-bit Morgan molecular fingerprint. The conversions were performed by utilizing the Rdkit.[51] Moreover, the ML in this work was performed using the Scikit-learn[52] package.

III. RESULTS AND DISCUSSION

A. Understanding of the database

Figure 1 gives the experimental PCE distribution of non-fullerene acceptors based OSCs in the data set. To facilitate the predictive power of the ML model, the database was classified into high performance, medium performance, and low performance according to the PCE. For the whole database, the minimum PCE was 0.03%, the maximum PCE was 18.50%, and the average PCE was 6.95%. Furthermore, to obtain an unbiased model, we choose PCE thresholds of 4% and 9% as criteria to classify low, medium, and high performance. That is, PCE < 4% for low performance, 4% ≤ PCE < 9% for medium performance, and PCE ≥ 9% for high performance. The data sorted in ascending order based on PCE was divided into three equal parts, and then the nearest integer at the boundary points was selected as the PCE thresholds, corresponding to 4% and 9%.

The analyzing RMSR and D/A ratio distribution can offer guidance for OSC manufacturing. By observing the shaded regions in the graph, we can easily and visually understand the relationship between the D/A ratio, RMSR, and PCE. Figure 2 shows the RMSR and the D/A ratio correlated with OSC performance parameters PCE, V_{OC}, J_{SC}, and FF. Figure 2(a) shows the RMSR value in the range of 4–7 nm and the D/A ratio in the range of 0.75–0.92 correspond to high PCE. The PCE can reach more than 15%. Figure 2(b) indicates that the D/A ratio in the range of 1.1–1.3 and the RMSR value at 1–3 nm correspond to V_{OC} about 1.01–1.35 V. As shown in Fig. 2(c), the larger J_{SC} correspond to the range from RMSR value in 1–2 nm and D/A ratio in 0.56–0.68, as well as RMSR value in 1–8 nm and D/A ratio in 1.35–1.75. It can be found from Fig. 2(d) that the high FF corresponds to RMSR value in 10–23 nm with the D/A ratio of 0.45–1.70, and also the RMSR value in 1.5–5.0 nm with a D/A ratio of 1.25–1.50, as well as RMSR value in 5–10 nm with a D/A ratio of 1.6–2.0. Therefore, it helps determine the range of values for D/A ratio and RMSR. The contour color fill plot shown in this study is generated using real experimental data and provides practical guidance. These can be approved by experimental studies. For example, Zhou *et al.* reported that when RMSR value = 1.52 nm and D/A ratio = 1.2, the corresponding PCE, V_{OC}, J_{SC}, and FF are 8.64%, 1.21 V, 11.36 mA/cm^2, and 62.86%, respectively.[53]

B. Performance of machine learning models

The optimal hyperparameter adjustment is performed by selecting the hyperparameters that have the greatest influence on the ML model, which is helpful to avoid the overfitting of the ML model and, thus, improve model generalization, thus achieving the purpose of improving prediction accuracy. The optimally tuned

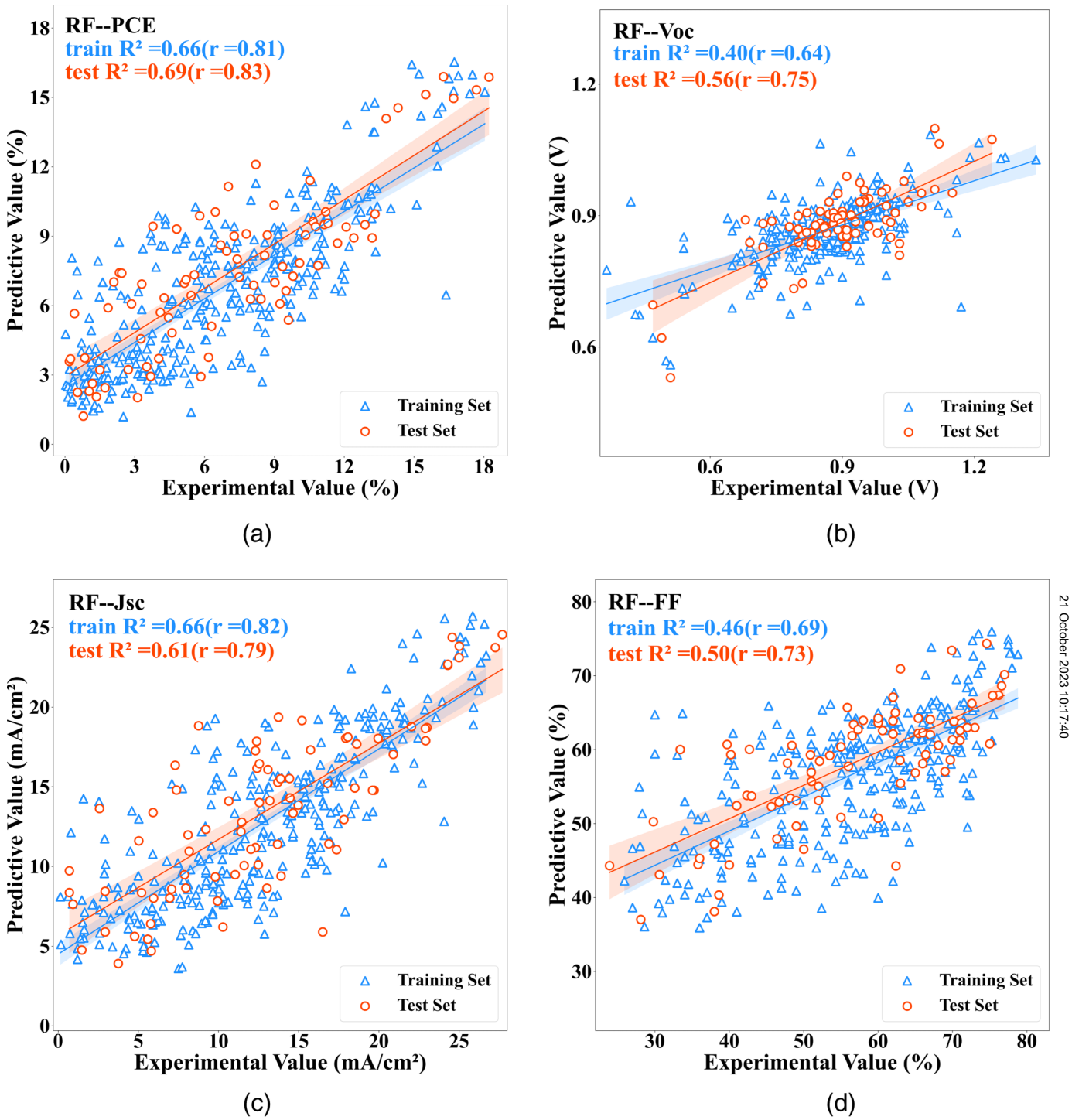


FIG. 3. The device parameters PCE, V_{OC}, J_{SC}, and FF were predicted using the selected RF algorithm on the training set (317 data points) and test set (80 data points), respectively. The blue indicates the prediction result of the training sets, whereas the red indicates the prediction results of the test sets. The corresponding coefficient of determination (R^2) and Pearson's correlation coefficient (r) were given in the upper left corner. The blue and orange areas indicate the error range of the corresponding fitted lines.

TABLE II. Instance predictions for donor–acceptor pairs in different regions of the database.

Performance regions	Donor	Acceptor	Experimental PCE (%)	Predictive PCE (%)	Reference
Low	PBDB-T	th-TDI	2.47	3.06	54
	PTB7-Th	F2B-T2PDI	2.51	2.58	55
	PTB7-Th	t-BPTI-3	3.68	2.99	56
Medium	PBDB-T	T4B-SePDI4	5.10	5.85	57
	J61	IDT-IC	6.95	8.26	58
	PTB7-Th	para-TrBRCN	8.29	7.50	59
High	PM6	NOCIC	9.89	10.32	60
	P2F-Ehp	QIP-4Cl	13.30	13.50	61
	PM6	BTP-C6Ph	15.50	15.61	62

hyperparameters for RF, AdaBoost, ETR, and GBRT ML models are given in Table S1 in the supplementary material. ML models with optimally tuned hyperparameters are used in the subsequent work.

Table I lists the prediction scores R^2, r, RMSE, and MAE of RF, AdaBoost, ETR, and GBRT models. Because R^2 indicates the goodness of fit between the predicted and experimental values, and the more the R^2 value tends to 1, the better the fit is. On the contrary, the closer the R^2 value tends to 0, the worse the fit is. According to the Table I results, the highest R^2 value for PCE is given by the RF model. For V_{OC}, the AdaBoost model provides the best R^2 values. For J_{SC}, the R^2 of the GBRT model is larger than that of other models on the test set, and the R^2 of the GBRT model is the same as that of the RF model on the training set. As to FF, RF and ETR models scored the highest R^2 value of 0.50 on the test set. The large r value indicates a strong correlation between the experimental and predicted results of photovoltaic parameters. The RF model gives the largest r of PCE and FF, while the AdaBoost model generates the largest r of V_{OC} and the GBRT model provides the largest r of J_{SC}. The RMSE measures the degree of deviation between the predicted and experimental values, and the smaller RMSE indicates the higher accuracy of the ML model. Similarly, MAE indicates the absolute difference between the predicted and experimental value, and the small MAE indicates the high accuracy of the ML model. In terms of RMSE and MAE, the performances of these ML models are similar to those of measures using R^2 and r.

To further facilitate the interpretation of ML model performances, the scatterplots of the photovoltaic parameters for the training and test sets are given in Fig. 3. The results of other ML models are given in Fig. S1 in the supplementary material. In Fig. 3, the data in training and test sets are presented as blue triangles and orange circles, respectively, with each triangle or circle representing a data point. The light-colored range around fit line indicates the error range. Based on the slope and the range of errors of the fit line, the RF model trained on the descriptor set is the best model among the four selected models for predicting PCE.

To check the reliability of the ML model, three donor–acceptor pairs are randomly selected from each region of three predefined performance regions (low, medium, and high) in the constructed database, and the experimental values are compared with the predicted results of the RF model in which the experimental RMSR and D/A ratio were adopted. The results are shown in Table II. As shown in the table, in the low performance region, the PCE errors between the experimental and predicted values for the three devices are 0.59%, 0.07%, and 0.69%, respectively. In the medium performance region, the PCE errors between the experimental and predicted values for the three devices are 0.75%, 1.31%, and 0.79%, respectively. In the high-performance region, the PCE errors between the experimental and predicted values for the three devices are 0.43%, 0.20%, and 0.11%, respectively. Therefore, it can be inferred that the trained RF model exhibits strong predictive accuracy.

To verify the generalization ability of the trained RF model, we selected three donor–acceptor pairs PM6:L8-BO, PBDB-TF:BTP-ec9, and PBDB-TF:AQx-2 beyond the database of this work. To determine if the selected structure exists in the database, we search the SMILES string of the selected three donor–acceptor pairs within the database to ensure that the donor–acceptor pairs do not present in the database. Furthermore, the principal component analysis method is also an option. Through the analysis of principal components, it can assist researchers in understanding the relative substructure relationships within the chemical space. The experimental and predictive PCE are provided in Table III. The D/A ratio, RMSR, molecular structures, and photovoltaic parameters were obtained from the corresponding experimental references. In terms of the data in Table III, the errors of PCE between experiment and prediction are 1.53%, 1.06%, and 1.56% for PM6:L8-BO, PBDB-TF:BTP-ec9, and PBDB-TF:AQx-2 OSCs, respectively, which are smaller than RMSE and MAE of PCE from the RF model (Table I). The small predicted PCE errors confirm the good generalization ability of the trained RF model.

C. New molecular design and performance prediction

The OSCs composed of Y-series acceptors show excellent photovoltaic performances since their inception.[66–70] For example, for

TABLE III. The high-performance donor and acceptor materials for OSCs.

Donor	Acceptor	Experimental PCE (%)	Predictive PCE (%)	Reference
PM6	L8-BO	18.32	16.79	63
PBDB-TF	BTP-ec9	17.80	16.74	64
PBDB-TF	AQx-2	16.64	15.08	65

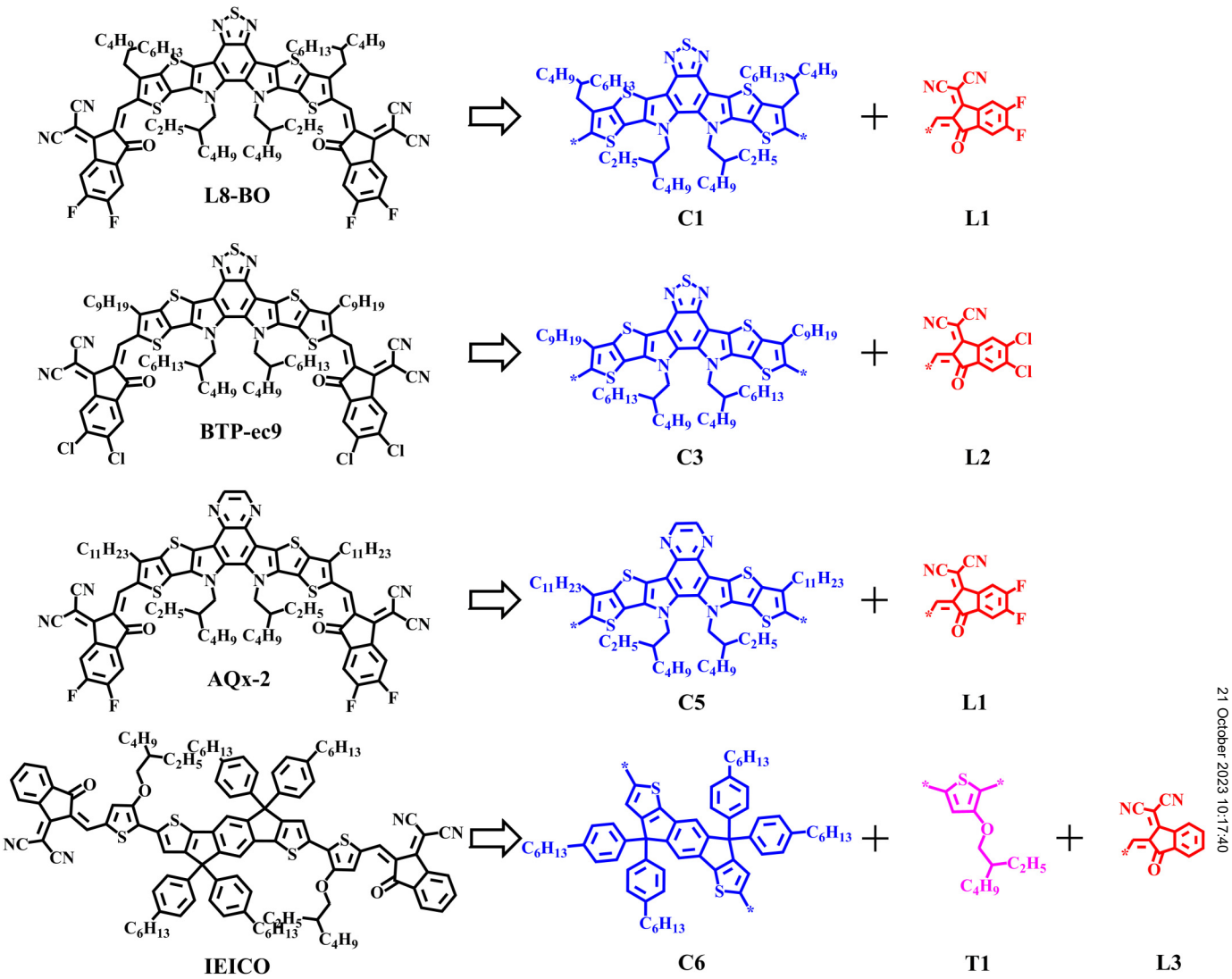


FIG. 4. Molecular structure of high-performance non-fullerene acceptors and structural fragments.

OSCs which consist of PM6/Y6 as an active layer, the PCE reached 18.04%.[71] Furthermore, the acceptor molecules we used to demonstrate the RF model generalization ability are also based on Y-series acceptors, and the OSCs fabricated with corresponding donors exhibit high PCE. Hence, developing novel electron donor and acceptor materials plays the critical role in improving PCE.

Due to the local character of the chemical bond, the molecular design usually conducts with the aid of a combination of molecular fragments which are building blocks. The high-performance non-fullerene acceptors L8-BO, BTP-ec9, AQx-2, and IEICO were selected for fragmentation in order to design non-fullerene acceptors. L8-BO, BTP-ec9, and AQx-2 are Y-series acceptors. Although the IEICO acceptor does not belong to the Y-series, it enriches the abundance of the fragment for molecular design.[72] Considering the A–D–A architecture of L8-BO, BTP-ec9, and AQx-2 and A–π–D–π–A architecture of IEICO, the molecular structural decomposition is depicted in Fig. 4. Hence, we obtained four central donor fragments C1, C2, C3, and C4, three acceptor fragments L1, L2, and L3, and one π-spacer T1 for molecular design. The C1 and C2 units are the same when their different side chains are ignored. The side chains can influence the physicochemical properties of molecules in different ways, and their design and manipulation for donors and acceptors in OSCs are important for optimizing photovoltaic performance. Through random combinations of donor, acceptor, and π-spacer fragments, we designed 20 acceptor molecules with symmetric A–D–A and A–π–D–π–A

TABLE IV. The RF predicted photovoltaic parameters of the designed NFAs with P3HT and PM6 as donor materials.

Donor	Acceptor	PCE (%)	V_{OC} (V)	J_{SC} (mA/cm^2)	FF (%)
P3HT	A1	13.01	0.81	23.82	63.65
	A3	12.98	0.82	24.00	65.13
	A5	11.88	0.81	22.40	62.88
	A8	12.96	0.78	23.55	63.93
	A9	13.09	0.78	23.45	64.20
	A10	6.19	0.82	10.35	54.83
	A11	12.90	0.78	23.77	64.40
	A12	12.96	0.78	23.63	64.37
	A14	11.77	0.78	22.05	62.39
	A15	11.88	0.78	21.99	62.67
PM6	A1	16.23	0.85	24.88	74.78
	A3	16.10	0.84	24.99	75.46
	A5	14.89	0.84	23.53	72.63
	A8	16.07	0.79	24.61	73.57
	A9	16.12	0.79	24.43	73.80
	A10	10.94	0.85	15.68	60.61
	A11	16.01	0.79	24.87	73.76
	A12	15.96	0.79	24.67	73.76
	A14	14.68	0.79	23.23	71.05
	A15	14.69	0.80	23.12	71.59

architectures. Figure S2 in the supplementary material shows 20 molecular structures of the designed non-fullerene acceptors, coded as A1–A20. Searching the SMILES string of the designed 20 acceptor molecules in the database confirms that the designed acceptors are new.

To predict the photovoltaic performance of the designed new acceptors, it is necessary to select donor molecules to form photovoltaic heterojunctions. Thus, two different donor materials, P3HT and PM6, were selected. For P3HT, it has the potential to be used in high-performance OSCs due to its better efficiency, ease of synthesis, low cost, and favorable industrialization.[73] Similarly, PM6 is a highly efficient OSCs donor material, and the PCE of PM6/Y6 OSC achieved over 18%.[71]

In order to predict the photovoltaic performance of OSCs that contain 20 new acceptors and P3HT/PM6 donors, it is required to determine the D/A ratio, RMSR, Fp-D, and Fp-A descriptors. The optimal parameters of RMSR and D/A ratio can be obtained through the analysis of the parameter distribution of high-performance OSCs in the contour color fill plot of Fig. 2. The RMSR and D/A ratio were set as 4 nm and 0.9 for ML prediction, respectively. Furthermore, Fp-D and Fp-A are, respectively, represented as 1024-bit Morgan fingerprints for the donor and acceptor, which were then used as inputs. Considering the good performance of the RF model, it was selected for predicting photovoltaic

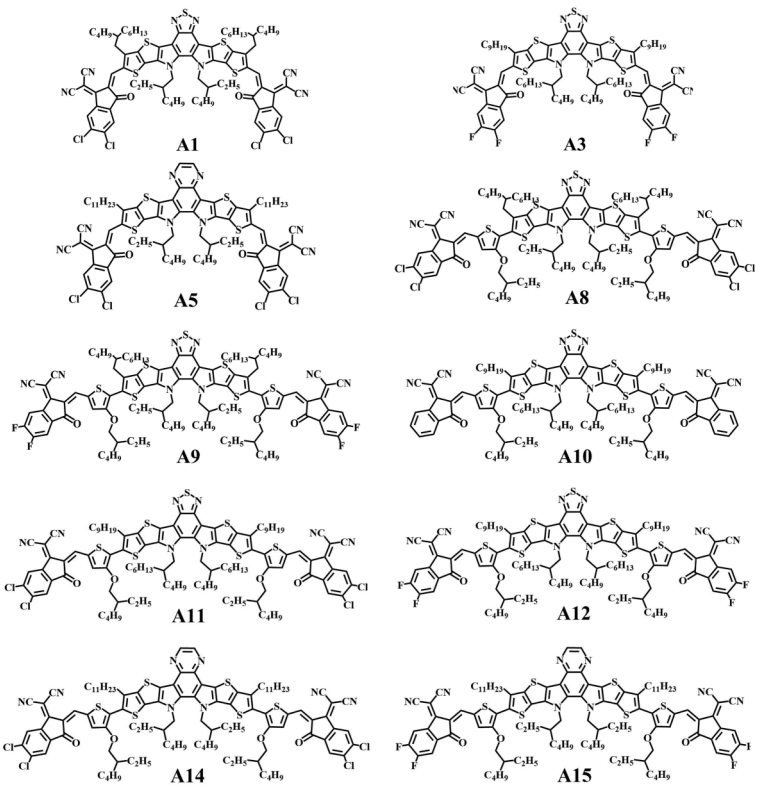


FIG. 5. The selected 10 molecular structures with high predicted PCE among 20 designed NFAs.

parameters. On the basis of the trained RF model, the predicted photovoltaic parameters are listed in Table S2 in the supplementary material. According to the predicted PCE, ten of the designed acceptors with high PCE and their corresponding predicted photovoltaic parameters are given in Table IV, and the corresponding molecular structures are presented in Fig. 5. When P3HT was adopted as a donor, six designed acceptors (A1, A3, A8, A9, A11, and A12) can reach the predicted PCE higher than 12%. While PM6 was applied as donor, five designed acceptors (A1, A3, A8, A9, and A11) can achieve the predicted PCE higher than 16%. Therefore, the potential high-performance acceptors for OSCs can be obtained through the virtual screening based on PCE predicted by the trained ML model.

IV. CONCLUSIONS

In summary, we constructed a database of 397 donor–acceptor pairs of OSCs along with device performance parameters and descriptor sets that include D/A ratio, RMSR, Fp-D, and Fp-A. The ML models RF, AdaBoost, ETR, and GBRT were trained based on the descriptor set for predicting PCE, V_{OC}, J_{SC}, and FF. The metrics R^2, r, RMSE, and MAE were selected to evaluate the prediction performance of ML models. The results showed that the RF model had the highest accuracy and stability for PCE prediction, with the r values of 0.83 and 0.81 on the test and training sets, respectively. Furthermore, based on the molecular decomposition of high-performance non-fullerene acceptors L8-BO, BTP-ec9, AQx-2, and IEICO, 20 acceptor molecules with symmetric A–D–A and A–π–D–π–A architectures were designed by combination of molecular donor, acceptor, and π-spacer fragments. The photovoltaic parameters of the designed acceptors combined with donors P3HT and PM6 were predicted using the trained RF model, and the virtual screening of designed acceptors was conducted based on the predicted PCE. The results indicate that six designed acceptors can reach the predicted PCE higher than 12% when P3HT was adopted as a donor. While PM6 was applied as a donor, five designed acceptors can achieve the predicted PCE higher than 16%. This work contributes to accelerate the developing of new materials for OSCs and enables the design and virtual screening of acceptor molecules with high potential performance.

SUPPLEMENTARY MATERIAL

See the supplementary material for further detailed analysis on data and analysis of other ML models.

ACKNOWLEDGMENTS

This work was supported by the National Natural Science Foundation of China (NNSFC) (Grant Nos. 11964016 and 12264025). The authors are grateful for the help of Mr. Bing Yang and Misses Hai-Yuan Yu, Jin-Hong Li, and Li Ma.

AUTHOR DECLARATIONS

Conflict of Interest

The authors have no conflicts to disclose.

Author Contributions

Cai-Rong Zhang: Conceptualization (equal); Formal analysis (equal); Funding acquisition (equal); Investigation (equal); Methodology (equal); Project administration (equal); Resources (equal); Supervision (equal); Validation (equal); Writing – review & editing (equal). **Ming Li:** Data curation (equal); Investigation (equal); Visualization (equal); Writing – original draft (equal). **Miao Zhao:** Data curation (equal); Investigation (equal); Visualization (equal); Writing – original draft (equal). **Ji-Jun Gong:** Formal analysis (equal). **Xiao-Meng Liu:** Formal analysis (equal). **Yu-Hong Chen:** Formal analysis (equal). **Yu-Hong Chen:** Formal analysis (equal). **You-Zhi Wu:** Formal analysis (equal). **Hong-Shan Chen:** Formal analysis (equal); Resources (equal).

DATA AVAILABILITY

The data that support the findings of this study are available within the supplementary material.

REFERENCES

[1]D. J. Burke and D. J. Lipomi, "Green chemistry for organic solar cells," Energy Environ. Sci. **6**, 2053 (2013).
[2]N. Yeh and P. Yeh, "Organic solar cells: Their developments and potentials," Renew. Sustain. Energy Rev. **21**, 421–431 (2013).
[3]J. Gao, G. Yu, J. C. Hummelen, F. Wudi, and A. J. Heeger, "Polymer photovoltaic cells: Enhanced efficiencies via a network of internal donor-acceptor heterojunctions," Science **270**, 1789–1791 (1995).
[4]B. Kan, M. Li, Q. Zhang, F. Liu, X. Wan, Y. Wang, W. Ni, G. Long, X. Yang, H. Feng, Y. Zuo, M. Zhang, F. Huang, Y. Cao, T. P. Russell, and Y. Chen, "A series of simple oligomer-like small molecules based on oligothiophenes for solution-processed solar cells with high efficiency," J. Am. Chem. Soc. **137**, 3886–3893 (2015).
[5]C. Yan, H. Tang, R. Ma, M. Zhang, T. Liu, J. Lv, J. Huang, Y. Yang, T. Xu, Z. Kan, H. Yan, F. Liu, S. Lu, and G. Li, "Synergy of liquid-crystalline small-molecule and polymeric donors delivers uncommon morphology evolution and 16.6% efficiency organic photovoltaics," Adv. Sci. **7**, 2000149 (2020).
[6]L. H. Han, C. R. Zhang, J. W. Zhe, N. Z. Jin, Y. L. Shen, W. Wang, J. J. Gong, Y. H. Chen, and Z. J. Liu, "Understanding the electronic structures and absorption properties of porphyrin sensitizers YD2 and YD2-o-C8 for dye-sensitized solar cells," Int. J. Mol. Sci. **14**, 20171–20188 (2013).
[7]M. Zhang, X. Guo, W. Ma, H. Ade, and J. Hou, "A large-bandgap conjugated polymer for versatile photovoltaic applications with high performance," Adv. Mater. **27**, 4655–4660 (2015).
[8]X. Song, N. Gasparini, M. M. Nahid, S. H. K. Paleti, J.-L. Wang, H. Ade, and D. Baran, "Dual sensitizer and processing-aid behavior of donor enables efficient ternary organic solar cells," Joule **3**, 846–857 (2019).
[9]Q. Liu, Y. Jiang, K. Jin, J. Qin, J. Xu, W. Li, J. Xiong, J. Liu, Z. Xiao, K. Sun, S. Yang, X. Zhang, and L. Ding, "18% efficiency organic solar cells," Sci. Bull. **65**, 272–275 (2020).
[10]Z. Zheng, J. Wang, P. Bi, J. Ren, Y. Wang, Y. Yang, X. Liu, S. Zhang, and J. Hou, "Tandem organic solar cell with 20.2% efficiency," Joule **6**, 171–184 (2022).
[11]R. M. Neal, "Pattern recognition and machine learning," Technometrics **49**, 366 (2007).
[12]O. Isayev, C. Oses, C. Toher, E. Gossett, S. Curtarolo, and A. Tropsha, "Universal fragment descriptors for predicting properties of inorganic crystals," Nat. Commun. **8**, 15679 (2017).
[13]G. Pilania, J. E. Gubernatis, and T. Lookman, "Multi-fidelity machine learning models for accurate bandgap predictions of solids," Comput. Mater. Sci. **129**, 156–163 (2017).

[14]L. Frye, S. Bhat, K. Akinsanya, and R. Abel, "From computer-aided drug discovery to computer-driven drug discovery," Drug Discov. Today Technol. **39**, 111–117 (2021).
[15]L. Liang, M. Liu, C. Martin, J. A. Elefteriades, and W. Sun, "A machine learning approach to investigate the relationship between shape features and numerically predicted risk of ascending aortic aneurysm," Biomech. Model. Mechanobiol. **16**, 1519–1533 (2017).
[16]N. Meftahi, M. Klymenko, A. J. Christofferson, U. Bach, D. A. Winkler, and S. P. Russo, "Machine learning property prediction for organic photovoltaic devices," npj Comput. Mater. **6**, 166 (2020).
[17]Z.-W. Zhao, M. del Cueto, Y. Geng, and A. Troisi, "Effect of increasing the descriptor set on machine learning prediction of small molecule-based organic solar cells," Chem. Mater. **32**, 7777–7787 (2020).
[18]H. Sahu, F. Yang, X. Ye, J. Ma, W. Fang, and H. Ma, "Designing promising molecules for organic solar cells via machine learning assisted virtual screening," J. Mater. Chem. A **7**, 17480–17488 (2019).
[19]Y. Zheng, W. Sun, K. Yang, Q. Zhang, A. A. Shah, Z. Wu, Y. Sun, L. Feng, D. Chen, Z. Xiao, S. Lu, Y. Li, and K. Sun, "Machine learning-assisted molecular design and efficiency prediction for high-performance organic photovoltaic materials," Sci. Adv. **5**, eaay4275 (2019).
[20]D. Padula, J. D. Simpson, and A. Troisi, "Combining electronic and structural features in machine learning models to predict organic solar cells properties," Mater. Horiz **6**, 343–349 (2019).
[21]S. Nagasawa, E. Al-Naamani, and A. Saeki, "Computer-aided screening of conjugated polymers for organic solar cell: Classification by random forest," J. Phys. Chem. Lett. **9**, 2639–2646 (2018).
[22]Y. Wu, J. Guo, R. Sun, and J. Min, "Machine learning for accelerating the discovery of high-performance donor/acceptor pairs in non-fullerene organic solar cells," npj Comput. Mater. **6**, 120 (2020).
[23]A. Mahmood, A. Irfan, and J.-L. Wang, "Machine learning for organic photovoltaic polymers: A minireview," Chin. J. Polym. Sci. **40**, 870–876 (2022).
[24]B. Yang, C. R. Zhang, Y. Wang, M. Zhao, H. Y. Yu, Z. J. Liu, X. M. Liu, Y. H. Chen, Y. Z. Wu, and H. S. Chen, "Open-circuit voltage loss and dielectric constants as new descriptors in machine learning study on organic photovoltaics," Int. J. Quantum Chem. **123**, e27039 (2023).
[25]J. Wang, Y. Cui, Z. Chen, J. Zhang, Y. Xiao, T. Zhang, W. Wang, Y. Xu, N. Yang, H. Yao, X. T. Hao, Z. Wei, and J. Hou, "A wide bandgap acceptor with large dielectric constant and high electrostatic potential values for efficient organic photovoltaic cells," J Am. Chem. Soc. **145**, 13686–13695 (2023).
[26]L. Fu, H. Hu, Q. Zhu, L. Zheng, Y. Gu, Y. Wen, H. Ma, H. Yin, and J. Ma, "Machine learning assisted prediction of charge transfer properties in organic solar cells by using morphology-related descriptors," Nano Res. **16**, 3588–3596 (2023).
[27]G. Han and Y. Yi, "Singlet-triplet energy gap as a critical molecular descriptor for predicting organic photovoltaic efficiency," Angew. Chem. Int. Ed. **61**, e202213953 (2022).
[28]C. Tong, "Refinement strategies for stratified sampling methods," Reliab. Eng. Syst. Saf. **91**, 1257–1265 (2006).
[29]G. Biau and E. Scornet, "A random forest guided tour," Test **25**, 197–227 (2016).
[30]R. E. Schapire and Y. Freund, "A decision-theoretic generalization of on-line learning and an application to boosting," J. Comput. Syst. Sci. **55**, 119–139 (1997).
[31]P. Geurts, D. Ernst, and L. Wehenkel, "Extremely randomized trees," Mach. Learn. **63**, 3–42 (2006).
[32]J. H. Friedman, "Stochastic gradient boosting," Comput. Stat. Data Anal. **38**, 367–378 (2002).
[33]L. Breiman, "Random forests," Mach. Learn. **45**, 5–32 (2001).
[34]J. Liang, S. Xu, L. Hu, Y. Zhao, and X. Zhu, "Machine-learning-assisted low dielectric constant polymer discovery," Mater. Chem. Front. **5**, 3823–3829 (2021).
[35]J. C.-W. Chan and D. Paelinckx, "Evaluation of Random Forest and Adaboost tree-based ensemble classification and spectral band selection for ecotope mapping using airborne hyperspectral imagery," Remote Sens. Environ. **112**, 2999–3011 (2008).
[36]A. Paul, A. Furmanchuk, W. K. Liao, A. Choudhary, and A. Agrawal, "Property prediction of organic donor molecules for photovoltaic applications using extremely randomized trees," Mol. Inf. **38**, e1900038 (2019).
[37]Q. Zhao, Y. Shan, C. Xiang, J. Wang, Y. Zou, G. Zhang, and W. Liu, "Predicting power conversion efficiency of binary organic solar cells based on Y6 acceptor by machine learning," J. Energy Chem. **82**, 139–147 (2023).
[38]H. Sahu, W. Rao, A. Troisi, and H. Ma, "Toward predicting efficiency of organic solar cells via machine learning and improved descriptors," Adv. Energy Mater. **8**, 1801032 (2018).
[39]P. Willett, *Similarity and Clustering in Chemical Information Systems* (John Wiley & Sons, Inc., 1987), Vol. 32, pp. 1–5.
[40]J. Yao, T. Kirchartz, M. S. Vezie, M. A. Faist, W. Gong, Z. He, H. Wu, J. Troughton, T. Watson, D. Bryant, and J. Nelson, "Quantifying losses in open-circuit voltage in solution-processable solar cells," Phys. Rev. Appl. **4**, 014020 (2015).
[41]Y. Cui, H. Yao, J. Zhang, T. Zhang, Y. Wang, L. Hong, K. Xian, B. Xu, S. Zhang, J. Peng, Z. Wei, F. Gao, and J. Hou, "Over 16% efficiency organic photovoltaic cells enabled by a chlorinated acceptor with increased open-circuit voltages," Nat. Commun. **10**, 2515 (2019).
[42]H. Sahu and H. Ma, "Unraveling correlations between molecular properties and device parameters of organic solar cells using machine learning," J. Phys. Chem. Lett. **10**, 7277–7284 (2019).
[43]Y. Kim, S. Cook, S. M. Tuladhar, S. A. Choulis, J. Nelson, J. R. Durrant, D. D. C. Bradley, M. Giles, I. McCulloch, C.-S. Ha, and M. Ree, "A strong regioregularity effect in self-organizing conjugated polymer films and high-efficiency polythiophene:fullerene solar cells," Nat. Mater. **5**, 197–203 (2006).
[44]H.-J. Seo, K.-M. Yoo, M. Song, J. S. Park, S.-H. Jin, Y. I. Kim, and J.-J. Kim, "Deep-blue phosphorescent iridium complexes with picolinic acid N-oxide as the ancillary ligand for high efficiency organic light-emitting diodes," Org. Electron. **11**, 564–572 (2010).
[45]Ö Güllü and A. Türüt, "Photovoltaic and electronic properties of quercetin/p-InP solar cells," Sol. Energy Mater. Sol. Cells **92**, 1205–1210 (2008).
[46]S. P. Peng and Y. Zhao, "Convolutional neural networks for the design and analysis of non-fullerene acceptors," J. Chem. Inf. Model. **59**, 4993–5001 (2019).
[47]D. Padula and A. Troisi, "Concurrent optimization of organic donor–acceptor pairs through machine learning," Adv. Energy Mater. **9**, 1902463 (2019).
[48]W. Sun, Y. Zheng, Q. Zhang, K. Yang, H. Chen, Y. Cho, J. Fu, O. Odunmbaku, A. A. Shah, Z. Xiao, S. Lu, S. Chen, M. Li, B. Qin, C. Yang, T. Frauenheim, and K. Sun, "Artificial intelligence designer for highly-efficient organic photovoltaic materials," J. Phys. Chem. Lett. **12**, 8847–8854 (2021).
[49]P. Malhotra, S. Biswas, F.-C. Chen, and G. D. Sharma, "Prediction of non-radiative voltage losses in organic solar cells using machine learning," Sol. Energy **228**, 175–186 (2021).
[50]M. B. Noel, M. O'Boyle, C. A. James, C. Morley, T. Vandermeersch, and G. R. Hutchison, "Open Babel: An open chemical toolbox," J. Cheminf. **3**, 1–14 (2011).
[51]G. Landrum, "RDKit documentation," Release **1**, 1–79 (2013).
[52]F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, and V. Dubourg, "Scikit-learn: Machine learning in Python," J. Mach. Learn. Res. **12**, 2825–2830 (2011).
[53]T. Dai, P. Lei, B. Zhang, A. Tang, Y. Geng, Q. Zeng, and E. Zhou, "Fabrication of high V(OC) organic solar cells with a non-halogenated solvent and the effect of substituted groups for 'same-A-strategy' material combinations," ACS Appl. Mater. Interfaces **13**, 21556–21564 (2021).
[54]N. Liang, K. Sun, J. Feng, Y. Chen, D. Meng, W. Jiang, Y. Li, J. Hou, and Z. Wang, "Near-infrared electron acceptors based on terrylene diimides for organic solar cells," J. Mater. Chem. A **6**, 18808–18812 (2018).
[55]W. T. Hadmojo, S. Y. Nam, T. J. Shin, S. C. Yoon, S.-Y. Jang, and I. H. Jung, "Geometrically controlled organic small molecule acceptors for efficient fullerene-free organic photovoltaic devices," J. Mater. Chem. A **4**, 12308–12318 (2016).
[56]H. C. Chen, B. H. Jiang, C. P. Hsu, Y. Y. Tsai, R. J. Jeng, C. P. Chen, and K. T. Wong, "The twisted benzo[ghi]-perylenetriimide dimer as a 3D electron

acceptor for fullerene-free organic photovoltaics," Chem. Eur. J. **24**, 17590–17597 (2018).
[57]J. Qu, Z. Mu, H. Lai, M. Xie, L. Liu, W. Lu, W. Chen, and F. He, "Effect of the molecular configuration of perylene diimide acceptors on charge transfer and device performance," ACS Appl. Energy Mater. **1**, 833–840 (2018).
[58]X. Li, T. Yan, H. Bin, G. Han, L. Xue, F. Liu, Y. Yi, Z.-G. Zhang, T. P. Russell, and Y. Li, "Insertion of double bond π-bridges of A–D–A acceptors for high performance near-infrared polymer solar cells," J. Mater. Chem. A **5**, 22588–22597 (2017).
[59]W. Wu, G. Zhang, X. Xu, S. Wang, Y. Li, and Q. Peng, "Wide bandgap molecular acceptors with a truxene core for efficient nonfullerene polymer solar cells: Linkage position on molecular configuration and photovoltaic properties," Adv. Funct. Mater. **28**, 1707493 (2018).
[60]J. Gao, Y. Li, S. Li, X. Xia, X. Lu, M. Shi, and H. Chen, "Non-fullerene acceptors with nitrogen-containing six-membered heterocycle cores for the applications in organic solar cells," Sol. Energy Mater. Sol. Cells **225**, 111046 (2021).
[61]C. Zhu, K. An, W. Zhong, Z. Li, Y. Qian, X. Su, and L. Ying, "Design and synthesis of non-fullerene acceptors based on a quinoxalineimide moiety as the central building block for organic solar cells," Chem. Commun. **56**, 4700–4703 (2020).
[62]G. Chai, Y. Chang, Z. Peng, Y. Jia, X. Zou, D. Yu, H. Yu, Y. Chen, P. C. Y. Chow, K. S. Wong, J. Zhang, H. Ade, L. Yang, and C. Zhan, "Enhanced hindrance from phenyl outer side chains on nonfullerene acceptor enables unprecedented simultaneous enhancement in organic solar cell performances with 16.7% efficiency," Nano Energy **76**, 105087 (2020).
[63]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, "Non-fullerene acceptors with branched side chains and improved molecular packing to exceed 18% efficiency in organic solar cells," Nat. Energy **6**, 605–613 (2021).
[64]Y. Cui, H. Yao, J. Zhang, K. Xian, T. Zhang, L. Hong, Y. Wang, Y. Xu, K. Ma, C. An, C. He, Z. Wei, F. Gao, and J. Hou, "Single-junction organic photovoltaic cells with approaching 18% efficiency," Adv. Mater. **32**, e1908205 (2020).
[65]Z. Zhou, W. Liu, G. Zhou, M. Zhang, D. Qian, J. Zhang, S. Chen, S. Xu, C. Yang, F. Gao, H. Zhu, F. Liu, and X. Zhu, "Subtle molecular tailoring induces significant morphology optimization enabling over 16% efficiency organic solar cells with efficient charge generation," Adv. Mater. **32**, e1906324 (2020).
[66]J. Yuan, T. Huang, P. Cheng, Y. Zou, H. Zhang, J. L. Yang, S. Y. Chang, Z. Zhang, W. Huang, R. Wang, D. Meng, F. Gao, and Y. Yang, "Enabling low voltage losses and high photocurrent in fullerene-free organic photovoltaics," Nat. Commun. **10**, 570 (2019).
[67]J. Yuan, Y. Zhang, L. Zhou, C. Zhang, T. K. Lau, G. Zhang, X. Lu, H. L. Yip, S. K. So, S. Beaupré, M. Mainville, P. A. Johnson, M. Leclerc, H. Chen, H. Peng, Y. Li, and Y. Zou, "Fused benzothiadiazole: A building block for n-type organic acceptor to achieve high-performance organic solar cells," Adv. Mater. **31**, e1807577 (2019).
[68]C. Zhang, J. Yuan, K. L. Chiu, H. Yin, W. Liu, G. Zheng, J. K. W. Ho, S. Huang, G. Yu, F. Gao, Y. Zou, and S. K. So, "A disorder-free conformation boosts phonon and charge transfer in an electron-deficient-core-based non-fullerene acceptor," J. Mater. Chem. A **8**, 8566–8574 (2020).
[69]S. Liu, J. Yuan, W. Deng, M. Luo, Y. Xie, Q. Liang, Y. Zou, Z. He, H. Wu, and Y. Cao, "High-efficiency organic solar cells with low non-radiative recombination loss and low energetic disorder," Nat. Photonics **14**, 300–305 (2020).
[70]M. Luo, L. Zhou, J. Yuan, C. Zhu, F. Cai, J. Hai, and Y. Zou, "A new non-fullerene acceptor based on the heptacyclic benzotriazole unit for efficient organic solar cells," J. Energy Chem. **42**, 169–173 (2020).
[71]Z. Chen, Q. Li, Y. Jiang, H. Lee, T. P. Russell, and Y. Liu, "Multi-site functional cathode interlayers for high-performance binary organic solar cells," J. Mater. Chem. A **10**, 16163–16170 (2022).
[72]C. Wang, Y. Bai, Q. Guo, C. Zhao, J. Zhang, S. Hu, T. Hayat, A. Alsaedi, and Z. Tan, "Enhancing charge transport in an organic photoactive layer via vertical component engineering for efficient perovskite/organic integrated solar cells," Nanoscale **11**, 4035–4043 (2019).
[73]Y. Kim, S. A. Choulis, J. Nelson, D. D. C. Bradley, S. Cook, and J. R. Durrant, "Composition and annealing effects in polythiophene/fullerene solar cells," J. Mater. Sci. **40**, 1371–1376 (2005).

Supplementary Information for

Machine Learning Study on Organic Solar Cells and Virtual Screening of Designed Non-Fullerene Acceptors

Cai-Rong Zhang [a*], Ming Li [a], Miao Zhao [a], Ji-Jun Gong [a], Xiao-Meng Liu [a], Yu-Hong Chen [a], Zi-Jiang Liu [b], You-Zhi Wu [c], Hong-Shan Chen [d]

a Department of Applied Physics, Lanzhou University of Technology, Lanzhou, Gansu 730050, China;

b School of Mathematics and Physics, Lanzhou Jiaotong University, Lanzhou 730070, China;

c School of Materials Science and Engineering, Lanzhou University of Technology, Lanzhou, Gansu 730050, China;

d College of Physics and Electronic Engineering, Northwest Normal University, Lanzhou, Gansu 730070, China.

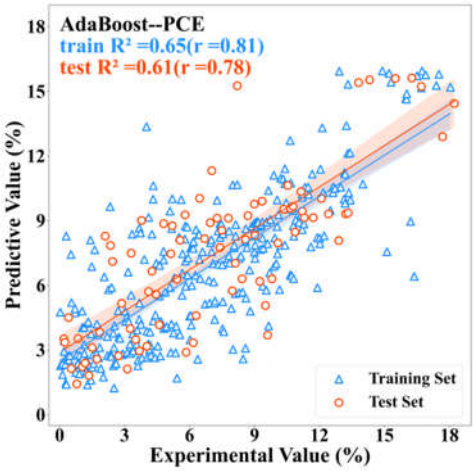
TABLE S1. Adjustment of four machine learning algorithms to obtain the optimal hyperparameters through 10-fold cross validation.

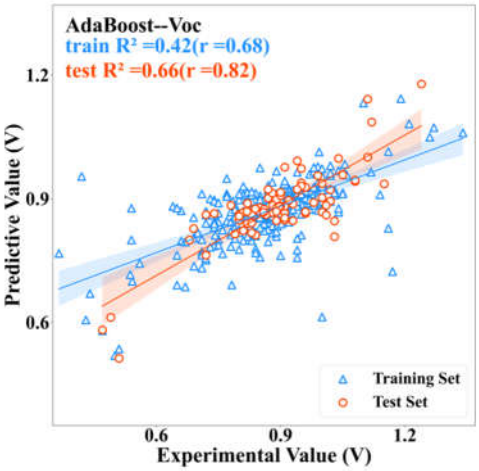
Model	hyperparameters
RF	bootstrap= True max_depth=14 max_features='auto' min_samples_leaf=3 min_samples_split=3 n_estimators=673
AdaBoost	max_depth=29 min_samples_split=13 min_samples_leaf=1 max_features='sqrt' n_estimators=452 learning_rate=0.01
ETR	n_estimators=231 max_features='auto' max_depth=41 min_samples_split=3 min_samples_leaf=3 bootstrap=True
GBRT	n_estimators=540 learning_rate=0.15 max_features='log2' max_depth=37 min_samples_split=5 min_samples_leaf=1

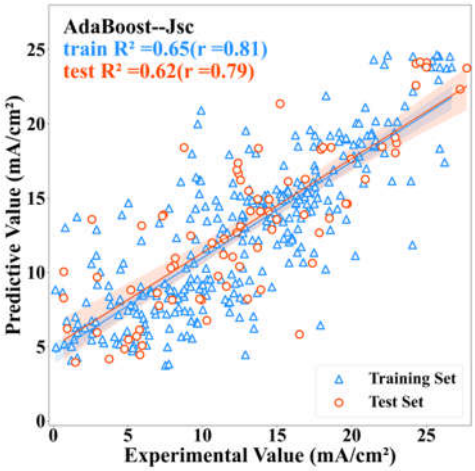
TABLE S2. The RF algorithm with the best prediction accuracy among the four ML algorithms was used to predict the device performance parameters of 20 new acceptor molecules designed with P3HT and PM6 as donor materials.

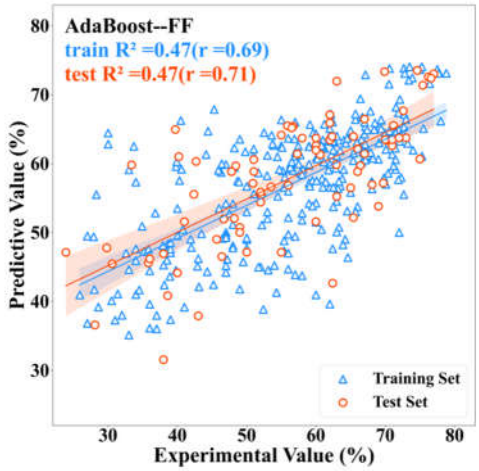
Donor	Acceptor	PCE (%)	V_{OC} (V)	J_{SC} (mA/cm^2)	FF (%)
P3HT	A1	13.01	0.81	23.82	63.65
	A2	6.03	0.84	10.66	51.92
	A3	12.98	0.82	24.00	65.13
	A4	5.94	0.85	10.46	53.74
	A5	11.88	0.81	22.40	62.88

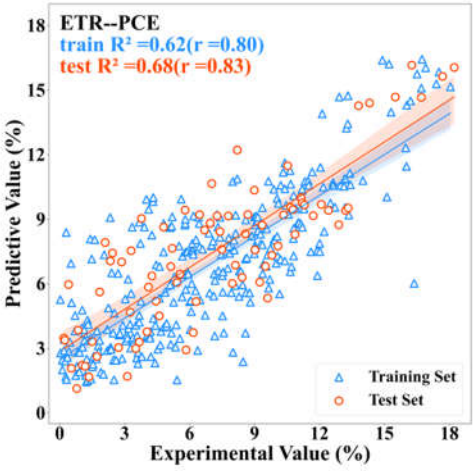
	A6	5.82	0.84	10.34	52.88
	A7	6.23	0.81	10.53	54.39
	A8	12.96	0.78	23.55	63.93
	A9	13.09	0.78	23.45	64.20
	A10	6.19	0.82	10.35	54.83
	A11	12.90	0.78	23.77	64.40
	A12	12.96	0.78	23.63	64.37
	A13	5.97	0.82	10.16	54.12
	A14	11.77	0.78	22.05	62.39
	A15	11.88	0.78	21.99	62.67
	A16	6.48	0.77	15.97	58.74
	A17	5.83	0.77	14.97	57.43
	A18	3.82	0.89	7.57	56.55
	A19	5.91	0.82	15.28	60.42
	A20	5.84	0.82	14.34	58.69
PM6	A1	16.23	0.85	24.88	74.78
	A2	10.77	0.89	15.83	59.27
	A3	16.10	0.84	24.99	75.46
	A4	10.79	0.89	15.83	60.18
	A5	14.89	0.84	23.53	72.63
	A6	10.41	0.88	15.76	58.90
	A7	10.81	0.85	15.59	60.09
	A8	16.07	0.79	24.61	73.57
	A9	16.12	0.79	24.43	73.80
	A10	10.94	0.85	15.68	60.61
	A11	16.01	0.79	24.87	73.76
	A12	15.96	0.79	24.67	73.76
	A13	10.37	0.85	15.46	59.26
	A14	14.68	0.79	23.23	71.05
	A15	14.69	0.80	23.12	71.59
	A16	8.84	0.79	17.36	61.03
	A17	7.77	0.79	16.49	58.71
	A18	6.86	0.93	11.84	57.76
	A19	9.58	0.86	17.01	64.62
	A20	8.33	0.86	16.21	61.91

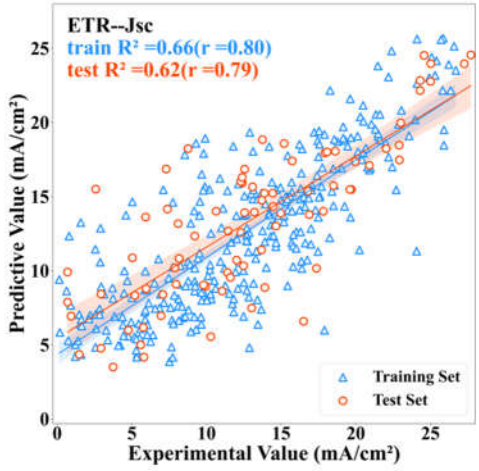

AdaBoost--PCE
train R² =0.65(r =0.81)
test R² =0.61(r =0.78)
Predictive Value (%)
Experimental Value (%)
Training Set
Test Set

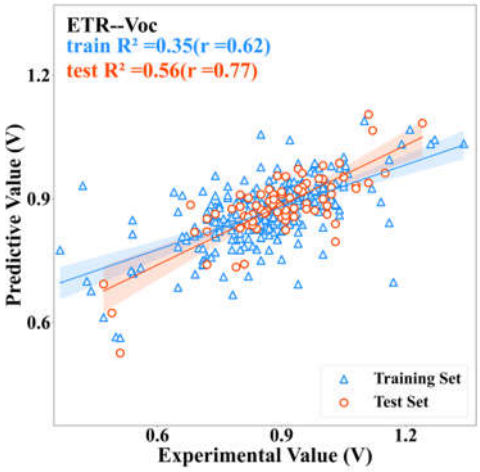

AdaBoost--Voc
train R² =0.42(r =0.68)
test R² =0.66(r =0.82)
Predictive Value (V)
Experimental Value (V)
Training Set
Test Set

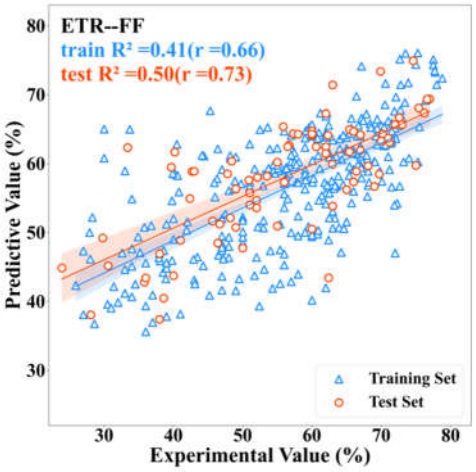

AdaBoost--Jsc
train R² =0.65(r =0.81)
test R² =0.62(r =0.79)
Predictive Value (mA/cm²)
Experimental Value (mA/cm²)
Training Set
Test Set

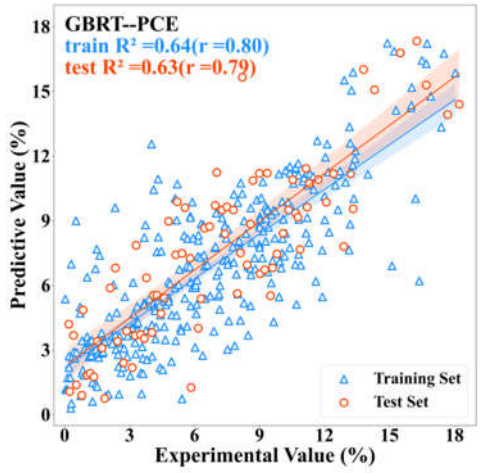

AdaBoost--FF
train R² =0.47(r =0.69)
test R² =0.47(r =0.71)
Predictive Value (%)
Experimental Value (%)
Training Set
Test Set

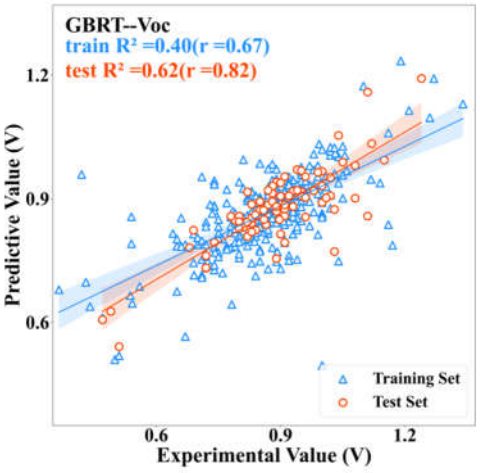

ETR--PCE
train R² =0.62(r =0.80)
test R² =0.68(r =0.83)
Predictive Value (%)
Experimental Value (%)
Training Set
Test Set

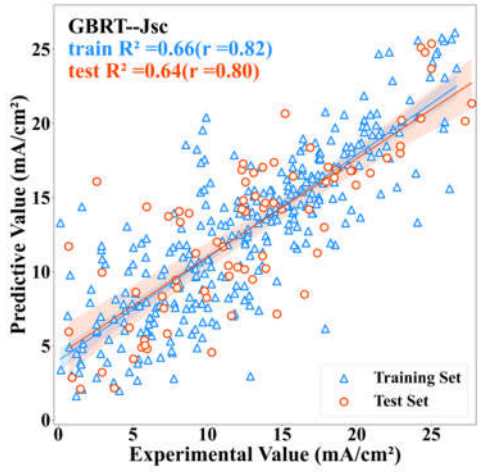

ETR--Jsc
train R² =0.66(r =0.80)
test R² =0.62(r =0.79)
Predictive Value (mA/cm²)
Experimental Value (mA/cm²)
Training Set
Test Set


ETR--Voc
train R² =0.35(r =0.62)
test R² =0.56(r =0.77)
Predictive Value (V)
Experimental Value (V)
Training Set
Test Set


ETR--FF
train R² =0.41(r =0.66)
test R² =0.50(r =0.73)
Predictive Value (%)
Experimental Value (%)
Training Set
Test Set


GBRT--PCE
train R² =0.64(r =0.80)
test R² =0.63(r =0.79)
Predictive Value (%)
Experimental Value (%)
Training Set
Test Set


GBRT--Voc
train R² =0.40(r =0.67)
test R² =0.62(r =0.82)
Predictive Value (V)
Experimental Value (V)
Training Set
Test Set


GBRT--Jsc
train R² =0.66(r =0.82)
test R² =0.64(r =0.80)
Predictive Value (mA/cm²)
Experimental Value (mA/cm²)
Training Set
Test Set

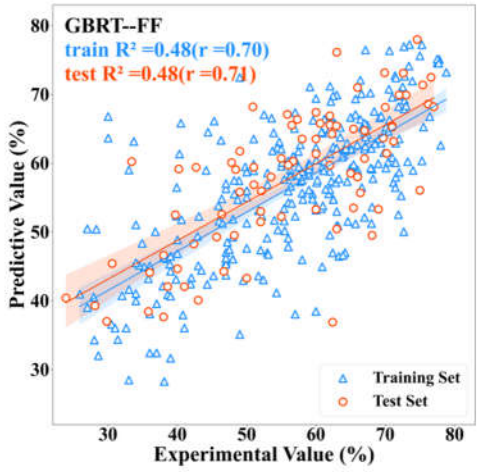
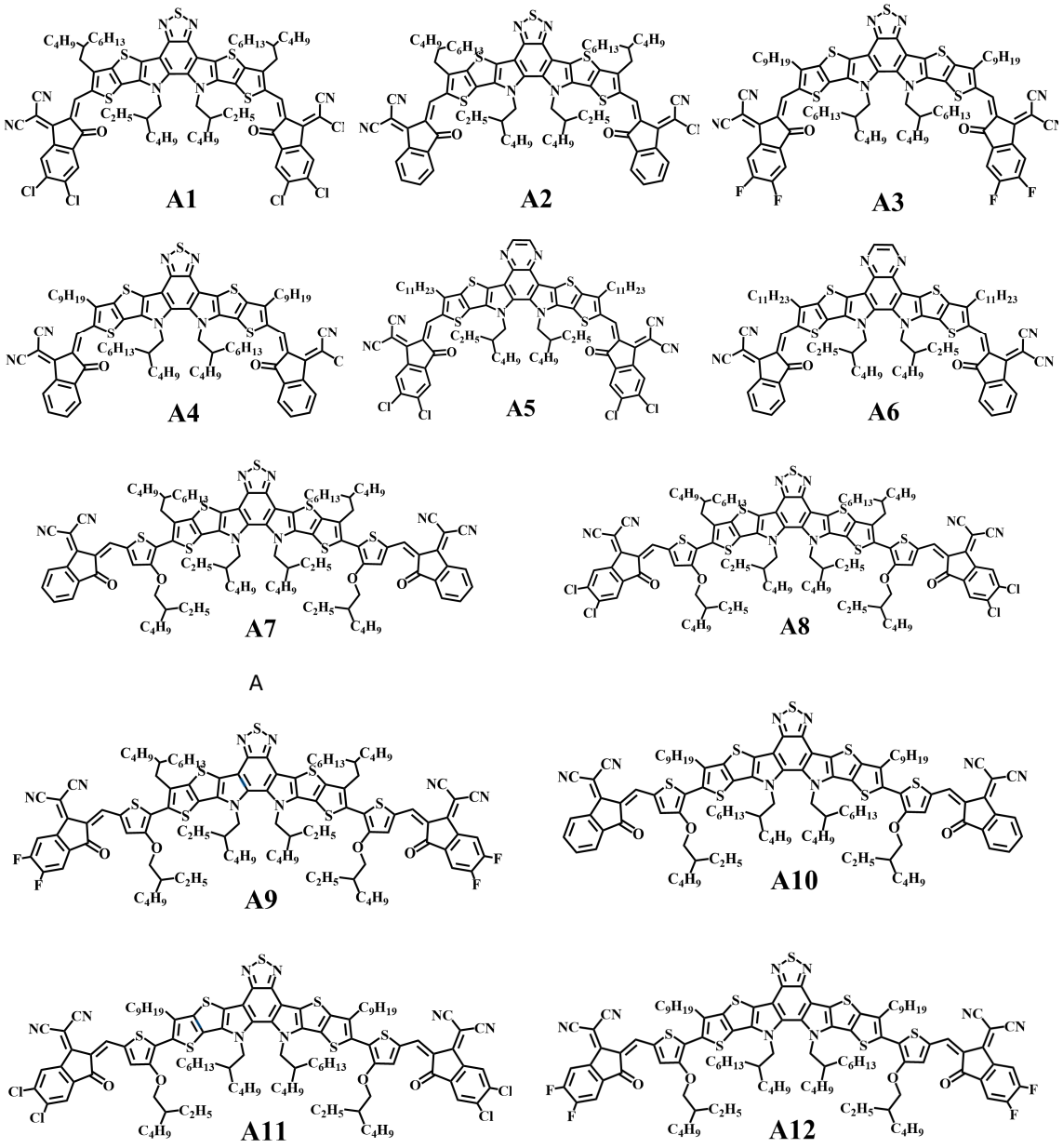

GBRT--FF
train R² =0.48(r =0.70)
test R² =0.48(r =0.71)
Predictive Value (%)
Experimental Value (%)
Training Set
Test Set

FIG. S1. The device parameters PCE, V_{OC}, J_{SC} and FF were predicted using four algorithms on the training set (317 data points) and test set (80 data points). The blue indicates the prediction result of the training sets, whereas the red indicates the prediction results of the test sets. The corresponding the coefficient of determination (R^2) and Pearson's correlation coefficient (r) were given in the upper left corner. The blue and orange areas indicate the error range of the corresponding fitted lines.



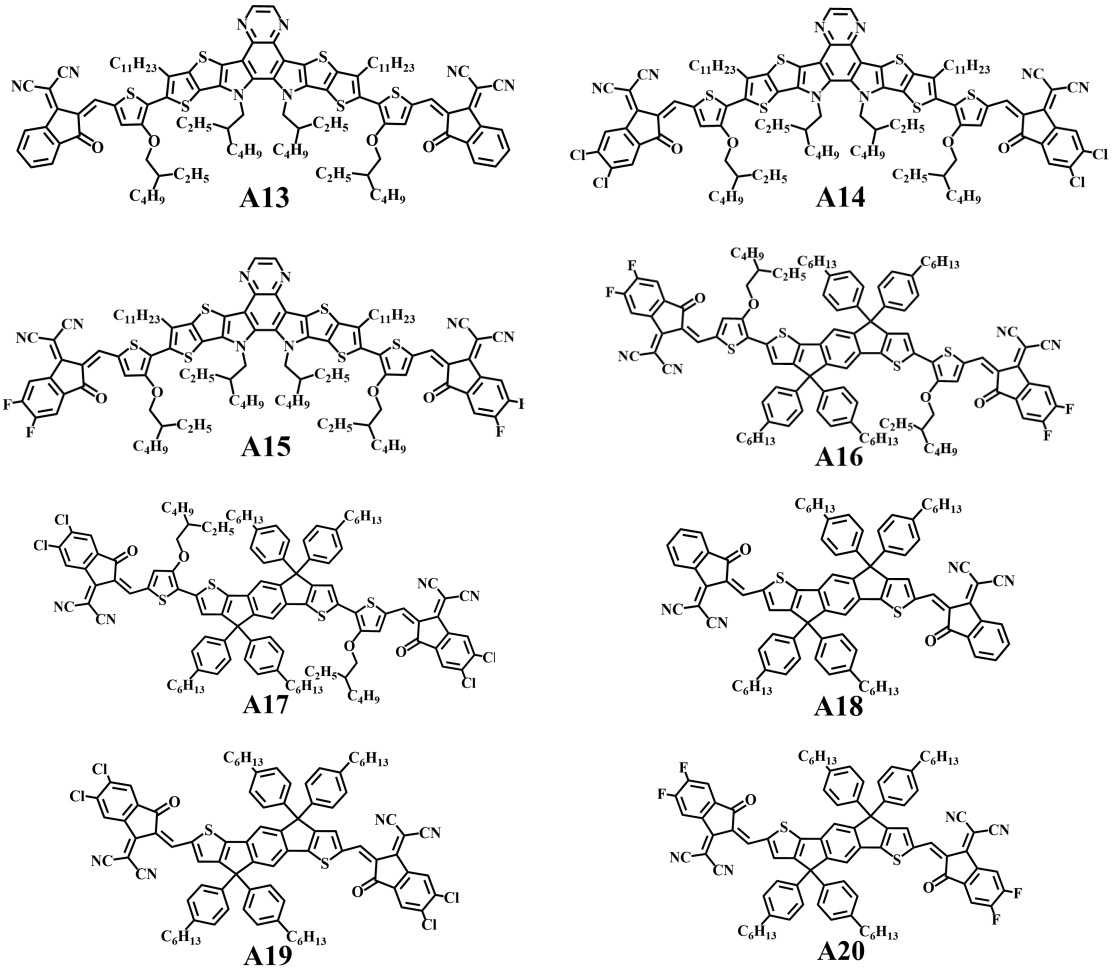


FIG. S2. Molecular structures of the 20 designed NFAs.