Machine Learning Study on the Virtual Screening of Donor–Acceptor Pairs for Organic Solar Cells

Ming Li, Cai-Rong Zhang, Mei-Ling Zhang, Ji-Jun Gong, Xiao-Meng Liu, Yu-Hong Chen, Zi-Jiang Liu, You-Zhi Wu, and Hong-Shan Chen*

The selection of electron donors and nonfullerene acceptors (NFAs) in organic solar cells (OSCs) is crucial for improving photovoltaic performance. Machine learning (ML) has brought a breakthrough solution. Herein, 292 donor-NFA pairs with experimental OSC parameters from the reported articles are collected. The ML descriptors include device processing parameters, molecular properties, and molecular structure. The five ML regression models, random forest (RF), extra tree regression, gradient boosting regression tree, adaptive boosting, and artificial neural network (ANN) are trained. GridSearchCV is used for hyperparameter optimization of ML regression models. The SHapley Additive exPlanation approach is employed to analyze descriptor importance. Among the trained five ML models, the RF model shows superior performance, achieving Pearson's correlation coefficient (*r*) of 0.81 on the test set. Based on the donors and NFAs in constructed dataset, the 9779 donor–NFA pairs for OSCs are generated by randomly combining donor and acceptor molecules. The trained RF model is utilized to predict the power conversion efficiency (PCE) of new donor–acceptor pairs for OSCs. The results indicate that the OSC composed of PBDB-TF as donor and L8-BO as acceptor can achieve the remarkable PCE of 17.9%.

1. Introduction

Clean energy is becoming increasingly important in today's world because traditional methods of energy production, such as coal and oil, produce large amounts of greenhouse gas emissions that lead to climate change and environmental problems.[1,2] Therefore, the search for more sustainable energy solutions has become critical. Organic solar cells (OSCs) as a novel category of solar cells possess numerous unique merits. Compared to traditional solar cells, OSCs have the advantages of being more affordable, highly flexible, and environmental friendly, along with ease of fabrication.[3–9] However, improving OSC performance remains a formidable challenge, as their performances are frequently affected by the intricacy of molecular structures for OSC materials.[10–12]

OSCs have been gaining increasing prominence, with their essential bulk heterojunction (BHJ) structures of active layer.[13–16] The key of BHJ OSC lies in their structure, where donor and acceptor materials are intimately mixed to form a large number of interpenetrating networks. This structure has advantages such as high efficiency, wide optical absorption, strong stability, and multimaterial selection, making it an important route for developing OSC. Therefore, the study of the active layer materials in BHJ OSCs is of paramount importance.[17–19]

Both polymer materials and small-molecule materials can be used as donor materials. The acceptor materials of the active layer are typically divided into fullerene acceptors (FAs) and nonfullerene acceptors (NFAs). In the novel OSC, scientists have shifted their research focus from FA to NFA materials due to the significant potential of NFA materials in expanding absorption range, enhancing photoelectric conversion efficiency, and reducing manufacturing costs. The combination of a donor with NFA usually offers greater design flexibility and superior performance compared to using FAs. The donors and NFA exhibit rich structural diversity, allowing for better topology and energy-level matching.[20–22] However, FA is constrained by its lower electron affinity and limitations in optical and electrical properties.[23–27] Through the relentless efforts of scientists, the highest power conversion efficiency (PCE) of NFA-based OSCs has now surpassed 20%.[28] Therefore, we focus on NFA-based BHJ OSCs.

Finding suitable and efficient donor and acceptor materials often requires a substantial amount of trial-and-error experimentation. Conducting a full experimental procedure is not only time-consuming but also costly, making many research endeavors impractical. However, approaches based on machine learning

M. Li, C.-R. Zhang, M.-L. Zhang, J.-J. Gong, X.-M. Liu, Y.-H. Chen
Department of Applied Physics
Lanzhou University of Technology
Lanzhou, Gansu 730050, China
E-mail: zhcrxy@lut.edu.cn

Z.-J. Liu
School of Mathematics and Physics
Lanzhou Jiaotong University
Lanzhou 730070, China

Y.-Z. Wu
School of Materials Science and Engineering
Lanzhou University of Technology
Lanzhou, Gansu 730050, China

H.-S. Chen
College of Physics and Electronic Engineering
Northwest Normal University
Lanzhou, Gansu 730070, China

The ORCID identification number(s) for the author(s) of this article can be found under https://doi.org/10.1002/pssa.202400008.

DOI: 10.1002/pssa.202400008

(ML) algorithms offer a cost-effective and rapid alternative strategy. This method can quickly and inexpensively screen and predict materials within a short timeframe, without relying solely on experimental procedures.[29–33]

Compared with traditional rule-based computing methods, ML can automatically discover patterns and laws hidden in data by learning from large amounts of data and can make predictions and decisions about new data.[34–39] To design efficient OSC devices, it's essential to identify the key factors that influence their target performance and utilize these factors for model training and performance prediction. ML methods can process different property and structure information into numerical features and establish the quantitative structure–performance relationship (QSPR) model by combining and exploring these numerical features. QSPR models are quantitative models that can predict molecular properties by established relationships between molecular structure and physical–chemical properties. Through QSPR models, we can predict the physical–chemical properties of unknown compounds, providing valuable guidance in fields such as materials science, drug design, environmental science, and more. For example, ML techniques can be employed to predict the optoelectronic properties of OSC materials, effectively aiding in the design and optimization of OSC materials.[40–46]

Over the past few years, numerous research teams have done important work exploring the use of ML for OSCs. In 2018, Saeki and colleagues utilized the extended connectivity fingerprints and molecular access system fingerprints to train artificial neural network (ANN) and random forest (RF) models and demonstrated that the RF model exhibits superior accuracy in predicting PCE.[47] In 2019, Troisi and colleagues computed geometric and electronic property descriptors for 249 donor–acceptor pairs and employed them to train ML models. They separately investigated the influence of molecular properties or structural parameters on the performance of ML models. It was observed that ML models based on molecular property descriptors and those based on structural parameter descriptors yielded similar results. However, when the descriptors used both types simultaneously used to train the ML models, there was a significant improvement in performance.[48] In 2020, Min and colleagues utilized ML to determine the optimal donor–acceptor pairs for OSCs. They trained ML models on a dataset consisting of 565 samples, achieving Pearson's correlation coefficients (r) of 0.71 and 0.70 for enhanced regression tree and RF models, respectively.[49] In recent years, Wang and others reiterated the importance of ML algorithms and discussed their applications in materials design.[45] By incorporating V_{oc} loss and dielectric constants into the descriptor set, our ML model proposed for predicting OSC performance outperforms models based on previous descriptor sets.[50] Ma and colleagues trained a gradient boosting regression model using morphological descriptors to predict charge transfer state energy ($r = 0.96$).[51] Yi and colleagues trained a ML model using three highly correlated molecular property descriptors, namely, the single-to-triple exciton energy gap, optical gap, and driving force, achieving a good prediction accuracy ($r = 0.81$).[52] Although there are more relevant studies, it is still necessary to improve the prediction accuracy of PCE from different descriptors and to develop novel donors and acceptors for OSCs.

In this work, we collected experimental articles on NFA-based OSCs and extracted and screened 292 OSC samples of molecular structures, energy levels, and morphology characterizations, as well as photovoltaic performance parameters. The descriptor set used for training ML model includes the mass ratio of donor-to-acceptor materials in OSC active layer, root mean square of roughness (RMSR), thickness of the active layer, maximum absorption wavelength of acceptors, energy levels, and molecular structures. We conducted feature importance analysis using SHapley Additive exPlanations (SHAP) approach to assess the significance of descriptors on the target property PCE. We used five ML algorithms, RF, extra tree regression (ETR), gradient boosting regression tree (GBRT), adaptive boosting (Adaboost), and ANN, to predict the photovoltaic performance parameters of OSCs.[53–55] According to the model evaluation results, the RF model demonstrated the best predictive capability and stability. The r for the PCE on the training set and test set are 0.78 and 0.81, respectively. Additionally, based on the collected donor and acceptor molecules, we reshuffled and combined them and obtained the 9779 unique donor–acceptor pairs. Using the selected best ML model, we made predictions for these new combinations. Among them, 42 donor–acceptor pairs show the predicted PCE higher than 16%, and three donor–acceptor pairs of them can achieve the predicted PCE greater than 17%.

2. Experimental Section

2.1. Data Collection

In this project, the first step involved data collection. We collected data from the experimental OSC articles published between 2019 and 2021. These articles were searched from the "Web of Science" website using "Organic Solar Cell" and "Bulk Heterojunction" as keywords. Totally, we collected 402 sets of experimental data for NFA OSCs. However, the completeness and value of the gathered data were limited due to the differences in authors' objectives. Some required parameters were not available in the articles. We took several steps to clean, filter, annotate, and prepare the data to ensure its quality. After this series of data cleaning procedures, we ultimately retained 292 sets of different donor–acceptor pairs for OSCs in the constructed database.

2.2. Descriptor Sets Selection

Molecular descriptors in mathematical terms can represent both the physical and chemical properties of molecules. Ma et al. used 13 key molecular features as descriptors to construct RF and GBRT models and obtained a predictive performance with r of 0.7.[56] We use the collected and calculated descriptors as well as the photovoltaic performance parameters PCE, open-circuit voltage (V_{oc}), short circuit current density (J_{sc}), and fill factor (FF) as inputs for training the ML model. In this work, ten kinds of descriptors were selected to be combined into composite descriptors. Compared to molecular property descriptors, the composite descriptor contains more molecular information and features. The descriptors selected in this work can be categorized into three groups: device processing parameter descriptors, molecular property descriptors, and molecular structure descriptors. The composite descriptor includes the following quantities. 1) Blending ratio of active layer donor-to-acceptor material (D/A ratio). A suitable blending ratio can make the active layer film have a wide absorption spectrum as well as good crystallinity and phase separation; 2) RMSR of active layer in OSC, which is a parameter to characterize the degree of surface roughness. The active layer with good nanostructure morphology can improve the effective exciton separation and the efficient transport of charges; 3) Active layer thickness (gauge). The appropriate thickness enables optimal light absorption and electron transport, thereby improving PCE. Thin- or thick-film thickness can lead to reduced light absorption or electron transport obstruction, reducing the performance of the OSCs; 4) The highest occupied molecular orbital energy level (HOMO-D/HOMO-A) and the lowest unoccupied molecular orbital energy level (LUMO-D/LUMO-A) for donor and acceptor. Energy level determines the matching of electron level between the light-absorbing material and the electron transport material, and appropriate energy level regulation can promote the separation and transmission of photogenerated charge; 5) Wavelength corresponding to the absorption peak in acceptor's absorption spectrum (MAX-A). Compared to the donor, the redshift of the acceptor facilitates the complementary absorption of active layers, enhancing photon capture efficiency. Among them, the three descriptors D/A ratio, RMSR, and gauge belong to the device process parameter descriptors, and the five descriptors HOMO-D, LUMO-D, HOMO-A, LUMO-A, and MAX-A belong to the molecular property descriptors. The SMILES string is a linear, text-based string that accurately conveys the chemical information regarding molecular structure. Utilizing SMILES strings, molecular chemical structures can be represented as text, making it easier for computers to process and analyze. This approach offers a feasible strategy for collecting valuable data, as these descriptors affect the properties of OSCs.[57–60] The 1024-bit Morgan fingerprints of donor and acceptor molecules were calculated using the Python-based RDKit package.[61] The drawn molecular structure file was converted to SMILES strings using the Open Babel software.[62] These SMILES strings were then used to compute the Morgan fingerprints. In this work, considering that it was not feasible to generate Morgan molecular fingerprints for infinitely connected polymers, and the photovoltaic performance differences among polymer donors were mainly determined by their repeating units, the SMILES of polymer donors were generated from the corresponding one repeating unit of donors, with the dangling bonds saturated using hydrogen atoms. This approach allows for the effective representation of molecular structure, which is crucial for training ML models in the context of OSC research. It's worth noting that we concatenate the descriptors into a vector for training ML model. The purpose is to input the 1024-bit Morgan fingerprint as a whole. The advantage of this approach lies in its ability to capture not only local features within a molecule but also the global characteristics of the molecule. This enables to model the similarities and differences among molecules, facilitating a deep understanding of the crucial structural features that influence OSC performances.

2.3. Shapley Additive Explanations

The SHAP method utilizes game theory to explain the output of an ML model. It can assess the accuracy of an ML model and determine the impact of descriptors on ML model's output. To understand the influence of descriptor importance on an ML model, it is essential to comprehend how descriptor change affects ML model's output and the descriptor distribution within the dataset. In SHAP value analysis, each descriptor was considered as a "contributor", and their importance can be assessed based on their contributions to the target quantity. This allows us to determine the degree of descriptor importance in relation to the target photovoltaic parameters. By evaluating the impact of each descriptor on the model's output, SHAP enables a nuanced understanding of the relative importance of different features. Understanding how changes in descriptors influence the model's predictions is crucial for interpreting the real-world implications of the ML model.

The SHAP method, grounded in game theory, offers a powerful tool for unraveling the black box of ML models. By delving into the contributions of each descriptor and considering their distribution, we gain a nuanced perspective on the intricate relationship between input features and model predictions, ultimately enhancing the interpretability and reliability of our findings.

2.4. ML Model Training

To construct the ML model, we used a series of ML algorithms that can be accessed from the Scikit-learn package.[63] These ML models took the feature descriptor as input and the performance parameter PCE of the OSC device as output.[64] We constructed five feasible ML models. Building ML models involves four main processes: database construction, descriptor selection, model training, and model evaluation. The general workflow for building ML models is illustrated in **Figure 1**. The filtered dataset includes 292 data points for ML training. The distribution intervals of their samples regarding device performance parameters are shown in **Figure 2**. The intervals 10–12% for PCE and 0.8–0.9 V for V_{OC} corresponded to the largest number of OSCs in dataset. The data samples were divided using stratified sampling method, considering various ratio splits and finding that an 80%:20% split ratio between the training and test sets is the most appropriate. Subsequently, the filtered data was used to train five ML models: RF, ETR, GBRT, Adaboost, and ANN. It's worth noting that these five ML models were employed for regression analysis. These five models represented different types of ML algorithms, including ensemble learning and neural networks. Each ML model offers unique capabilities in terms of handling different types of data structures, capturing complex relationships, and addressing specific challenges inherent in the dataset. Moreover, our previous work and that of other researchers demonstrated the advantages of these ML models.[22,39,46] This diversity helps in finding the best model for various data structures and properties.

On the technical side, a completely random division of a finite dataset may lead to overcrowding or no sample presence in a certain PCE region. In order to avoid this inconsistency, we adopted stratified sampling method with PCE as the target to ensure that the selected test set samples were evenly distributed. The tenfold crossvalidations were applied to avoid overfitting of ML models. Moreover, in order to find the best parameters for the ML model, we chose an exhaustive search tuning tool using the "GridSearchCV" function, passing in the hyperparameters "param_grid" and specifying the range of hyperparameters to be tuned. The purpose was to select the best hyperparameters and train the ML model, which was actually a training and comparison process.

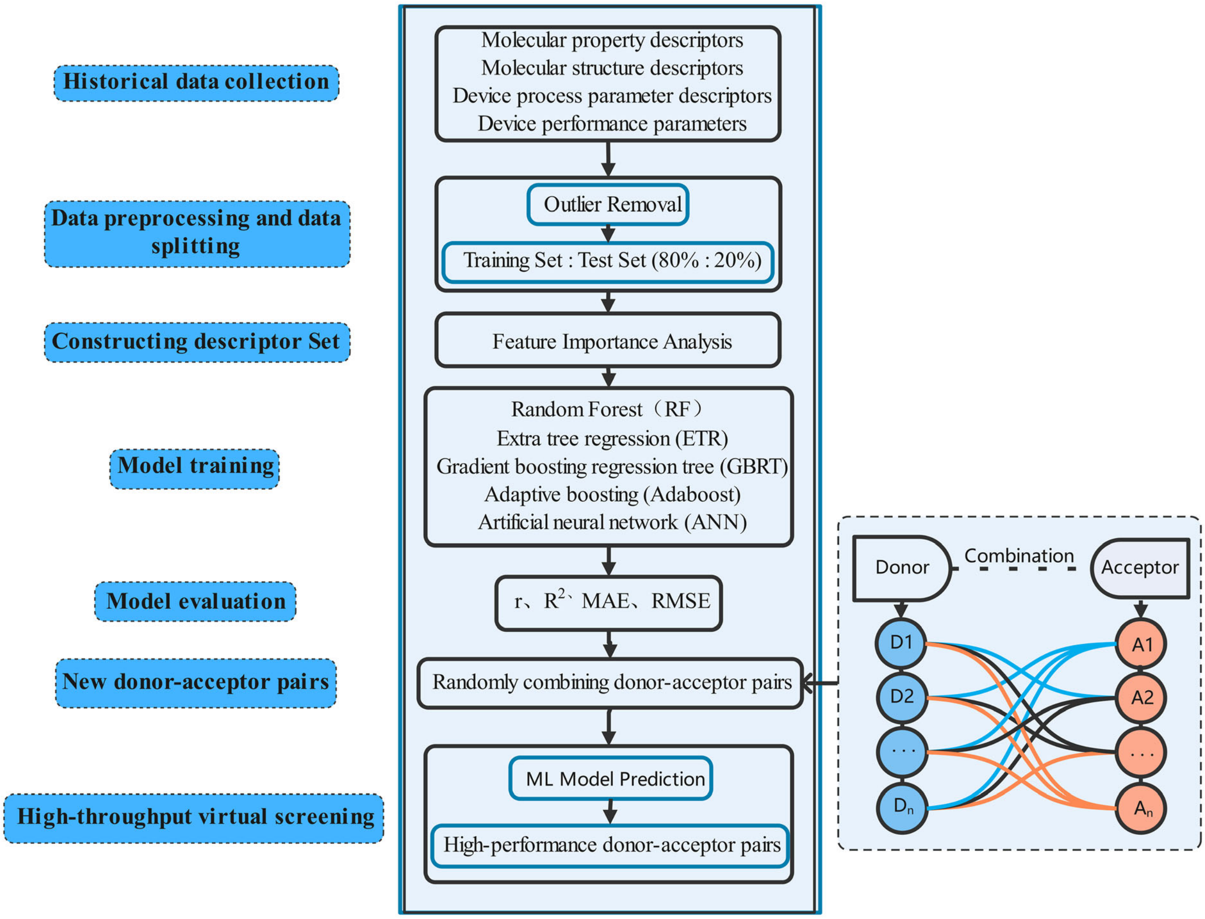
Historical data collection
Data preprocessing and data splitting
Constructing descriptor Set
Model training
Model evaluation
New donor-acceptor pairs
High-throughput virtual screening
Molecular property descriptors
Molecular structure descriptors
Device process parameter descriptors
Device performance parameters
Outlier Removal
Training Set : Test Set (80% : 20%)
Feature Importance Analysis
Random Forest（RF）
Extra tree regression (ETR)
Gradient boosting regression tree (GBRT)
Adaptive boosting (Adaboost)
Artificial neural network (ANN)
r、R2、MAE、RMSE
Randomly combining donor-acceptor pairs
ML Model Prediction
High-performance donor-acceptor pairs
Donor
Combination
Acceptor
D1
D2
···
Dn
A1
A2
···
An

Figure 1. Scheme of model design framework for developing high-performance OSCs donor-acceptor pairs.

After hyperparameter was optimally tuned, the ML model had good prediction ability and generalization ability.

2.5. ML Model Evaluation

To assess the performance of ML models and its ability to predict target properties, we used regression evaluation metrics commonly used for regression models. These metrics included the Pearson correlation coefficient (r), the determination coefficient (R^2), mean absolute error (MAE), and root mean squared error (RMSE) to evaluate the model. The model performance metrics r, R^2, MAE, and RMSE were used to test the prediction accuracy of ML models. The definitions are as follows.

$$r = \frac{\sum_{i=1}^{N}(R_i - \overline{R})(P_i - \overline{P})}{\sqrt{\sum_{i=1}^{N}(R_i - \overline{R})^2}\sqrt{\sum_{i=1}^{N}(P_i - \overline{P})^2}} \tag{1}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(R_i - P_i)^2}{N \times \mathrm{var}(R_i)} \tag{2}$$

$$\mathrm{MAE} = \frac{\sum_{i=1}^{N}|R_i - P_i|}{N} \tag{3}$$

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{N}(R_i - P_i)^2}{N}} \tag{4}$$

where N is the number of data points in the data set, R_i and P_i represent the actual and predicted values of the sample, respectively, $\overline{R_i}$ and $\overline{P_i}$ denote the mean of the actual and predicted values of the sample, respectively, and $\mathrm{var}(R_i)$ is the variance of the sample data. The values of r and R^2 ranged from 0 to 1. The closer it was to 1, the higher the prediction accuracy of the model. MAE indicates the mean of the absolute error between the actual and predicted values of the sample, and its value ranged from 0 to positive infinity. The smaller the MAE value, the smaller the error between the actual and predicted values of the sample, and the higher the prediction accuracy of ML model.

3. Results and Discussion

3.1. Data Visualization and Analysis

The performance of OSC devices composed of different donor and acceptor materials varies greatly because the materials properties greatly affect the device's performance. Chemical data can

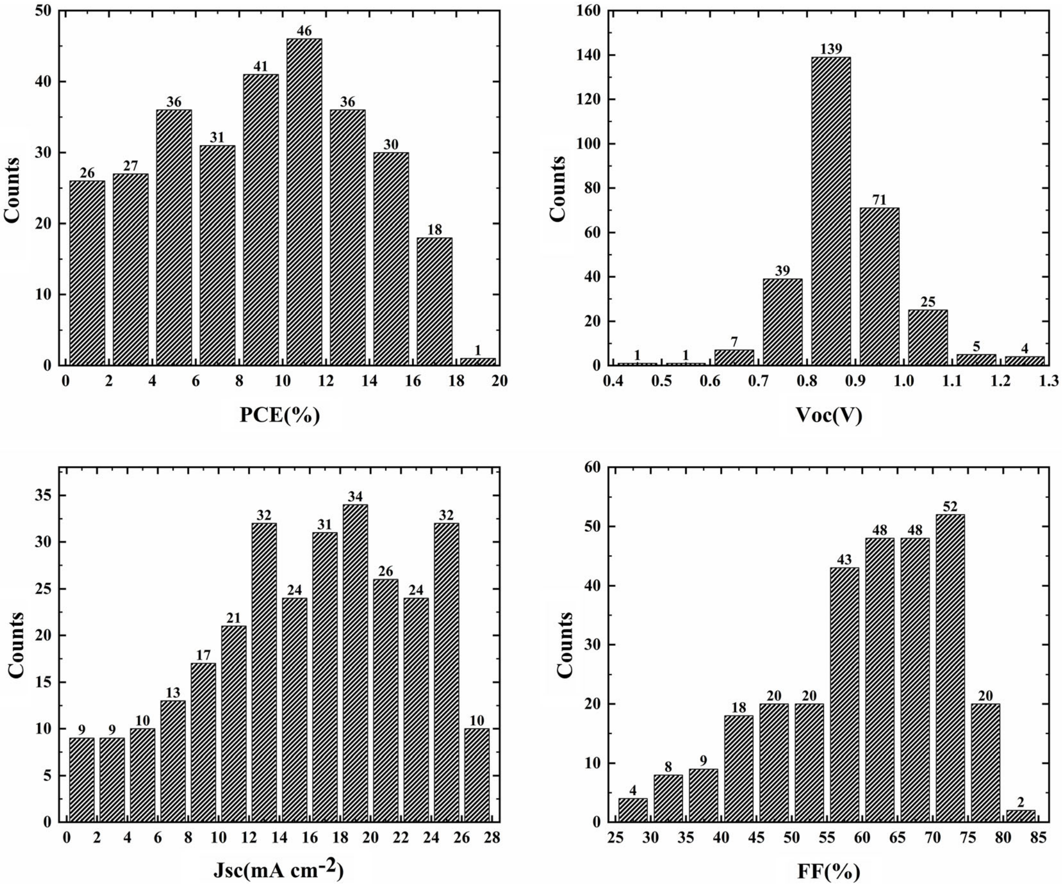

Figure 2. Experimental PCE, J_{SC}, V_{OC}, and FF distribution of nonfullerene based OSCs in the dataset.

help us understand the properties of material molecules and the patterns hidden behind data that can provide a lot of useful information. The detailed data visualization analysis was conducted. Apart from the molecular Morgan fingerprints, **Figure 3** presents the correlation heat map between the device performance parameter PCE and other parameters. The correlation between PCE and MAX-A is the highest among the descriptors presented in Figure 3. The figure shows the medium correlation of PCE with other descriptors, suggesting that each selected descriptor contributes to PCE. The figure also indicates that LUMO-A and MAX-A are the most important descriptors to determine V_{OC} and J_{SC}, respectively.

Efficient OSCs not only require careful selection of donor or acceptor molecules but also necessitate fine tuning of the experimental manufacturing conditions for OSCs. In the work of Ma et al. a grid search approach based on ML was employed to screen favorable device characteristics such as RMSR of the D/A blend and the D/A weight ratio.[65] The efficiency of the OSCs was enhanced by optimizing device specifications. In this study, a visual analysis of the datasets was conducted, and **Figure 4** presents the correlation of PCE with D/A ratio, RMSR, and gauge. The dependences of PCE on molecular property descriptors are shown in Figure S1, Supporting Information. As observed in Figure 4a, it's evident that PCE exceeds 15.13% when the D/A ratio falls within the range of 0.8–1.6 with the RMSR range of 6.5–7.5 nm. Figure 4b shows that PCE can exceed 15.13% when the D/A ratio falls within the range of 0.76–1.1 with the gauge range of 130–145 nm. Figure 4c indicates that PCE can exceed 15.13% when the RMSR falls within the range of 1–1.7 with the gauge range of 139–149 nm, and the RMSR falls within the range of 5.2–7.5 with the gauge range of 113–138 nm. By trying the best combination many times, it appears that the optimal PCE can be achieved when the D/A ratio is about 0.9, the gauge value is 140 nm, and the RMSR value is 1 nm.

3.2. Descriptor Importance Analysis

SHAP value analysis is a ML model interpretation method used to assess the importance of each feature in determining ML model outcomes. **Figure 5** displays the SHAP analysis of feature importance except molecular Morgan fingerprints. The vertical axis ranks the importance of all samples, while the horizontal axis represents the SHAP values of the features. Each point represents a sample, with red and blue indicating high and low feature values, respectively. As shown in the figure, the higher importance features are HOMO-D and MAX-A, suggesting that changes in these features can generate significant impact on the ML model prediction.

It is evident that the red points in HOMO-D mainly correspond to negative SHAP values, indicating that higher values of HOMO-D generate a negative impact on the PCE. It is

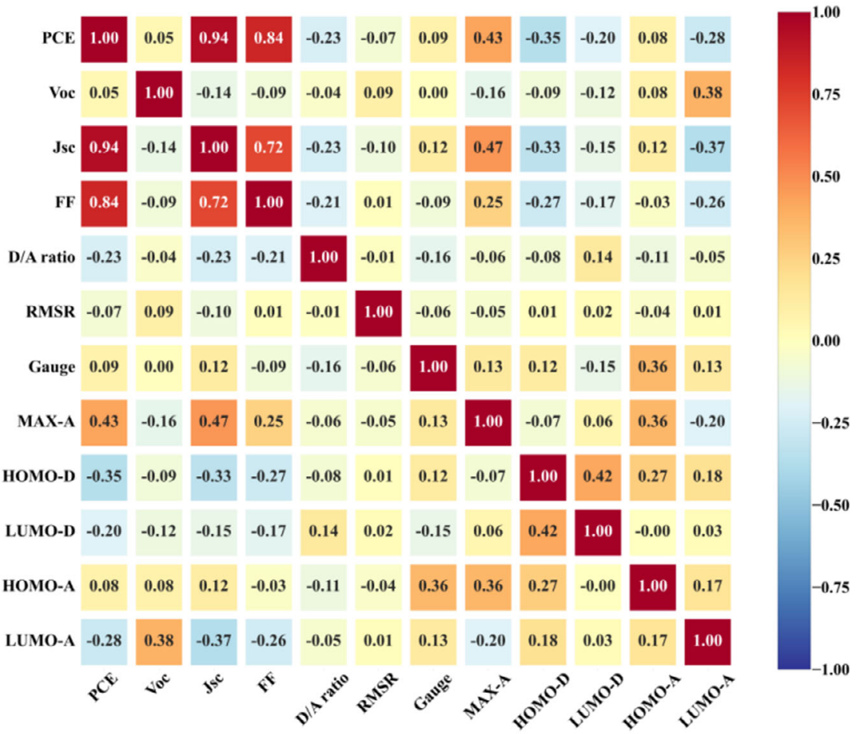

Figure 3. Pearson correlation matrix of the selected descriptors and OSCs parameters.

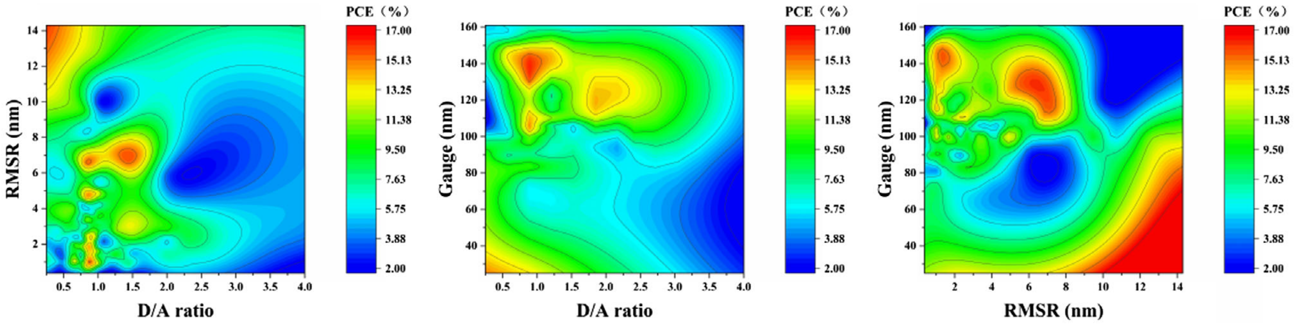

Figure 4. Contour color fill plot of the device parameters PCE with donor–acceptor weight ratio (D/A ratio), RMSR, and active layer film thickness (Gauge) for the 292 donor–acceptor pairs.

reasonable because the higher value of HOMO-D usually reduces charge transfer energy and then induces the reduction of V_{OC}.[16,21,66–69] The red points in MAX-A focus on positive SHAP values, indicating that a higher MAX-A value generates a positive effect on PCE. This can be understood from that the higher MAX-A value corresponds to the redshift of acceptor absorption spectrum, meaning the broaden absorption range that it is favorable to improve J_{SC}.[70,71] Similarly, the gauge importance is the lowest among these analyzed features, implying minimal influence on PCE. Actually, the D/A ratio, RMSR, and gauge involves the optimization of OSC device for the giving donor and acceptor. We did not incorporate the analysis of Morgan fingerprints in SHAP analysis. First, in materials science, the molecular structure of donors and acceptors is undoubtedly the most important, making SHAP analysis of Morgan fingerprints unnecessary. Second, the intrinsic characteristics of Morgan fingerprints (discreteness, high dimensionality, and locality) make their intuitive interpretation in SHAP analysis very complicated.

3.3. ML Model Regression Analysis

The training and test sets are divided according to the ratio of 4:1, including 233 and 59 samples in training and test sets, respectively. **Table 1** provides the r, R^2, MAE, and RMSE values for predicting PCE, while predicting V_{OC}, J_{SC}, and FF can be found in Table S1 in, Supporting Information. The data from the table indicates that the RF model gives the highest r and R^2 values. The r and R^2 values of RF model on the training set are 0.78 and 0.63, and the corresponding values on the test set are 0.81 and 0.65, respectively. Furthermore, the MAE and RMSE

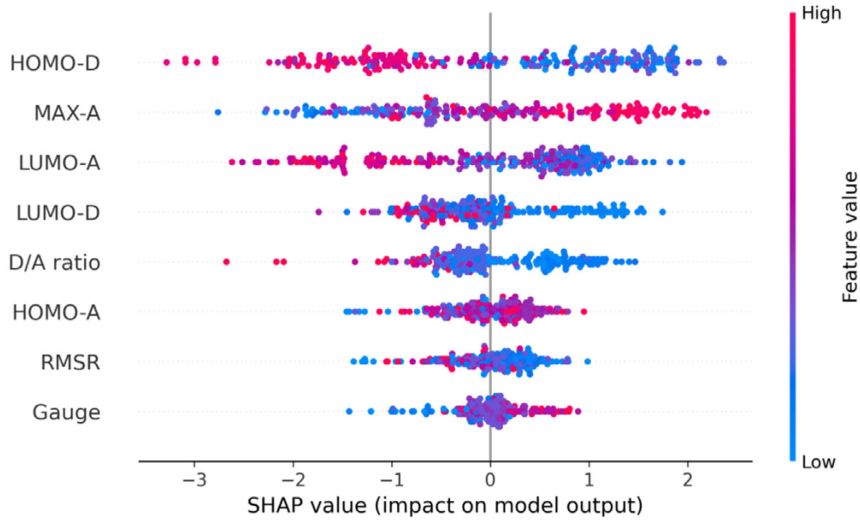

Figure 5. SHAP importance of the selected feature parameters.

Table 1. The corresponding Pearson's correlation coefficients (*r*) and coefficients of determination (R^2), mean absolute errors (MAE), and root mean square errors (RMSE) of the predicted device parameters PCE (%) using five ML algorithms on the train set (233 data points) and the test set (59 data points).

Model[a)]	*r*	R^2	MAE	RMSE
RF	0.81(0.78)	0.65(0.63)	1.68(1.94)	2.58(2.54)
ETR	0.78(0.76)	0.60(0.58)	1.93(2.16)	2.76(2.71)
GBRT	0.81(0.76)	0.62(0.57)	1.76(1.94)	2.67(2.67)
Adaboost	0.70(0.71)	0.46(0.48)	2.59(2.61)	3.19(3.03)
ANN	0.78(0.73)	0.59(0.54)	1.84(2.27)	2.77(3.05)

[a)] The tenfold crossvalidation was performed on the train set. The data outside and inside parentheses are the results predicted using test set and training set, respectively.

values of RF model were the smallest among the five ML models, with MAE and RMSE values of 1.94 and 2.54 on the training set and 1.68 and 2.58 on the test set, respectively. In the work of Ma and Muhammad et al. the RF model also achieved the best predictive performance.[42,56] In the work by Ma et al. the *r* value of predicting photovoltaic performance is 0.7. In the work of Muhammad et al. the predicted PCE on the test set achieved the R^2 value of 0.625. **Figure 6** displays scatter plot illustrating the predicted and experimental values of device performance parameters based on the RF model. Details of other ML models for comparing predicted and experimental values are presented in Figure S2, Supporting Information. In the figures, blue triangles represent training set samples, and red circles represent test set samples. The blue and red lines are fit regression lines for training and test sets, respectively. If the slope of the regression line equals 1, it indicates a perfect match between the predicted and experimental values. Although the slopes of RF model for device performance parameters are less than 1, it is closer to 1 than that of other ML models. Therefore, the R^2, *r*, RMSE, MAE, and regression lines indicate that the performance of the RF model is the best among these four ML models. Hence, we chose to utilize the trained RF model for further predicting OSC performance parameters.

In order to verify the reliability of ML model more intuitively, nine OSCs are randomly selected from the database of this work, and the experimental and predicted photovoltaic parameters are shown in **Table 2**. It can be found that the predicted results agree well with that of experiment, underlining the accuracy of the trained RF model in prediction ability.

In addition, we performed generalization tests for the trained RF model. **Table 3** shows the experimental and predicted results of three OSC devices selected beyond the database of this work. Two methods were employed for confirming that the dataset of this work does not include the three OSCs. Firstly, we searched the SMILES string of the corresponding donor and acceptor. By comparing the SMILES strings, we can determine whether identical donor–acceptor pairs exist in the database. However, subtle variations in molecular structure may result in changes to the SMILES string. Therefore, further validation was performed using principal component analysis to examine the distribution of the selected three OSCs in the principal component space. The results are presented in Figure S3, Supporting Information. The data generated two principal components by dimensionality reduction, namely, PC1 and PC2, corresponding to the *x*-axis and *y*-axis in the Figure S3, respectively. The black points in the Figure represent the projection of the original database OSC samples onto the given principal component axes, while the red points indicate the projection of the selected external OSC samples on the same axes. Figure S3, Supporting Information indicates that there is no complete overlap between the red and black samples. Therefore, there are no identical samples based on PC1 and PC2, meaning the selected three OSCs are different from those in datasets. Moreover, the experimental PCEs for D18:IT-M, PBDB-T:Y6, and BTR-Cl:BTP-FBr-FBr OSCs are 8.13%,[72] 11.70%,[73] and 14.79%,[74] respectively. The corresponding PCEs predicted using the trained RF model are 8.95%, 12.70%, and 13.21%, respectively. The predicted PCEs are very close to the corresponding experimental values, indicating the good generalization ability of the trained RF model.

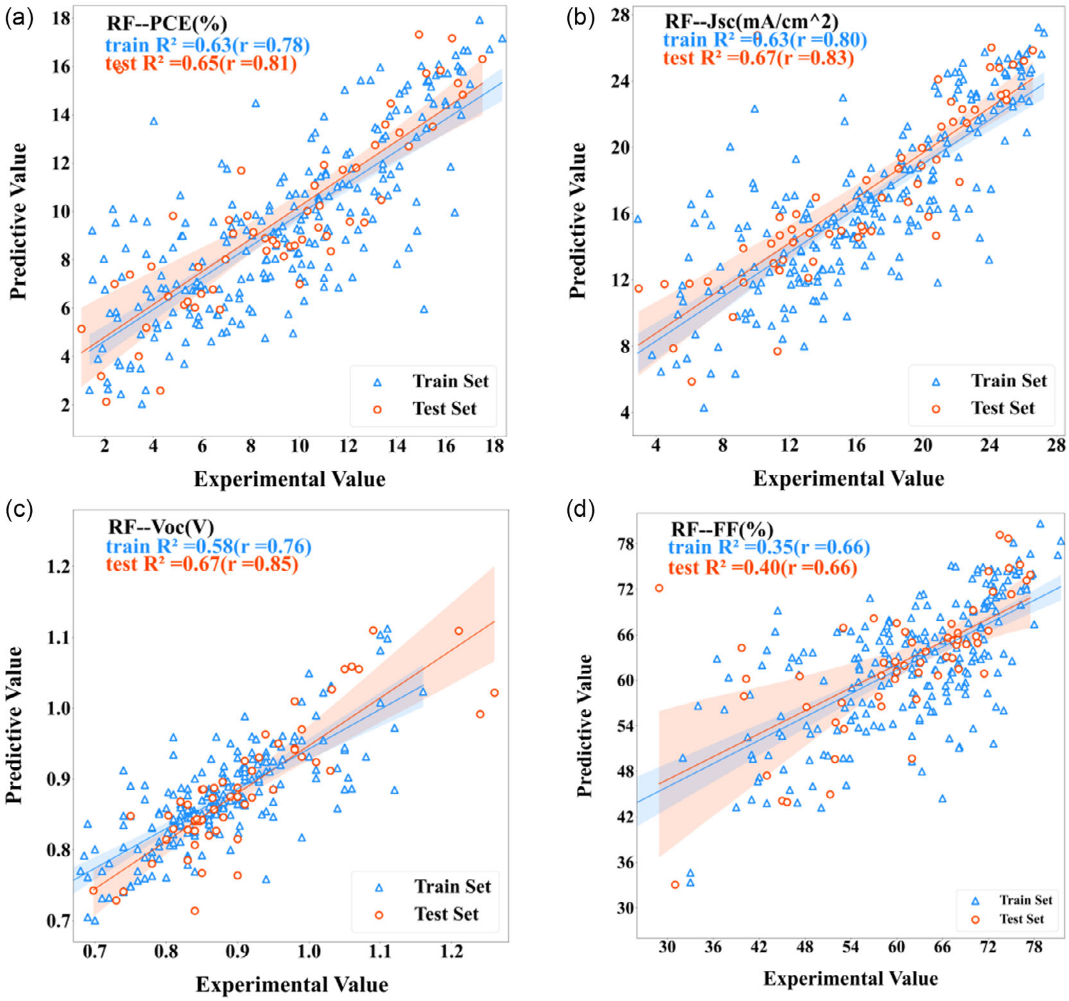

Figure 6. The device parameters a) PCE, b) J_{sc}, c) V_{oc}, and d) FF were predicted using the selected RF algorithm on the training set (233 data points) and test set (59 data points), respectively. The blue indicates the prediction result of the training sets, whereas the red indicates the prediction 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 red areas indicate the error range of the corresponding fitted lines.

Table 2. The experimental and predicted photovoltaic parameters of 9 donor–acceptor pairs randomly selected from the database.

Donor:Acceptor	Experimental value				Predictive value			
	PCE [%]	V_{OC} [V]	J_{SC} [mA cm^{-2}]	FF [%]	PCE [%]	V_{OC} [V]	J_{SC} [mA cm^{-2}]	FF [%]
PTB7-Th:PTbPDI	4.65	0.71	11.73	54.00	4.66	0.71	11.73	54.08
J52-2F:CRIC	6.12	0.93	13.41	48.70	6.11	0.93	13.42	48.70
PBDB-T:IDTT2OT	6.46	0.93	12.30	56.70	6.51	0.92	12.37	56.74
BO:IT-4F	8.44	0.72	19.20	58.50	8.47	0.72	19.20	58.65
PBDB-PSF:ITIC	9.84	0.99	16.99	58.73	9.82	0.99	16.92	58.85
PBDB-T:ITIC	10.06	0.91	16.10	68.70	10.01	0.91	16.01	68.35
PBDB-T:IT-2F	12.08	0.83	18.62	71.00	12.00	0.82	18.56	70.79
PM6:BTP-C6Ph	15.50	0.83	24.30	76.20	15.51	0.83	24.29	76.18
PBTATBT-4f:Y6	15.72	0.81	26.92	72.30	15.90	0.80	27.08	72.50

Table 3. The experimental and predicted photovoltaic parameters of donor–acceptor pairs outside the database.

D/A	Experimental value				Predictive value			
	PCE [%]	V_{OC} [V]	J_{SC} [mA cm^{-2}]	FF [%]	PCE [%]	V_{OC} [V]	J_{SC} [mA cm^{-2}]	FF [%]
D18:IT-M	8.13	1.02	14.41	55.40	8.95	0.91	16.84	62.38
PBDB-T:Y6	11.70	0.69	26.00	65.40	12.70	0.84	22.95	67.01
BTR-Cl:BTP-FBr-FBr	14.79	0.84	23.56	74.57	13.21	0.85	22.60	66.93

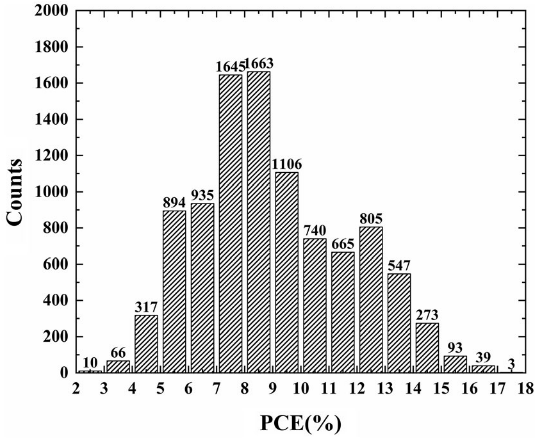

Figure 7. Interval distribution of PCE for 9779 new donor–acceptor pairs predicted using the RF model.

3.4. New Donor–Acceptor Pairs' Performance Prediction

The key of improving PCE is to find suitable donor–acceptor pairs for OSC. We filtered out the donor and acceptor molecules from the dataset, ensuring that the molecules are unique in the dataset. By randomly combining the donor and acceptor molecules in dataset, 10 004 donor–acceptor pairs were obtained for OSCs. After that, we eliminated the donor–acceptor pairs that are identical with those in the database. Then we generated 9779 novel donor–acceptor pairs.

We used the trained RF model to predict the photovoltaic parameters of the newly formed donor–acceptor pairs. Regarding the descriptor values for the newly generated donor–acceptor pairs, the optimal values of D/A ratio, RMSR, and gauge obtained from the analysis in Figure 4 were adopted as 0.9, 1, and 140 nm, respectively. The predicted PCE value distribution is shown in **Figure 7**. The distribution of predicted results for other device performance parameters is shown in Figure S4, Supporting Information. Within the predicted PCE distribution, there are 42 donor–acceptor pairs with the predicted PCE > 16%. Among them, three donor–acceptor pairs corresponded with the predicted PCE > 17%. The donor and acceptor molecular structures and the predicted PCE values are presented in **Figure 8**. The other donor–acceptor pairs with predicted PCE > 16% are given in Figure S5, Supporting Information. The predicted PCEs for the PBDB-TF:L8-BO, PBDB-TF:L8-HD, and PBDB-TF:L8-OD OSCs are 17.9%, 17.53%, and 17.53%, respectively. After the efforts of scientific researchers, PBDB-TF:L8-BO as the most promising OSC device, experimental PCE reached an impressive 18.5%.This is a difference of 0.6%

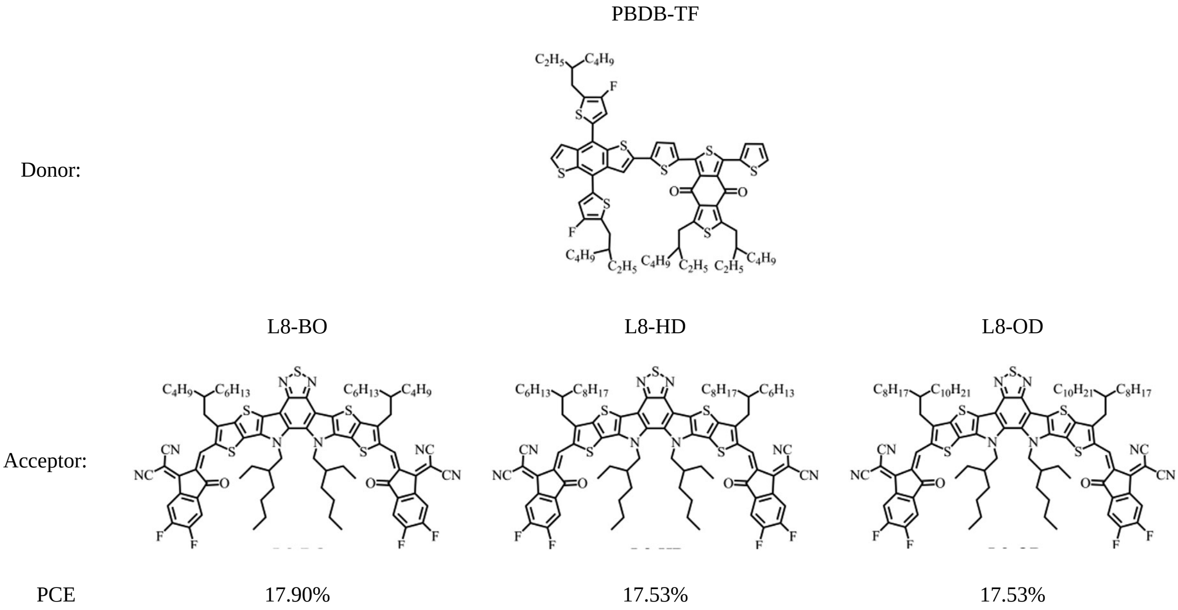

Figure 8. Molecular structures and the predicted PCEs for the selected donor–acceptor pairs.

from the predicted PCE value, which to some extent validates the accuracy of the trained ML model.[75] Through virtual screening of donor–acceptor pairs, high-performance OSC materials can be quickly screened.

4. Conclusion

In this study, we constructed the dataset consisting of 292 sets of experimental OSCs, including donor and acceptor molecular properties, molecular structures, device process parameters, and OSC performance parameters. To train ML models with robust predictive capabilities, we use molecular property descriptors, molecular structure descriptors, and device process parameters as combined inputs. The evaluation was performed using the Pearson correlation coefficient and SHAP analysis. Our findings revealed that the molecule property descriptors HOMO-D and MAX-A exhibited the highest correlation with PCE. We tested five regression ML models, and the results indicated that the RF model achieved the highest accuracy and stability, with the r value of 0.80 and the R^2 value of 0.65 on the test sets. Furthermore, the 9779 donor and acceptor pairs were generated by randomly combination of donor–acceptor pairs in the constructed database. Using the trained RF model, the photovoltaic parameters of 9779 donor and acceptor pairs were predicted. The predicted PCE values for PBDB-TF:L8-BO, PBDB-TF:L8-HD, and PBDB-TF:L8-OD OSCs are 17.9%, 17.53%, and 17.53%, respectively. This framework allows for the rapid and cost-effective design and screening of potential new donor–acceptor pairs combinations with the ability to predict their PCE.

Supporting Information

Supporting Information is available from the Wiley Online Library or from the author.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (grant no. 12264025). The authors are grateful for the help by Mr. Bing Yang, Miao Zhao, Long-Fei Lv and Rui Cao, Misses Hai-Yuan Yu, Li Ma, Jin-Hong Li, and Yu-Tong Ren.

Conflict of Interest

The authors declare no conflict of interest.

Author Contributions

M.L. took care of investigation; data curation; writing the original draft; and visualization. C.-R.Z. took care of conceptualization; methodology; validation, formal analysis, writing the review and editing, investigation, resources, supervision, project administration, and funding acquisition. M.-L.Z. took care of formal analysis. J.-J.G. took care of formal analysis. X.-M.L. took care of formal analysis. Y.-H.C. formal analysis. Z.-J.L. took care of formal analysis and resources. Y.-Z.W. took care of formal analysis. H.-S.C. took care of formal analysis and resources.

Data Availability Statement

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

Keywords

donor and acceptors, machine learning, organic photovoltaics, organic solar cells, virtual screenings

Received: January 13, 2024
Revised: February 22, 2024
Published online: March 11, 2024

[1] Q. Nie, A. Tang, Q. Guo, E. Zhou, *Nano Energy* **2021**, *87*, 106174.
[2] Y. Zhang, Y. Ji, Y. Zhang, W. Zhang, H. Bai, M. Du, H. Wu, Q. Guo, E. Zhou, *Adv. Funct. Mater.* **2022**, *32*, 2205115.
[3] A. Mahmood, A. Tang, X. Wang, E. Zhou, *Phys. Chem. Chem. Phys.* **2019**, *21*, 2128.
[4] Y. Wang, P. Liang, H. Yang, W. Li, Z. Wang, Z. Liu, J. Wang, X. Shen, *Mater. Today Energy* **2020**, *17*, 100423.
[5] J. Hou, O. Inganas, R. H. Friend, F. Gao, *Nat. Mater.* **2018**, *17*, 119.
[6] Y.-J. Lin, Y.-C. Li, C.-C. Yeh, S.-F. Chung, L.-M. Huang, T.-C. Wen, Y.-H. Wang, *Appl. Phys. Lett.* **2006**, *89*, 223518.
[7] L. Bian, E. Zhu, J. Tang, W. Tang, F. Zhang, *Prog. Polym. Sci.* **2012**, *37*, 1292.
[8] H. Phillips, S. Zheng, A. Hyla, R. Laine, T. Goodson, E. Geva, B. D. Dunietz, *J. Phys. Chem. A* **2012**, *116*, 1137.
[9] L. Ma, C. R. Zhang, M. L. Zhang, X. M. Liu, J. J. Gong, Y. H. Chen, Z. J. Liu, Y. Z. Wu, H. S. Chen, *Adv. Theory Simul.* **2023**, 2300624.
[10] Z. Yang, M. Gao, W. Wu, X. Yang, X. W. Sun, J. Zhang, H.-C. Wang, R.-S. Liu, C.-Y. Han, H. Yang, H. C.-Y. Han, H. Yang, W. Li, *Mater. Today* **2019**, *24*, 69.
[11] L. Duan, N. K. Elumalai, Y. Zhang, A. Uddin, *Sol. Energy Mater. Sol. Cells* **2019**, *193*, 22.
[12] W. Qiu, S. Zheng, *Sol. RRL* **2021**, *5*, 2100023.
[13] G. Forti, A. Nitti, P. Osw, G. Bianchi, R. Po, D. Pasini, *Int. J. Mol. Sci.* **2020**, *21*, 8085.
[14] S. Alam, M. S. Akhtar, E.-B. Kim, H.-S. Shin, S. Ameen, *Appl. Sci.* **2020**, *10*, 5743.
[15] A. R. Mohd Yusoff, D. Kim, F. K. Schneider, W. J. da Silva, J. Jang, *Energy Environ. Sci.* **2015**, *8*, 1523.
[16] M. Zhao, C. R. Zhang, M. L. Zhang, X. M. Liu, J. J. Gong, Z. J. Liu, Y. H. Chen, H. S. Chen, *Int. J. Quantum Chem.* **2022**, *123*, e27047.
[17] G. Dennler, M. C. Scharber, C. J. Brabec, *Adv. Mater.* **2009**, *21*, 1323.
[18] S. D. Collins, N. A. Ran, M. C. Heiber, T.-Q. Nguyen, *Adv. Energy Mater.* **2017**, *7*, 1602242.
[19] Z. Xu, F. Pan, C. Sun, S. Hong, S. Chen, C. Yang, Z. Zhang, Y. Liu, T. P. Russell, Y. Li, D. Wang, *ACS Appl. Mater. Interfaces* **2020**, *12*, 9537.
[20] H.-Y. Yu, C.-R. Zhang, M.-L. Zhang, X.-M. Liu, J.-J. Gong, Z.-J. Liu, Y.-Z. Wu, H.-S. Chen, *New J. Chem.* **2022**, *46*, 20204.
[21] C.-R. Zhang, H.-Y. Yu, M.-L. Zhang, X.-M. Liu, Y.-H. Chen, Z.-J. Liu, Y.-Z. Wu, H.-S. Chen, *Phys. Chem. Chem. Phys.* **2023**, *25*, 25465.
[22] B. Yang, C.-r. Zhang, Y. Wang, M.-l. Zhang, Z.-j. Liu, Y.-z. Wu, H.-s. Chen, *Chin. J. Chem. Phys.* **2023**, *36*, 199.
[23] C.-D. Park, T. A. Fleetham, J. Li, B. D. Vogt, *Org. Electron.* **2011**, *12*, 1465.
[24] C. Yan, S. Barlow, Z. Wang, H. Yan, A. K. Y. Jen, S. R. Marder, X. Zhan, *Nat. Rev. Mater.* **2018**, *3*, 72.
[25] V. A. Trukhanov, D. Y. Paraschuk, *Polym. Sci. Ser. C* **2014**, *56*, 72.

[26] R. Hussain, F. Hassan, M. U. Khan, M. Y. Mehboob, R. Fatima, M. Khalid, K. Mahmood, C. J. Tariq, M. N. Akhtar, *Opt. Quantum Electron.* **2020**, *52*, 364.
[27] M. Khalid, M. U. Khan, E. T. Razia, Z. Shafiq, M. M. Alam, M. Imran, M. S. Akram, *Sci. Rep.* **2021**, *11*, 19931.
[28] D. Zhou, J. Wang, Z. Xu, H. Xu, J. Quan, J. Deng, Y. Li, Y. Tong, B. Hu, L. Chen, *Nano Energy* **2022**, *103*, 107802.
[29] N. Meftahi, M. Klymenko, A. J. Christofferson, U. Bach, D. A. Winkler, S. P. Russo, *NPJ Comput. Mater.* **2020**, *6*, 166.
[30] X. Rodriguez-Martinez, E. Pascual-San-Jose, M. Campoy-Quiles, *Energy Environ. Sci.* **2021**, *14*, 3301.
[31] M. R. S. A. Janjua, *Synth. Met.* **2021**, *279*, 116865.
[32] Z. Zhu, Z. Rahman, M. Aamir, S. Z. A. Shah, S. Hamid, A. Bilawal, S. Li, M. Ishfaq, *RSC Adv.* **2023**, *13*, 2057.
[33] M.-Y. Sui, Z.-R. Yang, Y. Geng, G.-Y. Sun, L. Hu, Z.-M. Su, *Sol. RRL* **2019**, *3*, 1900258.
[34] Y. Cui, P. Zhu, X. Liao, Y. Chen, *J. Mater. Chem. C* **2020**, *8*, 15920.
[35] Q. Wu, M. Pan, S. Zhang, D. Sun, Y. Yang, D. Chen, D. A. Weitz, X. Gao, *Energies* **2022**, *15*, 6666.
[36] B. Zhu, R. Wu, X. Yu, *Acta Chim. Sin.* **2020**, *78*, 1366.
[37] K. V. Park, K. H. Oh, Y. J. Jeong, J. Rhee, M. S. Han, S. W. Han, J. Choi, *Clin. Exp. Otorhinolar.* **2020**, *13*, 148.
[38] E. Hossain, I. Khan, F. Un-Noor, S. S. Sikander, M. S. H. Sunny, *IEEE Access* **2019**, *7*, 13960.
[39] C.-R. Zhang, M. Li, M. Zhao, J.-J. Gong, X.-M. Liu, Y.-H. Chen, Z.-J. Liu, Y.-Z. Wu, H.-S. Chen, *J. Appl. Phys.* **2023**, *134*, 153104.
[40] A. Mahmood, J.-L. Wang, *Energy Environ. Sci.* **2021**, *14*, 90.
[41] K. T. Butler, D. W. Davies, H. Cartwright, O. Isayev, A. Walsh, *Nature* **2018**, *559*, 547.
[42] K. M. Katubi, M. Saqib, M. Maryam, T. Mubashir, M. H. Tahir, M. Sulaman, Z. A. Alrowaili, M. S. Al-Buriahi, *Inorg. Chem. Commun.* **2023**, *151*, 110610.
[43] N. Alwadai, S. U. Khan, Z. M. Elqahtani, S. Ud-Din Khan, *Molecules* **2022**, *27*, 5905.
[44] W. Liu, Y. Lu, D. Wei, X. Huo, X. Huang, Y. Li, J. Meng, S. Zhao, B. Qiao, Z. Liang, Z. Xu, D. Song, *J. Mater. Chem. A* **2022**, *10*, 17782.
[45] A. Mahmood, A. Irfan, J.-L. Wang, *Chin. J. Polym. Sci.* **2022**, *40*, 870.
[46] J.-H. Li, C.-R. Zhang, M.-L. Zhang, X.-M. Liu, J.-J. Gong, Y.-H. Chen, Z.-J. Liu, Y.-Z. Wu, H.-S. Chen, *Org. Electron.* **2024**, *125*, 106988.
[47] S. Nagasawa, E. Al-Naamani, A. Saeki, *J. Phys. Chem. Lett.* **2018**, *9*, 2639.
[48] D. Padula, J. D. Simpson, A. Troisi, *Mater. Horiz.* **2019**, *6*, 343.
[49] Y. Wu, J. Guo, R. Sun, J. Min, *NPJ Comput. Mater.* **2020**, *6*, 120.
[50] B. Yang, C. R. Zhang, Y. Wang, M. Zhao, H. Y. Yu, Z. J. Liu, X. M. Liu, Y. H. Chen, Y. Z. Wu, H. S. Chen, *Int. J. Quantum Chem.* **2023**, *123*, e27039.
[51] L. Fu, H. Hu, Q. Zhu, L. Zheng, Y. Gu, Y. Wen, H. Ma, H. Yin, J. Ma, *Nano Res.* **2023**, *16*, 3588.
[52] G. Han, Y. Yi, *Angew. Chem. Int. Ed.* **2022**, *61*, e202213953.
[53] G. Sastre, F. Daeyaert, in *AI-Guided Design and Property Prediction for Zeolites and Nanoporous Materials*, John Wiley & Sons **2023**.
[54] Z. Ciğeroğlu, G. Küçükyıldız, A. Haşimoğlu, F. Taktak, N. Açıksöz, *Korean J. Chem. Eng.* **2020**, *37*, 1975.
[55] M. Jamei, M. Karbasi, I. Adewale Olumegbon, M. Mosharaf-Dehkordi, I. Ahmadianfar, A. Asadi, *J. Mol. Liq.* **2021**, *335*, 116434.
[56] H. Sahu, H. Ma, *J. Phys. Chem. Lett.* **2019**, *10*, 7277.
[57] A. Mahmood, S. U.-D. Khan, F. U. Rehman, *J. Saudi Chem. Soc.* **2015**, *19*, 436.
[58] A. Mahmood, A. Irfan, *J. Comput. Electron.* **2020**, *19*, 931.
[59] M. U. Khan, R. Hussain, M. Yasir Mehboob, M. Khalid, Z. Shafiq, M. Aslam, A. A. Al-Saadi, S. Jamil, M. Janjua, *ACS Omega* **2020**, *5*, 24125.
[60] A. Mahmood, S. U.-D. Khan, U. A. Rana, M. H. Tahir, *Arabian J. Chem.* **2019**, *12*, 1447.
[61] A. Capecchi, D. Probst, J.-L. Reymond, *J. Cheminf.* **2020**, *12*, 43.
[62] N. M. O'Boyle, M. Banck, C. A. James, C. Morley, T. Vandermeersch, G. R. Hutchison, *J. Cheminf.* **2011**, *3*, 33.
[63] G. E. V. Fabian Pedregosa, A. Gramfort, V. Michel, B. Thirion, *J. Mach. Learn. Res.* **2011**, *12*, 2825.
[64] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. J. Dubourg, *J. Mach. Learn. Res.* **2011**, *12*, 2825.
[65] Y. Wen, Y. Liu, B. Yan, T. Gaudin, J. Ma, H. Ma, *J. Phys. Chem. Lett.* **2021**, *12*, 4980.
[66] M. Wei, Z. Wang, Z. Wen, X. Hao, W. Qin, *Appl. Phys. Lett.* **2018**, *113*, 093301.
[67] T. Fritsch, J. Kurpiers, S. Roland, N. Tokmoldin, S. Shoaee, T. Ferron, B. A. Collins, S. Janietz, K. Vandewal, D. Neher, *Adv. Energy Mater.* **2022**, *12*, 2200641.
[68] J. Gao, N. Yu, Z. Chen, Y. Wei, C. Li, T. Liu, X. Gu, J. Zhang, Z. Wei, Z. Tang, X. Zhang, H. Huang, *Adv. Sci.* **2022**, *9*, 2203606.
[69] S. Zheng, E. Geva, B. D. Dunietz, *J. Chem. Theory Comput.* **2013**, *9*, 1125.
[70] H. Xu, Y. Yang, C. Zhong, X. Zhan, X. Chen, *J. Mater. Chem. A* **2018**, *6*, 6393.
[71] X. Li, T. Yan, H. Bin, G. Han, L. Xue, F. Liu, Y. Yi, Z.-G. Zhang, T. P. Russell, Y. Li, *J. Mater. Chem. A* **2017**, *5*, 22588.
[72] M. Yanxian, L. Quanbin, W. Hongbin, C. Yong, *J. Mater. Chem. A* **2023**, *11*, 6237.
[73] K. D. Hui, C. Huijeong, G. Peddaboodi, L. Dongchan, K. BongSoo, C. Shinuk, *Sol. RRL* **2023**, *7*, 2201012.
[74] Y. Qianguang, H. Dingqin, M. Kumar, H. Dong, S. Ahmed, P. Huang, Z. Xiao, S. Lu, *Sol. RRL* **2023**, *7*, 2201062.
[75] Q. Li, X. Liao, Y. Sun, Y. Xu, S. Liu, L. M. Wang, Z. Cao, X. Zhan, T. Zhu, B. Xiao, Y.-P. Cai, F. Huang, *Small* **2023**, 2308165.