

A Structure-Based Machine Learning Framework for Polymer Solar Cells: Prediction, Interpretation, and Virtual Screening

Yuan-Fang Lee

f10549022@ntu.edu.tw

National Taiwan University <https://orcid.org/0000-0003-1154-0443>

Yu-Hung Hsiao

Te-An Chen

Min-Hsuan Lee

Hsuan-Hung Liu

Ching-I Huang

Research Article

Keywords: Polymer solar cells, Machine learning, Explainable artificial intelligence, Random forest, Molecular fingerprints

Posted Date: September 23rd, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10968603/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

A Structure-Based Machine Learning Framework for Polymer Solar Cells: Prediction, Interpretation, and Virtual Screening

Yuan-Fang Lee^{1*}, Yu-Hung Hsiao¹, Te-An Chen¹, Min-Hsuan Lee², Hsuan-Hung Liu³,

Ching-I Huang^{1*}

¹ Institute of Polymer Science and Engineering, National Taiwan University, No. 1, Section 4, Roosevelt Road, Taipei 10617, Taiwan

² PhD Program in Biomedical Artificial Intelligence, National Tsing Hua University, No.101, Section 2, Kuang-Fu Road, Hsinchu 30013, Taiwan

³ Department of Chemical and Materials Engineering, National University of Kaohsiung, No. 700, Kaohsiung University Road, Nanzih District, Kaohsiung 811726, Taiwan

*Corresponding author: f10549022@ntu.edu.tw

*Corresponding author: chingih@ntu.edu.tw

Abstract

Polymer solar cells (PSCs) have attracted extensive attention due to their lightweight, flexibility, and solution-processability. However, the performance of PSCs arises from the intricate coupling of molecular structure, energy-level alignment, charge generation, transport, recombination, and morphology, posing a significant challenge for rational materials design. With the accumulation of experimental data and advances in polymer informatics, machine learning has emerged as a powerful tool for accelerating PSC research. In this work, we construct the largest experimental database of 5,797 binary bulk-heterojunction PSC devices reported from 2010 to 2025 and develop an interpretable machine learning framework that predicts device performance directly from donor and acceptor structures. Random forest models using Morgan fingerprints achieve a strong prediction of power conversion efficiency (PCE), with $R = 0.85$ and $RMSE = 1.94\%$. Statistical and Shockley-Queisser limit analyses show that historical PCE improvements have been driven primarily by short-circuit current density, whereas further gains increasingly depend on open-circuit voltage. Fragment-level SHapley Additive exPlanations reveal distinct structural roles: Y6-derived acceptors with ADA'DA architectures and INCN-type end groups contribute strongly to PCE and photocurrent, while donor fragments mainly regulate voltage. Screening more than 2.2 million virtual donor and acceptor combinations further indicates that high-performance systems occupy a restricted chemical space dominated by optimized Y6-derived acceptors paired with halogenated

benzodithiophene-based donors. These results demonstrate how experimentally grounded, interpretable machine learning can clarify structure-performance relationships and guide the rational design and screening of organic photovoltaic materials.

Keywords: Polymer solar cells, Machine learning, Explainable artificial intelligence, Random forest, Molecular fingerprints

1. Introduction

In recent years, the emergence of polymer informatics has transformed the paradigm of polymeric materials development [1]. By integrating big datasets with machine learning (ML) techniques, this data-driven approach enables the discovery of hidden correlations and accelerates the exploration of processing-structure-property-performance (PSPP) relationships [2]. Owing to the ability of ML to handle high-dimensional and nonlinear problems, polymer informatics has been widely applied across diverse research domains. Representative applications include the prediction of glass-transition temperature and mechanical properties, assessment of polymer blend miscibility, and identification of high-performance polymers for dielectric, membrane, and optoelectronic applications [3, 4].

Among various material applications, polymer solar cells (PSCs) represent one of the most challenging PSPP systems because device performance arises from the complex coupling of molecular structure, energy-level alignment, charge generation, transport, recombination, morphology, and processing conditions [5]. Over the past three decades, continuous advances in donor and acceptor materials have steadily improved device performance [6-8]. In particular, the emergence of non-fullerene acceptors (NFAs), represented by ITIC [9] and later Y6 [10] derivatives, has revolutionized PSC development owing to their superior light absorption, tunable energy levels, and favorable morphology, enabling power conversion efficiencies (PCE) approaching 20% [11, 12]. Despite these remarkable advances, the discovery of high-

performance materials still relies largely on empirical trial-and-error experiments, resulting in extensive experimental costs and slow development cycles [13]. These challenges have driven increasing interest in AI-assisted PSC research.

Early studies are largely inspired by the Harvard Clean Energy Project (CEP) [14-17]. They combine high-throughput quantum-chemical calculations with cheminformatics methods to explore millions of virtual donor materials for fullerene-based PSCs, establishing a framework for data-driven virtual screening in organic photovoltaics. As PSC materials have evolved toward modern NFA systems, numerous ML studies have increasingly relied on experimentally measured device databases [18-43]. Diverse molecular representations, including descriptor-based, fingerprint-based, graph-based, and hybrid representations that combine structural and electronic information, have been investigated. These representations are combined with various regression models, such as random forests (RF), Gaussian process regression (GPR), kernel ridge regression (KRR), support vector machine (SVM), gradient boosting (GB), and neural networks, to predict photovoltaic properties and performance. For example, Lee [18, 19] uses donor and acceptor frontier orbital energy descriptors derived from experimentally reported PSC devices with RF models to predict PCE and open-circuit voltage (V_{OC}). The acceptor bandgap and donor highest occupied molecular orbital (HOMO) are identified as the dominant descriptors governing PCE, whereas the acceptor lowest occupied molecular orbital (LUMO) and donor HOMO are identified as the key descriptors for V_{OC} prediction. Sun *et al.*

[20] systematically compare different molecular representations and machine-learning algorithms using a database of 1,719 experimentally reported donor materials. They show that molecular fingerprints longer than 1000 bits combined with RF models achieve the best predictive performance. Recently, Zhu *et al.* [21] systematically compared molecular descriptors and fingerprint representations using four machine-learning algorithms on an experimental database of 1,586 NFA-based PSC devices. Although descriptor-based models achieve slightly higher prediction accuracy, they show that Morgan fingerprints provide superior generalization and computational efficiency for high-throughput screening. SHapley Additive exPlanations (SHAP) analysis identifies the D unit of Y6-series acceptors, the cyano group in the acceptor, and BDD-based donor fragments as key structural motifs associated with high PSC performance. Guided by these insights, they synthesized three new NFAs that achieve experimental PCEs up to 15.86%, with prediction errors below 3%. Zhang *et al.* [22] constructed an experimental database of 547 donor and acceptor pairs and developed an RF model using Morgan fingerprints for PCE prediction, while employing MACCS fingerprints for SHAP-based substructure interpretation. SHAP analysis reveals that six-membered-ring donor fragments and halogen-containing acceptor fragments are frequently observed in high-performance PSC materials. They further recombined donor, acceptor, and π -spacer substructures to generate over 3.4 billion donor–acceptor combinations, identifying 14,296 candidate devices with predicted PCEs above 14%, including 123 combinations exceeding

15.5%. Despite these advances, current ML studies remain limited by the lack of comprehensive experimental databases, interpretable frameworks that connect molecular structures with photovoltaic performance, and systematic assessment of model applicability for materials discovery. These limitations hinder the translation of ML predictions into reliable molecular design principles and practical materials discovery.

In this work, we construct the largest experimental database of binary bulk-heterojunction (BHJ) PSCs reported between 2010 and 2025 and develop a structure-based machine learning framework for prediction, interpretation, and materials discovery. Using only donor and acceptor molecular structures as input, RF models achieve strong predictive performance for PCE, J_{SC} , V_{OC} , and FF. SHAP analysis further reveals the distinct donor and acceptor fragments governing individual photovoltaic parameters, enabling the establishment of parameter-specific molecular design guidelines. We further demonstrate that the proposed framework reliably predicts previously unreported donor and acceptor combinations within the known chemical space and identify a series of promising PSC candidates with predicted PCEs exceeding 18.47% through large-scale virtual screening.

2. Methodology

2.1 Dataset Construction and Preparation

In this study, the database of binary BHJ PSCs is constructed through a systematic literature survey. Publications are retrieved from SciFinderⁿ using the keyword “polymer solar cell”. The resulting records are manually curated, and only studies of binary BHJ PSC devices containing extractable donor and acceptor structures, energy level information, and complete photovoltaic performance data are retained for database construction. To ensure consistency with experimentally accessible properties, the energy levels of materials are collected primarily from cyclic voltammetry (CV) measurements as reported in the original literature. Due to limitations in structural representation and Simplified Molecular Input Line Entry System (SMILES) conversion, polymeric donors and acceptors are approximated using their repeating unit structures. For copolymers, representative structures are constructed by connecting one unit of each constituent monomer, without explicitly considering molecular weight, sequence distribution, or compositional ratio. Certain material systems, such as modified fullerenes with complex side groups and metal–organic compounds, are excluded from the dataset due to difficulties in SMILES encoding and Morgan fingerprint generation, as well as their limited relevance to mainstream PSC material design.

2.2 Molecular Representation

The chemical structures of donor and acceptor materials are manually reconstructed from the reported experimental literature using ChemDraw. These structures are subsequently converted into SMILES strings and molecular fingerprints using the RDKit package in Python. SMILES representations are used to record molecular structures, while Morgan fingerprints (also known as extended-connectivity fingerprints, ECFPs) [44] are employed as input descriptors for machine learning models. Morgan fingerprints encode atom-centered local chemical environments into fixed-length numerical vectors. Two key parameters govern the structural information captured by Morgan fingerprints: the radius (r) and the fingerprint dimensionality (nBits). The radius determines the spatial extent of each fragment, controlling the level of structural detail from local atomic environments to larger functional motifs. The fingerprint dimensionality defines the size of the feature space. To systematically evaluate the impact of these parameters, a grid of radius values ($r = 3-8$) and fingerprint dimensionalities (nBits = 512–131,072) is explored. The results indicate that models with nBits ≥ 8192 consistently achieve higher and more stable test performance, while smaller fingerprint sizes (nBits ≤ 1024) lead to significant degradation due to increased feature collision. Considering both computational efficiency and model generalization, a fingerprint size of nBits = 8192 is selected. Within this feature space, model performance is relatively insensitive to the radius in the range of 3–8. Therefore, the choice of radius is guided by chemical interpretability. A radius of $r = 5$ is found to effectively capture chemically meaningful substructures, including

functional groups, representative monomer units, aromatic systems, and key fragment interactions. Each donor and acceptor molecule is independently transformed into a binary fingerprint vector using $r = 5$ and $n\text{Bits} = 8192$. The final descriptor for each device is constructed by concatenating the donor and acceptor fingerprints, resulting in a 16,384-dimensional vector representation.

2.3 Model Training and Evaluation

Following relevant studies, we use a random forest (RF) model due to its ability to capture complex non-linear relationships and its robustness in high-dimensional feature spaces [20, 22]. As an ensemble learning method, it aggregates predictions from multiple decision trees trained on bootstrapped samples, effectively reducing variance while maintaining strong predictive performance. In this work, RF regression models are constructed using the scikit-learn library in Python to predict device performance metrics, including PCE, J_{SC} , V_{OC} , and FF, as well as material electronic properties such as HOMO, LUMO, and bandgap for both donor and acceptor materials. Model performance is evaluated on an independent test set using Pearson's correlation coefficient (R) and root mean squared error (RMSE). These metrics are calculated according to the following equations:

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

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

where y_i and $\hat{y}_i$ denote the experimental and predicted values, respectively, $\bar{y}$ and $\bar{\hat{y}}$ are their corresponding mean values, and n is the number of samples in the test set.

For model optimization, several hyperparameters are tuned using grid search with 5-fold cross-validation, including the number of trees (`n_estimators`), maximum tree depth (`max_depth`), and node splitting criteria (`min_samples_split` and `min_samples_leaf`). The optimal hyperparameter combination for the PCE prediction model is determined to be `n_estimators = 1000`, `max_depth = 50`, `min_samples_split = 2`, and `min_samples_leaf = 1`.

2.4 Model Explanation via SHAP

To understand the effect of chemical structure on device performance, we apply SHapley Additive exPlanations (SHAP) [45, 46] to compute feature-level contributions for each input chemical fragment derived from the Morgan Fingerprint. For a given feature i , the SHAP value ϕ_i is defined as:

$$\phi_i = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|! (|F| - |S| - 1)!}{|F|!} [f_{S \cup \{i\}}(x_{S \cup \{i\}}) - f_S(x_S)] \quad (3)$$

where F denotes the set of all features, S is a subset of features excluding i , and $f()$ represents the model prediction. The SHAP value quantifies the marginal contribution of feature i to the prediction relative to the model baseline. For each predicted property, global feature

importance is evaluated using the mean absolute SHAP value across all samples, while SHAP value distributions are analyzed to identify fragments with positive or negative effects on the target property.

2.5 Shockley–Queisser Limit Analysis

To benchmark the theoretical efficiency limits of PSCs, the experimental device performance is compared with the Shockley–Queisser (SQ) [47] detailed balance limit. The SQ limit represents the maximum attainable photovoltaic performance of an ideal single-junction solar cell, assuming unity absorption above the bandgap, no absorption below the bandgap, and radiative recombination as the only unavoidable loss pathway. In this work, the effective bandgap (E_g) of each donor and acceptor pair is defined as the smaller value of the donor and acceptor bandgap [36]. This approximation represents the lowest-energy absorption onset in a binary bulk-heterojunction active layer and therefore determines the long-wavelength limit of photon harvesting. The theoretical SQ-limit values of PCE, J_{SC} , and V_{OC} in our work are obtained from the tabulated values reported by Rühle under AM1.5G illumination and standard test conditions [48]. The resulting relationships between experimental PSC performance and the corresponding SQ limits for PCE, J_{SC} , and V_{OC} are presented in Figure 3.

3. Results and Discussion

3.1 Data Collection and Statistical Overview

A total of 5,797 experimental data points of binary bulk heterojunction polymer solar cells are collected from 2,058 publications reported between 2010 and 2025. For model training and prediction capability evaluation, the dataset is split by year: data from 2010 to 2023 are used for model training, and newly reported devices from 2024 to 2025 are reserved for prediction. The 2010-2023 dataset contains 5,628 device entries, corresponding to 3,842 unique donor and acceptor combinations, including 2,474 donor materials and 832 acceptor materials. The 2024-2025 dataset includes 169 new device entries, covering 121 previously unseen donor and acceptor combinations, with 59 new donors and 48 new acceptors. Table 1 exhibits an example of one dataset recorded in the database, including material names, experimental HOMO/LUMO/bandgap values, SMILES representations, and device performance metrics (P_{CE} , J_{SC} , V_{OC} , and FF). Figure 1 shows the statistical distributions of the photovoltaic performance metrics (P_{CE} , J_{SC} , V_{OC} , and FF) in the 2010-2023 dataset. In particular, Figure 1(a) shows that both P_{CE} and J_{SC} exhibit a right-skewed distribution, reflecting the historical evolution of PSC materials. Early devices, primarily based on fullerene acceptors, manifest relatively low performance, with maximum values of approximately 11.5% for P_{CE} [49] and 22 mA cm^{-2} for J_{SC} [50]. The introduction of non-fullerene acceptors (NFAs), such as ITIC and Y-series systems, significantly improves performance, reaching up to 18.6% for P_{CE} [51] and

29.83 mA cm⁻² for J_{SC} [52]. The strong correlation between PCE and J_{SC} can be seen in Figure 2(a). The correlation coefficient between PCE and J_{SC} is 0.95, which implies the J_{SC} might be the dominant factor of improving PCE. In contrast, the relationships between PCE and V_{OC} , as well as PCE and FF, are more scattered. Both V_{OC} and FF can correspond to either high- or low-performance devices, indicating that their contributions are strongly coupled with other factors rather than acting as independent performance determinants.

Table 1 A representative entry from our database includes molecular energy levels (HOMO, LUMO, and bandgap), chemical structure descriptors (SMILES), and device performance metrics (V_{OC} , J_{SC} , FF, and PCE), as illustrated using data from a reference [53]

Materials	HOMO (eV)	LUMO (eV)	Bandgap (eV)	SMILES
PM6 (donor)	-5.45	-3.63	1.82	CCCCC(CC)Cc1sc(-c2c3cc(-c4ccc(-c5sc(-c6cc cs6)c6c5C(=O)c5c(CC(CC)CCCC)sc(CC(CC) CCCC)c5C6=O)s4)sc3c(-c3cc(F)c(CC(CC)CC CC)s3)c3ccsc23)cc1F
Y6 (acceptor)	-5.65	-4.10	1.55	CCCCCCCCCCCCc1c(/C=C2\C(=O)c3cc(F)c(F))cc3C2=C(C#N)C#N)sc2c1sc1c3c4nsnc4c4c5s c6c(CCCCCCCCCC)c(/C=C7\C(=O)c8cc(F) c(F)cc8C7=C(C#N)C#N)sc6c5n(CC(CC)CCC C)c4c3n(CC(CC)CCCC)c21
D-A combination	V_{OC} (V)	J_{SC} (mA/cm ²)	FF	PCE (%)
PM6:Y6	0.839	25.011	0.734	14.741

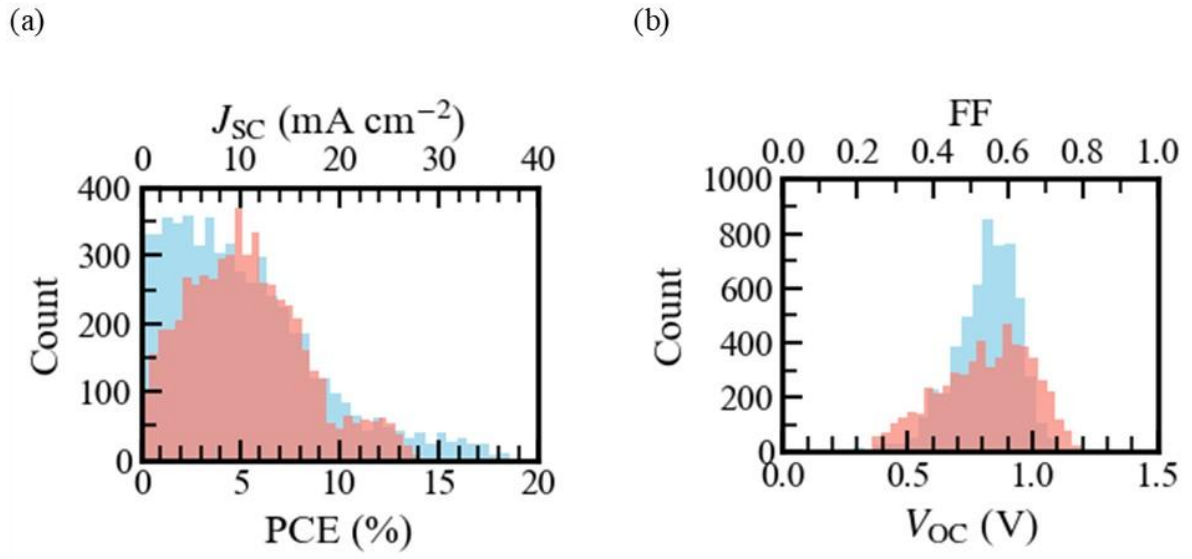


Fig. 1 Statistical distribution histogram of (a) PCE (blue) and J_{SC} (red), (b) V_{OC} (blue) and FF (red) in the 2010-2023 experimental literature

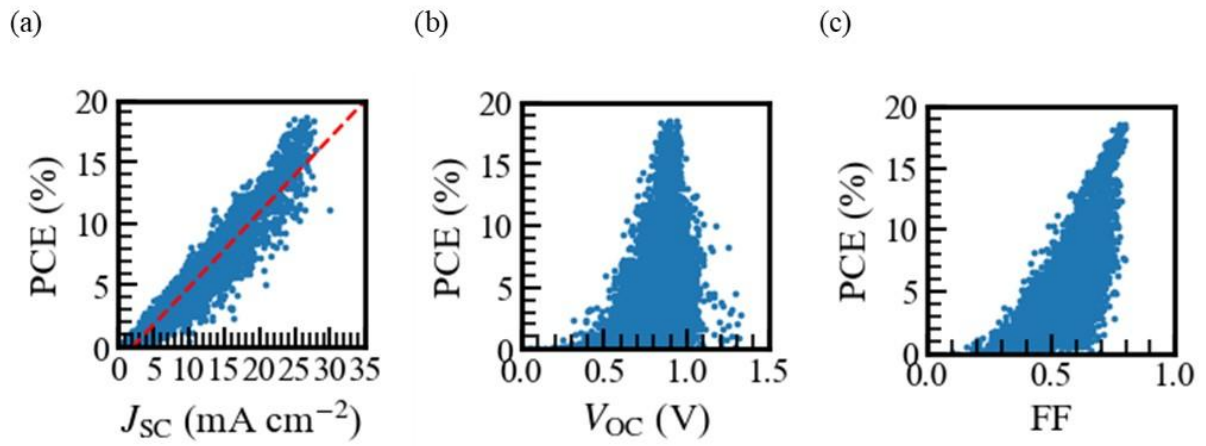


Fig. 2 Plots of PCE vs. (a) J_{SC} , (b) V_{OC} , and (c) FF in the 2010-2023 experimental literature

To further evaluate the development of device performance, we compare the experimental values of PCE, J_{SC} , and V_{OC} with the theoretical Shockley–Queisser limit [47] as a function of the bandgap in Figure 3 (a-c), respectively. Here, the E_g is chosen as the smaller one of the donor and acceptor bandgaps. The results show that current devices remain significantly below the theoretical maximum PCE. Notably, many devices already approach the theoretical limit of J_{SC} , whereas V_{OC} still exhibits a gap from its corresponding limit. This indicates that although J_{SC} has historically been the dominant factor in improving PCE, its enhancement is approaching saturation. In contrast, V_{OC} remains the primary limiting factor for further efficiency gains. Furthermore, the SQ trends reveal an intrinsic trade-off: both PCE and J_{SC} favor smaller bandgaps, while higher V_{OC} requires larger bandgaps. This opposing tendency highlights the need to identify an optimal V_{OC} within this trade-off space. As mentioned before in Figure 2(b), the observed V_{OC} range of 0.8-1.0 V likely reflects an optimized operating regime. This is consistent with the approximated SQ limit, where the maximum PCE (~33%) is achieved at an E_g of around 1.3 eV, corresponding to $V_{OC} \approx 1.0$ V and $J_{SC} \approx 34$ mA cm⁻². An important question that arises from this observation is whether these performance trends are encoded in molecular structure and can therefore be learned by structure-based machine learning models. To address this question, we next examine the predictive capability of random forest models trained directly on molecular fingerprints.

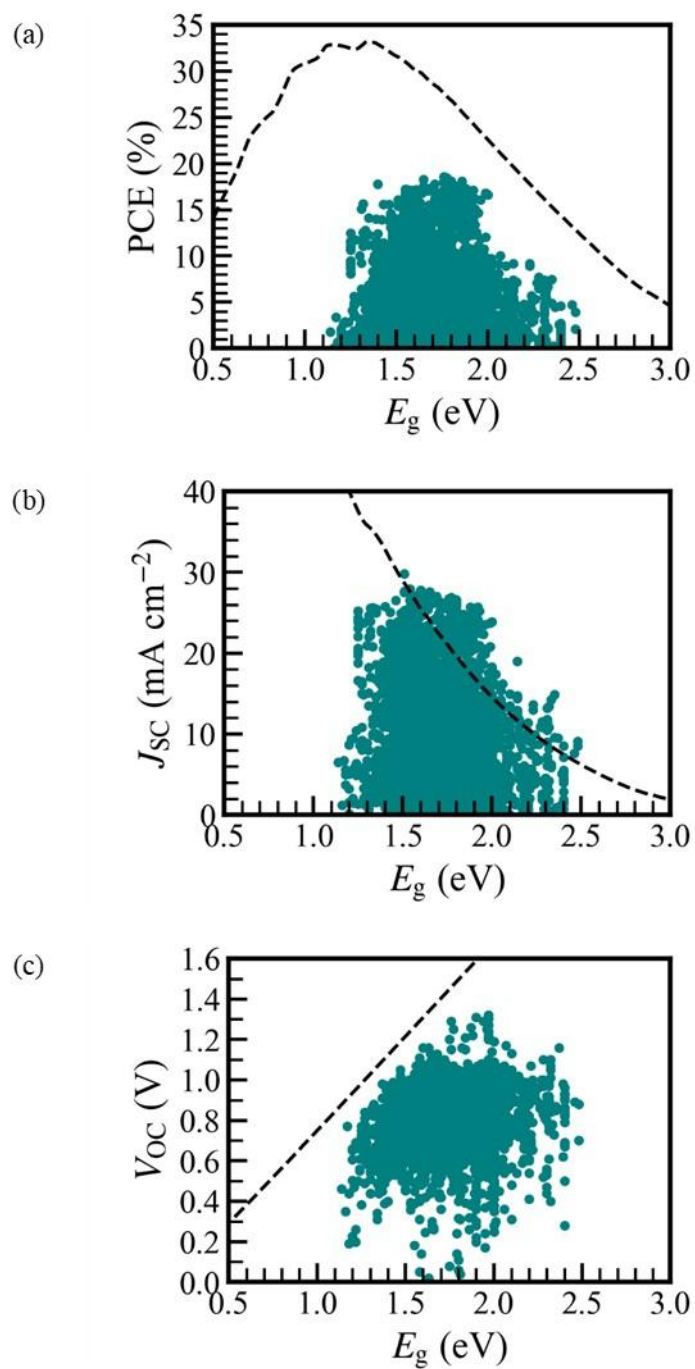


Fig. 3 A comparison of experimental PSC performance, (a) PCE, (b) J_{SC} , and (c) V_{OC} with the Shockley–Queisser (SQ) limit as a function of the bandgap (E_g). The black dashed lines indicate the corresponding theoretical SQ limits

3.2 Structure-based prediction of PSC performance

Figure 4 displays the predicted and experimental values for four key device parameters: (a) PCE, (b) J_{sc} , (c) V_{oc} , and (d) FF via the optimized random forest models. In all cases, the data points are largely distributed along the diagonal reference line, indicating that the models successfully capture the underlying relationships between molecular structure and device performance. This tendency is quantitatively evaluated with Pearson’s coefficient (R) and root mean square error (RMSE) in Table 2. The models for PCE, J_{sc} , and V_{oc} prediction exhibit comparable performance, with R reaching 0.97 for the training set and 0.85 for the test set, as reflected in Figure 4 with training (blue) and test (red) data scattering. FF remains reasonably predictable (test R = 0.738), but its lower accuracy suggests contributions beyond molecular structure. This likely arises from its sensitivity to morphological and device-level factors, such as charge transport balance, molecular packing, and intermolecular interactions [54]. The RMSE further provides a quantitative measure of prediction uncertainty. For PCE, the RMSE corresponds to an approximate error range of $\pm 2\%$ for newly reported materials. The moderate gap in R value between training and test performance suggests mild overfitting, while still indicating acceptable generalization within the learned chemical space [55]. The extent of this limitation will be further examined in Section 3.3 through predictions on newly reported materials.

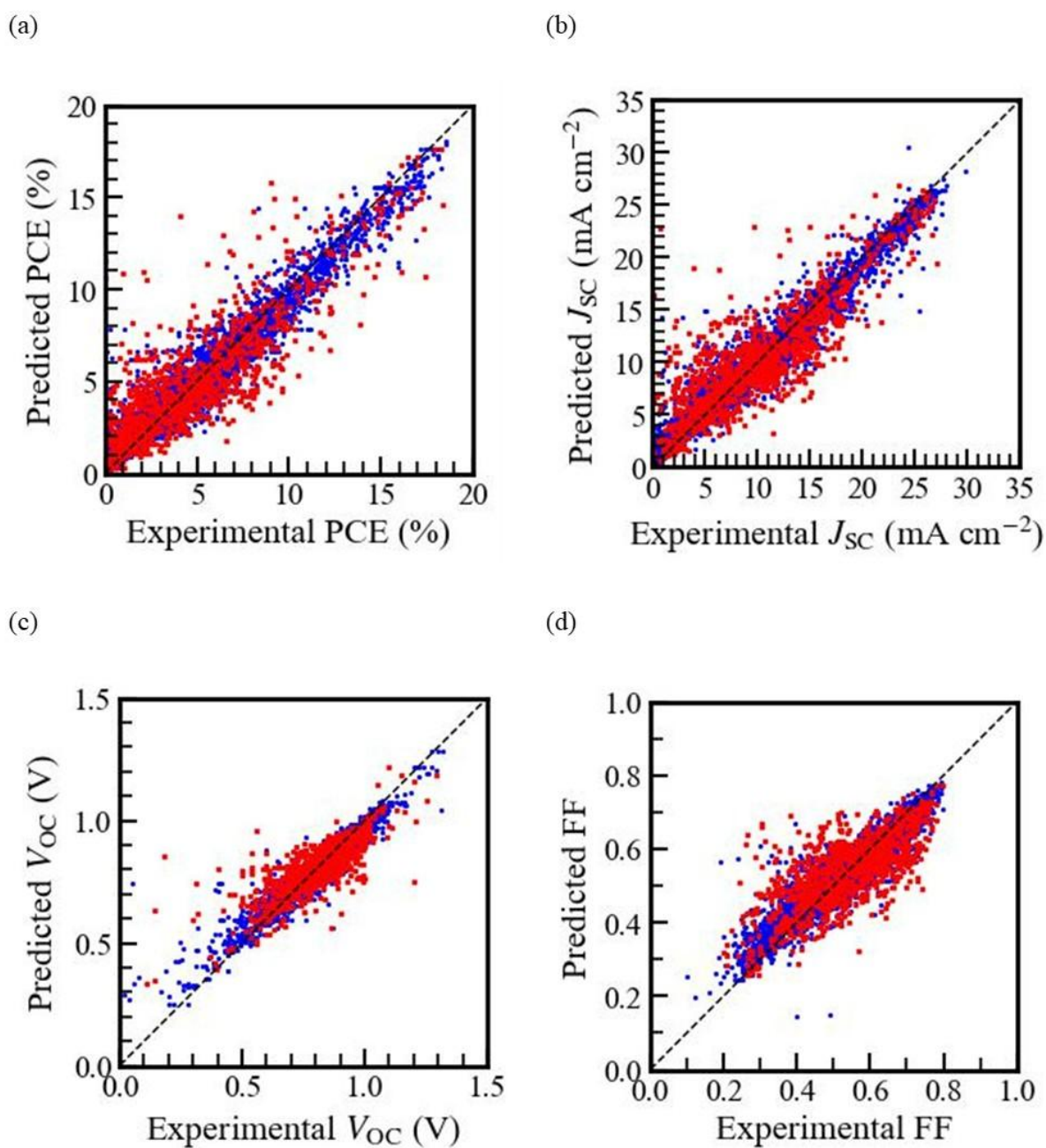


Fig. 4 Predicted versus experimental values of PSC device performance from the random forest model: (a) PCE (%), (b) J_{SC} (mA cm^{-2}), (c) V_{OC} (V), (d) FF. Blue and red dots represent training and test set data, respectively. The diagonal dashed line represents perfect prediction

Table 2 Performance of random forest models on different prediction targets, including PCE, J_{SC} , V_{OC} , FF, and energy levels of donor and acceptor

Prediction target	Training set		Test set	
	R	RMSE	R	RMSE
PCE (%)	0.968	0.971	0.853	1.939
J_{SC} (mA/cm ²)	0.970	1.481	0.848	3.050
V_{OC} (V)	0.965	0.038	0.825	0.077
FF	0.938	0.045	0.738	0.084
HOMO of Donor (eV)	0.910	0.076	0.766	0.121
LUMO of Donor (eV)	0.902	0.113	0.806	0.160
HOMO of Acceptor (eV)	0.917	0.085	0.870	0.103
LUMO of Acceptor (eV)	0.713	0.116	0.645	0.119

Table 3 compares the present model with previous studies that employed machine learning for PCE prediction. The advantages of our work can be summarized as follows. First, this study utilizes one of the largest experimentally collected PSC datasets, covering both fullerene and non-fullerene acceptors, enabling the model to capture broader material evolution trends. Second, the use of Morgan fingerprints provides a simple yet effective molecular representation without requiring additional quantum-chemical calculations or manual feature engineering, allowing the analysis to focus directly on structural effects. Third, the model achieves competitive performance for PCE prediction. In addition, the model is further evaluated using an external dataset derived from the Clean Energy Project Database (CEPDB) [15]. A randomly sampled subset (10% of the database) is used for prediction, yielding a strong correlation ($R = 0.82$) between predicted and reference values. This result highlights the model’s ability to capture generalizable structure–performance relationships beyond the experimental dataset.

These advantages facilitate further investigation into model interpretability and the prediction of new materials.

Table 3 Comparison of PCE prediction capabilities of binary bulk heterojunction polymer solar cells among various studies, including our results and related literature, highlighting different methods and achieved R of test set

Data set	Data Source	Model Input	ML models	R values of test set	Ref.
5,628	Literature	Morgan Fingerprint	RF	0.853	Our results
221,367	CEPDB ²⁷	Morgan Fingerprint	RF	0.820	
320	Literature	Morgan fingerprint and energy levels	KRR	0.78	Ref. [23]
566	Literature	Material properties and fingerprint	RF	0.85	Ref. [24]
290	Literature	Descriptor of quantum simulation and statistical analysis	GBRT	0.80	Ref. [25]
500	Literature	Morgan fingerprint	RF	0.83	Ref. [22]
565	Literature	Fingerprints based on rational designs	BRT	0.71	Ref. [26]
1586	Literature	Morgan fingerprint	XGBoost	0.8444	Ref. [21]
		Mordred descriptor	XGBoost	0.8603	
2.3 million	CEPDB	Descriptor of quantum simulation, statistical analysis, and molecular mass	Linear regression	0.916	Ref. [14]

To further investigate the key fragment signals captured by the RF models, SHAP analysis is performed for the device metrics. Figure 5 presents the top 20 influential fragments for predicting (a) PCE, (b) J_{sc} , (c) V_{oc} , and (d) FF. Acceptor-related fragments are labeled with (A), and donor-related fragments are labeled with (D). The corresponding acceptor and donor structural fragments are summarized in Tables 4 and 5, respectively. To facilitate chemical interpretation, fragments with clearly identifiable origins are represented by complete functional motifs or representative molecular substructures, such as the INCN-class end group (A) and PCBM-related structures (A). In contrast, fragments without specific molecular origins are retained in their original fragment forms, such as a linear unsaturated double bond (A) and a ketone group (D). The presence of each fragment in the materials is indicated in red, whereas its absence is indicated in blue. Fragments with concentrated red distributions in the positive SHAP region and blue distributions in the negative region can be regarded as key structural features that strongly promote the predicted property.

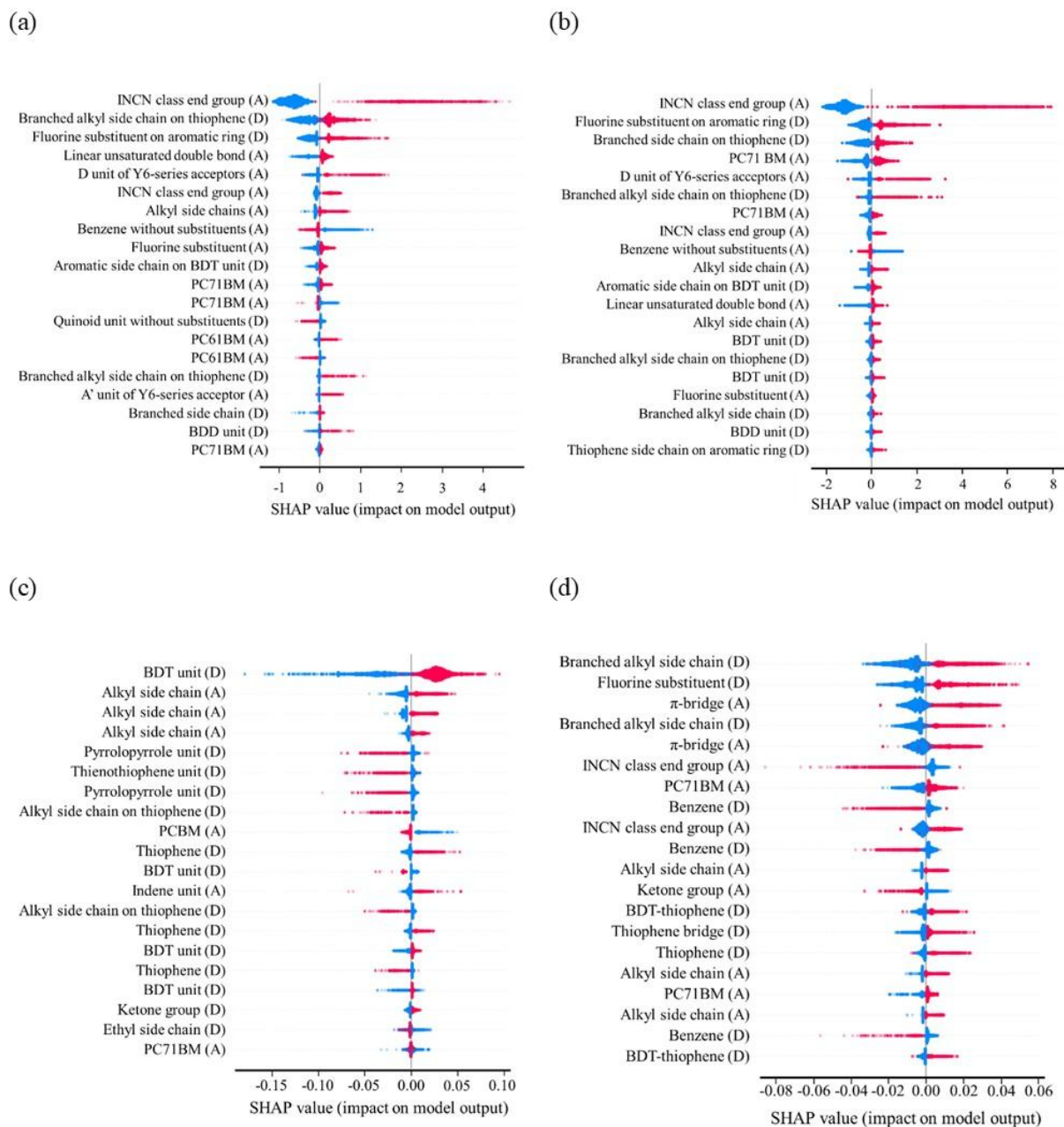


Fig. 5 SHAP summary plot showing the top 20 most important chemical fragments in the prediction model of (a) PCE, (b) J_{SC} , (c) V_{OC} , and (d) FF. Each point represents the SHAP value of one sample for a given feature. Red: presence of the fragment; Blue: absence. The further from zero, the stronger the influence of the fragment on the predicted PCE

Table 4 Representative acceptor structural fragments identified from SHAP analysis of the RF models

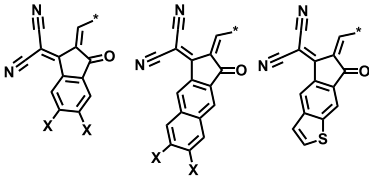
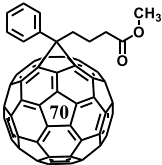
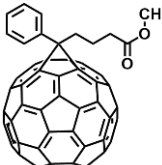
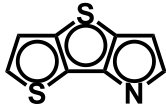
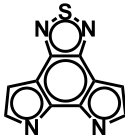
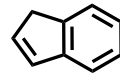
Fragment name	Diagram	Fragment name	Diagram
INCN class end group	 X = H, F, Cl, Br	PC71BM	
Linear unsaturated double bond		PC61BM	
D unit of Y6-series acceptors		A' unit of Y6-series acceptor	
Alkyl side chains		Indene unit	
Benzene without substituents		Ketone group	
Fluorine substituent			

Table 5 Representative donor structural fragments identified from SHAP analysis of the RF models

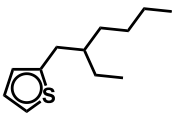
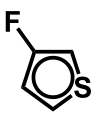
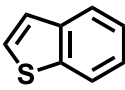
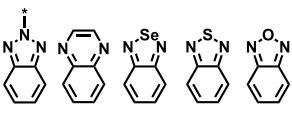
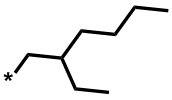
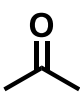
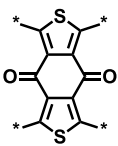
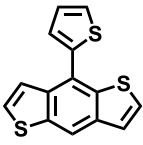
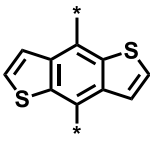
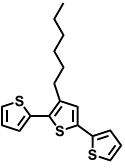
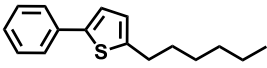
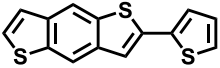
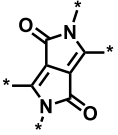
Fragment name	Diagram	Fragment name	Diagram
Branched alkyl side chain on thiophene		Thienothiophene unit	
Fluorine substituent on aromatic ring		Thiophene fused ring	
Quinoid unit without substituents		Thiophene	
Branched side chain		Ketone group	
BDD unit		Fluorine substituent	
Aromatic side chain on BDT unit		Benzene	
BDT unit		Thiophene bridge with side chain	
Thiophene side chain on aromatic ring		BDT-thiophene	
Pyrrolopyrrole unit			

Figure 5(a) presents the SHAP analysis of the PCE model, in which several fragments exhibit distinct positive and negative contributions to device performance. Among them, the INCN-class end group (A) shows the broadest and strongest positive SHAP distribution, indicating that it is one of the most critical structural motifs for achieving high PCE performance. Other fragments, including branched alkyl side chains on thiophene (D), fluorine substituents on aromatic rings (D), and the D unit of Y6-series acceptors (A), exhibit moderately positive contributions. Additional fragments, such as linear unsaturated double bonds (A), alkyl side chains (A), fluorine substituents (A), the A' unit of Y6-series acceptors (A), branched side chains (D), and the BDD unit (D), show relatively weaker but still favorable contributions. Notably, all three key structural motifs of the Y6 architecture (ADA'DA) [56] are identified as positive contributors, with their importance ranked as the INCN end group (A unit), the fused D unit, and the A' unit. In contrast, fullerene-related fragments, including PC71BM (A) and PC61BM (A), exhibit both positive and negative SHAP distributions, indicating that their contributions are partially offset across different device systems and have lower overall importance in determining high-PCE performance. Furthermore, unsubstituted benzene (A) and quinoid units without substituents (D) exhibit negative SHAP contributions. This trend suggests that the observed effect does not originate from the benzene or quinoid structures themselves, but rather from the absence of substituents. Together with the positive contributions of fluorine substituents on both donor and acceptor fragments, these results

highlight the critical role of fluorination in enhancing the PCE, consistent with numerous experimental studies [57-59].

Figure 5(b) shows that the overall SHAP contribution pattern of the J_{SC} model closely resembles that of the PCE model. Among these fragments, the INCN-class end group and Y6-series-related motifs remain dominant positive contributors, while PC71BM exhibits a relatively more positive distribution compared with the PCE model, implying that fullerene acceptors still retain certain advantages in charge transport and photocurrent generation.

Figure 5(c) presents a markedly different SHAP contribution pattern for the V_{OC} model, with fragment importance becoming more strongly dominated by donor-related structures. The BDT unit (D) shows the strongest positive contribution, indicating its important role in maintaining favorable energy-level alignment and suppressing voltage loss [60]. Conversely, pyrrolopyrrole units (D), thienothiophene units (D), and alkyl side chains on thiophene (D) exhibit negative SHAP contributions. Thiophene fragments (D) exhibit both positive and negative distributions, indicating that their influence strongly depends on the surrounding molecular environment and substitution pattern. On the acceptor side, alkyl side chains (A) and indene units (A) exhibit notable positive contributions, suggesting that branched side chains weaken excessive intermolecular coupling, suppress non-radiative recombination, and maintain favorable morphology [61].

Figure 5(d) shows a more diverse SHAP fragment contribution in the FF model than those of the PCE, J_{SC} , and V_{OC} models. In addition to previously identified important motifs, the FF model places greater emphasis on branched alkyl side chains (D), fluorine substituents (D), and π -bridge-related acceptor fragments (A). These fragments are closely associated with intermolecular interactions, molecular packing, phase separation, and charge transport balance, reflecting the strong dependence of FF on morphology-related characteristics [54].

Overall, the SHAP analysis reveals two dominant structure–performance relationships in PSCs: acceptor-centered photocurrent generation governing J_{SC} and donor-mediated electronic regulation governing V_{OC} . These observations are consistent with recent SHAP-based studies by Zhu *et al.* [21] and Zhang *et al.* [22], which likewise identified Y-series acceptors and fluorinated donor motifs as key structural features associated with high-performance PSCs. However, previous studies primarily focused on interpreting PCE or applied SHAP for virtual screening, whereas our work systematically resolves the structural determinants of PCE, J_{SC} , V_{OC} , and FF individually. Our results establish a direct structure–performance framework linking molecular fragments to individual photovoltaic processes, providing mechanistic insights that can guide future molecular designs.

3.3 Prediction of newly reported PSC materials

To further evaluate the applicability of our structure-based prediction model, predictions are performed on 121 newly reported PSC devices from 2024 to 2025 and compared with experimental values in Figure 6. When categorizing the new devices into combinations composed of known materials (red) and those involving newly introduced donor or acceptor structures (blue), a clear distinction emerges: systems constructed from previously reported materials scatter closer to the diagonal line and exhibit significantly higher prediction accuracy. Quantitative analysis in Table 6 shows that approximately 90% of these “known material” combinations fall within acceptable relative error ranges ($RE \leq 20\%$), whereas systems involving new materials exhibit a substantially larger deviation, with 40% below $RE \leq 20\%$. This result suggests that the model is well-suited for large-scale combinatorial screening within the chemical space of known materials, rather than for directly predicting the performance of a device with unseen materials.

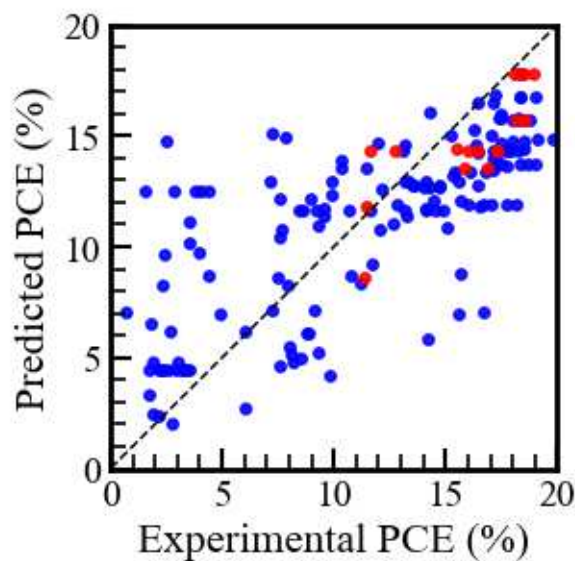


Fig. 6 Predicted versus experimental PCE values for newly reported PSC materials, categorized by combination type. Systems composed of previously known donor and acceptor materials but forming new combinations (red), systems involving new donors or acceptors (blue)

Table 6 Prediction accuracy distribution categorized by the combination type of donor–acceptor systems. Model performance is evaluated using relative error (RE) of PCE prediction, classified into three ranges: accurate ($RE < 10\%$), moderate ($10\% \leq RE < 20\%$), and poor ($RE \geq 20\%$)

combination type	Accurate ($RE < 10\%$)	Moderate ($RE\ 10\text{-}20\%$)	Poor ($RE > 20\%$)
Known materials	33.3%	55.6%	11.1%
Unknown materials	16.5%	25.2%	58.3%

Therefore, in order to explore the potential of full data from 2010 to 2025, we retrain the model with the full database and screen all possible combinations of 2,533 known donor and 880 known acceptor materials. In total, the device performance of 2,229,040 hypothetical

donor and acceptor pairs are predicted. Based on the predicted PCE, Table 7 shows the top 10 donor and acceptor combinations ($\text{PCE} \geq 18.47\%$). First, the model demonstrates good predictive consistency for previously reported high-performance systems, where the relative prediction errors are generally below 5%. More importantly, the predicted high-performance region is not randomly distributed throughout the chemical space. Instead, three representative acceptors, DPz-F, LC301, and L8-BO, repeatedly emerge in combination with a diverse range of donor materials, as shown in Figure 7. Notably, these three acceptors all originate from modified Y6-series molecular architectures. Among them, L8-BO mainly involves side-chain engineering at the outer (β -position) of the Y6 framework. The introduction of branched alkyl side chains partially weakens excessive acceptor–acceptor aggregation while improving the formation of interconnected charge-transport pathways and strengthening donor–acceptor intermolecular interactions [62]. This structural optimization enables L8-BO systems to maintain balanced morphology while simultaneously improving charge transport and photocurrent generation. LC301 further introduces asymmetric modification on one side of the INCN-class end group. Such asymmetric molecular design enhances intramolecular charge redistribution and dipole characteristics, thereby improving molecular packing and reducing energy loss [63]. Compared with symmetric Y6-type acceptors, this strategy provides more flexible regulation of intermolecular interactions and electronic structure, allowing LC301-based systems to achieve improved V_{OC} while maintaining efficient charge separation. In

contrast, DPz-F adopts a stronger electron-withdrawing A' unit together with side-chain engineering at the inner (N-position), where two molecular subunits are connected through a benzene-core-linked dimer architecture. This type of dimerized acceptor design enhances intermolecular interactions, charge delocalization, and molecular packing, while simultaneously suppressing non-radiative recombination losses [12]. As a result, DPz-F-based systems exhibit excellent balance among photocurrent generation, charge transport, and voltage retention, making DPz-F the most dominant acceptor appearing in the top-ranked candidate systems. Overall, these three acceptors exhibit strong advantages in charge generation, intermolecular interaction, energy-loss suppression, and morphology regulation. Therefore, it is reasonable that the RF model repeatedly identifies them as representative high-performance acceptor frameworks within the explored chemical space.

The donor materials paired with these acceptors can be broadly categorized into three representative groups: D18 derivatives (D18, D18-Cl, D18-ChCl, and DTTz-40), PM6-derived copolymers (PQB-2, D3, and PM6-CNBT-15), and PFBZ, as summarized in Figure 8. The donor D units in these systems are predominantly based on fluorinated or chlorinated benzodithiophene (BDT) frameworks combined with various acceptor units to construct conjugated copolymer backbones. These donor systems provide balanced charge transport capability, suitable energy-level alignment, and favorable intermolecular interactions, allowing them to effectively cooperate with highly optimized Y6-derived acceptors.

Table 7 Top 10 donor–acceptor combinations with the highest predicted PCE identified from large-scale combinatorial screening using the retrained random forest model

Donor	Acceptor	Predicted PCE (%)	Experimental PCE (%)
D18	DPz-F	19.01	19.8
D18-Cl	DPz-F	18.95	-
DTTz-40	DPz-F	18.84	-
PQB-2	DPz-F	18.60	-
D18-ChCl	DPz-F	18.57	-
D3	LC301	18.53	-
D3	L8-BO	18.53	18.37
PM6-CNBT-15	DPz-F	18.50	-
PFBZ	DPz-F	18.47	-
D18	L8-BO	18.47	18.35

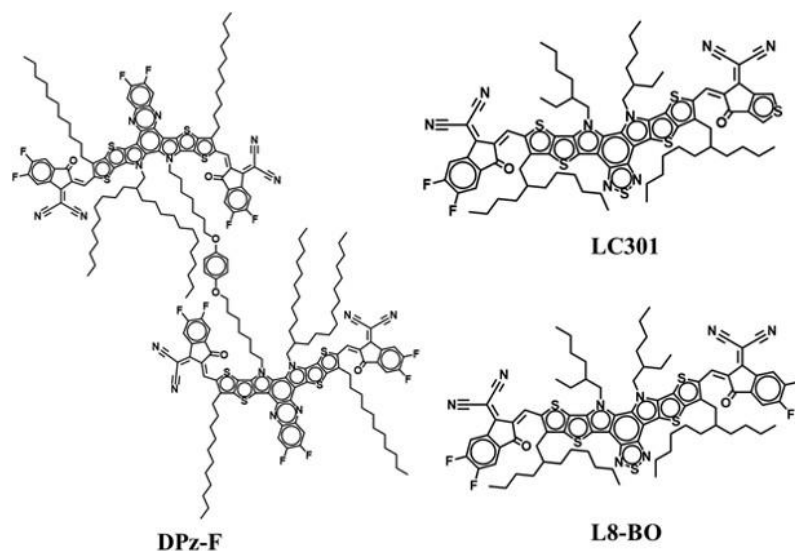


Fig. 7 Molecular structures of representative acceptors (DPz-F, LC301, and L8-BO) appearing in the top-ranked donor–acceptor combinations identified from large-scale combinatorial screening

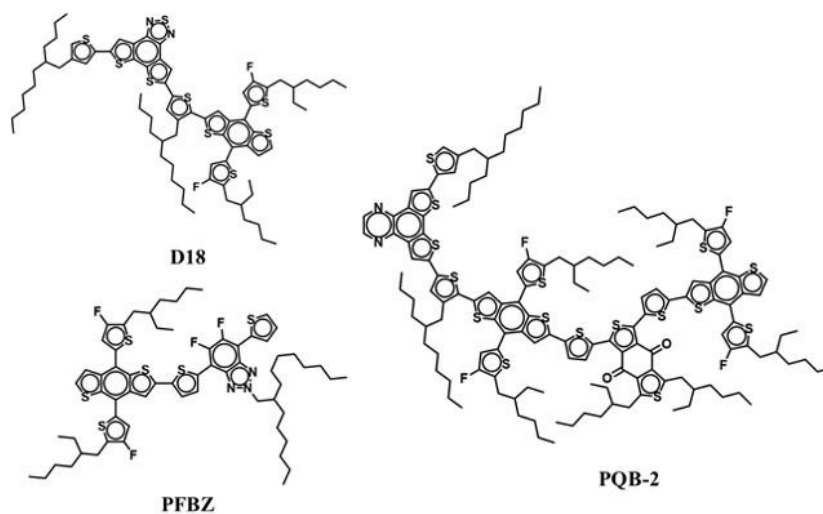


Fig. 8 Molecular structures of representative donors (D18, PFBZ, and PQB-2) appearing in the top-ranked donor–acceptor combinations identified from large-scale combinatorial screening

Overall, these results indicate that the random forest model is most suitable for screening new combinations within the known chemical space. Analysis of the top 10 predicted

combinations further reveals that high-efficiency PSC systems are not randomly distributed, but are instead concentrated around a limited number of well-optimized Y6-series acceptors with ADA'DA architecture. The results are strongly consistent with the SHAP analysis discussed in Section 3.2. These insights provide a practical design guideline:

- (1) prioritize Y6-derived acceptor frameworks with ADA'DA architectures, where INCN-class end groups together with optimized backbone and side-chain engineering consistently dominate high-PCE systems;
- (2) employ donor molecular design, particularly BDT-based copolymer backbones, to regulate energy-level alignment and suppress voltage loss while maintaining efficient charge transport;
- (3) both the SHAP analysis and experimental statistics indicate that halogenated, especially fluorinated, donor and acceptor materials are strongly associated with enhanced PSC performance;
- (4) utilize machine learning models for rapid combinatorial screening within this physically meaningful design space.

4. Conclusion

In this work, we establish the largest experimental database of 5,797 binary bulk-heterojunction polymer solar cell datasets reported between 2010 and 2025 and develop an interpretable structure-based machine learning framework. Random forest models using donor and acceptor Morgan fingerprints directly predict device performance, achieving $R = 0.85$ and

RMSE = 1.94% for PCE. Combined statistical and Shockley-Queisser limit analyses indicate that historical PCE gains are driven primarily by J_{SC} , while further improvements increasingly depend on V_{OC} and materials energy-level matching. Fragment-level SHAP analysis reveals specific structural roles. Y6-derived acceptors with ADA'DA architectures and INCN-type end groups contribute strongly to PCE and J_{SC} , whereas donor fragments primarily govern V_{OC} . Halogenated motifs, particularly fluorinated structures, on both donor and acceptor are consistently associated with favorable performance. These findings connect model interpretation with established molecular design strategies for high-performance PSCs.

Our model shows its reliability for new combinations of known materials, supporting its use for combinatorial screening within represented chemical space. Screening more than 2.2 million donor and acceptor pairs has identified a restricted high-performance region enriched in optimized Y6-derived acceptors and halogenated BDT-based polymer donors. Looking forward, model application and materials discovery could be further advanced by expanding new experimental data and incorporating geometry-aware molecular representations like three-dimensional molecular graphs. Overall, this work demonstrates that combining large-scale experimental data with interpretable machine learning enables not only accurate prediction but also physically meaningful understanding of structure-performance relationships, providing a practical framework for rational molecular design and accelerated discovery of next-generation polymer solar cells.

Acknowledgements This work was supported by the National Science and Technology Council (NSTC), Taiwan, under Grant Nos. MOST 110-2622-E-002-037, MOST 111-2221-E-002-025, NSTC 113-2221-E-002-017, NSTC 114-2221-E-002-046, and NSTC 114-2221-E-007-146; the Ministry of Agriculture, Taiwan (Grant No. 115AS-1.3.2-AS-13); and the National Tsing Hua University (NTHU) Interdisciplinary Research Project (Grant No. 115Q2717E1), supported by the Ministry of Education (MOE) Higher Education Sprout Project, Taiwan.

Author contributions Yuan-Fang Lee: Conceptualization, Data curation, Methodology, Software, Formal analysis, Validation, Visualization, Writing – original draft, Writing – review and editing. Yu-Hung Hsiao: Conceptualization, Methodology, Software. Te-An Chen: Conceptualization, Data curation, Methodology. Min-Hsuan Lee: Conceptualization, Methodology, Writing – review and editing. Hsuan-Hung Liu: Conceptualization, Methodology, Writing – review and editing. Ching-I Huang: Conceptualization, Formal analysis, Supervision, Project administration, Funding acquisition, Writing – review and editing.

Data availability The datasets constructed and analyzed during the current study are available in the Zenodo repository: <https://doi.org/10.5281/zenodo.16784519>

Code availability The underlying code for this study is available in OPV_RF_SHAP and can be accessed via this link: https://github.com/LeeYuanFang/OPV_RF_SHAP.git.

Declarations

Competing interests The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

- [1] Chen L, Pilania G, Batra R, Huan TD, Kim C, Kuenneth C, Ramprasad R (2021) Polymer informatics: current status and critical next steps. Mater Sci Eng: R: Rep 144:100595. <http://doi.org/10.1016/j.mser.2020.100595>
- [2] Olson GB (1997) Computational design of hierarchically structured materials. Science 277:1237-1242. <https://doi.org/10.1126/science.277.5330.1237>
- [3] Audus DJ, de Pablo JJ (2017) Polymer informatics: opportunities and challenges. ACS Macro Lett 6:1078-1082. <http://doi.org/10.1021/acsmacrolett.7b00228>

- [4] Sha W, Li Y, Tang S, Tian J, Zhao Y, Guo Y, Zhang W, Zhang X, Lu S, Cao YC, Cheng S (2021) Machine learning in polymer informatics. *InfoMat* 3:353-361. <http://doi.org/10.1002/inf2.12167>
- [5] Li G, Zhu R, Yang Y (2012) Polymer solar cells. *Nat Photonics* 6:153-161. <https://doi.org/10.1038/nphoton.2012.11>
- [6] Zhou H, Yang L, You W (2012) Rational design of high performance conjugated polymers for organic solar cells. *Macromolecules* 45:607-632. <https://doi.org/10.1021/ma201648t>
- [7] Lu B, Wang J, Zhang Z, Wang J, Yuan X, Ding Y, Wang Y, Yao Y (2021) Recent progress of Y-series electron acceptors for organic solar cells. *Nano Select* 2:2029-2039. <https://doi.org/10.1002/nano.202100036>
- [8] Mishra A, Sharma GD (2023) Harnessing the structure-performance relationships in designing non-fused ring acceptors for organic solar cells. *Angew Chem Int Ed* 62:e202219245. <https://doi.org/10.1002/anie.202219245>
- [9] Lin Y, Wang J, Zhang ZG, Bai H, Li Y, Zhu D, Zhan X (2015) An electron acceptor challenging fullerenes for efficient polymer solar cells. *Adv Mater* 27:1170-4. <https://doi.org/10.1002/adma.201404317>
- [10] Yuan J, Zhang Y, Zhou L, Zhang G, Yip HL, Lau TK, Lu X, Zhu C, Peng H, Johnson PA, Leclerc M, Cao Y, Ulanski J, Li Y, Zou Y (2019) Single-junction organic solar cell with over

15% efficiency using fused-ring acceptor with electron-deficient core. *Joule* 3:1140-1151.

<https://doi.org/10.1016/j.joule.2019.01.004>

[11] Fu J, Fong PWK, Liu H, Huang CS, Lu X, Lu S, Abdelsamie M, Kodalle T, Sutter-Fella CM, Yang Y, Li G (2023) 19.31% binary organic solar cell and low non-radiative recombination enabled by non-monotonic intermediate state transition. *Nat Commun* 14:1760.

<https://doi.org/10.1038/s41467-023-37526-5>

[12] Zeng L, Hu R, Zhang M, Lee S, Wang Q, Meng S, Chen Q, Liu J, Xue L, Mi L, Yang C, Zhang Z-G (2025) Halogen-substituted phenazine cores reduce energy losses and optimize carrier dynamics in tethered acceptors for 19.8% efficient and stable polymer solar cells.

Energy Environ Sci 18:6754-6763. <http://doi.org/10.1039/d5ee01686j>

[13] Shafian S, Mohd Salehin FN, Lee S, Ismail A, Mohamed Shuhidan S, Xie L, Kim K (2025) Development of organic semiconductor materials for organic solar cells via the integration of computational quantum chemistry and AI-powered machine learning. *ACS Appl Energy Mater*

8:699-722. <http://doi.org/10.1021/acsaem.4c02937>

[14] Olivares-Amaya R, Amador-Bedolla C, Hachmann J, Atahan-Evrenk S, Sánchez-Carrera RS, Vogt L, Aspuru-Guzik A (2011) Accelerated computational discovery of high-performance materials for organic photovoltaics by means of cheminformatics. *Energy Environ Sci* 4:4849.

<http://doi.org/10.1039/c1ee02056k>

- [15] Hachmann J, Olivares-Amaya R, Atahan-Evrenk S, Amador-Bedolla C, Sánchez-Carrera RS, Gold-Parker A, Vogt L, Brockway AM, Aspuru-Guzik A (2011) The Harvard Clean Energy Project: large-scale computational screening and design of organic photovoltaics on the world community grid. *J Phys Chem Lett* 2:2241-2251. <http://doi.org/10.1021/jz200866s>
- [16] Hachmann J, Olivares-Amaya R, Jinich A, Appleton AL, Blood-Forsythe MA, Seress LR, Román-Salgado C, Trepte K, Atahan-Evrenk S, Er S, Shrestha S, Mondal R, Sokolov A, Bao Z, Aspuru-Guzik A (2014) Lead candidates for high-performance organic photovoltaics from high-throughput quantum chemistry - the Harvard Clean Energy Project. *Energy Environ Sci* 7:698-704. <http://doi.org/10.1039/c3ee42756k>
- [17] Lopez SA, Sanchez-Lengeling B, de Goes Soares J, Aspuru-Guzik A (2017) Design principles and top non-fullerene acceptor candidates for organic photovoltaics. *Joule* 1:857-870. <http://doi.org/10.1016/j.joule.2017.10.006>
- [18] Lee MH (2020) Robust random forest based non-fullerene organic solar cells efficiency prediction. *Org Electron* 76:105465. <http://doi.org/10.1016/j.orgel.2019.105465>
- [19] Lee MH (2022) Identifying correlation between the open-circuit voltage and the frontier orbital energies of non-fullerene organic solar cells based on interpretable machine-learning approaches. *Sol Energy* 234:360-367. <http://doi.org/10.1016/j.solener.2022.02.010>
- [20] Sun W, Zheng Y, Yang K, Zhang Q, Shah AA, Wu Z, Sun Y, Feng L, Chen D, Xiao Z, Lu S, Li Y, Sun K (2019) Machine learning-assisted molecular design and efficiency prediction

for high-performance organic photovoltaic materials. *Sci Adv* 5:eaay4275.

<http://doi.org/doi:10.1126/sciadv.aay4275>

[21] Zhu Z, Zhu C, Tu Y, Shao T, Wang Y, Liu W, Liu Y, Zang Y, Wei Q, Yan W (2024) Machine-learning-assisted exploration of new non-fullerene acceptors for high-efficiency organic solar cells. *Cell Rep Phys Sci* 5:102316. <http://doi.org/10.1016/j.xcrp.2024.102316>

[22] Zhang CR, Lv LF, Li M, Liu X-M, Gong JJ, Liu ZJ, Wu YZ, Chen HS (2025) High throughput molecular design of electron donors and non-fullerene acceptors using machine learning combined with substructure importance. *J Mater Chem C* 13:14864-14874. <http://doi.org/10.1039/d5tc00931f>

[23] Padula D, Troisi A (2019) Concurrent optimization of organic donor–acceptor pairs through machine learning. *Adv Energy Mater* 9:1902463. <http://doi.org/10.1002/aenm.201902463>

[24] Kranthiraja K, Saeki A (2021) Experiment-oriented machine learning of polymer:non-fullerene organic solar cells. *Adv Funct Mater* 31:2011168. <http://doi.org/10.1002/adfm.202011168>

[25] Sahu H, Yang F, Ye X, Ma J, Fang W, Ma H (2019) Designing promising molecules for organic solar cells via machine learning assisted virtual screening. *J Mater Chem A* 7:17480-17488. <http://doi.org/10.1039/c9ta04097h>

- [26] Wu Y, Guo J, Sun R, Min J (2020) Machine learning for accelerating the discovery of high-performance donor/acceptor pairs in non-fullerene organic solar cells. *npj Comput Mater* 6:120. <http://doi.org/10.1038/s41524-020-00388-2>
- [27] Siddiqui H, Usmani T (2024) Interpretable AI and machine learning classification for identifying high-efficiency donor-acceptor pairs in organic solar cells. *ACS Omega* 9:34445-34455. <http://doi.org/10.1021/acsomega.4c02157>
- [28] Padula D, Simpson JD, Troisi A (2019) Combining electronic and structural features in machine learning models to predict organic solar cells properties. *Mater Horiz* 6:343-349. <http://doi.org/10.1039/c8mh01135d>
- [29] Sahu H, Ma H (2019) Unraveling correlations between molecular properties and device parameters of organic solar cells using machine learning. *J Phys Chem Lett* 10:7277-7284. <http://doi.org/10.1021/acs.jpcclett.9b02772>
- [30] Huang Y, Zhang J, Jiang ES, Oya Y, Saeki A, Kikugawa G, Okabe T, Ohuchi FS (2020) Structure–property correlation study for organic photovoltaic polymer materials using data science approach. *Phys Chem C* 124:12871-12882. <https://doi.org/10.1021/acs.jpcc.0c00517>
- [31] Zhao Z-W, del Cueto M, Geng Y, Troisi A (2020) Effect of increasing the descriptor set on machine learning prediction of small molecule-based organic solar cells. *Chem Mater* 32:7777-7787. <http://doi.org/10.1021/acs.chemmater.0c02325>

- [32] Miyake Y, Saeki A (2021) Machine learning-assisted development of organic solar cell materials: Issues, analyses, and outlooks. *J Phys Chem Lett* 12:12391-12401. <http://doi.org/10.1021/acs.jpcclett.1c03526>
- [33] Sun W, Zheng Y, Zhang Q, Yang K, Chen H, Cho Y, Fu J, Odunmbaku O, Shah AA, Xiao Z, Lu S, Chen S, Li M, Qin B, Yang C, Frauenheim T, Sun K (2021) Artificial intelligence designer for highly-efficient organic photovoltaic materials. *J Phys Chem Lett* 12:8847-8854. <http://doi.org/10.1021/acs.jpcclett.1c02554>
- [34] Wen Y, Liu Y, Yan B, Gaudin T, Ma J, Ma H (2021) Simultaneous optimization of donor/acceptor pairs and device specifications for nonfullerene organic solar cells using a qspr model with morphological descriptors. *J Phys Chem Lett* 12:4980-4986. <http://doi.org/10.1021/acs.jpcclett.1c01099>
- [35] Kranthiraja K, Saeki A (2022) Machine learning-assisted polymer design for improving the performance of non-fullerene organic solar cells. *ACS Appl Mater Interfaces* 14:28936-28944. <http://doi.org/10.1021/acsami.2c06077>
- [36] Suthar R, T A, Karak S (2023) Machine-learning-guided prediction of photovoltaic performance of non-fullerene organic solar cells using novel molecular and structural descriptors. *J Mater Chem A* 11:22248-22258. <http://doi.org/10.1039/d3ta04603f>

- [37] Zhang CR, Li M, Zhao M, Gong JJ, Liu XM, Chen YH, Liu ZJ, Wu YZ, Chen HS (2023) Machine learning study on organic solar cells and virtual screening of designed non-fullerene acceptors. *J Appl Phys* 134:153104. <http://doi.org/10.1063/5.0169284>
- [38] Das B, Mondal A (2024) Predictive modeling and design of organic solar cells: A data-driven approach for material innovation. *ACS Appl Energy Mater* 7:9349-9363. <http://doi.org/10.1021/acsaem.4c01847>
- [39] Khatua R, Das B, Mondal A (2024) Physics-informed machine learning with data-driven equations for predicting organic solar cell performance. *ACS Appl Mater Interfaces* 16:57467-57480. <http://doi.org/10.1021/acsami.4c10868>
- [40] Li M, Zhang CR, Zhang ML, Gong JJ, Liu XM, Chen YH, Liu ZJ, Wu YZ, Chen HS (2024) Machine learning study on the virtual screening of donor–acceptor pairs for organic solar cells. *Phys Status Solidi (A)* 221:2400008. <http://doi.org/10.1002/pssa.202400008>
- [41] Lv LF, Zhang CR, Cao R, Liu XM, Zhang ML, Gong JJ, Liu ZJ, Wu YZ, Chen HS (2024) Design and virtual screening of donor and non-fullerene acceptor for organic solar cells using long short-term memory model. *J Mater Chem A* 12:23859-23871. <http://doi.org/10.1039/d4ta04665j>
- [42] Cao R, Zhang CR, Liu XM, Gong JJ, Zhang ML, Liu ZJ, Wu YZ, Chen HS (2025) Molecular design of organic photovoltaic donors and non-fullerene acceptors: a combined

machine learning and genetic algorithm approach. *J Mater Chem C* 13:12150-12168.

<http://doi.org/10.1039/d5tc02014j>

[43] Chen LQ, Zhang CR, Sang CC, Liu XM, Gong JJ, Zhang ML, Chen HS (2025) High-throughput molecular design of donors and non-fullerene acceptors for organic solar cells based on convolutional neural networks. *J Chem Inf Model* 65:10107-10123.

<http://doi.org/10.1021/acs.jcim.5c01634>

[44] Rogers D, Hahn M (2010) Extended-connectivity fingerprints. *J Chem Inf Model* 50:742-754. <https://doi.org/10.1021/ci100050t>

[45] Lundberg SM, Lee S-I (2017) A unified approach to interpreting model predictions. *Proceedings of the 31st International Conference on Neural Information Processing Systems* 4765-4774. <https://doi.org/10.48550/arXiv.1705.07874>

[46] Lundberg SM, Erion G, Chen H, DeGrave A, Prutkin JM, Nair B, Katz R, Himmelfarb J, Bansal N, Lee S-I (2020) From local explanations to global understanding with explainable AI for trees. *Nat Mach Intell* 2:56-67. <https://doi.org/10.1038/s42256-019-0138-9>

[47] Shockley W, Queisser HJ (1961) Detailed balance limit of efficiency of p-n junction solar cells. *J Appl Phys* 32:510-519. <http://doi.org/10.1063/1.1736034>

[48] Rühle S (2016) Tabulated values of the Shockley–Queisser limit for single junction solar cells. *Sol Energy* 130:139-147. <http://doi.org/10.1016/j.solener.2016.02.015>

- [49] Rasool S, Hoang QV, Van Vu D, Trang Bui TT, Jin S-M, Ho TT, Song CE, Lee HK, Lee SK, Lee JC, Moon SJ, Lee E, Shin WS (2019) High-efficiency non-halogenated solvent processable polymer/pcbm solar cells via fluorination-enabled optimized nanoscale morphology. *J Mater Chem A* 7:24992-25002. <http://doi.org/10.1039/c9ta08960h>
- [50] Zhang W, Hu R, Zeng X, Su X, Chen Z, Zou X, Peng J, Zhang C, Yartsev A (2019) Effect of post-thermal annealing on the performance and charge photogeneration dynamics of PffBT4T-2OD/PC(71)BM solar cells. *Polymers (Basel)* 11:408. <http://doi.org/10.3390/polym11030408>
- [51] Pang B, Liao C, Xu X, Yu L, Li R, Peng Q (2023) Benzo[d]thiazole based wide bandgap donor polymers enable 19.54% efficiency organic solar cells along with desirable batch-to-batch reproducibility and general applicability. *Adv Mater* 35:e2300631. <http://doi.org/10.1002/adma.202300631>
- [52] Lee SW, Hussain MW, Lee J, Whang DR, Jeong WH, Choi H, Chang DW (2023) Effect of chlorine substituents on the photovoltaic properties of monocyanated quinoxaline-based D-A-type polymers. *ACS Appl Mater Interfaces* 15:5547-5555. <http://doi.org/10.1021/acsami.2c19702>
- [53] Sun H, Liu T, Yu J, Lau TK, Zhang G, Zhang Y, Su M, Tang Y, Ma R, Liu B, Liang J, Feng K, Lu X, Guo X, Gao F, Yan H (2019) A monothiophene unit incorporating both fluoro and

ester substitution enabling high-performance donor polymers for non-fullerene solar cells with 16.4% efficiency. *Energy Environ Sci* 12:3328-3337. <https://doi.org/10.1039/c9ee01890e>

[54] Kim MS, Kim BG, Kim J (2009) Effective variables to control the fill factor of organic photovoltaic cells. *ACS Appl Mater Interfaces* 1:1264-9. <http://doi.org/10.1021/am900155y>

[55] Kumagai M, Ando Y, Tanaka A, Tsuda K, Katsura Y, Kurosaki K (2022) Effects of data bias on machine-learning-based material discovery using experimental property data. *Sci Technol Adv Mater: Methods* 2:302-309. <http://doi.org/10.1080/27660400.2022.2109447>

[56] Yuan J, Zhang H, Zhang R, Wang Y, Hou J, Leclerc M, Zhan X, Huang F, Gao F, Zou Y, Li Y (2020) Reducing voltage losses in the A-DA'D-A acceptor-based organic solar cells. *Chem* 6:2147-2161. <http://doi.org/10.1016/j.chempr.2020.08.003>

[57] Ma Y, Liang Q, Wu H (2025) Role of donor fluorination in voltage losses of organic solar cells. *Next Materials* 8:100546. <http://doi.org/10.1016/j.nxmater.2025.100546>

[58] Deng M, Xu X, Duan Y, Qiu W, Yu L, Li R, Peng Q (2024) 19.32% efficiency polymer solar cells enabled by fine-tuning stacking modes of Y-type molecule acceptors: synergistic bromine and fluorine substitution of the end groups. *Adv Mater* 36:e2308216. <http://doi.org/10.1002/adma.202308216>

[59] Leclerc N, Chavez P, Ibraikulov OA, Heiser T, Leveque P (2016) Impact of backbone fluorination on pi-conjugated polymers in organic photovoltaic devices: a review. *Polymers (Basel)* 8: <https://doi.org/10.3390/polym8010011>

- [60] Li J, Lin M, Wang H, Iqbal J, Wang K (2025) New advances in benzodithiophene (BDT)-based small molecules in non-fullerence organic solar cells. *Dyes Pigm* 233:112508. <http://doi.org/10.1016/j.dyepig.2024.112508>
- [61] Sharma VV, Landep A, Lee SY, Park SJ, Kim YH, Kim GH (2024) Recent advances in polymeric and small molecule donor materials for Y6 based organic solar cells. *Next Energy* 2:100086. <http://doi.org/10.1016/j.nxener.2023.100086>
- [62] Lei Y, Meng X, Guo Q, Sun X, Zhou E (2025) Recently progress of aromatic side chain engineering on Y6 for high performance organic solar cells. *Nano Res* 18:94907513. <http://doi.org/10.26599/nr.2025.94907513>
- [63] Li X, Duan X, Liang Z, Yan L, Yang Y, Qiao J, Hao X, Zhang C, Zhang J, Li Y, Huang F, Sun Y (2022) Benzo[1,2-b:4,5-b']difuran based polymer donor for high-efficiency (>16%) and stable organic solar cells. *Adv Energy Mater* 12:2103684. <http://doi.org/10.1002/aenm.202103684>