Uncovering Structure–Performance Relationships in Organic Photovoltaics: Interpretable Machine Learning Model for Predicting the Power Conversion Efficiency

Yi Yang, Arowa Yasmeen, and Ovidiu Daescu

***Abstract*—Organic photovoltaics (OPVs) represent a promising photovoltaic technology, but the design of candidate molecules has traditionally followed a trial-and-error approach, which is inefficient. However, machine learning provides a data-informed strategy by learning from large OPV material datasets, supporting the accelerated discovery and optimization of high-performance OPV materials. In this study, we use an extreme gradient boosting (XGBoost) model to predict the density functional theory-calculated power conversion efficiency (PCE) of OPV donor materials using structural features extracted from the Harvard Photovoltaic Dataset (HOPV15) dataset. To enhance predictive performance, we select the most informative molecular fingerprints based on the averaged feature importance scores from both random forest and XGBoost. Our XGBoost model achieves state-of-the-art predictive accuracy on HOPV15 with $R^2 = 0.918$ and RMSE $=$ 0.302%, outperforming prior methods. Using SHapley Additive exPlanations, we identify key Morgan and PubChem substructures that influence PCE, offering interpretable insights. This framework supports accurate, explainable OPV prediction and holds promise for high-throughput screening.**

***Index Terms*—Extreme gradient boosting (XGBoost), feature selection, machine learning (ML), molecular fingerprints, organic photovoltaics (OPVs), power conversion efficiency (PCE), random forest (RF), SHapley Additive exPlanations (SHAP) analysis.**

I. Introduction

With the rapid population growth of global energy demand, organic photovoltaics (OPVs), due to their low-cost production and structural flexibility, have emerged as a promising alternative to traditional silicon-based solar cells [1]. OPVs utilize π-conjugated organic molecules as active layer materials, where power conversion efficiency (PCE) is mainly decided by the energy gap between the highest occupied molecular orbital (HOMO) of the donor and the lowest unoccupied molecular orbital (LUMO) of the acceptor [2], [3]. PCE can be measured through experimental setups evaluating V_{OC} (open-circuit potential), J_{SC} (short-circuit density), P_{in} (input power density, and fill factor, but these methods suffer from inconsistencies due to variations in fabrication conditions and measurement protocols [4], [5]. Computational approaches, such as density functional theory (DFT) and the Scharber model (see Section B, Eq. S1, Supplementary), offer theoretical estimates for PCE but are computationally expensive and limited in accuracy for complex OPV systems [6], [7].

Machine learning (ML) has emerged as a powerful alternative, capturing nonlinear relationships between molecular structures and photovoltaic properties. ML-based predictions of OPV efficiency have been remarkably successful, yielding R^2 values (see Section B, Supplementary) ranging from 0.6 to 0.93, depending on the quality of datasets, the ML algorithms, and choice of features [7], [8]. However, the training of ML models on experimental datasets can still be challenging, as they are prone to noise, missing values, or dataset consistency issues [9], [10].

In this study, we propose an ML framework for predicting computationally generated PCE at high accuracy, thereby addressing gaps in experimental data, as outlined previously. Using the Harvard Photovoltaic Dataset (HOPV15) [11], we applied extreme gradient boosting (XGBoost) and random forest (RF) for feature selection, prioritizing molecular descriptors such as LUMO, HOMO, Morgan fingerprints, and PubChem fingerprints. SHapley Additive exPlanation (SHAP) analysis reveals that LUMO is the dominant predictor of PCE, followed by HOMO.

Our XGBoost model achieved state-of-the-art predictive performance on the test set (R^2=0.918, RMSE=0.302%, MAE=0.148%) and demonstrated robust transferability on donor molecules from the Clean Energy Project Database (CEPDB), yielding a strong Pearson correlation of 0.76. SHAP analysis revealed LUMO and HOMO as dominant predictors, while structural features encoded by Morgan and PubChem fingerprints provided additional predictive value. Unlike prior studies focused solely on accuracy or feature ranking, we uniquely applied SHAP interaction values alongside Spearman correlation to uncover synergistic effects between structural and electronic descriptors. This approach identified substructures—Morgan bits 790, 722, 870 (linked to high PCE) and PubChem bit 473 (linked to low PCE)—which we statistically validated

Received 25 March 2025; revised 15 May 2025 and 25 June 2025; accepted 19 July 2025. Date of publication 12 August 2025; date of current version 23 October 2025. *(Corresponding author: Ovidiu Daescu.)*

The authors are with the Department of Computer Science, University of Texas at Dallas, Richardson, TX 75080 USA (e-mail: Ovidiu.Daescu@utdallas.edu).

The source code is available at https://github.com/AI4Science2025/OPV_HOPV15.

This article has supplementary downloadable material available at https://doi.org/10.1109/JPHOTOV.2025.3592683, provided by the authors.

Digital Object Identifier 10.1109/JPHOTOV.2025.3592683

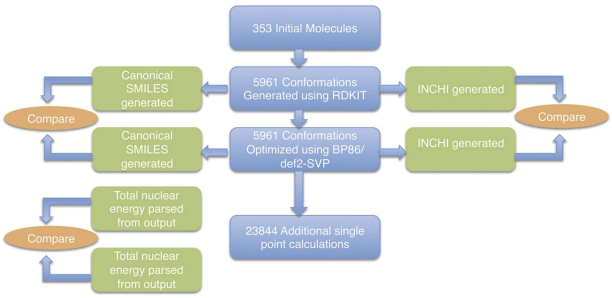

Fig. 1. Computational workflow for the generation of the HOPV15 dataset. Reproduced from López et al.'s [12], *Scientific Data*, 2016, under a CC BY license.

in high-performing donor versus low-performing donor. These findings offer a chemically grounded, interpretable framework for guiding future OPV material design.

II. Methodology

A. HOPV15 Dataset

The HOPV15 dataset, originally developed and published by López et al. [12], comprises computational and experimental photovoltaic properties of organic solar cell donors. Due to inconsistencies and missing values in experimental data, computational data—where HOMO and LUMO energy levels are derived from DFT and PCE is estimated using the Scharber model—provides a more reliable resource for ML. In the computational HOPV15 dataset, each donor is represented by its simplified molecular input line entry system (SMILES) string [13], along with details on molecular type (small molecule/polymer), solar cell architecture (bulk, bilayer, DSSC), and fullerene-based acceptor materials [11]. It serves as a complete resource for ML modeling, including molecular descriptors (HOMO, LUMO, HOMO–LUMO gap) and photovoltaic properties (V_{OC}, J_{SC}, and PCE). The computational pipeline for generating the HOPV15 is shown in Fig. 1. The workflow begins with SMILES-based conformer generation via RDKit [14], producing 1500 initial 3-D structures, minimized via force-field methods. After removing duplicates, up to 20 low-energy conformers per molecule were optimized using the BP86 functional [15] and the def3-SVP basis set, followed by single-point energy calculations with PBE0 [16], B3LYP [17], M06-2X [18], and BP86 functionals. For each functional, the molecular properties were averaged across conformers, as conformational effects on electronic properties were found to be negligible [7]. For ML modeling, we use HOMO and LUMO energy levels calculated at the B3LYP/def2-SVP level of theory, consistent with previous studies on this dataset [7]. B3LYP is a widely benchmarked hybrid functional, offering reliable performance across organic systems, making it a suitable choice when the optimal functional is uncertain.

B. Data Processing

Outliers were removed using z-score thresholding, where data points with feature values exceeding three standard deviations (z-score > 3) from the mean were excluded. A z-score of 3 indicates that the data point lies in the extreme 0.3% of a normal distribution, making it a likely outlier. After this step, 347 data points out of 350 remained. Next, data points with PCE values below zero were removed, as negative PCE values are nonphysical, reducing the dataset to 345 data points. To eliminate redundant molecular representations, isomeric molecules were identified and excluded. This process involved canonicalizing SMILES representations using RDKit, ensuring that structurally equivalent molecules were assigned the same canonical SMILES string. Duplicate isomeric molecules were grouped, and only one representative molecule per group was retained, leading to the removal of three molecules. After this step, the dataset contained 342 unique donor molecules. To refine the feature set and prevent redundancy, features exhibiting high linear correlation were removed. Specifically, pairs of features with a Pearson correlation coefficient exceeding 0.8 were identified, and one feature from each highly correlated pair was excluded. To ensure a balanced distribution of target values during model training and evaluation, the dataset was stratified by binning the continuous target values into quantile-based categories, which divides the target values into bins of equal frequency. This approach allowed the stratified cross-validation method to maintain proportional representation of different ranges of the target variable across the training and validation splits, thereby improving the robustness and generalizability of the model evaluation. For the train-test split, the dataset was divided into training/validation (80%) and test sets (20%) while preserving the stratification. In addition, during cross-validation, the training data was further stratified using the same binning strategy, ensuring that each fold maintained a consistent target value distribution.

C. Generation of Fingerprints

The features used in this study fall into two categories: structural fingerprints and physical/electronic descriptors. Molecular fingerprints serve as abstract representations of chemical structures and are commonly used in substructure analysis and similarity searches. Here, we employ MACCS keys, Morgan fingerprints, and PubChem fingerprints.

MACCS keys consist of 166 predefined binary structural features widely used in cheminformatics [19]. PubChem fingerprints, an 881-bit binary vector, encode the presence or absence of predefined substructure templates [20]. Morgan fingerprints—derived from the extended-connectivity fingerprint algorithm—use a circular, atom-centered approach to encode molecular structure [21]. Each atom's environment is iteratively expanded up to a specified radius, incorporating neighboring atoms and bonds, and then hashed into a fixed-length bit vector. This makes Morgan fingerprints more flexible and data-driven than fragment-based methods, such as PubChem and MACCS. In this work, we used a radius of 5 and a vector size of 2048 bits to capture detailed topological features.

In addition, RDKit descriptors (210 bits) capture physicochemical (e.g., molecular weight), topological (e.g., polar surface area), and fragment-based (e.g., aromatic rings) properties. These were computed from SMILES strings using RDKit version 2024.09.4, while PubChem fingerprints were generated

via the online platform ChemDes [22]. The combination of fingerprints and descriptors ensures comprehensive molecular characterization for ML model development.

D. Feature Screening

The fingerprints and descriptors described in Section II-C contribute to a large, high-dimensional feature set. Excessive features in ML models increase complexity and redundancy, causing overfitting [23] and escalating computational costs [24]. Feature selection was based on significance scores derived from RF and XGBoost, selected for their strong predictive performance (see Section III-A) and complementary learning strategies: RF's ensemble learning and XGBoost's gradient boosting. Building on prior work combining CatBoost and XGBoost for feature selection [8], we adopt a similar approach using RF and XGBoost. Importance scores from both models were independently normalized using MinMaxScaler [(1), normalization metric] and then averaged [(2), aggregation metric] to ensure a more robust ranking, leveraging RF's ensemble evaluation and XGBoost's refinement for reliable feature selection

$$F_{\text{norm}} = \frac{F - F_{\text{min}}}{F_{\text{max}} - F_{\text{min}}} \tag{1}$$

$$F_{\text{combined}} = \frac{F_{\text{norm, RF}} + F_{\text{norm, XGB}}}{2}. \tag{2}$$

E. ML Models and Validation Techniques

A range of ML models, including RF, decision tree, XGBoost, support vector regression (SVR), and ElasticNet, were evaluated to capture diverse modeling paradigms (see Supplementary Section A for an overview). RF and XGBoost were chosen for their effectiveness in handling high-dimensional, nonlinear relationships using ensemble learning. Decision tree served as a simple baseline, while SVR (RBF kernel) was selected for its ability to model complex feature interactions. ElasticNet was included for its capability to address collinearity and perform automatic feature selection. All models were implemented using Scikit-learn 1.6 for reproducibility. Hyperparameter tuning was performed using nested cross-validation with an exhaustive grid search. The outer loop applied stratified binning to split the dataset into training and test sets, ensuring a balanced distribution of the target variable. In the inner loop, models were trained and validated using cross-validation across different hyperparameter sets, selecting the best configuration based on averaged R^2, RMSE, and MAE. The final model was then retrained on the full training dataset and evaluated on an independent test set to assess generalization and minimize overfitting [25]. To ensure stability, the fine-tuned model was tested across 30 random states, with prediction variance measuring robustness. *Y*-randomization further validated model reliability by shuffling response values—poor performance on scrambled data confirmed that predictions were based on real patterns rather than chance [26].

F. SHAP-Based Analysis of Feature Importance

SHAP [27] is a game-theoretic approach that interprets ML predictions by quantifying each feature's contribution. It decomposes a model's output into a baseline value (mean target value) and feature-specific SHAP values

$$n_i = n_{\text{base}} + f(m_{i1}) + f(m_{i2}) + \cdots + f(m_{ik})$$

where $f(m_{ij})$ represents the SHAP value of feature j for sample i. Positive values indicate an increase in prediction, while negative values suggest a decrease. Unlike traditional feature importance methods, SHAP provides both global (overall importance) and local (individual impact) explanations, clarifying not just which features matter but also their directional influence. In this study, SHAP was used to assess how HOMO, LUMO, and molecular fingerprints contribute to OPV property predictions, enhancing interpretability and feature selection insights.

TABLE I
PERFORMANCE COMPARISON OF ML MODELS

Model	Train R^2	Train RMSE%	Test R^2	Test RMSE%
RF	0.787± 0.012	0.458± 0.014	0.775	0.502
Decision tree	0.509± 0.089	0.686± 0.059	0.491	0.755
XGBoost	0.889± 0.021	0.305± 0.028	0.825	0.443
SVR	0.677± 0.035	0.555± 0.031	0.712	0.570
ElasticNet	−0.002± 0.001	1.010± 0.001	−0.001	1.058

III. RESULTS AND DISCUSSION

A. Model Development

This study aims to develop an ML-based approach to predict Scharber PCE using molecular properties, providing a data-driven alternative to Scharber model that enhances interpretability and quantifies key feature contributions. To achieve this, we evaluated RF, decision tree, XGBoost, SVR, and ElasticNet (see Section II-E), selecting a diverse range of linear and nonlinear models. Given the high dimensionality of molecular fingerprints, we addressed overfitting risks by incorporating HOMO and LUMO energies alongside MACCS keys as the base feature set. Prior study [28] has demonstrated that combining electronic properties, such as HOMO and LUMO, with structural fingerprints improves ML models' predictive accuracy for organic solar cells. Following this insight, we prioritized these features to establish a compact yet informative starting point for model comparison. In addition, MACCS keys, with a lower dimensionality, reduce computational complexity and overfitting, making them ideal for initial feature selection. Table I summarizes model performance. XGBoost achieved the highest test $R^2 = 0.825$ and lowest test RMSE $= 0.443\%$, with minimal deviation from the training R^2 (0.889 ± 0.021), indicating strong generalization. RF also performed well (R^2=0.775, RMSE $= 0.502\%$), demonstrating robustness in handling high-dimensional data. Decision tree, SVR, and ElasticNet showed weaker performance, with higher RMSE and lower R^2, making them less suitable. Given their predictive power and interpretability, RF and XGBoost were chosen for feature selection, with importance scores combined to identify key features. Due to XGBoost's superior performance, subsequent experiments in this study utilize this model exclusively.

TABLE II
PERFORMANCE METRICS FOR DIFFERENT FEATURE SETS IN PCE PREDICTION

Feature set	R^2	RMSE%	MAE%
LUMO	0.7810	0.4950	0.1380
+HOMO	0.8190	0.4350	0.1390
+MACCS50	0.7920	0.4810	0.2440
+MACCS30	0.7870	0.4640	0.2400
+MACCS10	0.7860	0.4890	0.2750
+MACCS5	0.7770	0.5050	0.2570
+PubChem50	0.8120	0.4570	0.2440
+PubChem30	0.8010	0.4710	0.2450
+PubChem10	0.8290	0.4370	0.2330
+PubChem5	0.8000	0.4720	0.2560
+RDKit Desc50	0.7740	0.5020	0.2570
+RDKit Desc30	0.7670	0.5090	0.2360
+RDKit Desc10	0.7600	0.5170	0.2540
+RDKit Desc5	0.7490	0.5280	0.2550
+Morgan50	0.8390	0.4230	0.2570
+Morgan30	0.7940	0.4500	0.2440
+Morgan10	0.8490	0.4090	0.2200
+Morgan5	0.7610	0.5100	0.2580

B. Feature Selection

Using the feature selection method from Section II-D, we formed and evaluated subsets of LUMO, HOMO, and either a fingerprint type (MACCS, PubChem, Morgan) or RDKit descriptors. Top features were identified based on averaged importance scores, with subsets selected from the top 50 to top 5. LUMO consistently emerged as the most influential feature across all thresholds and all feature combinations. The top 5 features across feature combinations are shown in Supplementary Fig. S1. LUMO's consistent dominance aligns with the research of Lee et al. [29], who found donor LUMO to be the most important feature when using RF for PCE prediction. To optimize predictive accuracy while minimizing feature count, we established donor LUMO as the baseline, comparing its performance against feature sets incorporating additional structural information.

We trained XGBoost models with different feature combinations, including LUMO alone, HOMO+LUMO, and LUMO with selected structural descriptors. Table II shows that HOMO+LUMO improves performance over LUMO only (R^2 = 0.819 versus 0.781, RMSE = 0.436% versus 0.495%), underscoring their complementary roles in capturing electronic properties. Among structural descriptors, PubChem and Morgan fingerprints enhance performance the most, while RDKit and MACCS fail to significantly improve the model performance.

Morgan fingerprints outperform PubChem, with the subset including the top 9 Morgan fingerprints (+Morgan10) subset achieving the highest R^2 (0.849) and lowest RMSE (0.409%) and MAE (0.22%). Larger Morgan sets show diminishing returns, while overly reducing features (e.g., +Morgan5) lowers accuracy. PubChem fingerprints also contribute significantly, with the +PubChem10 subset achieving an R^2 of 0.829, RMSE of 0.437%, and MAE of 0.233%. However, increasing the PubChem feature set to +PubChem30 results in a performance decline (R^2 = 0.801), indicating that a carefully selected smaller subset provides the best predictive performance.

TABLE III
PERFORMANCE METRICS OF THE XGBOOST MODEL FOR PCE PREDICTION

Metric	Training set (Mean ± Std)	Test set
RMSE%	0.319 ± 0.020	0.302
MAE%	0.159 ± 0.010	0.148
R^2	0.874 ± 0.020	0.918

The table reports the average performance (mean ± std) on the training set and the corresponding values on the test set.

Morgan fingerprints, adjustable in radius and bit size, encode varying levels of structural details. In this research, we used the commonly chosen 2048-bit radius to maximize structural information, guided by the findings of Zhu et al. [30], who showed that this configuration outperformed MACCS and smaller radius Morgan fingerprints, achieving an RMSE of 2.20% and an R^2 of 0.7121 on experimentally verified donor–acceptor combinations.

While PubChem fingerprints contribute to PCE predictions, their combination with LUMO only does not significantly surpass the predictive power of LUMO and HOMO, likely due to their shorter bit length (881 bits), which encodes less structural information. Similarly, MACCS keys and RDKit descriptors fail to enhance predictions. MACCS keys, with a significantly shorter bit length, provide limited structural detail. RDKit descriptors primarily encode global molecular properties, including physical descriptors such as molecular weight and topological polar surface area, rather than fine-grained structural variations essential for PCE modeling. Similarly, Zhao et al. [31] found that physical descriptors alone lack the predictive power of fingerprints. These findings highlight the critical role of fingerprint length and structural detail in model performance, with Morgan fingerprints proving most effective for PCE prediction.

C. Model Performance Enhancement and Generalization Evaluation

To further improve performance, we integrated top-performing subsets from PubChem and Morgan fingerprints with HOMO and LUMO. Given the impact of acceptor materials on OPV efficiency [32], [33], [34], acceptor-type information was added to better capture donor–acceptor interactions and electronic dynamics [35].

The unified feature set, consisting of one-hot encoded acceptor type, HOMO, LUMO, the top 9 Morgan and the top 9 PubChem fingerprints, was prepared for training and evaluation. Pearson correlation analysis was conducted to remove highly correlated features. The resulting correlation heatmap (see Fig. 2) confirms linear independence among features with Pearson coefficients $r < 0.8$, ensuring they provide complementary information for accurate PCE prediction. The unified feature set effectively captures both structural and electronic properties, improving PCE prediction. As shown in Table III, the XGBoost model trained on the unified feature set demonstrates strong predictive performance. On the test set, it achieves an impressive R^2 of 0.918, surpassing the training set (R^2 =0.874), indicating

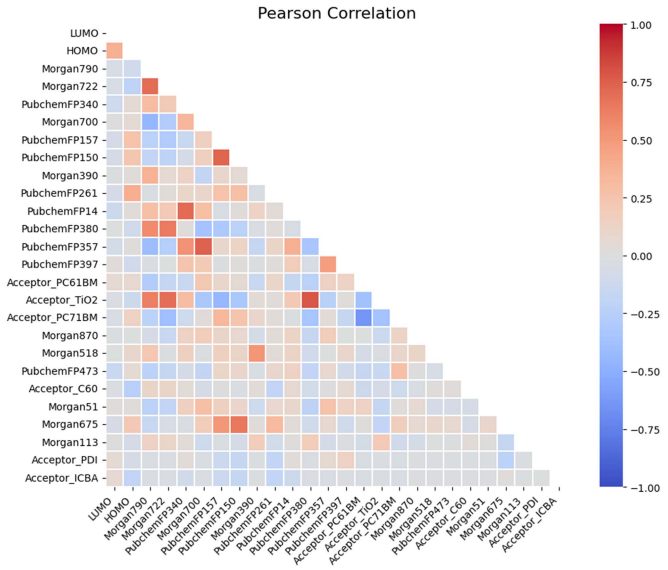

Fig. 2. Pearson correlation heatmap of the unified feature set, including HOMO, LUMO, one-hot encoded acceptor type, and top-selected Morgan and PubChem fingerprints.

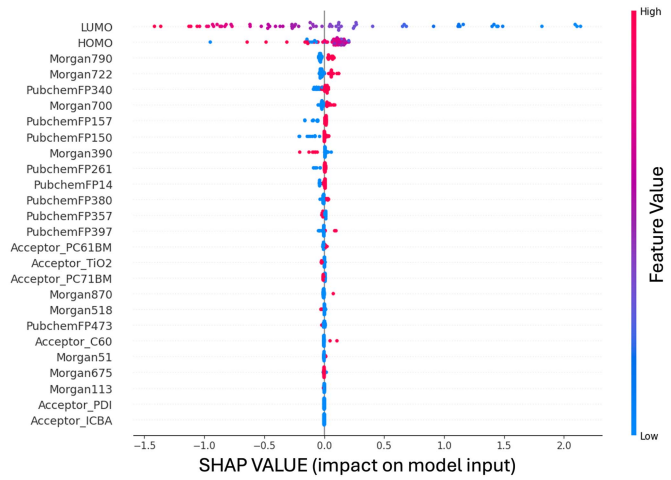

Fig. 3. SHAP summary plot illustrating the impact of features on XGBoost model predictions, including their direction and magnitude.

excellent generalization. Comparable RMSE and MAE values between the training (0.319/0.159) and test sets (0.302/0.148) further highlight the model's stability. We also compared our predicted PCE values with experimental measurements and found a higher Pearson correlation ($r = 0.21$) than the Scharber model ($r = 0.13$). Although the improvement is modest, it shows that our model better captures structure–performance relationships and is more in line with real-world results, despite challenges such as variation in device setups and processing conditions (see Supplementary Fig. S5). To the best of our knowledge, this represents the highest reported accuracy for interpretable, fingerprint-based PCE prediction on this dataset, surpassing prior models such as BRANNLP by Meftahi et al. [7] (R^2=0.78) despite lower computational cost. Our model also matches or exceeds the performance of models trained on experimental datasets, which mostly report $R^2 < 0.8$ due to data noise and scarcity [8], [10], [30].

To evaluate generalization, we tested the trained XGBoost model on donor molecules from the CEPDB, a large-scale computational OPV dataset that are chemically diverse yet share structural motifs with HOPV15 [36]. We selected structurally similar CEPDB donors with a Tanimoto coefficient [37] > 0.8 to HOPV15 donors, based on Morgan fingerprints. The model achieved a Pearson correlation of $r = 0.76$ between predicted and reported PCE values, indicating strong predictive transferability to structurally related, unseen molecules.

D. Feature Analysis

Explaining feature contributions is essential for designing efficient OPV materials. Traditional feature importance methods in RF and XGBoost rank features by their contribution to the model's performance but do not reveal the direction, context, or interactions of features, nor how specific feature values influence individual predictions [38]. In contrast, SHAP quantify each feature's contribution to a model's prediction by considering its marginal impact across all possible feature combinations. SHAP provides both global insights, by aggregating feature contributions across multiple predictions, and local insights into individual predictions. To better understand how features contribute to PCE prediction, SHAP analysis was applied, with the results presented in Fig. 3.

LUMO is the most significant feature, negatively correlated with PCE (see Fig. 3), despite not being explicitly included in the Scharber PCE calculation. Its dominance likely stems from the LUMO offset between donor and acceptor molecules, which directly correlates with open-circuit voltage (V_{OC}) in bulk heterojunction (BHJ) solar cells [39]. The HOPV15 dataset, which is primarily composed of BHJ architectures, shows limited variation in acceptor LUMO values. As a result, donor LUMO becomes the main driver of the LUMO offset, influencing V_{OC}, ultimately, PCE. Although HOMO contributes to hole transport and recombination processes [40], it shows a weaker yet nonnegligible impact on PCE in our model (see Fig. 3, Supplementary Fig. S2). While V_{OC} is defined by the offset between donor HOMO and acceptor LUMO, variations in HOMO alone do not always lead to proportional changes in V_{OC} due to energetic disorder, interfacial recombination, and voltage losses. Furthermore, the narrower distribution of HOMO values compared to LUMO in HOPV15 (Supplementary Fig. S6) limits its statistical influence in the model. In data-driven algorithms such as XGBoost, features with low variance offer fewer informative split points and contribute less to reducing prediction error. As a result, even if HOMO is mechanistically relevant, its impact on PCE appears diminished in the model. In contrast, donor LUMO—through its impact on the optical gap, short-circuit current density (J_{SC}), and exciton dissociation—emerges as more predictive of PCE.

Among the other features, the highest ranking ones are mostly fingerprints rather than acceptor descriptors (see Fig. S2, Supplementary). Since fingerprints are binary features, their

SHAP contributions depend on whether the specific structural substructure is present (red dots) or absent (blue dots), as seen in Fig. 3. Most fingerprints positively influence PCE when present, except for Morgan390, which has a noticeable negative impact on model performance. In contrast, acceptor-type descriptors have minimal influence, likely due to the dataset's limited variability in fullerene-based acceptors (e.g., PC61 BM) with similar chemical and electronic properties, making them less informative for the model.

E. Feature Interactions and Correlations

Given the central role of LUMO in determining PCE, identifying features that interact with or influence LUMO can provide valuable insights for molecular design. To explore these relationships, a SHAP interaction bar plot was generated to illustrate the feature interaction strength with LUMO [see Fig. 4(a)], alongside a Spearman correlation plot to reveal monotonic associations between features and PCE [see Fig. 4(b)].

Most features exhibit synergistic interactions with LUMO, improving PCE prediction. Among them, HOMO shows the strongest interaction with LUMO, explaining the improved model performance when both are used as predictors instead of LUMO alone [see Fig. 4(a)]. Features with strong interactions with LUMO, such as Morgan790, PubChemFP340, PubChem150, and Morgan722, also exhibit relatively high positive correlation with PCE, as shown in Fig. 4(b). Notably, Morgan722 shows a strong negative interaction with LUMO, indicating that its impact on PCE prediction may arise either through direct contributions or interactions with other key molecular properties (i.e., HOMO). Features with weaker interactions with LUMO, such as Morgan700 and PubChem157, generally exhibit weaker correlations with PCE, as shown in Fig. 4(b), suggesting stronger interactions with LUMO generally enhance OPV efficiency. However, PubChemFP473 deviates from this trend despite its strong interaction with LUMO, its presence is associated with a significant decrease in PCE. This observation underscores the intricate interplay between molecular features and their collective influence on OPV performance.

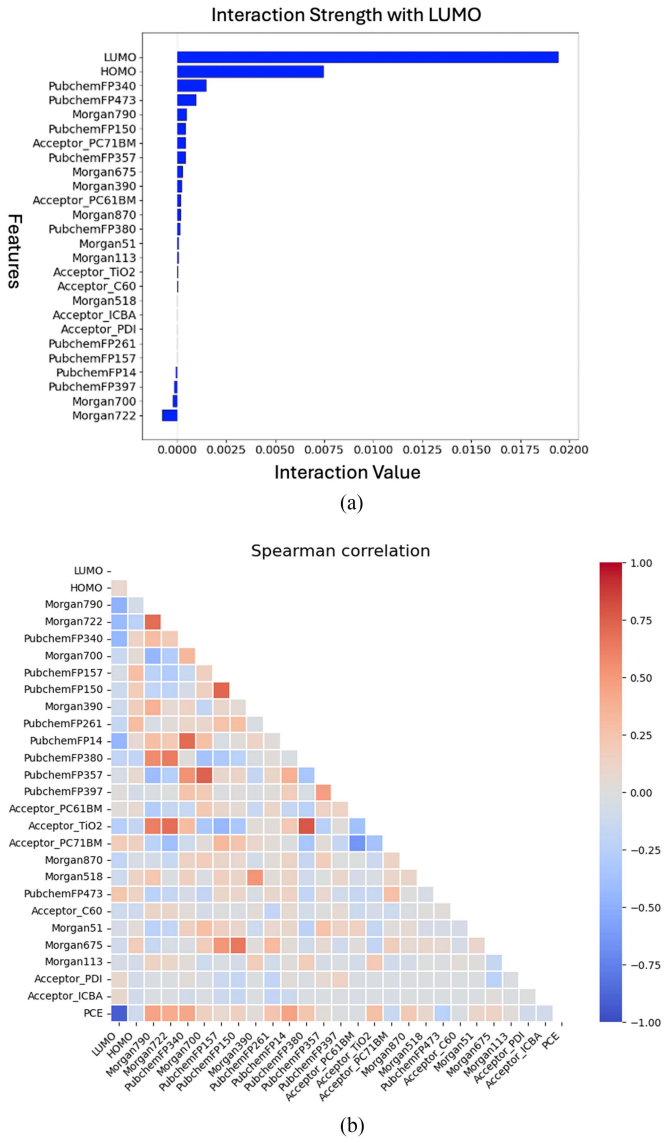

Fig. 4. Feature interaction strength and correlation analysis for PCE prediction. (a) Interaction strengths between LUMO and selected molecular descriptors evaluated by SHAP interaction values. (b) Spearman correlation heatmap illustrating the relationships between top molecular descriptors and PCE.

F. Substructure Visualization and Interpretation

The association between features and PCE is further validated by analyzing the percentage of top features present in donor molecules with either high or low PCE, as shown in Supplementary Fig. S3. Molecules in the top 20% of the PCE distribution are classified as the high PCE range, while those in the bottom 20% are classified as the low PCE range. The percentage of each fingerprint's presence within these two groups was calculated, and a significance threshold of $p = 0.05$ was applied to assess the statistical relevance of the differences.

Key features such as Morgan790, Morgan722, PubChemFP340, and PubChemFP14 show a significantly higher presence in high PCE molecules, along with notable percentage differences between the high PCE and low PCE groups, making them ideal structural candidates for donor molecular design. In addition, Morgan870 [see Fig. 5(c) and (d)], unique to the high PCE group, is also considered an ideal candidate despite its lower percentage.

As shown in Supplementary Fig. S3, features such as Morgan675 and PubChemFP261, which show high percentages in both the low PCE and high PCE groups, are less prioritized because they lack the discriminatory power to distinguish between high- and low-performing molecules. In contrast, PubChemFP473 (see Table V), being exclusively present in the low-PCE group, underscores the importance of excluding such features from molecular designs to prevent adverse impacts on photovoltaic efficiency. The structures of the remaining fingerprints are presented in Supplementary Table S1 (PubChem fingerprints) and Supplementary Fig. S4 (Morgan fingerprints).

Morgan fingerprints encode the local chemical environment within a defined radius by hashing similar features into the

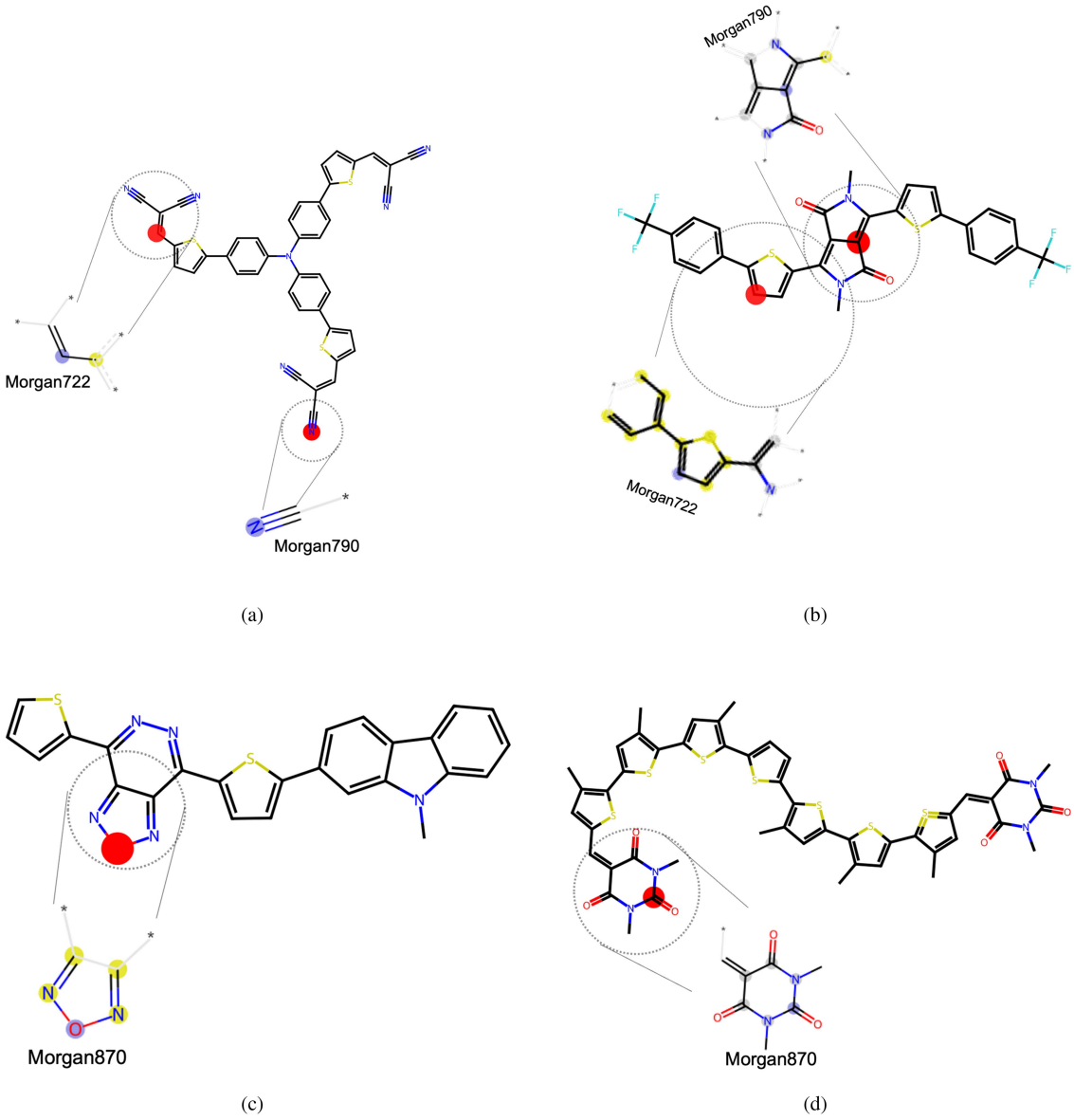

Fig. 5. Visualization of Morgan fingerprints 722, 790 (a) and (b), and 870 variants (c) and (d) in diverse molecular contexts. The circled atom represents the central atom within the substructure encoded by each Morgan fingerprint.

TABLE IV
SUMMARY OF PCE VALUES AND THE PRESENCE OF PUBCHEM FINGERPRINTS

Subfigure ref.	PCE value	PubChem FP340	PubChem FP14	PubChem FP473
a	4.17	Yes	Yes	No
b	3.86	No	Yes	No
c	4.90	Yes	Yes	No
d	3.54	No	Yes	No

TABLE V
DESCRIPTIONS AND PROPERTIES OF PUBCHEM FINGERPRINTS

PubChem fingerprint	Description
PubChemFP340	$C(\sim)(\sim C)(\sim N)$: Simple aliphatic chains with a nitrogen atom.
PubChemFP14	Presence of at least one nitrogen atom (N).
PubChemFP473	S-C:N:C: Sulfur bonded to a carbon in an aromatic system with nitrogen.

same bit. As a result, a single bit can represent substructures with minor differences. Morgan bits 790, 722, and 870 are visualized within their respective source donor molecules in Fig. 5. Information on their PCE values and the presence of PubChem fingerprint bits 340, 14, and 473 is provided in Table IV, while the structural details captured by PubChem fingerprints are summarized in Table V.

The most prevalent Morgan 790 structure features a central nitrogen (N) atom within a nitrile functional group (C $\equiv$ N), forming a strong electron-withdrawing group [see Fig. 5(a)]. This nitrile group is attached to an aromatic ring system, reducing the electron density of the aromatic system. Among donor molecules in the high PCE group, multiple C $\equiv$ N are frequently observed, as shown in Fig. 5(a). A rare variation of Morgan 790 is identified [see Fig. 5(b)], where the central atom

is carbon (C). In this case, the central carbon is embedded in a five-membered aromatic ring containing two nitrogen atoms (N) and two electron-withdrawing groups (C=O).

Morgan 722 has a central carbon (C) atom most commonly connected to a sulfur-containing thiophene ring or, less frequently, situated within it. This central carbon plays a crucial role as part of the conjugation pathway, facilitating efficient charge transport. In most cases, the central carbon is also directly bonded to another carbon atom, which is further connected to two nitrile groups (C ≡ N). This dual nitrile configuration further strengthens electron-withdrawing effects. The frequent coexistence of Morgan 722 and Morgan 790 in high PCE molecules highlights their synergistic impact on photovoltaic performance. Morgan 870 appears in two common formats [see Fig. 5(c) and (d)]. The first format has a central oxygen (O) atom directly bonded to two nitrogen (N) atoms within a heterocyclic ring [see Fig. 5(c)], forming a highly electron-deficient system. The second format features a central carbon (C) atom bonded to two nitrogen (N) atoms and an oxygen (O) atom, forming a carbonyl group (C=O) [see Fig. 5(d)].

The varied structures of Morgan 722, 790, and 870 all share a common structural property: they all contain highly electronegative oxygen and nitrogen atoms, which strongly withdraw electron density, resulting in lower LUMO and HOMO energy levels [41], [42], [43]. This electric modulation can potentially optimize both donor–acceptor LUMO alignment and overall HOMO-LUMO orbital alignment, both of which are crucial for improving PCE [44], [45]. An optimized LUMO energy alignment with the acceptor can facilitate efficient charge transfer and reducing recombination losses. As a result, these molecular features contribute to enhanced charge separation efficiency, ultimately improving PCE performance [39]. In addition, a narrower HOMO-LUMO band gap has been identified as a key factor in high-performance OPVs [46], as it also enhances charge transfer efficiency. PubChem fingerprints 340 and 14 (see Table V) highlight the role of nitrogen as a strong electron-withdrawing element, enhancing PCE performance.

In contrast, PubChemFP473, unique to the low PCE group, likely reduces performance due to the sulfur (S) disrupting conjugation within the aromatic system, hindering electron delocalization and charge transport, and ultimately lowering PCE. These structural features highlight the importance of integrating both conjugated systems and electron-withdrawing groups to achieve high-performance organic solar cells.

IV. Conclusion

In this study, we developed and evaluated ML models for predicting the PCE of OPV materials using the HOPV15 dataset. Among the algorithms tested, XGBoost and RF emerged as the top performers, achieving an RMSE of 0.302% and an R^2 of 0.918. This outperforms the previous best-reported benchmark (2020) [7]. This improvement can be attributed to our incorporation of molecular fingerprints, LUMO and HOMO into a streamlined feature set. This feature set was constructed using advanced feature selection techniques leveraged the complementary strengths of XGBoost and RF, thereby enhancing both predictive accuracy and interpretability while providing valuable insights into the structural features influencing OPV performance. SHAP value analysis further highlighted that LUMO consistently emerged as the most important molecular feature across various feature combinations. In addition, a subset of features from Morgan and PubChem fingerprints was found to interact with LUMO with a strong SHAP interaction strength, indicating that these features contribute significantly to PCE prediction, partially through their interactions with LUMO. These insights can guide the rational design of high-efficiency OPV materials.

Admittedly, a notable limitation of the HOPV15 dataset's computational data is the poor alignment of computed properties, such as HOMO, LUMO, and Scharber PCE, with their experimental counterparts [28], [47], [48]. Systematic computational errors in HOPV15 have been reported to correlate with molecular structure, making it possible to apply targeted corrections by mapping these errors in molecular space. Gaussian process regression (GPR), combined with Bayesian hyperparameter optimization, serves as an effective empirical correction method for computed values, addressing systematic errors by mapping molecular similarities through the Tanimoto kernel and providing calibrated predictions with quantified uncertainty [48]. In the same study, Pyzer-Knapp et al. [48] visualized these systematic errors using force-directed graphs based on molecular similarity, identifying clusters of molecules with similar deviations from experimental values that could then be corrected through GPR. Clusters of molecules exhibiting similar deviations from experimental values were identified, and these deviations were then corrected using GPR. Their calibration of DFT-calculated PCE predictions against the experimental HOPV15 dataset resulted in a significant improvement, increasing the correlation coefficient r to 0.65. The results of which demonstrate that GPR effectively corrects the systematic errors introduced by quantum chemical calculations and the Scharber model, aligning computational predictions more closely with experimental values. This significantly enhances the practical applicability of the HOPV15 dataset in real-world scenarios. Therefore, improving calibration methods to better align computational PCE values with experimental data could significantly enhance the model's generalizability to real-world applications. Incorporating more relevant features into the kernel function of GPR, particularly those identified in this study, may further refine predictive accuracy and improve adaptation to experimental conditions.

Furthermore, nonfullerene acceptors (NFAs) have become the leading class of electron acceptors in high-performance OPV research, offering greater stability and efficiency [49]. Given this shift toward more diverse acceptors, it is important to note that in this study, we relied on donor-side descriptors due to the limited variation in the fullerene-based acceptors used in the HOPV15 dataset, which made donor features sufficient to explain most of the variance in PCE. Nonetheless, we acknowledge this as a limitation for broader applicability. Future work will focus on expanding the dataset to include a more chemically diverse set of NFAs and on developing joint donor–acceptor embeddings using deep learning models, such as transformers. These learned representations, integrated with molecular fingerprints, can enhance

the expressiveness of our XGBoost framework while maintaining interpretability—enabling more comprehensive modeling of donor–acceptor interactions in diverse OPV systems.

By addressing these challenges, the ML method and the feature set we formed can be leveraged to bridge the gap between theoretical predictions and experimental outcomes, advancing the development of next-generation OPV materials.

Acknowledgment

The authors would like to thank Dr. Mihaela Stefan for introducing the research problem and for her valuable discussions related to PCE estimation, which helped initiate this study. This article has benefited from the use of OpenAI's ChatGPT for language editing and grammar enhancement. No part of the scientific content or conclusions was generated by the AI system.

References

[1] E. K. Solak and E. Irmak, "Advances in organic photovoltaic cells: A comprehensive review of materials, technologies, and performance," *RSC Adv.*, vol. 13, no. 18, pp. 12244–12269, 2023.

[2] D. López-Durán, E. Plésiat, M. Krompiec, and E. Artacho, "Gap variability upon packing in organic photovoltaics," *PLoS One*, vol. 15, no. 6, 2020, Art. no. e0234115.

[3] H. Abdulaziz, A. Gidado, A. Musa, and A. Lawal, "Electronic structure and non-linear optical properties of neutral and ionic pyrene and its derivatives based on density functional theory," *J. Mater. Sci. Res. Rev.*, vol. 2, no. 3, pp. 1–13, 2019.

[4] Z. Li, J. Yang, and P. A. N. Dezfuli, "Study on the influence of light intensity on the performance of solar cell," *Int. J. Photoenergy*, vol. 2021, no. 1, 2021, Art. no. 6648739.

[5] Y. Cui et al., "Accurate photovoltaic measurement of organic cells for indoor applications," *Joule*, vol. 5, no. 5, pp. 1016–1023, 2021.

[6] M. C. Scharber et al., "Design rules for donors in bulk-heterojunction solar cells—towards 10% energy-conversion efficiency," *Adv. Mater.*, vol. 18, no. 6, pp. 789–794, 2006, doi: 10.1002/adma.200501717.

[7] N. Meftahi et al., "Machine learning property prediction for organic photovoltaic devices," *npj Comput. Mater.*, vol. 6, no. 1, 2020, Art. no. 166.

[8] Q. Zhao, Y. Shan, H. Zhou, G. Zhang, and W. Liu, "Machine learning-assisted performance prediction and molecular design of all-small-molecule organic solar cells based on the Y6 acceptor," *Sol. Energy*, vol. 265, 2023, Art. no. 112115.

[9] A. Eibeck et al., "Predicting power conversion efficiency of organic photovoltaics: Models and data analysis," *ACS Omega*, vol. 6, no. 37, pp. 23764–23775, 2021.

[10] M. Seifrid et al., "Beyond molecular structure: Critically assessing machine learning for designing organic photovoltaic materials and devices," *J. Mater. Chem. A*, vol. 12, no. 24, pp. 14540–14558, 2024.

[11] S. A. Lopez et al., "The harvard organic photovoltaic dataset," *Sci. Data*, vol. 3, no. 1, pp. 1–7, 2016.

[12] S. A. Lopez et al., "The Harvard organic photovoltaic dataset," *Sci. Data*, vol. 3, no. 1, pp. 1–7, 2016.

[13] D. Weininger, "Smiles, a chemical language and information system. 1. introduction to methodology and encoding rules," *J. Chem. Inf. Comput. Sci.*, vol. 28, no. 1, pp. 31–36, 1988.

[14] G. Landrum, "RDKit: Open-source cheminformatics," 2006. [Online]. Available: http://www.rdkit.org

[15] A. D. Becke, "Density-functional exchange-energy approximation with correct asymptotic behavior," *Phys. Rev. A*, vol. 38, no. 6, 1988, Art. no. 3098.

[16] J. P. Perdew, M. Ernzerhof, and K. Burke, "Rationale for mixing exact exchange with density functional approximations," *J. Chem. Phys.*, vol. 105, no. 22, pp. 9982–9985, 1996.

[17] A. Becke, "Density-functional thermochemistry. III. the role of exact exchange (1993) j," *Chem. Phys.*, vol. 98, 1993, Art. no. 5648.

[18] Y. Zhao and D. G. Truhlar, "The M06 suite of density functionals for main group thermochemistry, thermochemical kinetics, noncovalent interactions, excited states, and transition elements: Two new functionals and systematic testing of four M06-class functionals and 12 other functionals," *Theor. Chem. Accounts*, vol. 120, pp. 215–241, 2008.

[19] L. David, A. Thakkar, R. Mercado, and O. Engkvist, "Molecular representations in AI-driven drug discovery: A review and practical guide," *J. Cheminformatics*, vol. 12, no. 1, 2020, Art. no. 56.

[20] P. Carracedo-Reboredo et al., "A review on machine learning approaches and trends in drug discovery," *Comput. Struct. Biotechnol. J.*, vol. 19, pp. 4538–4558, 2021.

[21] D. Rogers and M. Hahn, "Extended-connectivity fingerprints," *J. Chem. Inf. Model.*, vol. 50, no. 5, pp. 742–754, 2010.

[22] J. Dong et al., "Chemdes: An integrated web-based platform for molecular descriptor and fingerprint computation," *J. Cheminformatics*, vol. 7, pp. 1–10, 2015.

[23] V. Bolón-Canedo, N. Sánchez-Maroño, and A. Alonso-Betanzos, "Feature selection for high-dimensional data," *Prog. Artif. Intell.*, vol. 5, pp. 65–75, 2016.

[24] E. Debie and K. Shafi, "Implications of the curse of dimensionality for supervised learning classifier systems: Theoretical and empirical analyses," *Pattern Anal. Appl.*, vol. 22, pp. 519–536, 2019.

[25] G. C. Cawley and N. L. Talbot, "On over-fitting in model selection and subsequent selection bias in performance evaluation," *J. Mach. Learn. Res.*, vol. 11, pp. 2079–2107, 2010.

[26] C. Rücker, G. Rücker, and M. Meringer, "y-randomization and its variants in QSPR/QSAR," *J. Chem. Inf. Model.*, vol. 47, no. 6, pp. 2345–2357, 2007.

[27] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Proc. Adv. Neural Inf. Process. Syst.*, Curran Associates, Inc., 2017, pp. 4765–4774. [Online]. Available: http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf

[28] D. Padula, J. D. Simpson, and A. Troisi, "Combining electronic and structural features in machine learning models to predict organic solar cells properties," *Mater. Horiz.*, vol. 6, no. 2, pp. 343–349, 2019.

[29] M.-H. Lee, "Insights from machine learning techniques for predicting the efficiency of fullerene derivatives-based ternary organic solar cells at ternary blend design," *Adv. Energy Mater.*, vol. 9, no. 26, 2019, Art. no. 1900891.

[30] Z. Zhu et al., "Machine-learning-assisted exploration of new non-fullerene acceptors for high-efficiency organic solar cells," *Cell Rep. Phys. Sci.*, vol. 5, no. 12, 2024.

[31] Z.-W. Zhao, M. del Cueto, Y. Geng, and A. Troisi, "Effect of increasing the descriptor set on machine learning prediction of small molecule-based organic solar cells," *Chem. Mater.*, vol. 32, no. 18, pp. 7777–7787, 2020.

[32] L. Shi, C. K. Lee, and A. P. Willard, "The enhancement of interfacial exciton dissociation by energetic disorder is a nonequilibrium effect," *ACS Central Sci.*, vol. 3, no. 12, pp. 1262–1270, 2017.

[33] S. M. Menke, N. A. Ran, G. C. Bazan, and R. H. Friend, "Understanding energy loss in organic solar cells: Toward a new efficiency regime," *Joule*, vol. 2, no. 1, pp. 25–35, 2018.

[34] S. M. Menke and R. J. Holmes, "Exciton diffusion in organic photovoltaic cells," *Energy Environ. Sci.*, vol. 7, no. 2, pp. 499–512, 2014.

[35] R. Suthar, T. Abhijith, and S. Karak, "Machine-learning-guided prediction of photovoltaic performance of non-fullerene organic solar cells using novel molecular and structural descriptors," *J. Mater. Chem. A*, vol. 11, no. 41, pp. 22248–22258, 2023.

[36] J. Hachmann et al., "The Harvard Clean Energy Project: Large-scale computational screening and design of organic photovoltaics on the world community grid," *J. Phys. Chem. Lett.*, vol. 2, no. 17, pp. 2241–2251, 2011.

[37] D. Bajusz, A. Rácz, and K. Héberger, "Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations?," *J. Cheminformatics*, vol. 7, pp. 1–13, 2015.

[38] V. Hassija et al., "Interpreting black-box models: A review on explainable artificial intelligence," *Cogn. Comput.*, vol. 16, no. 1, pp. 45–74, 2024.

[39] G. T. Mola and N. Abera, "Correlation between LUMO offset of donor/acceptor molecules to an open circuit voltage in bulk heterojunction solar cell," *Physica B: Condens. Matter*, vol. 445, pp. 56–59, 2014.

[40] C. He et al., "Manipulating the D: A interfacial energetics and intermolecular packing for 19.2% efficiency organic photovoltaics," *Energy Environ. Sci.*, vol. 15, no. 6, pp. 2537–2544, 2022.

[41] Y. N. Luponosov et al., "Effects of electron-withdrawing group and electron-donating core combinations on physical properties and photovoltaic performance in D-π-A star-shaped small molecules," *Org. Electron.*, vol. 32, pp. 157–168, 2016.

[42] H. Watanabe, Y. Ito, K. Tanaka, and Y. Chujo, "Enhancement of the lowering effect on energy levels of LUMO by the formation of B-N dative bond for near-infrared light absorption properties based on 1, 3, 4, 6, 8, 9 b-hexaazaphenalene," *Asian J. Org. Chem.*, vol. 12, no. 6, 2023, Art. no. e202300156.

[43] Y.-Z. Dai et al., "Embedding electron-deficient nitrogen atoms in polymer backbone towards high performance n-type polymer field-effect transistors," *Chem. Sci.*, vol. 7, no. 9, pp. 5753–5757, 2016.
[44] D. Wang, X. Zhang, W. Ding, X. Zhao, and Z. Geng, "Density functional theory design and characterization of D–A–A type electron donors with narrow band gap for small-molecule organic solar cells," *Comput. Theor. Chem.*, vol. 1029, pp. 68–78, 2014.
[45] B.-G. Kim et al., "Energy level modulation of HOMO, LUMO, and band-gap in conjugated polymers for organic photovoltaic applications," *Adv. Funct. Mater.*, vol. 23, no. 4, pp. 439–445, 2013.
[46] P. Cheng and Y. Yang, "Narrowing the band gap: The key to high-performance organic photovoltaics," *Accounts Chem. Res.*, vol. 53, no. 6, pp. 1218–1228, 2020.
[47] S. A. Lopez, B. Sanchez-Lengeling, J. de Goes Soares, and A. Aspuru-Guzik, "Design principles and top non-fullerene acceptor candidates for organic photovoltaics," *Joule*, vol. 1, no. 4, pp. 857–870, 2017.
[48] E. O. Pyzer-Knapp, G. N. Simm, and A. A. Guzik, "A Bayesian approach to calibrating high-throughput virtual screening results and application to organic photovoltaic materials," *Mater. Horiz.*, vol. 3, no. 3, pp. 226–233, 2016.
[49] C. Yan et al., "Non-fullerene acceptors for organic solar cells," *Nature Rev. Mater.*, vol. 3, no. 3, pp. 1–19, 2018.