

## NRC Publications Archive Archives des publications du CNRC

### **Computational modeling of indoor organic photovoltaics: dataset curation, predictive analysis, and machine learning approaches**

Yu, Hung-Nien; Chen, Hsu-Yuan; Sharma, Ganesh D.; Cheng, Yen-Ju; Hsu, Chain-Shu; Chu, Ta-Ya; Lu, Jianping; Chen, Fang-Chung

This publication could be one of several versions: author's original, accepted manuscript or the publisher's version. / La version de cette publication peut être l'une des suivantes : la version prépublication de l'auteur, la version acceptée du manuscrit ou la version de l'éditeur.

For the publisher's version, please access the DOI link below. / Pour consulter la version de l'éditeur, utilisez le lien DOI ci-dessous.

#### **Publisher's version / Version de l'éditeur:**

<https://doi.org/10.1007/s11831-025-10310-y>

*Archives of Computational Methods in Engineering, 2025-07-11*

#### **NRC Publications Archive Record / Notice des Archives des publications du CNRC :**

<https://nrc-publications.canada.ca/eng/view/object/?id=d6b88486-0072-451e-bd89-d828993fe121>

<https://publications-cnrc.canada.ca/fra/voir/objet/?id=d6b88486-0072-451e-bd89-d828993fe121>

Access and use of this website and the material on it are subject to the Terms and Conditions set forth at

<https://nrc-publications.canada.ca/eng/copyright>

READ THESE TERMS AND CONDITIONS CAREFULLY BEFORE USING THIS WEBSITE.

L'accès à ce site Web et l'utilisation de son contenu sont assujettis aux conditions présentées dans le site

<https://publications-cnrc.canada.ca/fra/droits>

LISEZ CES CONDITIONS ATTENTIVEMENT AVANT D'UTILISER CE SITE WEB.

**Questions?** Contact the NRC Publications Archive team at

PublicationsArchive-ArchivesPublications@nrc-cnrc.gc.ca. If you wish to email the authors directly, please see the first page of the publication for their contact information.

**Vous avez des questions?** Nous pouvons vous aider. Pour communiquer directement avec un auteur, consultez la première page de la revue dans laquelle son article a été publié afin de trouver ses coordonnées. Si vous n'arrivez pas à les repérer, communiquez avec nous à PublicationsArchive-ArchivesPublications@nrc-cnrc.gc.ca.



# Computational Modeling of Indoor Organic Photovoltaics: Dataset Curation, Predictive Analysis, and Machine Learning Approaches

Hung-Nien Yu<sup>1</sup> · Hsu-Yuan Chen<sup>2</sup> · Ganesh D. Sharma<sup>3,4</sup> · Yen-Ju Cheng<sup>5,6</sup> · Chain-Shu Hsu<sup>5,6</sup> · Ta-Ya Chu<sup>7</sup> · Jianping Lu<sup>7</sup> · Fang-Chung Chen<sup>1,6</sup>

Received: 10 April 2025 / Accepted: 20 June 2025  
© The Author(s) 2025

## Abstract

This study presents a comprehensive dataset that encompasses the indoor device performance of organic photovoltaic (OPV) materials, their corresponding SMILES codes, and frontier molecular orbital (FMO) energy levels. This dataset comprises a total of 128 subsets and features 64 pairs of donors and acceptors. We demonstrate that traditional models, such as the Shockley–Queisser limit and Scharber’s model, are insufficient for accurately predicting the behavior of indoor OPVs based on the molecular orbitals of these materials. In contrast, we explore the predictive capabilities of four machine learning (ML) models for estimating the power conversion efficiencies (PCEs) of indoor OPVs, utilizing molecular structure information and FMO data from the dataset we compiled. The trained ML models exhibit strong predictive performance with high correlation coefficients ( $r > 0.8$ ) for indoor PCE values; notably, the support vector regression (SVR) model achieves the highest  $r$  of 0.878. The generalization capabilities of the models are also assessed using previously unseen materials, and the results demonstrate high accuracy rates. The SVR algorithm reaches the best average accuracy of 92.1%, underscoring its potential for efficiently screening materials for indoor applications. Our findings suggest that this dataset, with opportunities for future expansion, could significantly facilitate material design and accelerate computer-aided materials screening, reducing the need for extensive experimental testing in the development of indoor OPVs.

**Keywords** Data · Organics · Indoor · Photovoltaics · Machine learning

## 1 Introduction

Organic photovoltaics (OPVs) have attracted considerable attention owing to their advantages, such as being lightweight, mechanically flexible, semi-transparent, and having low fabrication costs [1–3]. Notably, the power conversion efficiencies (PCEs) of OPVs have seen substantial improvements, with the latest certified PCE reaching an impressive 19.2% under standard 1-sun conditions (AM1.5G, 100 mW cm<sup>-2</sup>), making them a practical option for solar energy harvesting [3]. While sunlight is not always available in every location, ambient indoor lighting serves as a readily accessible energy source in daily life. OPVs outperform silicon-based photovoltaics in converting indoor light into electricity due to their spectral tunability and higher optical absorptivity [4–9]. Furthermore, OPVs generally exhibit much higher shunt resistances, leading to reduced leakage currents. These features render OPVs highly promising for use in versatile distributed power sources under low-level lighting conditions. For example, OPVs are ideal for

✉ Fang-Chung Chen  
fchendop@nycu.edu.tw

<sup>1</sup> Department of Photonics, College of Electrical and Computer Engineering, National Yang Ming Chiao Tung University, Hsinchu 30010, Taiwan

<sup>2</sup> National Chutung Senior High School, Hsinchu 31046, Taiwan

<sup>3</sup> Department of Physics, The LNM Institute of Information Technology, Jamdoli, Jaipur 302031, Rajasthan, India

<sup>4</sup> Department of Electronics and Communication Engineering, The LNM Institute of Information Technology, Jamdoli, Jaipur 302031, Rajasthan, India

<sup>5</sup> Department of Applied Chemistry, National Yang Ming Chiao Tung University, Hsinchu 30010, Taiwan

<sup>6</sup> Center for Emergent Functional Matter Science, National Yang Ming Chiao Tung University, Hsinchu 30010, Taiwan

<sup>7</sup> Advanced Electronics and Photonics Research Centre, National Research Council Canada, Ottawa, ON K1A 0R6, Canada

powering off-grid portable electronics, wearable devices, and internet of things (IoT) items. Particularly, IoT systems comprise wireless interactions among numerous electronics and sensors, each having unique identifiers and real-time data. Considering the substantial energy demands for sustaining these IoT systems, OPVs offer a promising solution to enhance the sustainability of IoT products [4–9].

Significant efforts have been devoted to developing efficient OPVs in the last decade, including materials design and synthesis, morphology optimization, and device architecture engineering [1, 2]. However, identifying suitable organic compounds within the vast range of possibilities remains highly challenging, requiring extensive and costly experimental work to fully understand the performance of OPV materials. Although density functional theory (DFT) computations can provide insights into the electronic properties of organic semiconductors, computer-guided material design is still necessary to directly forecast device performance and PCEs from the structural features of the organic molecules [10–14].

Recently, data-driven approaches, including machine learning (ML), have been employed to develop quantitative structure-property relationship (QSPR) models that link material structures to their properties [10–16]. The use of computer-aided techniques has sparked considerable interest in the OPV community. In 2006, Scharber et al. created a model to estimate PCEs based on bandgaps and energy levels of conjugated polymers [17]. The Harvard Clean Energy Project (CEP) collected calculations and experimental data for thousands of organic molecules, predicting their PCEs using Scharber's model [18]. Furthermore, ML models have been proposed to identify high-performance non-fullerene acceptors and optimal donor/acceptor pairs [10, 19, 20]. This data-driven approach has proven to be highly effective in exploring the QSPR of materials and accelerating the discovery of new OPV materials.

Indoor photovoltaics function distinctively from conventional solar cells in various aspects. Although solar irradiation spans a broad spectrum of wavelengths, artificial light sources—such as white light-emitting diodes (LEDs) and fluorescent tubes (FTs)—primarily emit light within the visible range. This difference necessitates larger optimal bandgaps for the semiconductors when used under indoor lighting conditions [21–23]. Furthermore, the intensity of indoor light sources is significantly lower than that of sunlight, which leads to distinct design principles for materials utilized in indoor OPVs [21–23]. While many studies using data-driven and ML methods have explored material properties and device performance for OPVs, research focusing specifically on indoor OPVs remains relatively limited.

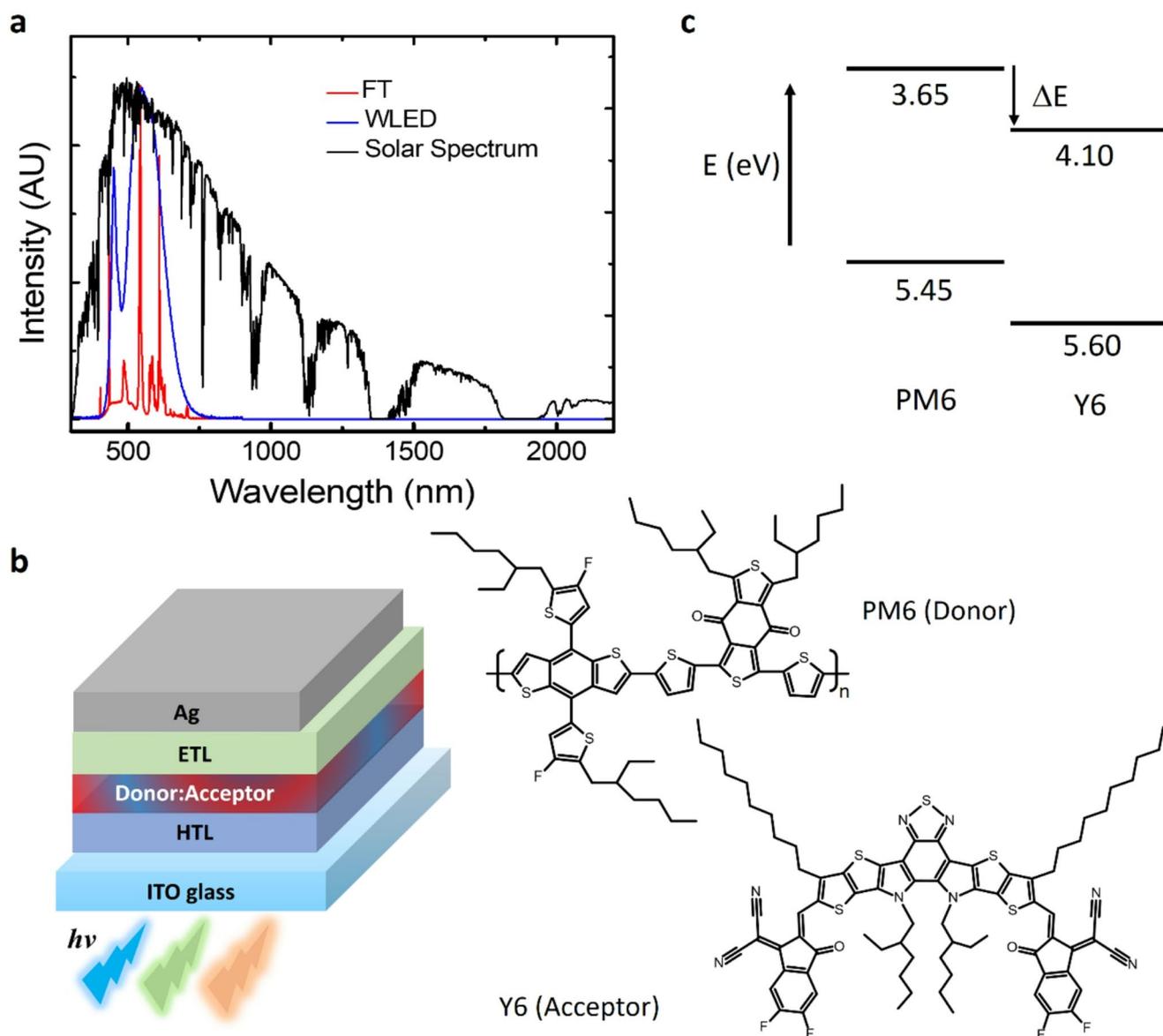
In this work, we established a dataset outlining the properties of materials and the device performance of indoor

OPVs. This dataset includes 128 subsets of data, derived from various combinations of 31 donor (D) materials and 32 acceptor (A) materials. Key photovoltaic performance indicators, including open-circuit voltage ( $V_{oc}$ ), short-circuit photocurrent ( $J_{sc}$ ), fill factor ( $FF$ ), and PCE were encompassed in the dataset, along with detailed information on the D and A materials utilized in these indoor OPVs. Furthermore, the dataset also consisted of the molecular orbitals of the materials, including their highest occupied molecular orbital (HOMO), lowest unoccupied molecular orbital (LUMO), and bandgap values. The chemical structures of the materials were represented using the Simplified Molecular Input Line Entry System (SMILES) notation. Our analysis indicated that traditional theories, such as the Shockley–Queisser limit and Scharber's model, fell short of accurately describing device behaviors based on the frontier molecular orbital (FMO) energy levels of these materials. Conversely, ML models developed from this dataset demonstrated high accuracy in predicting PCE values. To the best of our knowledge, this dataset is the first reported for indoor OPVs. We anticipate that this dataset can serve as a valuable resource for material design tools and that the outcomes of this work may ultimately improve computer-aided materials screening for indoor OPVs.

## 2 Results

### 2.1 Dataset Description and Analysis

Figure 1(a) illustrates a typical device structure for OPVs, with a blend of organic semiconductors positioned between two electrodes. To overcome the high binding energy of excitons in organic semiconductors, most high-efficiency OPVs are designed based on the concept of bulk-heterojunction (BHJ), which enables effective exciton separation at the interfaces between organic donors and acceptors [24, 25]. Although some ternary blends of indoor OPVs have demonstrated high efficiencies, studies in this area remain limited [26, 27]. Therefore, this work focuses on OPVs prepared using organic blends containing only one pair of donor and acceptor molecules. The indoor OPV dataset was manually collected from 45 literature sources (see Supplementary Information). It contains 128 subsets, with each subset providing fundamental properties of the donor/acceptor (D/A) pairs, such as their energy levels, bandgaps, and the corresponding photovoltaic parameters. Notably, the acceptors include both fullerene and non-fullerene molecules. Specifically, there are 56 subsets for fullerene-based acceptors, while non-fullerene acceptors (NFAs) account for 72 subsets. This higher number of NFAs reflects the increasing trend in research that emphasizes the use of NFAs as active



**Fig. 1** Indoor photovoltaics. **(a)** The solar irradiation spectrum and the emission spectra of the FT and white LED light sources. **(b)** A typical device structure of indoor OPVs and chemical structures of the

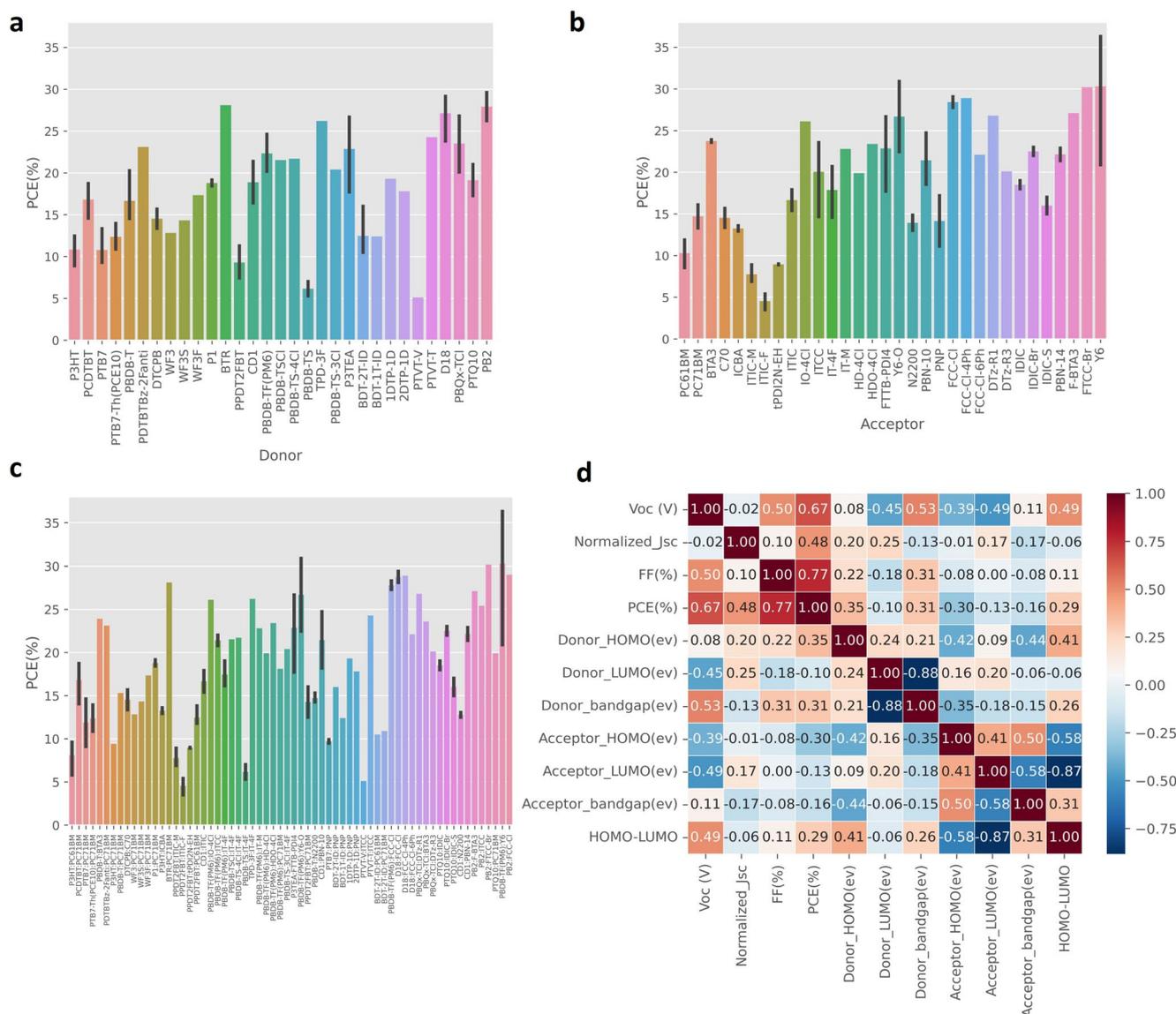
materials in OPVs. The relatively balanced numbers help to eliminate significant biases that may arise from the dataset itself.

Unlike the standard AM1.5G illumination condition for conventional solar cells, there is still no standardized illumination condition available for indoor photovoltaic devices. Therefore, the dataset also specifies the types of indoor light sources used, either white LEDs or FTs, along with their respective illuminance values [21–23]. While other artificial light sources like incandescent bulbs exist, white LEDs and FTs are the most commonly used indoor lighting options. Research results using other types of indoor light sources are still quite limited. Figure 2 shows the overview of the

representative donor (PM6) and acceptor (Y6) materials. **(c)** The corresponding energy levels of the two materials in **(b)**

dataset; it includes 31 donors and 32 acceptors. The combinations of these donors and acceptors result in 64 different D/A pairs. The maximum and minimum PCE values are 36.3 and 3.5%, respectively. The D/A pair of the best device consists of PM and Y6 as the donor and acceptor molecules, respectively; the chemical structures of this representative molecular pair are also illustrated in Fig. 1(a).

The feature selection process aims to identify the physical and chemical properties within the dataset that exhibit strong correlations with the target property for ML models. In general, the stronger the correlation between the features and the target, the more straightforward the learning task becomes. Pearson's correlation analysis was employed to



**Fig. 2** The dataset statistics. The PCE values of the (a) donors; (b) acceptors in the dataset. The black bars represent the range of PCE distributions for the single donor or acceptor used in different combinations. (c) The PCE values of the D/A combinations. The black bars

indicate the range of PCE distributions for the combinations reported from different literature. (d) Pearson's correlation coefficient for the descriptors of the donor and acceptor molecules

assess the importance of the selected features in relation to the target, and the heatmap results are illustrated in Fig. 2(d). According to Fig. 2(d), it is evident that the PCEs of indoor OPVs are strongly correlated with  $FF$  values. Because  $FF$  is typically linked to device quality, including thin-film defects, morphologies of the BHJ active layers, and device interfaces, it is not surprising that OPVs with high  $FF$  values tend to exhibit high PCEs [28]. Moreover, Pearson's correlation coefficient ( $r$ ) for short-circuit photocurrents was 0.48, considerably lower than that of open-circuit voltages ( $V_{oc}$ ), which was 0.67. It is important to note that the  $J_{sc}$  values were normalized using their corresponding luminance in this analysis due to the absence of a standard for

indoor lighting conditions. Additionally, two subsets from the literature were excluded from the analysis because they did not provide luminance values, reporting power density instead. Therefore, the higher Pearson's correlation coefficient of  $V_{oc}$  may be attributed to the narrow light spectral range of indoor light sources, which only covers the visible wavelengths from 400 to 800 nm. Consequently, photovoltage likely plays a more critical role in determining the performance of OPVs under artificial lighting conditions [8, 29]. This finding aligns with the conclusion of our previous study, which highlighted that increasing the  $V_{oc}$  value is the key factor in achieving high PCEs for indoor OPVs [29]. Different from sunlight irradiation, which has a wide

spectral regime from the ultraviolet to the infrared range, most indoor light sources only emit in the visible spectral range. Therefore, indoor OPVs can only absorb a very limited spectral range of photons, which contributes to the photocurrents.

Given the importance of  $V_{oc}$  values for the performance of OPVs, we further examined the Pearson correlation coefficients for  $V_{oc}$  [Fig. 2(d)]. For conventional inorganic photovoltaics,  $V_{oc}$  values are mostly affected by the bandgap of the semiconductors. For OPVs, interestingly,  $V_{oc}$  was found to be strongly correlated with the bandgap of donors ( $r=0.53$ ), the lowest unoccupied molecular orbital (LUMO) of the acceptors ( $r=-0.49$ ), and the LUMO of the donors ( $r=-0.45$ ). The results reflected the nature of the OPVs, which are based on the BHJ concept. The photons absorbed by the organic semiconductors of the OPVs become highly bound excitons. As illustrated in Fig. 1c, the strong binding energy of the resulting excitons can only be overcome at the interfaces between donors and acceptors. Subsequently, the separated electrons and holes are transported through acceptors and donors, respectively, and eventually collected by the opposite electrodes. Therefore, the photovoltaic properties are reasonably affected by the molecular orbitals involved in the processes.

Many literatures indicate that most observed  $V_{oc}$  values are closely related to the energy level offset between the highest occupied molecular orbital (HOMO) of the donors and the LUMO of the acceptors, as described by Eq. (1) [15, 17, 30–32]:

$$qV_{oc}=(|HOMO(D)|-|LUMO(A)|-\Delta) \quad (1)$$

where  $q$  is the elementary charge,  $\Delta$  is an empirical value ranging from 0.3 to 0.5 eV and  $|HOMO(D)|-|LUMO(A)|$  represents the effective energy gap of the donor/acceptor pair. For instance, in Scharber's model for fullerene-based OPVs, the empirical energy loss term ( $\Delta$ ) was reported to be 0.3 eV [17]. Therefore, we also calculated the energy difference ( $|HOMO(D)|-|LUMO(A)|$ ) for each subset, finding a strong correlation with  $V_{oc}$  values ( $r=0.49$ ) [Fig. 2(d)]. The results from the dataset confirmed that the energy levels of the D/A pairs influence the  $V_{oc}$  values significantly.

In contrast to conventional solar cells, the bandgap of the acceptors appeared to be less critical for indoor OPVs as the Pearson correlation coefficient was only 0.11. Typical acceptors, particularly non-fullerene molecules, are specifically designed to capture long-wavelength photons from solar irradiation in OPVs. However, due to the much narrower spectral range of indoor light sources, the role of these acceptors in absorbing photons with wavelengths longer than 780 nm was minimal. On the other hand, most

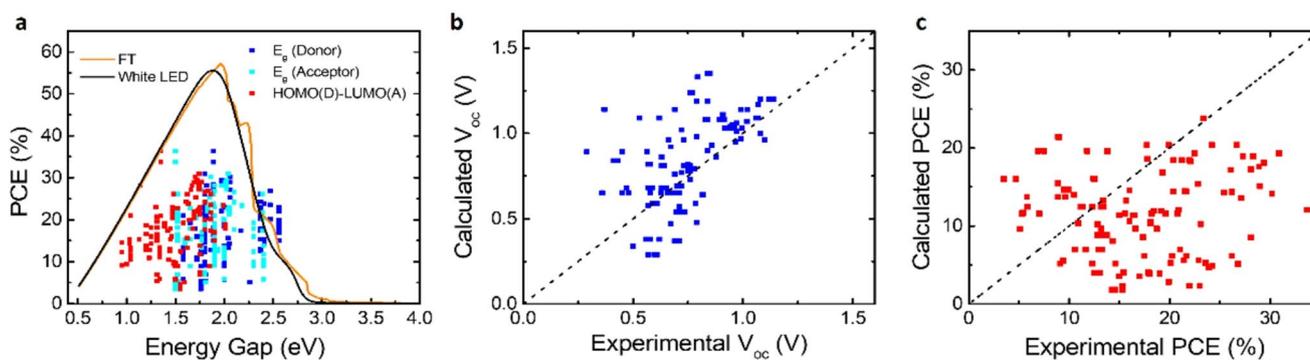
donors have wider bandgaps and are usually designed as the primary absorbing material, covering nearly the entire visible spectrum for indoor OPVs. In other words, the donors probably absorb more visible photons than acceptors, resulting in a higher Pearson correlation coefficient. Consequently, as indicated by the high Pearson correlation coefficient ( $r=0.53$ ) previously, the  $V_{oc}$  values are strongly correlated with the bandgaps of the donors. The results from the Pearson correlation analysis in this dataset will be used later to guide the selection of input features for the machine learning models during the learning process.

## 2.2 Traditional Models for Indoor OPVs: Shockley–Queisser Limit

In theory, ideal solar cells follow the principle of “detailed balance”, which states that the maximum extractable energy is determined by the difference between absorbed and emitted radiation [22, 33]. Shockley–Queisser limit (SQ limit) establishes the highest theoretical efficiency for a solar cell utilizing a single  $p-n$  junction under the assumption that radiative recombination is the only loss mechanism. Our previous study estimated the upper PCE limits for two types of artificial light sources [22]; the results are presented in Fig. 3a. However, in the case of OPVs, the presence of two distinct semiconductor materials complicates the computation and prediction of the PCEs. Figure 3a also reveals the PCEs from the dataset as a function of the bandgap ( $E_g$ ) values from various assumptions. As illustrated in Fig. 3a, some experimental outcomes surpassed the predicted SQ limits for certain  $E_g$  values of donor or acceptor materials. On the other hand, as shown in Eq. (1), when the  $E_g$  is defined as the effective energy gap of the donor/acceptor pair ( $E_g=|HOMO(D)|-|LUMO(A)|$ ), the results appear more reasonable. However, predicting PCEs still remains highly challenging for indoor OPVs.

## 2.3 Traditional Models for Indoor OPVs: Modified Scharber's Model

Scharber's model has been widely used for evaluating the performance of OPVs and has been applied in numerous virtual screening studies [17, 34, 35]. This model estimates the maximum efficiency based solely on the  $E_g$  values and frontier molecular orbitals of the polymer donors. As revealed in Eq. (1),  $V_{oc}$  is assumed to be the effective energy gap of the donor/acceptor pair minus an empirical energy loss (0.3 eV). Additionally, the EQE and  $FF$  values are both set at 65% [17]. To adapt Scharber's model for understanding the limit of PCEs of indoor OPVs, the empirical energy loss term ( $\Delta$ ) should be further adjusted. Because the light intensity of indoor light sources is significantly lower than



**Fig. 3** Results from traditional models. **(a)** SQ limits at various bandgaps; the solid curves are calculated from an FT or a white LED at 1000 lx [22]. **(b)** The correlation between the calculated  $V_{oc}$  values from the modified Scharber's model and the experimental  $V_{oc}$  data.

**(c)** The correlation between the calculated PCE values from modified Scharber's model and the experimental results data. The diagonal line indicates the perfect positive correlation ( $r=1$ )

that of solar irradiation, additional energy losses should be considered. Previously, Koster et al. demonstrated that the  $V_{oc}$  value of an ideal device with only bimolecular recombination and negligible leakage current can be expressed by Eq. (2) [36]:

$$V_{oc} = \frac{E_g}{q} - \frac{kT}{q} \ln \left( \frac{(1-P)\gamma N_c^2}{PG} \right) \quad (2)$$

where  $q$  is the elementary charge,  $k$  is Boltzmann's constant,  $T$  is temperature,  $P$  is the dissociation probability of a bound electron-hole pair,  $g$  is the bimolecular recombination rate coefficient,  $N_c$  is the effective density of states, and  $G$  is the photogeneration rate. Because  $G$  depends on light intensity,  $V_{oc}$  should exhibit a logarithmic dependence on light intensity with a slope of  $kT/q$ . Later, Proctor and Nguyen examined this dependence in nonideal devices and found that the logarithmic relationship between  $V_{oc}$  and light intensity is particularly sensitive to leakage currents, often resulting in slopes greater than  $kT/q$  [37]. Based on this, we calculated  $V_{oc}$  values by assuming a slope of  $2kT/q$  and considering an indoor light intensity that is 1000 times lower than one-sun illumination [37]. Figure 3b presents the results derived from incorporating the adjusted energy loss term. The Pearson correlation coefficient ( $r$ ) between the calculated  $V_{oc}$  values and the experimental results was 0.535. From Fig. 3b, the calculated  $V_{oc}$  values were overall overestimated, suggesting the presence of additional voltage loss mechanisms in the OPVs.

Figure 3c presents the predicted PCE values derived from the modified Scharber's model using the emission spectrum of the indoor light source FT. Notably, the majority of PCE values appear to be underestimated. The average  $FF$  in the dataset was 0.645, which indicates that the assumed  $FF$  value of 0.65 was a reasonable approximation. Because the

$V_{oc}$  values were underestimated (Fig. 3b), we suspected the photocurrent prediction could be inaccurate. More critically, the results exhibit considerable variability, as evidenced by a very low  $r$  value of 0.095. The low correlation suggests that Scharber's model is inadequate for accurately predicting the indoor performance of OPVs based solely on material properties. Scharber's model was developed initially using observations from fullerene-based OPVs, in which fullerene derivatives have limited absorption capabilities. Consequently, its accuracy diminishes when applied to OPVs utilizing NFA materials. While Scharber's model offers low computational cost, there is still a need for models that provide higher accuracy.

## 2.4 Machine Learning Models for Indoor OPVs

Machine learning presents alternative approaches for uncovering hidden QSPR relationships [16]. As material informatics advances, a new generation of paradigms for material research and development is emerging [16]. A material database is utilized to train ML models, which can subsequently be employed to discover new materials and/or predict their properties. Consequently, this section aims to evaluate the capability of ML models to predict the PCEs of indoor OPVs using the dataset we compiled for this study [38]. Herein, we intentionally select four ML models based on different principles: support vector regression (SVR) [31, 39], random forests (RF) [31, 40], k-nearest neighbors (KNN) [40], and artificial neural networks (ANN) [39, 41, 42]. Among them, three representative supervised learning algorithms include SVR, RF, and KNN, and ANN belongs to a neural network model. Particularly, SVR and RF have been reported for predicting the performance of OPVs operated under one-sun illumination with high accuracy.

Initially, we trained the models using only the chemical structures as the input features. Subsequently, we

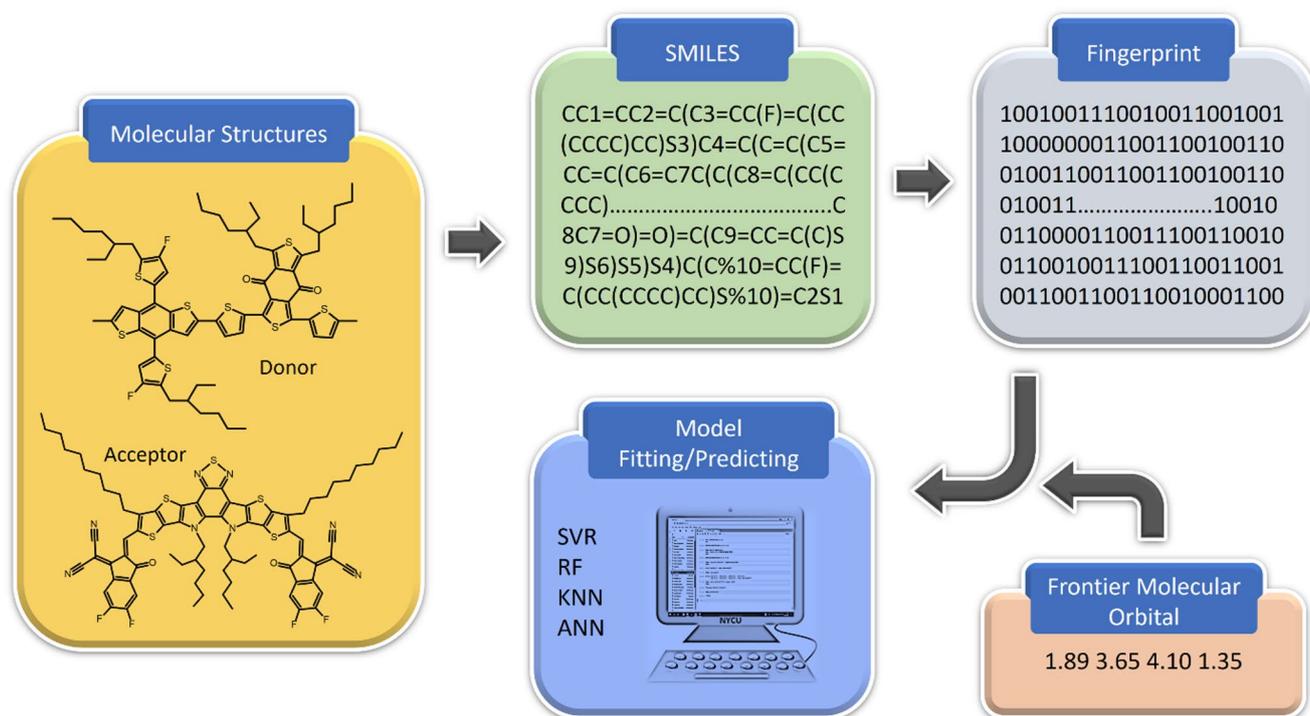
incorporated information regarding the FMOs into the inputs. The workflow for building and evaluating ML models is illustrated in Fig. 4. The modeling work involved utilizing two distinct sets of features to train the models. In the initial trial, we employed the fingerprints of both the donor and acceptor materials. For the second trial, these fingerprints were combined with data regarding the FMO information. Further, four models were trained to predict PCEs based on molecular structure descriptions obtained from molecular fingerprint analysis. The chemical structures of the donor/acceptor pairs were first converted into SMILES codes. For conjugated polymers, the repeating units were used as the inputs instead. Subsequently, using RDKit with a Python API [43], these codes were transformed into Morgan fingerprints. Cross-validation functions from the scikit-learn Python package were used to train and validate the models [44]. The training subset, along with the corresponding PCE values, was then fitted into the models. The test subsets were used to evaluate the models. The performance of the trained models was evaluated using several metrics, including the Pearson correlation coefficient ( $r$ ), mean squared error (MSE), root mean squared error (RMSE), coefficient of determination ( $R^2$ ), and mean absolute percentage error (MAPE). The hyperparameter settings for each model are listed in Table S1.

Figure 5(a-d) presents the correlation between the experimental and predicted PCE values derived from the four ML

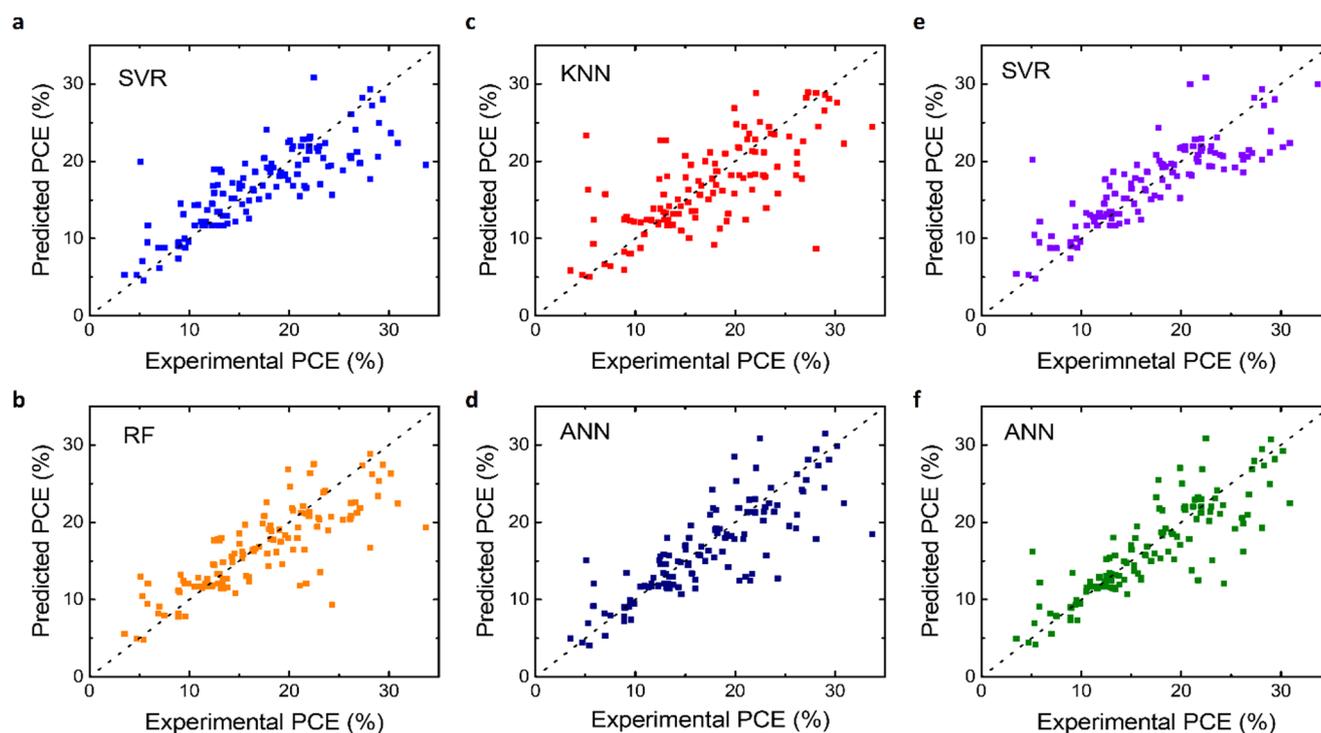
models. The results indicate that each model achieved a commendable fit for the indoor OPV dataset. Table 1 provides a summary of the evaluation metrics associated with these models. Notably, SVR exhibited the highest performance, attaining a high  $r$  value of 0.861 and an  $R^2$  value of 0.669. The regression-based ANN, which employed only three hidden layers, also demonstrated a high correlation coefficient of 0.847 and an  $R^2$  of 0.602. The KNN model, which exhibited a slightly lower  $r$  of 0.743, still reflected satisfactory performance.

Collectively, these results suggest that employing the fingerprints of the D/A pairs as input features to train the ML models could achieve much better performance than conventional physics-based models. While predicting the PCE values with the modified Scharber's model, a very low  $r$  of 0.095 was obtained (Fig. 3c). Because conventional theories rely only on FMO data and lack knowledge of chemical structures, insufficient information limits the performance in predicting the device behavior. Particularly, the material properties often affect the device efficiencies significantly in OPVs.

As revealed in the previous Pearson correlation analysis (Fig. 2d), the properties of frontier molecular orbitals showed a significant correlation with PCE values. Consequently, we selected four additional features for inclusion in the ML models: the bandgap of the donors ( $r=0.53$ ), the LUMO of the acceptors ( $r = -0.49$ ), the LUMO of the



**Fig. 4** The workflow for the ML modeling. The input features are either the fingerprints of the D/A materials or the combinations of the fingerprints with the energy levels of the frontier molecular orbitals



**Fig. 5** The performance of the ML models. The correlation between the true (experimental) and predicted PCE values in (a)SVR; (b)RF; (c)KNN; (d)ANN models. The correlation between the experimental

and predicted PCE values in (e)SVR; (f)ANN models after the FMO information of the D/A materials is considered as the input features. The diagonal lines indicate the perfect positive correlation ( $r=1$ )

**Table 1** Performance of the models trained with fingerprint only

Model	SVR	RF	KNN	ANN
Feature(s)	Fingerprint Only			
r	0.861±0.073	0.834±0.101	0.743±0.160	0.847±0.072
MSE	16.20±12.83	17.40±13.98	22.65±14.98	17.88±13.63
RMSE	3.74±1.49	3.93±1.46	4.51±1.51	3.98±1.42
R <sup>2</sup>	0.669±0.153	0.619±0.196	0.505±0.208	0.602±0.184
MAPE	0.190±0.089	0.180±0.068	0.236±0.108	0.068±0.064

**Table 2** Performance of the models trained with fingerprint and molecular orbitals

Model	SVR	RF	KNN	ANN
Feature(s)	Fingerprint+Molecular Orbitals <sup>a)</sup>			
r	0.878±0.063	0.832±0.111	0.790±0.112	0.796±0.184
MSE	14.75±12.08	17.84±15.91	22.23±14.93	21.59±23.11
RMSE	3.56±1.43	3.91±1.60	4.49±1.50	4.14±2.10
R <sup>2</sup>	0.690±0.164	0.605±0.226	0.514±0.149	0.546±0.353
MAPE	0.169±0.080	0.183±0.078	0.208±0.102	0.208±0.059

<sup>a)</sup>The StandardScaler utility class from scikit-learn was used to perform the scaling

donors ( $r = -0.45$ ), and the effective energy gap of the D/A pairs ( $r=0.49$ ). These features, derived from the energy levels of the D/A pairs, differ from the previous Morgan fingerprints, which consist of only binary numbers. In the absence of standardization, certain ML algorithms may exhibit varying performance characteristics. Consequently, we implemented feature scaling through the process of standardization to ensure consistency. Scikit-learn offers

a variety of methods for data preprocessing. As illustrated in Fig. 5(e, f) and detailed in Table 2, the best modeling results were obtained using standardized features, achieved by removing the mean and scaling to unit variance. We also noted that some prediction results are significantly different after adding FMO data to train the models. Unfortunately, we could not understand the reason currently. It is probably necessary to use some explainable AI techniques to uncover

why such differences arise. Among the four algorithms analyzed, KNN demonstrated significant improvement, with  $r$  and  $R^2$  values rising to 0.790 and 0.514, respectively. In contrast, there were no notable changes for the SVR and RF algorithms. However, the SVR model still exhibited the best performance, with  $r$  and  $R^2$  values achieving 0.878 and 0.690, respectively (Fig. 5e). The metrics of the ANN model, however, declined, showing  $r$  and  $R^2$  values of 0.796 and 0.546, respectively. Note that the network structure remained unchanged (three hidden layers) to allow for a fair comparison of performance after standardization. The ANN model was not well optimized, as the purpose of this study is not to pursue the best performance. Further tuning the hyperparameters should deliver better performance. Moreover, we believe that newer models, such as deep neural networks (DNNs), could yield better results. In principle, these models are suitable for large datasets. With the increasing amount of labeled data, DNNs usually can generalize better than ML models, including SVR and RF, which may not scale as effectively. Consequently, increasing the size of the dataset, or including other material features, such as molecular weights and polymer dispersity indexes, thin-film morphology, and manufacturing conditions, are possible approaches to improve the models.

Table S2 provides a detailed overview of the performance metrics associated with the other scaling techniques. When the features of the energy levels were rescaled to a range from 0 to 1, the performance of most models remained comparable (Table S2). Notably, the experimental PCE results and the predicted PCE values generated by the ANN methods exhibited a comparatively higher correlation, with  $r$  of 0.857 (Fig. 5f). These overall results indicated that some distinct scaling techniques are more appropriate for specific ML algorithms. Additionally, an examination of various performance metrics across different preprocessing approaches revealed that the rescaling processes had a negligible effect on the RF model. This decision-tree-based algorithm demonstrated considerable robustness against arbitrary scaling of the data<sup>44</sup>.

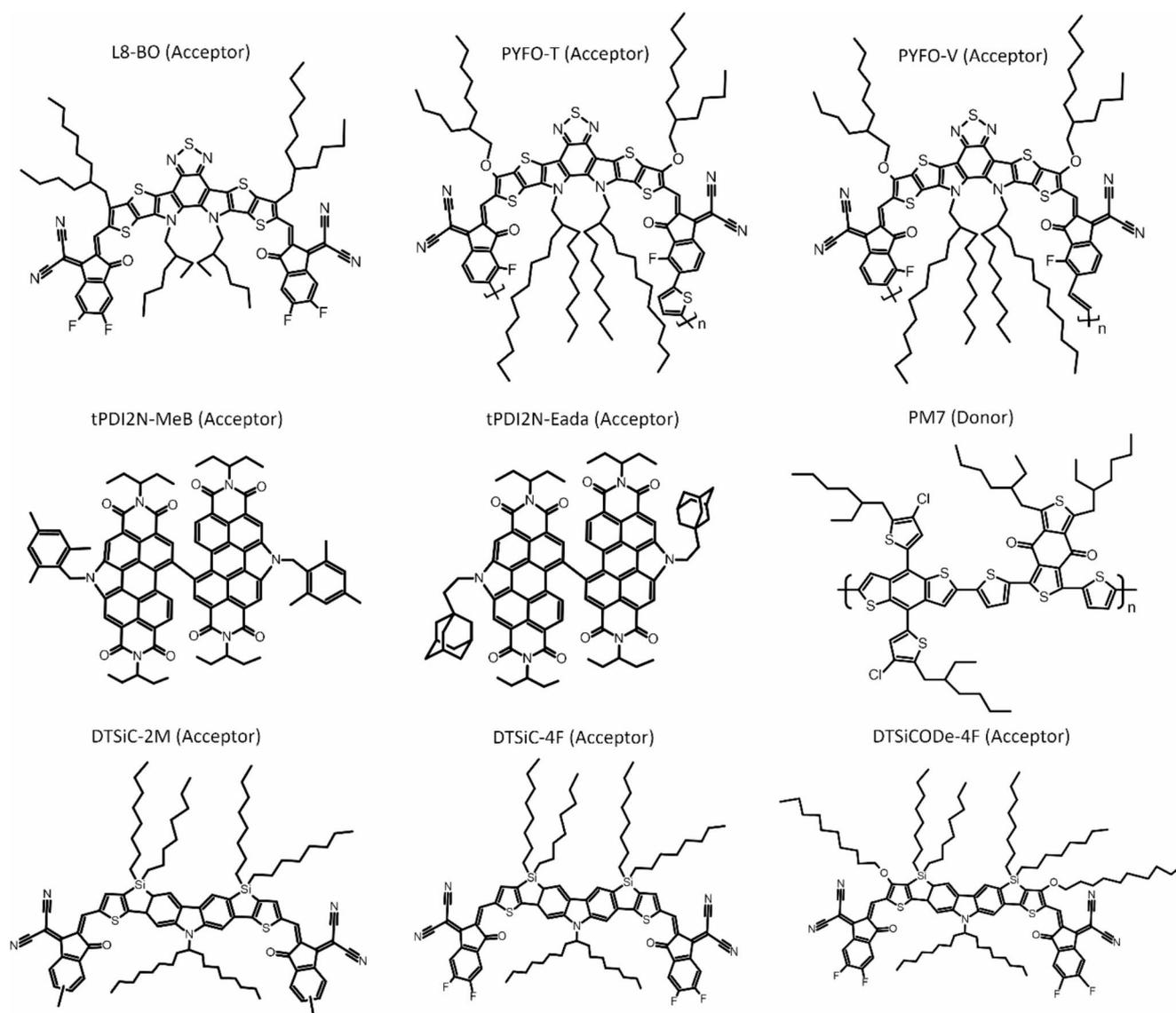
In comparison to the performance metrics of the models trained exclusively on the information of chemical structures, as shown in Table 2, the enhancement in the predictive capability of the ML methods that utilized additional features relative to energy levels was rather limited. Only the accuracy of the KNN model was improved slightly. Contrary to what one might intuitively expect, most of the ML prediction accuracies did not show significant improvement. This limitation may stem from the fact that structural fingerprints have already encapsulated much of the relevant information, including various chemical and physical parameters. In other words, the material properties are highly correlated to the chemical structures, which are coded in the molecular fingerprints. Consequently, these additional descriptors, such as the FMO data, may only provide

minimal information or contributions to the overall predictive capability [45].

The results of the ML models could also provide valuable physical insights and guide the rational design of new molecules. For example, we identified six donor-acceptor (D/A) pairs that exhibited the highest PCE values using the best SVR model. The chemical structures of these combinations are shown in Scheme S1. From this scheme, it is evident that derivatives from the PM6:Y6 family demonstrated the best performance, particularly for the modified medium-bandgap NFAs. Additionally, the P3TEA: FTTB-PDI4 pair showed promise, likely due to the larger bandgaps of both the donor and acceptor molecules [5]. Further detailed analysis of the ML models is ongoing to fully understand the design rules behind these findings.

## 2.5 Predictions for Unseen Materials

One of the challenges in ML modeling is developing models that can provide accurate predictions when faced with new and unseen data. The generalization of ML algorithms is a crucial aspect of developing practical tools, as it allows them to produce reliable predictions in real-world situations. Therefore, eight subsets of data sourced from recent literature were utilized to assess the generalization capabilities of the ML models developed in this study [27, 46–49]. Crucially, this data had not been included in the training process, meaning it represented entirely new information to these models. Figure 6 depicts the chemical structures of the “unseen” materials used for the generalization test. Among the testing materials, two donors, PM6 and PTQ10, were seen but paired with other “unseen” acceptors in the dataset during model testing, making them “unseen” D/A pairs to the models. As illustrated in Fig. 6, the first acceptor was L8-BO, which features a Y6-based dithienothiophen[3,2-b]-pyrrolobenzothiadiazole core with branched alkyl chains [48]. Additionally, PM6 was paired with two polymer acceptors, PYFO-T and PYFO-V [47]. The molecular design of these polymers was also inspired by Y-series NFAs, incorporating branched alkoxy side chains at the conjugated cores. Thiophene and vinylene were included as the polymer linkers at the end of PYFO-T and PYFO-V, respectively. In the second group of the “unseen” pairs, the core structure of the donor, PTQ10, is different from the PM6 family. However, it was paired with two new acceptors, tPDI2N-tMeB and tPDI2N-Eada, whose molecular designs are based on the N-annulated perylene diimide dimer.<sup>49</sup> Only one similar acceptor, tPDI2N-EH, appeared once in the dataset, but was paired with a very different donor, PPDT2FBT. Finally, in the third group of the D/A combinations, both the donors and acceptors were new to the training dataset. The donor in these pairs, PM7, is characterized by the substitution



**Fig. 6** The chemical structures of the unseen organic molecules for the ML models. The materials include eight new acceptors and one donor, delivering eight different D/A combinations

of the two F atoms in the repeating unit of PM6 with Cl atoms (Fig. 6) [27]. The three acceptors in this group were silicon-bridged carbazole-based NFAs, which featured F atoms and methyl groups at the end-groups for DTSiC-4 F and DTSiC-2 M, respectively. Furthermore, alkoxy groups were introduced in the  $\beta$ -positions of the thiophene units in DTSiC-4 F to create DTSiCODE-4 F [27]. Notably, no silicon-bridged compounds were included in the training data for the previous four ML models.

Table 3 summarizes the prediction results of various D/A combinations from each model trained using chemical structures. The accuracy values were notably high, with average accuracies of 92.1%, 86.4%, 84.1%, and 87.6% for the SVR, RF, KNN, and ANN models, respectively. A high average accuracy of 95.5% was reached using the

SVR model for the third unseen D/A pairs, in which both donors and acceptors were new to the models. Individually, the highest accuracy achieved was 99.5% using the SVR algorithm when predicting the PM6:PYFO-T combination. On the other hand, the KNN model showed less consistency in their predictions. While some predictions were close to the experimental PCEs, others deviated significantly from the actual values. Among the eight pairs analyzed, the accuracies for the PM6:PYFO-V combination were consistently lower across all four models. While the chemical structure of PYFO-V is comparable to that of PYFO-T, the prediction results varied significantly. This discrepancy was likely attributed to its tighter  $\pi$ - $\pi$  stacking and greater crystallinity, which enhanced its PCEs<sup>47</sup>. Consequently, predicting outcomes for the PM6:PYFO-V combination has proven more

**Table 3** Results of predictions from the ML models for the unseen D/A combinations

D/A Combination (Experimental PCE) [%]		SVR	RF	KNN	ANN
PM6:L8-BO (24.17)	Prediction	22.55	22.59	30.3	26.68
	Accuracy [%]	93.3	93.5	79.8	90.6
PM6:PYFO-T (19.9)	Prediction	20.01	20.80	22.26	22.73
	Accuracy [%]	99.5	95.7	89.4	87.5
PM6:PYFO-V (25.7)	Prediction	20.12	20.77	22.26	21.53
	Accuracy [%]	78.3	80.8	86.6	83.8
PTQ10: tPDI2N-tMeB (15.3)	Prediction	18.40	19.52	12.60	14.01
	Accuracy [%]	83.2	78.4	82.4	91.6
PTQ10: tPDI2N- Eada (16.7)	Prediction	18.43	19.52	12.60	14.68
	Accuracy [%]	96.1	85.6	75.4	87.9
PM7:DTSiC-4 F (21.17)	Prediction	19.31	23.05	15.95	20.31
	Accuracy [%]	91.2	91.8	75.3	95.9
PM7:DTSiC-2 M (19.53)	Prediction	19.34	22.79	18.19	23.23
	Accuracy [%]	99.1	85.7	93.1	84.1
PM7:DTSiCDe-4 F (20.14)	Prediction	19.38	25.17	22.26	25.31
	Accuracy [%]	96.2	80.0	90.5	79.6

complex, as chemical structure codes, such as SMILES strings, may not sufficiently represent the pertinent physical properties.

The generalization of the ML models trained using fingerprints and molecular orbitals was also investigated, and the prediction results for the D/A pairs are summarized in Table S3. The average accuracy values were 80.6%, 85.5%, 84.7%, and 87.4% for the SVR, RF, KNN, and ANN models, respectively. The accuracies became lower than those previously reported in Table 3 for the SVR, RF, and ANN models. Only the average accuracy of the KNN increased slightly from 84.1 to 84.7%. Surprisingly, the accuracy was significantly lower than that previously reported in Table 3 for the SVR model. Additionally, some accuracies were even lower than 70%, indicating the robustness of the models had declined. As previously discussed, the enhancements in the performance metrics of the ML methods utilizing additional data from the FMOs were marginal (see Table 2). The results indicate that incorporating FMO data may have even compromised the generalizability of the models, especially for the SVR model. While the underlying reason for this observation remains unclear, we suspect that the energy levels were too closely spaced, causing critical information to be obscured.

While FMO data may sometimes be considered redundant or even detrimental to model generalizations, we cannot overlook the potential for other physical descriptors to influence our models positively [45]. Numerous experimental parameters can significantly affect device efficiencies. For instance, various factors such as the molecular weights of polymers, the ratio of donor to acceptor materials, thin-film morphologies, and device structures are likely to have

considerable impacts on device performance. Furthermore, process parameters, including annealing temperatures, spin-coating rates, and the incorporation of processing additives, play a crucial role in determining the PCE values. Therefore, we plan to include at least some of these features, such as the annealing temperatures and durations, and interfacial layers on both sides of the electrodes, to further improve the models in the near future. Moreover, newer models, including DNNs, could also be adopted once the dataset is extended. DNNs are possibly rather suited for leveraging a more diverse set of input features compared to traditional ML models like RF and SVR. Beyond the chemical structures, integrating information about these parameters could enhance the accuracy of model predictions from various perspectives.

### 3 Conclusion

In conclusion, we have compiled a dataset that outlines the indoor device performance of OPVs, the SMILES codes of the materials, and the FMO energy levels. To the best of our knowledge, the dataset in this work is the first one reported so far for indoor OPVs. The data analysis shows that photovoltage likely plays a more critical role in determining the performance of indoor OPVs. Conventional theories, such as the Shockley–Queisser limit and Scharber's model, prove insufficient in predicting the device behaviors based on the FMO information of these materials. Conversely, we evaluated the capability of the four ML models to predict the PCE values of indoor OPVs, utilizing molecular structure information and FMOs derived from the dataset. Overall, the trained ML models demonstrate a strong correlation coefficient ( $r > 0.8$ ) when predicting the indoor PCE values for OPVs. The necessity of using FMO information was further examined by comparing the performance metrics of the models trained exclusively on molecular fingerprints with those incorporating FMO data. This comparison clearly demonstrates that the frontier molecular orbitals can be regarded as redundant in the development of the ML models. Additionally, the models can accurately predict the performance of indoor OPVs made with previously unseen materials, showcasing their strong capability to screen potential candidates for indoor applications. We foresee the potential of this dataset for further expansion, making it an asset for material design decision-making tools. Ultimately, we believe the findings of this study could significantly advance computer-aided materials screening while minimizing the need for extensive experimental measurements for indoor OPVs.

## 4 Method

The dataset was collected manually from the literature. During the data analysis, various values for photovoltaic parameters, including  $V_{oc}$ ,  $J_{sc}$ ,  $FF$ , and PCE values, were obtained for certain pairs of materials. These values were averaged for Pearson's correlation analysis. It is important to note that  $J_{sc}$  was normalized with respect to the corresponding luminance for this analysis, as these values were particularly sensitive to lighting levels. All the ML models considered in this study were implemented in python using the scikit-learn library. Chemical structures were converted to SMILES strings using ChemSketch software [50]. and subsequently transformed into molecular fingerprints using the RDKit library. The Morgan algorithm was employed during the conversion, with the default set to 2048 bits per hash. The optimal hyperparameters for each algorithm were determined by minimizing the mean squared errors. Cross-validation from the scikit-learn python package was utilized to train and validate the models for ML applications. The training subset, along with the corresponding PCE values, was fitted into the models. The test subsets were then utilized to evaluate the model performance. Performance metrics for each model were calculated by averaging the results from 10-fold cross-validation. The function "cross\_val\_predict" from the scikit-learn library was employed for model visualization ( $cv=10$ ). The training and validation loss curves of the ANN model are shown in Fig. S1 and S2. It is important to note that the new and "unseen" subset of data was not utilized in the training of the models.

**Supplementary Information** The online version contains supplementary material available at <https://doi.org/10.1007/s11831-025-10310-y>.

**Acknowledgements** The authors would like to thank the support from National Science and Technology Council, Taiwan (grant no. NSTC 112-2221-E-A49-072-MY3 and 114-2218-E-A49-016) and the Featured Areas Research Center Program within the framework of the Higher Education Sprout Project by the Ministry of Education (MOE) in Taiwan. This work was also financially supported by the AI for Design Challenge Program at National Research Council of Canada.

**Author Contributions** HNY and FCC conceptualized and initiated this project. HNY and HNC collected and built the datasets. HNY constructed the ML models. HNY and FCC optimized the ML models. HNY wrote the draft of the manuscript. GDS, YJC, CSH, TYC, and TYC contributed to the fruitful discussions of the project and the editing manuscript. All authors read and approved the final manuscript.

**Funding** Open Access funding enabled and organized by National Yang Ming Chiao Tung University

**Data Availability** Data is provided within the manuscript or supplementary information files.

## Declarations

**Competing Interests** The authors declare no competing interests.

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

## References

1. Yi J, Zhang G, Yu H, Yan H (2024) Advantages, challenges and molecular design of different material types used in organic solar cells. *Nat Rev Mater* 9:46–62
2. Brabec CJ, Distler A, Du X, Egelhaaf HJ, Hauch J, Heumueller T, Li N (2020) Material strategies to accelerate OPV technology toward a GW technology. *Adv Energy Mater* 10:2001864
3. Best research-cell efficiencies <https://www.nrel.gov/pv/assets/pdf/best-research-cell-efficiencies.pdf>. Accessed 8 March 2025
4. Kim TH, Park NW, Saeed MA, Jeong SY, Woo HY, Park J, Shim JW (2023) Record indoor performance of organic photovoltaics with long-term stability enabled by self-assembled monolayer-based interface management. *Nano Energy* 112:108429
5. Ma LK, Chen Y, Chow PCY, Zhang G, Huang J, Ma C, Zhang J, Yin H, Cheung AMH, Wong KS, So SK, Yan H (2020) High-efficiency indoor organic photovoltaics with a band-aligned interlayer. *Joule* 4:1486–1500
6. Huang CL, Kumar G, Sharma GD, Chen FC (2020) Plasmonic effects of copper nanoparticles in polymer photovoltaic devices for outdoor and indoor applications. *Appl Phys Lett* 116:253302
7. Kumar G, Chen FC (2023) A review on recent progress in organic photovoltaic devices for indoor applications. *J Phys D: Appl Phys* 56:353001
8. Chen FC (2019) Emerging organic and organic/inorganic hybrid photovoltaic devices for specialty applications: low-level-lighting energy conversion and biomedical treatment. *Adv Opt Mater* 7:1800662
9. Lee C, Lee JH, Lee HH, Nam M, Ko DH (2022) Over 30% efficient indoor organic photovoltaics enabled by morphological modification using two compatible non-fullerene acceptors. *Adv Energy Mater* 12:2200275
10. 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
11. Mahmood A, Irfanb A, Wang JL (2022) Machine learning and molecular dynamics simulation-assisted evolutionary design and discovery pipeline to screen efficient small molecule acceptors for PTB7-Th-based organic solar cells with over 15% efficiency. *J Mater Chem A* 10:4170–4180
12. 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
13. Zhang S, Li S, Song S, Zhao Y, Gao L, Chen H, Li H, Lin J (2025) Deep learning-assisted design of novel donor–acceptor

- combinations for organic photovoltaic materials with enhanced efficiency. *Adv Mater* 37:2407613
14. 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
  15. Greenstein BL, Hutchison GR (2022) Organic photovoltaic efficiency predictor: data-driven models for non-fullerene acceptor organic solar cells. *J Phys Chem Lett* 13:4235–4243
  16. Butler KT, Davies DW, Cartwright H, Isayev O, Walsh A (2018) Machine learning for molecular and materials science. *Nature* 559:547–555
  17. Scharber MC, Mühlbacher D, Koppe M, Denk P, Waldauf C, Heeger AJ, Brabec CJ (2006) Design rules for donors in bulk-heterojunction solar cells—towards 10% energy-conversion efficiency. *Adv Mater* 18:789–794
  18. 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
  19. 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
  20. Zhang Q, Zheng YJ, Sun W, Ou Z, Odunmbaku O, Li M, Chen S, Zhou Y, Li J, Qin B, Sun K (2022) High-efficiency non-fullerene acceptors developed by machine learning and quantum chemistry. *Adv Sci* 9:2104742
  21. Freunek M, Freunek M, Reindl LM (2013) Maximum efficiencies of indoor photovoltaic devices. *IEEE J Photovoltaics* 3:59–64
  22. Wu MJ, Kuo CC, Jhuang LS, Chen PH, Lai YF, Chen FC (2019) Bandgap engineering enhances the performance of mixed-cation perovskite materials for indoor photovoltaic applications. *Adv Energy Mater* 9:1901863
  23. Burwell G, Zeiske S, Caprioglio P, Sandberg OJ, Kay AM, Farrar MD, Kim YR, Snaith HJ, Meredith P, Armin A (2024) Wide-gap perovskites for indoor photovoltaics. *Solar RRL* 8:2400180
  24. Wadsworth A, Hamid Z, Kosco J, Gasparini N, McCulloch I (2020) The bulk heterojunction in organic photovoltaic, photodetector, and photocatalytic applications. *Adv Mater* 32:2001763
  25. Chen FC, Tseng HC, Ko CJ (2008) Solvent mixtures for improving device efficiency of polymer photovoltaic devices. *Appl Phys Lett* 92:103316
  26. Singh R, Shin SC, Lee H, Kim M, Shim JW, Cho K, Lee JJ (2019) Ternary blend strategy for achieving high-efficiency organic photovoltaic devices for indoor applications. *Chem Eur J* 25:6154–6161
  27. Busireddy MR, Huang SC, Su YJ, Lee ZY, Wang CH, Scharber MC, Chen JT, Hsu CS (2023) Eco-friendly solvent-processed dithienosilicon-bridged carbazole-based small-molecule acceptors achieved over 25.7% PCE in ternary devices under indoor conditions. *ACS Appl Mater Interfaces* 15:24658–24669
  28. Wang G, Wang J, Cui Y, Chen Z, Wang W, Yu Y, Zhang T, Ma L, Xiao Y, Qiao J, Xu Y, Hao XT, Hou J (2024) Achieving high fill factor in organic photovoltaic cells by tuning molecular electrostatic potential fluctuation. *Angew Chem Int Ed* 63:e202401066
  29. Yang SS, Hsieh ZC, Keshtov ML, Sharma GD, Chen FC (2017) Toward high-performance polymer photovoltaic devices for low-power indoor applications. *Solar RRL* 1:1700174
  30. 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
  31. Eibeck A, Nurkowski D, Menon A, Bai J, Wu J, Zhou L, Mosbach S, Akroyd J, Kraft M (2021) Predicting power conversion efficiency of organic photovoltaics: models and data analysis. *ACS Omega* 6:23764–23775
  32. Qia B, Wang J (2012) Open-circuit voltage in organic solar cells. *J Mater Chem* 22:24315–24325
  33. Shockely W, Queisser HJ (1961) Detailed balance limit of efficiency of p-n junction solar cells. *J Appl Phys* 32:510–519
  34. Imamura Y, Tashiro M, Katouda M, Hada M (2017) Automatic high-throughput screening scheme for organic photovoltaics: estimating the orbital energies of polymers from oligomers and evaluating the photovoltaic characteristics. *J Phys Chem C* 121:28275–28286
  35. Zanlorenzi C, Akcelrud L (2017) Theoretical studies for forecasting the power conversion efficiencies of polymer-based organic photovoltaic cells. *J Polym Sci Part B: Polym Phys* 55:919–927
  36. Koster LJA, Mihailtchi VD, Ramaker R, Blom PWM (2005) Light intensity dependence of open-circuit voltage of polymer:fullerene solar cells. *Appl Phys Lett* 86:123509
  37. Proctor CM, Nguyen TQ (2015) Effect of leakage current and shunt resistance on the light intensity dependence of organic solar cells. *Appl Phys Lett* 106:083301
  38. Malhotra P, Khandelwal K, Biswas S, Chen FC, Sharma GD (2022) Opportunities and challenges for machine learning to select combination of donor and acceptor materials for efficient organic solar cells. *J Mater Chem C* 10:17781–17811
  39. Chen FC (2019) Virtual screening of conjugated polymers for organic photovoltaic devices using support vector machines and ensemble learning. *Int J Polym Sci* 2019:4538514
  40. Nagasawa S, Al-Naamani E, Saeki A (2018) Computer-aided screening of conjugated polymers for organic solar cell: classification by random forest. *J Phys Chem Lett* 9:2639–2646
  41. Abadi EAJ, Sahu H, Javadpour SM, Goharimanesh M (2022) Interpretable machine learning for developing high-performance organic solar cells. *Mater Today Energy* 25:100969
  42. Malhotra P, Biswas S, Chen FC, Sharma GD (2021) Prediction of non-radiative voltage losses in organic solar cells using machine learning. *Sol Energy* 228:175–186
  43. RDKit Open-Source Cheminformatics Software. <https://www.rdkit.org/> Accessed 8 March 2025
  44. scikit-learn. <https://scikit-learn.org/stable/> Accessed 8 March 2025
  45. Zhao ZW, 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
  46. Oh S, Kang Y, Kim TH, Kim SJ, Lee MJ, Lee GM, Saeed MA (2024) Shim JW enhancing the indoor performance of organic photovoltaic devices: interface engineering with an amino-benzoic-acid-based self-assembled monolayer. *J Phys Energy* 6:025015
  47. Zou B, Ng HM, Yu H, Ding P, Yao J, Chen D, Pun SH, Hu H, Ding K, Ma R, Qammar M, Liu W, Wu W, Lai JYL, Zhao C, Pan M, Guo L, Halpert JE, Ade H, Li G, Yan H (2024) Precisely controlling polymer acceptors with weak intramolecular charge transfer effect and superior coplanarity for efficient indoor all-polymer solar cells with over 27% efficiency. *Adv Mater* 36:2405404
  48. Li C, Zhou J, Song J, Xu J, Zhang H, Zhang X, Guo J, Zhu L, Wei D, Han G, Min J, Zhang Y, Xie Z, Yi Y, Yan H, Gao F, Liu F, Sun Y (2021) Non-fullerene acceptors with branched side chains and improved molecular packing to exceed 18% efficiency in organic solar cells. *Nat Energy* 6:605–613
  49. Nazari M, Cieplichowicz E, Welch GC (2024) Air processed, high open-circuit voltage indoor organic photovoltaic cells based on side chain modified N-annulated perylene diimides. *Can J Chem Eng* 102:4120–4128

50. ChemSketch freeware. <https://www.acdlabs.com/resources/free-chemistry-software-apps/chemsketch-freeware/> Accessed 8 March 2025

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