npj computational materials

www.nature.com/npjcompumats

ARTICLE OPEN

Efficient screening framework for organic solar cells with deep learning and ensemble learning

Hongshuai Wang[1,2,6], Jie Feng[1,2,6], Zhihao Dong[1,2], Lujie Jin[1,2], Miaomiao Li[3], Jianyu Yuan[1,4]✉ and Youyong Li[1,2,5]✉

Organic photovoltaics have attracted worldwide interest due to their unique advantages in developing low-cost, lightweight, and flexible power sources. Functional molecular design and synthesis have been put forward to accelerate the discovery of ideal organic semiconductors. However, it is extremely expensive to conduct experimental screening of the wide organic compound space. Here we develop a framework by combining a deep learning model (graph neural network) and an ensemble learning model (Light Gradient Boosting Machine), which enables rapid and accurate screening of organic photovoltaic molecules. This framework establishes the relationship between molecular structure, molecular properties, and device efficiency. Our framework evaluates the chemical structure of the organic photovoltaic molecules directly and accurately. Since it does not involve density functional theory calculations, it makes fast predictions. The reliability of our framework is verified with data from previous reports and our newly synthesized organic molecules. Our work provides an efficient method for developing new organic optoelectronic materials.

npj Computational Materials (2023)9:200 ; https://doi.org/10.1038/s41524-023-01155-9

INTRODUCTION

Organic semiconducting materials exhibit great synthetic flexibility, which allows for excellent tunability over the bandgap, energy level, and carrier mobility, offering great potential in the design of efficient optoelectronic devices like organic solar cells (OSCs). In comparison with inorganic counterparts, OSCs show unique advantages like light weight, good flexibility, semi-transparency, etc.[1–3]. Advances in the last decade in functional materials design, morphology optimization, and device architecture engineering have led to certified power conversion efficiencies (PCEs) of over 19%, demonstrating great potential for emerging photovoltaic technology. However, exploring suitable organic molecules in the vast organic compound space is extremely difficult, and efficiency breakthrough in the lab needs the constant input of intensive labor and time.

Although DFT calculations allow us to acquire many electronic structural properties of organic molecules without complex organic synthesis, we still lack an effective mathematical model to calculate the PCEs directly from the physical properties of the molecules[4–6]. In addition, although DFT calculations save economic costs, the huge time cost still limits their application in the high-throughput screening of molecules. Therefore, it's an urgent problem to establish a quantitative structure–property relationship (QSPR) model that can conduct the high-throughput screening of the organic compound space to find more suitable molecules.

As a powerful technology for mining relationships hidden in big data, artificial intelligence has brought great development and prosperity to the field of machine learning[7–11]. With the development of material informatics, a new generation of material research and development paradigms is gradually formed: (1) Use the material database to train machine learning models. (2) Use the model to predict new materials. (3) Verify the results with experiments or calculations[12–14]. Machine learning shows an excellent performance in accelerating the discovery of new materials, guiding the design of new materials, and exploring the QSPR of materials[15–19].

The recent application of machine learning in the field of OSC shows its potential in performing the high-throughput screening of organic molecules effectively[20–25]. Scharber et al. built a model that can calculate the PCEs with a function of the bandgap and the energy levels of the conjugated polymer[26]. Harvard Clean Energy Project (CEP) collected the calculations and experimental data of thousands of organic photoelectric molecules and predicted their PCEs using Scharber's model[27,28]. The same team later used the Gaussian process regression (GPR) method in machine learning to correct Scharber's model, which increased its Pearson correlation coefficient (r) from 0.3 to 0.43 and made a rough estimate of the PCEs[29]. However, due to the low accuracy and the time-consuming quantum mechanics calculations of the energy levels, Scharber's model is not competent for the fast and accurate high-throughput screening[30].

By removing the expensive input of quantitative microscopic properties, Sun et al. established a model with deep learning that can quickly classify photoelectric molecules[25,31]. This model can use molecular graphs or fingerprint information as input to predict the PCEs interval (0–3% or 3–14.6%). Since the acquisition of molecular fingerprints does not require additional quantum mechanics calculations, this model can achieve a fast classification of molecules but cannot predict the value of PCEs. Moreover, the accuracy of this model is not satisfactory (69.41%) since the existing data cannot meet the high demand for deep learning. The input of the molecular and microscopic properties is very helpful to improve the accuracy of the model. For example, Alessandro et al. trained a KRR model that can predict the PCEs better ($r = 0.68$) by combining both structural and electronic descriptors, and such accuracy has met the requirements of the high-

[1]Institute of Functional Nano & Soft Materials (FUNSOM), Soochow University, 199 Ren'ai Road, Suzhou 215123 Jiangsu, PR China. [2]Jiangsu Key Laboratory for Carbon-Based Functional Materials and Devices, Soochow University, 199 Ren-Ai Road, Suzhou Industrial Park, Suzhou, Jiangsu 215123, PR China. [3]School of Materials Science and Engineering, and Tianjin Key Laboratory of Molecular Optoelectronic Science, Tianjin University, Tianjin 300072, China. [4]Jiangsu Key Laboratory of Advanced Negative Carbon Technologies, Soochow University, Suzhou 215123 Jiangsu, PR China. [5]Macao Institute of Materials Science and Engineering, Macau University of Science and Technology, Taipa, 999078 Macau, SAR, China. [6]These authors contributed equally: Hongshuai Wang, Jie Feng. ✉email: jyyuan@suda.edu.cn; yyli@suda.edu.cn

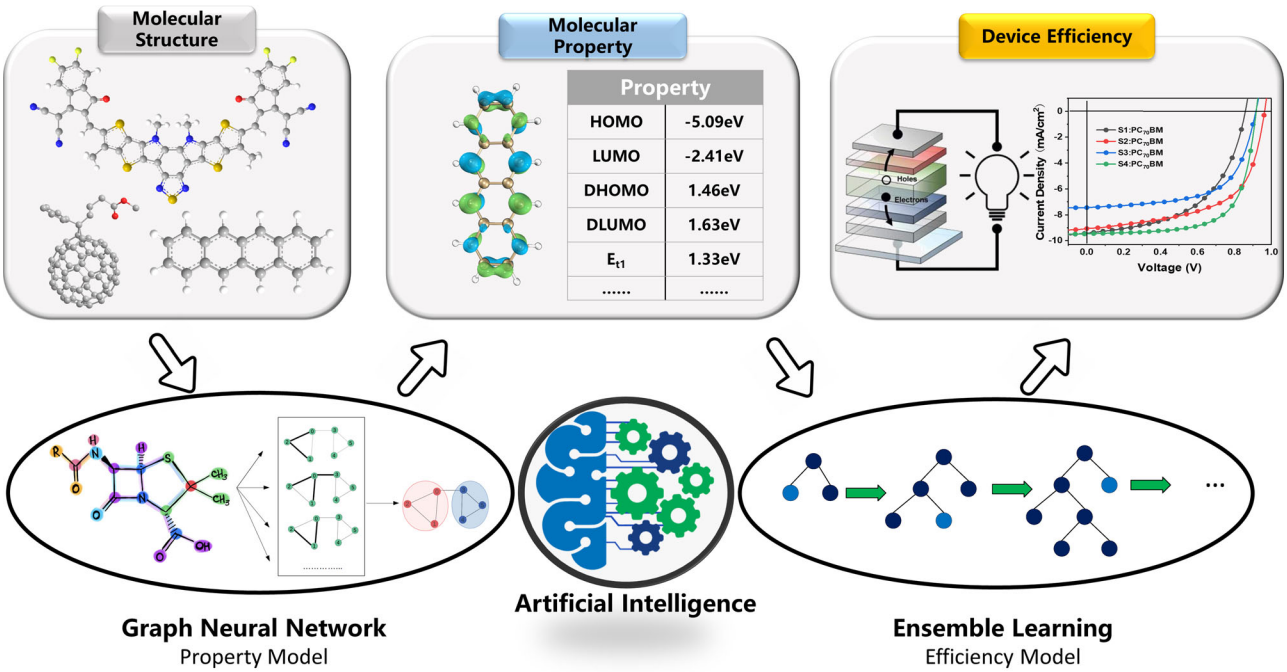

Fig. 1 The workflow of framework. Convert molecules to graphs as input to Graph Neural Network models (Property Model). Property Model predicts the molecular properties of molecules as input to the Light Gradient Boosting Machine (Efficiency Model). Efficiency Model predicts the final power conversion efficiencies (PCEs) of the organic solar cells.

throughput screening[30]. With the data collection from published literature and quantitative calculations, Ma et al. trained a model that can directly predict the PCEs using the GBRT method[32,33]. This model takes the 13 microscopic properties of molecules as input and shows a high accuracy ($r = 0.79$). Obviously, the addition of molecular, microscopic properties improves the accuracy of existing models, which meets the requirements of high-throughput screening.

However, many expensive calculations of microscopic properties (especially excited states) greatly limited the high-throughput screening of the large-scale organic compound space for suitable molecules. Therefore, we have to train an accurate machine-learning model for the high-throughput screening of organic optoelectronic molecules with input that can be easily obtained.

In this work, we established an automated framework that can quickly predict the PCEs of OSCs. First, a small dataset containing high-quality experimental data was used to train an ensemble learning model that can predict the PCEs based on the physical and chemical properties of molecules. Then we trained a deep learning model that can predict the molecular properties accurately by using a graph neural network (GNN) architecture and a dataset containing a large number of molecular structures and properties. Specifically, we used self-learning input (SLI)-GNN, which was recently developed by ourselves[34]. Based on these two models, we designed this framework that can directly predict the PCEs based on the molecular structure. Finally, the performance of our framework in high throughput screening is verified by our experimental results. By a combination of deep learning and ensemble learning, we achieve direct, fast, and accurate prediction of PCEs based on molecular structure.

RESULTS

Workflow

Figure 1 shows our efficient prediction workflow for OSCs, which includes two parts: (1) Predict molecular properties based on molecular structures with the GNNs model (Property Model). (2) Train an ensemble learning model to realize the prediction of the PCEs based on the molecular physical properties (Efficiency Model). Here we call the GNN model as Property Model and the ensemble learning model as Efficiency Model.

We collected data and performed DFT calculations to build a database including the experimental device efficiencies and the physicochemical properties of molecules. After that, the Efficiency Model was trained using this database, and the relationship between the molecular microscopic properties and macroscopic properties of devices was established owing to the ability of machine learning to extract complex relationships. In order to avoid the time-consuming DFT calculations for the microscopic properties of molecules, we established the Property Model with a database containing a large number of molecular structures and properties to predict molecular microscopic properties based on molecular structures. With the input of molecular structures, the Property Model can be used to quickly predict desired molecular microscopic properties, and the Efficiency Model can be used to predict the PCEs of the device based on microscopic properties predicted by the Property Model. In our framework, the Efficiency Model is critical for accuracy, and the Property Model is critical for efficiency. The combination of these 2 models leads to a direct, fast, and accurate prediction of OSCs from molecular structures.

Dataset and feature selection

Our database contains 440 small molecule/fullerene pairs and their corresponding PCEs. These data come from published literature. Since one molecule was synthesized by multiple experiments and may correspond to multiple PCEs, the highest PCEs were chosen as the criterion. In order to improve the Property Model by transfer learning, 200,000 pieces of data from the Clean Energy Project Database (CEPDB) were used for pre-training[35]. These data only include structures and microscopic properties of the molecule without the experimental PCEs. The details of data selection are given in the Supplementary Note 1.

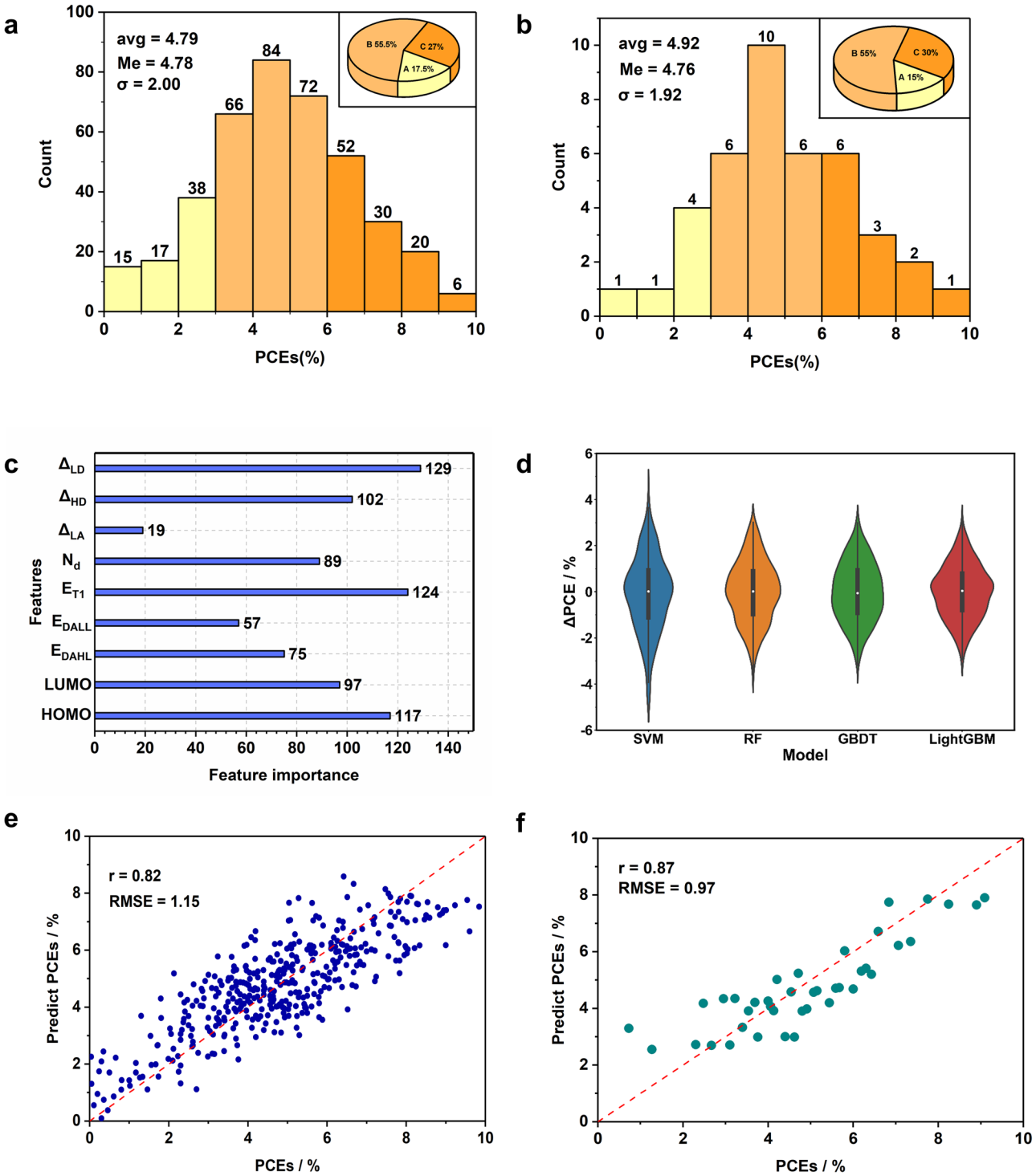

Fig. 2 The performance of the Efficiency Model. a The distribution of PCEs in the training set. **b** The distribution of PCEs in the testing set. **c** The feature importance for the LightGBM model. **d** The violin plots of predictive errors (ΔPCE) trained by different ML techniques. **e** The predicted PCEs for the LightGBM model versus experimental PCEs for the training set. **f** The predicted PCEs for the LightGBM model versus experimental PCEs for the test set.

The 440 pieces of data are divided into the training set and the test set according to the ratio of 10:1, and the 2 datasets are used to train and test model, respectively. The PCEs distributions of the training set and the test set are shown in Fig. 2a, b, respectively. The ratio of OSCs with the low (0–3%), medium (3–6%), and high (6%~) PCEs is about 2:5:3. From Fig. 2a, b, we can see there is little difference between the training set and the test set for the PCEs distributions, as the average (avg), the median (Me), and the dispersion coefficient (σ) is close, which ensures that the test set is suitable for examining the prediction accuracy of the models.

Feature selection is an important step in the process of building machine learning models. Generally, the stronger the correlation between the features and the targets, the less difficult the learning

Table 1. The selected features and their physical interpretations.

Feature	Physical interpretations
HOMO	The highest occupied molecular orbital energy level of donors
LUMO	The lowest unoccupied molecular orbital energy level of donors
E_{DAHL}	The difference between HOMO of donor and LUMO of acceptor
E_{DALL}	The difference between LUMO of the donor and LUMO of the acceptor
E_{T1}	The energy of the electronic transition to the lowest-lying triplet state
N_d	Number of unsaturated atoms in the donor molecules
Δ_{LA}	The difference between LUMO and LUMO + 1 of acceptor
Δ_{HD}	The difference between HOMO and HOMO − 1 of donor
Δ_{LD}	The difference between LUMO and LUMO + 1 of donor

task would be. The feature selection process is to find suitable physical and chemical properties of molecules showing a strong correlation with the PCEs for the Efficiency Model. Earlier studies proved that the microscopic properties of molecules are helpful in improving the accuracy of the model[36,37]. Since our ultimate aim is to achieve high-throughput screening, we need to consider the availability of data when selecting microscopic properties. Although the final DFT calculation will be skipped, these properties still need to be calculated to generate the training set for the training Property Model. Considering some of the descriptors previously proposed and the availability of data[33], we finally selected nine properties shown in Table 1 as the learning features for the Efficiency Model. The selection of these features was made with a careful consideration of several factors, including their availability, relevance to the target property (PCE), and prior knowledge from the literature. To achieve high-throughput screening, the availability of data for these features is an important criterion. While some features could yield useful insights, their inclusion would depend on whether they could be reliably calculated across a wide range of molecules. To capture the complexity of the system. The features chosen include ground-state properties, excited-state properties, and characteristics of molecular structure. The Pearson's correlation coefficient between all features was considered. Supplementary Fig. 1 shows the Pearson's correlation coefficient for all features that help us obtain an initial insight into the data. Most of these features show a weak correlation with each other. Since the addition of redundant features can reduce the difficulty of learning, these 9 selected features are not independent of each other, and some features, such as N_d (number of unsaturated atoms in the donor molecules), can be obtained easily without DFT calculation. These features of 440 data are calculated by DFT and constitute the dataset for training the model. Considering the consistency of the data distribution, the same calculation method as that of CEPDB[35] was used. The DFT calculation details are given in the section "Methods".

Efficiency Model: from molecular property to device performance

Support vector machine (SVM), random forest (RF), gradient boosting decision tree (GBDT), and LightGBM are common machine learning models that have been widely used in the field of materials science. Based on the training set of 400 small molecule donors, 4 different models (SVM, RF, GBDT, and LightGBM) were used to capture the correlation between the physicochemical properties of molecules and the PCEs of devices. Leave-One-Out validation was used to assess how well each model generalizes to another dataset and ensure they are not over-fitting[38]. Hyperparameters were optimized with grid searches in the training and the Leave-One-Out validation. The implementation details and hyperparameters of 4 models are given in the "Methods" and Supplementary Table 1. The PCEs distributions corresponding to the prediction of the models and the experimental determinations, as well as the error distributions on the dataset, are shown in Fig. 2 and Supplementary Fig. 2. We employed the violet plot to visualize the error distributions of four models in Fig. 2d and in the violet plot, the fatter part means the more centralized distribution of the data. For the SVM and RF model, there are several abnormal points with absolute errors exceeding 4%, which deteriorate the performance of the model. For GBDT and LightGBM models, the errors are basically distributed, ranging from −2% to ~2%, and most of the points are concentrated from −1% to 1%. The comparison of these four models shows that the 2 integrated models using LightGBM and GBDT have better performance than SVM and RF. From Fig. 2e, f, it can be seen that the performances of LightGBM on the training set and test set are similar ($r = 0.82$, $r = 0.87$), which indicates that the model is not overfitting. Figure 2e, f and Supplementary Table 2 show that the RMSE and the r of LightGBM are also better than SVM and RF models. Since the principles of GBDT and LightGBM are similar, the performance of the two models is equivalent. LightGBM has better generalization ability than traditional boosting models like GBDT, it was selected for the Efficiency Model to predict the PCEs in the end.

The influence of each feature on PCEs is compared by analyzing the weight of each feature. The computational details used to analyze these parameters' importance are given in the section "Methods". Figure 2c shows the weight ranking of each feature used in the trained LightGBM model. It can be seen that the molecular orbital energy level has a significant effect on the PCEs, especially the difference between LUMO and LUMO + 1 and the difference between HOMO and HOMO-1. The physical meaning of these differences is the degeneracy of the HOMO and the LUMO of a molecule, and their important influences on the photoelectric properties of the molecule have been proven[39,40]. In addition, the energy of the electronic transition to the lowest-lying triplet state is quite informative among the features[41,42]. It is clear that the excited-state properties have an influence on the photo-physical processes, thus having a relation with the short-circuit current density[43–45]. The distinction between the LUMO and LUMO + 1 energy levels of the acceptor appears to be less significant for the model. This is likely because, in our dataset, the acceptors are fullerenes, and there are only two distinct molecules present ($PC_{61}BM$ and $PC_{71}BM$).

Property Model: from molecular structure to molecular property

The Efficiency Model realizes the prediction of the PCEs based on the molecular properties. However, molecular microscopic properties that require time-consuming DFT calculations are inapplicable to high-throughput calculations. Therefore, the Property Model is trained to establish the relationship between the structures and the molecular properties. Classical machine learning methods cannot capture the molecular structure information well, and it is difficult to achieve molecular structure-to-property prediction. As a new deep learning architecture, GNN has a wide range of applications in chemistry due to its natural adaptability to molecular structures. We constructed a Property Model using a GNN to capture the relationship between molecular structure and properties. Figure 3 shows the framework of the Property Model.

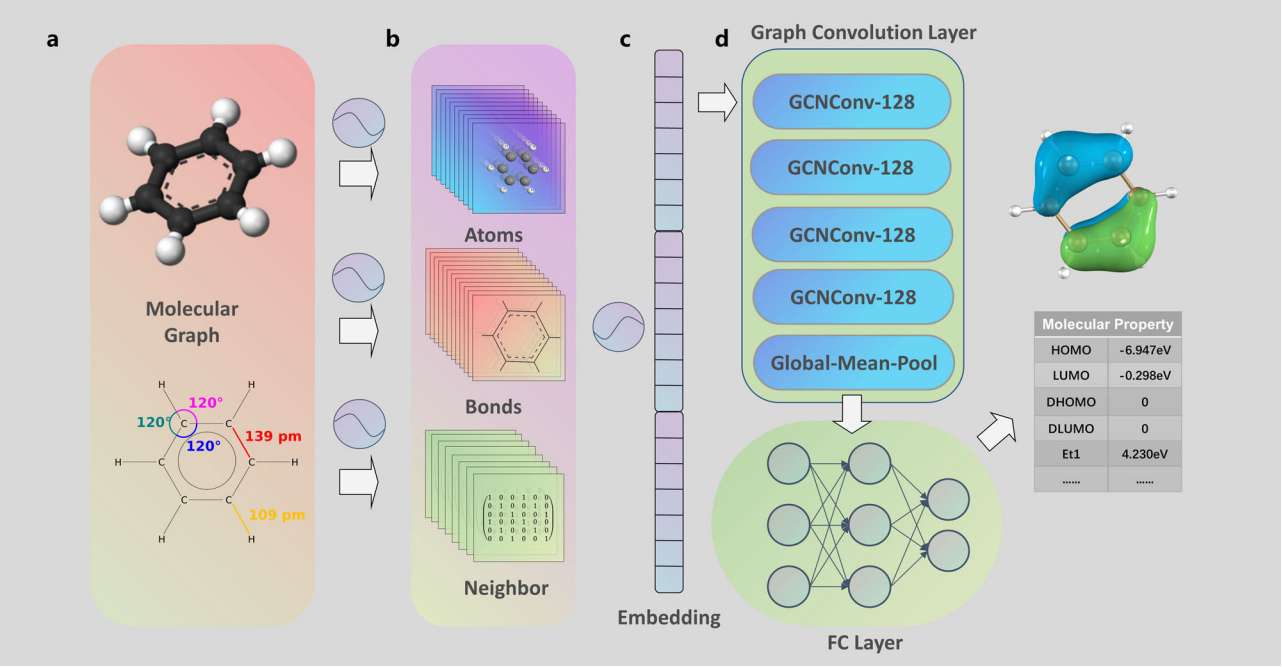

Fig. 3 Illustration of the graph convolution neural networks. a The molecular structure was converted to the molecular graph. **b** Extract three features of atoms, bonds, and neighbor matrix from molecular graphs. **c** The feature matrix embedding operation is input to the graph convolution layer. **d** Use the output of the graph convolutional layer as the input of the fully connected layer (FC Layer), and the FC layer predicts molecular properties.

First, we convert the input molecule structure into a graph, with atoms as nodes and chemical bonds as edges. Three matrices are used to record the properties of the molecular graph. Matrix 1 describes the node properties of atoms, matrix 2 describes the properties of edges, and matrix 3 is an adjacency matrix that describes the adjacency properties of nodes. The three matrices are passed to the graph convolution layer after the embedding operation. The message-passing mechanism of graph convolution can well describe the influence of connections between nodes. The output of the graph convolutional layer is pooled as input to the fully connected (FC) layer. The FC layers give the final prediction of molecular properties. More details on model building are given in "Methods".

The complexity of the graphical model enables it to handle more complex and variable molecular structures but, at the same time, increases the data volume requirements. We endeavored to supplant the Efficiency Model with this model, utilizing the identical dataset that had been previously employed for the Efficiency Model's training. We have presented the outcomes of this training results in Supplementary Fig. 3. It demonstrates that the model has a tendency to predict the outcome as the mean of the power conversion efficiency (PCE). This observation indicates that the model's training is suboptimal. It further suggests that the current volume of data is insufficient for the model to grasp the relationship between structural attributes and device efficiency. In order to improve the accuracy of the model, the strategy of transfer learning was adopted[46–48]. Based on the original 400 training set, 200,000 CEP data containing molecular structures and physicochemical properties are used to train the Property Model, which can predict properties based on molecular structures. This dataset is used to pre-train the Property Model, a process in which the model learns to generalize the relationship between molecular structures and their properties. Essentially, the model is learning a mapping from molecular structure to molecular properties. Once this pre-training phase is completed, the model is then fine-tuned on the original 400-molecule dataset. This step adjusts the parameters of the model specifically to the task of predicting the properties of molecules in the smaller but more focused dataset. The fine-tuning process is where the "transfer" in transfer learning occurs knowledge gained from the larger dataset is applied to improve performance on a smaller, related dataset. We utilized the 5-fold cross-validation (5-CV) method to assess the generalization ability and robustness of our models. This approach is effective for estimating the performance of the model on unseen data and provides a good understanding of the variability of the model predictions[49]. Supplementary Fig. 4 and Fig. 4 show the error distributions corresponding to the direct training with 400 molecular training data and the transfer learning, respectively. It can be seen that transfer learning improves training accuracy effectively. The results of the 5 features to be predicted are given in Fig. 4. It can be clearly seen that the prediction errors of the 5 features are almost distributed in the range of −0.2 eV and −0.2 eV, which are close to the errors of DFT calculations, therefore Property Model can accurately predict the microscopic properties of the molecules we need. The prediction accuracy of the Property Model for the molecular ground state properties (HOMO, LUMO, Δ_{HD}, Δ_{LD}) is better than E_{T1}, which is consistent with the noise distribution of our dataset. The reason is the accuracy of the DFT calculations of the ground state energy level is better than that of the excited state energy.

Efficient and generalization ability verification

The Property Model and Efficiency Model utilize a combination of GNNs and LightGBM to implement the framework from molecular structure to device efficiency prediction. Forty test data from previous reports were used to verify the entire framework. The 40 test data have never been used in the training progress of the above 2 models, so they are 40 "unseen" data for this framework. First, molecular structures were used as the input into the Property Model to obtain the corresponding physical and chemical properties. After that, we used the predicted physical and chemical properties as the input into the Efficiency Model to obtain the PCEs. The predicted PCEs and the error distributions are shown in Fig. 5a, b. From Fig. 5a, b, it can be seen that the

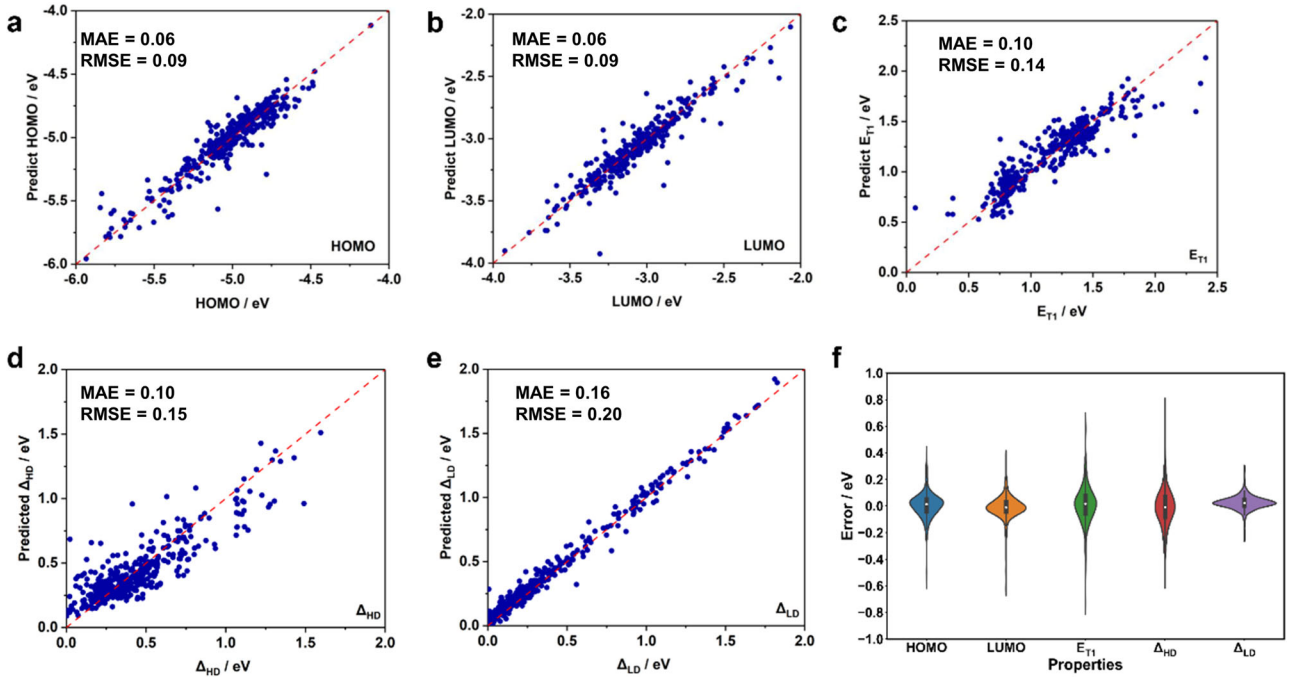

Fig. 4 The 5-CV results for the SLI-GNN model (Property Model) versus the calculated values from DFT. a The 5-CV results for HOMO. **b** The 5-CV results for LUMO. **c** The 5-CV results for E_{T1}. **d** The 5-CV results for Δ_{HD}. **e** The 5-CV results for Δ_{LD}. **f** The violin plots of errors for five properties by Property Model.

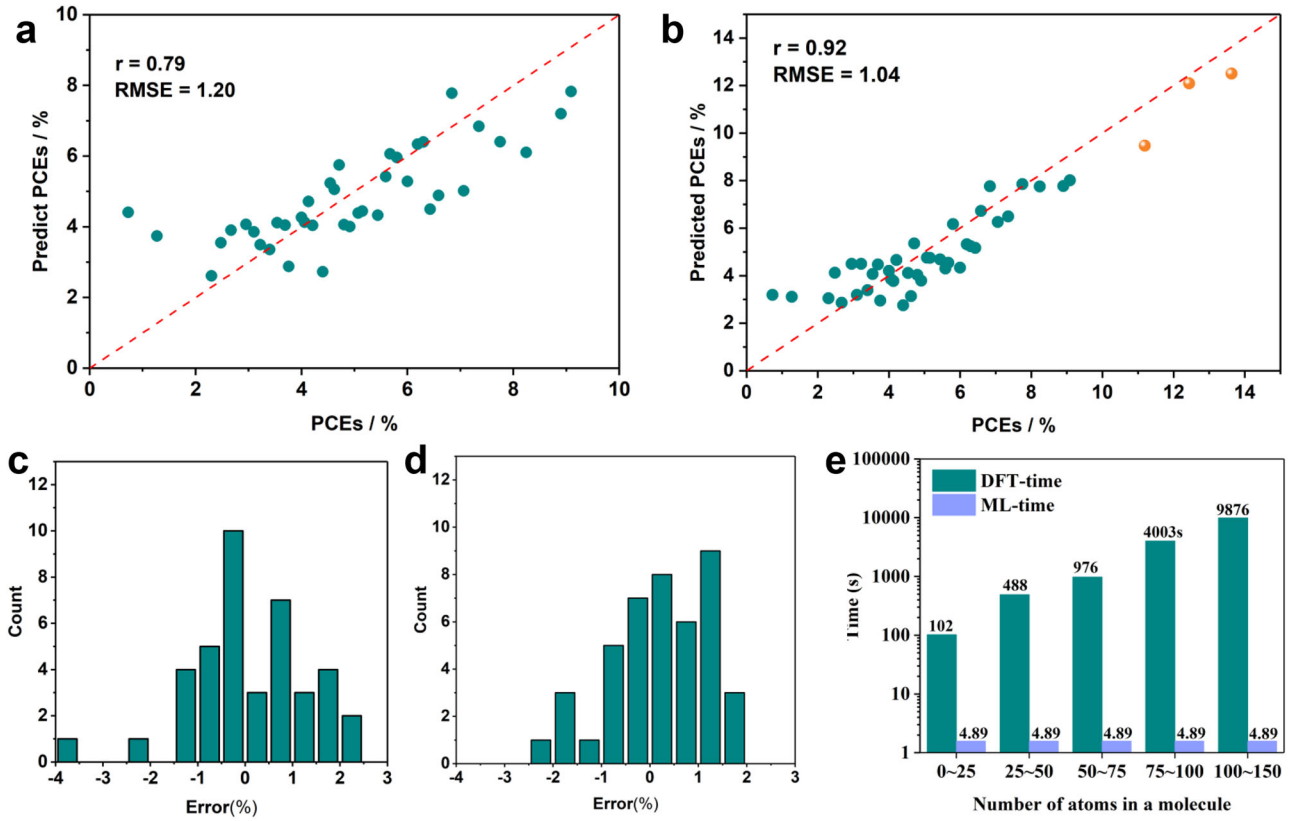

Fig. 5 The performance of the whole framework. a The predicted PCEs by machine learning versus experimental PCEs for the test set. **b** The predicted PCEs by machine learning versus experimental PCEs for the test set include the Y6 acceptor. **c** The corresponding error between the predicted PCEs and experimental PCEs for the test set. **d** The corresponding error between the predicted PCEs and experimental PCEs for the test set includes the Y6 acceptor. **e** Comparison of time consumption for predicting molecular properties by DFT and ML.

prediction results are still accurate ($r = 0.79$, RMSE = 1.20). Due to the bias of the Property Model, the RMSE increases and r decreases versus the PCEs that are directly predicted based on the properties calculated by DFT. For these 40 data, we only fed their molecular structures into the framework and skipped the DFT calculations. Therefore, we have a full progress in approaching the accurate prediction of the efficiency from the molecular structures directly. We contrast the time consumption of DFT calculation of molecular properties with the time consumption of ML prediction in Fig. 5e. It can be clearly seen that the ML method is significantly more time-saving than DFT, and the time-consuming of DFT is exponentially related to the number of atoms. The prediction progress takes only a few minutes, and the whole framework is compiled into an easy-to-use Python package, which is described in detail in Supplementary Note 3.

The construction of the Property Model and Efficiency Model is based on the fullerene acceptor OSC dataset. Recently, non-fullerene acceptors received more attention due to various advantages[50–55]. The Efficiency Model considered the characteristics of the molecular acceptors during construction and thus can be easily generalized to non-fullerene acceptors. We reconstructed the Efficiency Model by adding 11 non-fullerene acceptor device efficiencies and their molecular structure of donor and acceptor to the original dataset as new datasets. Three molecules that did not participate in training and the original test set were used as a new test set to test the new model. Details of the non-fullerene acceptor device data are given in Supplementary Table 4. The tested PCEs and errors are given in Fig. 5c, d. RMSE decrease and r increase in Fig. 5c compared to Fig. 5a. For non-fullerene acceptors devices, the model shows excellent generalization ability.

Novel molecular screening and experimental verification

Novel molecules were designed to test the accuracy and rapid screening capabilities of our framework. Conjugated molecules used in OSCs are composed of different types of building blocks such as donor (D), p-spacer (S), acceptor (A), and end-capping (C) units shown in Fig. 6a. Through the arrangement and combination of the 20 fragments in Fig. 6a, 375 molecules of 3 configurations (CADAC, CSASC, CSDSC) were designed as input for our framework. Their PCEs were predicted by our model in a few minutes with a personal computer. This demonstrates the capability of the model to rapidly screen molecules. The predicted results for all designed molecules are given in Supplementary Table 5. The distribution of PCEs is shown in Fig. 6b. As can be seen from Fig. 6b, the PCEs of the molecule in the CADAC configuration are higher than that of the other two configurations.

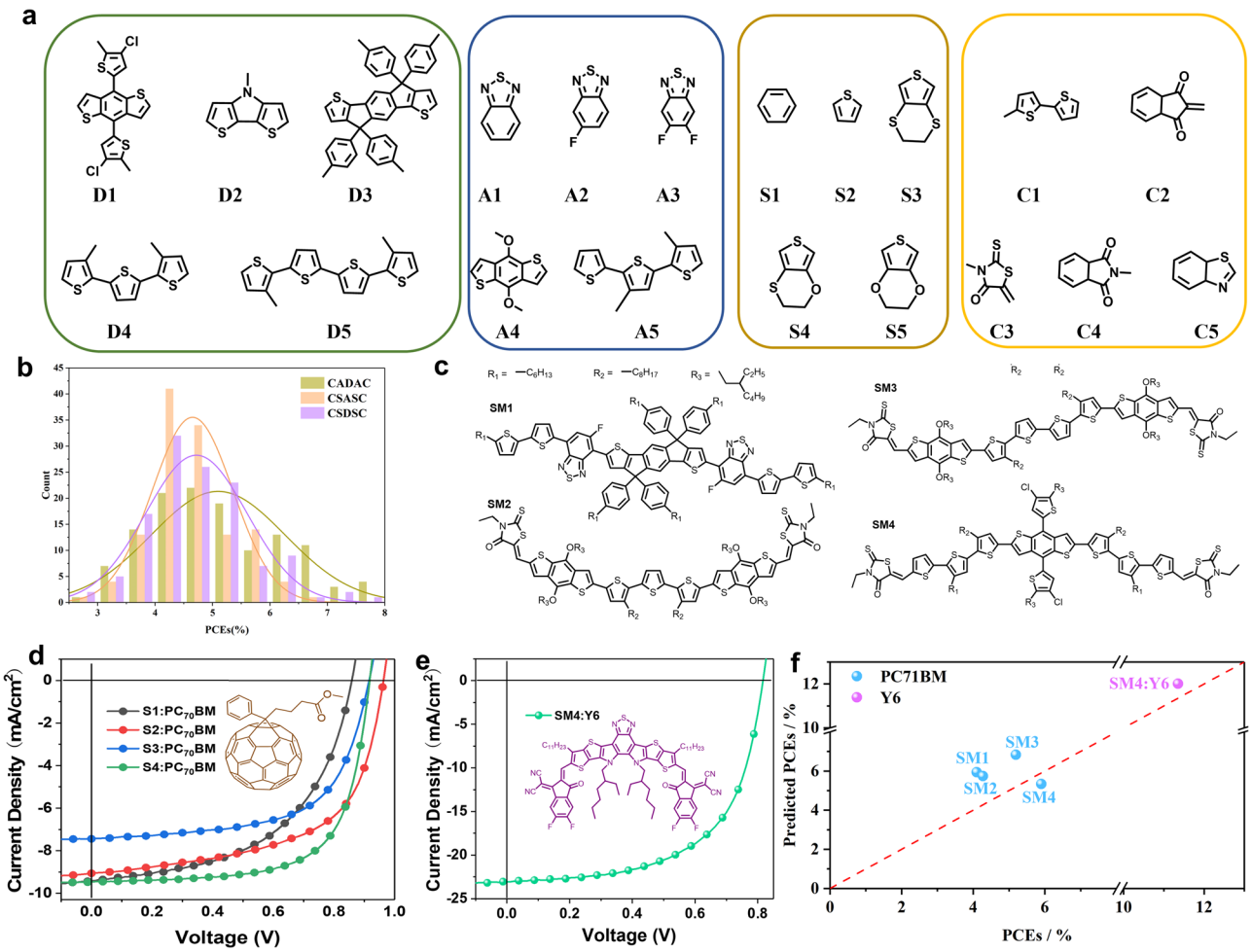

Fig. 6 The results of experimental verification. a Twenty molecular fragments of four types used to splice molecules. **b** PCEs distribution of 375 designed molecules. **c** The formula of four small molecules (SM1–SM4) that are synthesized and manufactured into photovoltaic devices to verify this framework. **d** *I–V* curves of four synthetic molecules with two type acceptors $PC_{71}BM$ manufactured into photovoltaic devices. **e** *I–V* curves of four synthetic molecules with two type acceptors Y6 manufactured into photovoltaic devices. **f** The predicted PCEs by machine learning versus experimental PCEs for devices manufactured with SM1–SM4.

Table 2. The device performances of SM1–SM4 and their corresponding PCEs predicted from our model.

Donor	Acceptor	V_{oc} (V)	J_{sc} (mA/cm^2)	FF (%)	PCEs (%)	Predicted PCEs (%)
SM1	$PC_{71}BM$	0.85	9.40	0.54	4.29	5.94
SM2	$PC_{71}BM$	0.97	9.06	0.59	5.19	6.83
SM3	$PC_{71}BM$	0.92	7.42	0.63	4.27	5.74
SM4	$PC_{71}BM$	0.90	9.71	0.62	5.44	5.34
SM4	Y6	0.81	23.05	0.61	11.35	12.01

Table 3. The calculated and predicted features of SM1–SM4.

	SM1/eV	SM2/eV	SM3/eV	SM4/eV
HOMO_cal	−4.771	−4.987	−4.908	−4.826
HOMO_pre	−4.780	−4.917	−4.882	−4.847
LUMO_cal	−2.979	−2.967	−2.929	−2.923
LUMO_pre	−3.007	−2.954	−2.919	−2.967
E_{T1}_cal	1.224	1.410	1.403	1.449
E_{T1}_pre	1.277	1.554	1.558	1.469
Δ_{HD}_cal	0.464	0.247	0.159	0.261
Δ_{HD}_pre	0.323	0.404	0.355	0.316
Δ_{LD}_cal	0.158	0.134	0.088	0.075
Δ_{LD}_pre	0.197	0.113	0.105	0.101

The subscript 'cal' means the value calculated from DFT calculations, while the subscript 'pre' means the value predicted from the Property Model.

In order to validate our framework, 4 small molecules (SM1–SM4) with CADAC configuration that have not been reported were synthesized in this work. The formulas of four molecules are shown in Fig. 6c, with a detailed synthetic routine displayed in Supplementary Fig. 5 according to the previous report. They are the representatives of the first generation of organic photovoltaics molecules (SM1)[56], the second generation of organic photovoltaics molecules (SM2 and SM3)[57], and the third generation of organic photovoltaics molecules (SM4)[58]. These molecules were manufactured into OSC devices with $PC_{71}BM$ (SM1–SM4) and Y6(SM-4) acceptors. We further investigated the photovoltaic performance of these functional molecules using the conventional device structure described in the experimental part. In addition, we adopted both conventional fullerene acceptor $PC_{71}BM$ as well as recently emerged non-fullerene molecule Y6 as the electron acceptor. The *I*–*V* curve of the optimal device for each molecule is shown in Fig. 6d, e. The PCEs determined by experiments and predicted by our framework are shown in Fig. 6f and Table 2. It can be seen that the predicted PCEs of $PC_{71}BM$-based devices are very close to the actual experimental PCEs, with an error below 2%. The predicted PCEs of the Y6-based device (12.01%) are also very close to the actual experimental PCEs (11.35%). The trend of the predicted value is also consistent with that of the experimental value, while the predicted PCEs are higher than the experimental PCEs. In order to analyze the predicted results, the features predicted by the Property Model for the 4 molecules are given in Table 3. As shown in Table 3, the HOMO and the LUMO of the molecules predicted by the Efficiency Model are very accurate, as well as the predicted E_{T1} and Δ_{LD}. The slight deviation from the calculated value of Δ_{HD} should be the major factor that influences the predicted PCEs.

DISCUSSION

By using machine learning, a general approach to predict the PCEs of organic solar cells was developed, which shows excellent performance ($r = 0.79$) without any DFT calculation. Efficiency Model built a quantitative model from molecular properties to device performance by using the ability of machine learning to derive relationships from large amounts of data. This model solves the problem that traditional calculation cannot directly obtain accurate device efficiency theoretically. The Property Model established the corresponding relationship between molecular structure and properties utilizing the natural adaptability between GNN and molecular structure. The problem in the Efficiency Model that large-scale screening cannot be achieved due to the high computational cost of molecular properties is solved by the Property Model. An OCSs database which has nearly 500 molecules with the calculated properties (HOMO, LUMO, E_{T1} et al.) was established. The workflow was developed into an easy-to-use Python software package, which is open access. Our work is expected to significantly assist the development of OCSs and the design of new efficient materials like OLEDs or organic catalysts with high-throughput screening.

The combination of the two models makes it possible to predict device performance from molecular structure. The reliability of the model has been verified by experiments. However, since the feature selection of the Efficiency Model only considers some properties of the electronic structure of a single molecule, it ignores some factors that can affect the device efficiency, such as molecular side chains, which are highly correlated with solubility. Therefore, the accuracy of this working framework can be further improved if more features at larger scales are considered. But it also increases the difficulty of building the model.

Although we attempted to assess the generalization ability of their model by incorporating 11 non-fullerene acceptor devices in the Efficiency Model and testing it on 3 additional non-fullerene acceptor devices, the number of testing systems appears to be insufficient. As we are aware of the significance of this new class of materials, we are in the process of expanding our dataset to include more non-fullerene acceptors for both training and testing purposes. We believe that as we gather more data, the accuracy and applicability of our model will further improve.

In addition, since the space of organic molecules is almost infinite, how to construct the molecular structure that needs to be screened is another problem that needs to be explored. This approach involves generating completely new molecular structures from scratch using computational algorithms, often guided by desired properties or specific design criteria. More active learning strategies, such as generative adversarial networks and reinforcement learning combined with our model, could enable the transition from high throughput screening to automatic design. This will help researchers efficiently explore the vast chemical landscape and identify promising candidates for organic solar cells.

METHOD

Quantum chemical calculations

The ground state structures of all molecules were optimized by DFT/BP86 with the def2svp basis set[59–61]. The energy levels of the molecules were calculated with B3LYP/def2svp based on the optimized geometries[59,61,62]. The energy of the electronic transition to the lowest-lying triplet state was calculated by TD-DFT/M062X with the 6–311 g(d) basis set[63,64]. To avoid a very high computational cost, side alkyl chains were not considered in the calculations though they have negligible influences on electronic properties. All the calculations above were performed using Gaussian 09 software package[65].

Machine learning models

Scikit-Learn was used to obtain the SVM, GBDT, and RF algorithms, which were chosen to build models for the control study[66]. The details of the model evaluation metrics are given in Supplementary Note 2. SVM is an algorithm that learns the importance of each training data point for representing the decision boundary between the 2 categories[67,68]. RF and GBDT are 2 sub-models of Decision Tree (DT) models[69]. To address the problem of overfitting that limits the application of DT, RF utilizes randomness injection into the tree building while GBDT tries to correct the mistakes of the previous tree continuously[70–72]. The Light Gradient Boosting Machine (LightGBM) model is a tree-based learning algorithm improved from GBDT, hence this algorithm has the advantages in regularization and multiple loss functions[73]. Instead of using all the sample points to calculate the gradient, GOSS (Gradient-based one-side sampling) was used to calculate the gradient. Exclusive feature bundling combines certain features together to reduce the dimensionality of the features and find the best segmentation point to reduce consumption. Hyperparameters of these models are given in Supplementary Table 1.

Feature importance in Fig. 2c is computed with the feature_importances_ in the Scikit Learn package. It is fundamentally a measure of how much a particular feature contributes towards improving the prediction accuracy of the model. This is achieved by analyzing how each feature influences the splitting decisions within the ensemble of decision trees that form the gradient-boosting model. The "split" method was used in this study; the importance of a feature is determined by the number of times it is used to split the data across all trees in the model. This provides an indication of the frequency of a feature's usage in generating the decision trees.

Property Model is implemented in PyTorch Geometric, a library built upon PyTorch to easily write and train GNNs. The atom, bond, and neighbor feature embedding layers produce 64–256 dimensional inputs to the graph convolution layers. The main body of the network consists of 3–9 graph convolution (GCN) layers, each with hidden dimension 128. The final atom representations are reduced by global mean pooling and mapped to regression outputs by the full connection layer. Hyperparameters of this model are given in Supplementary Table 3.

Devices fabrication

First, PEDOT:PSS was spin-coated onto the pre-cleaned ITO substrate at 4500 rpm for 40 s after the filtration through a 0.45 μm filter, and then the substrates were baked at 150 °C for 10 min under ambient conditions. Subsequently, for $PC_{71}BM$-based device, a blend of the small molecule and $PC_{71}BM$ was dissolved in chloroform (the total concentration: 14 mg mL^{-1}, D/A weight ratio: 1/0.8, 0.5% DIO as additive), and the blend solutions were spin-cast onto the PEDOT:PSS layer at a spin-coating rate of 2500 rpm, after the active layers were treated with thermal annealing at 100 °C for 10 min; for Y6 based device, a blend of the small molecule and Y6 was dissolved in chloroform (the total concentration: 16 mg/mL, D/A weight ratio: 2/1). The blend solutions were spin-cast onto the PEDOT:PSS layer at a spin-coating rate of 3000 rpm, and the active layers were treated with thermal annealing at 150 °C for 5 min. Then, 8 nm PFN-Br was spin-coated onto the active layer. Finally, Al at a speed of 2 Å/s (100 nm) was thermally evaporated to accomplish the device fabrication. The *J*–*V* characterization was performed by Keithley 2400 digital source meter under simulated AM 1.5 G solar irradiation at 100 mW cm^{-2}. The device area is 0.0725 cm^2, and solar cell devices were measured in forward scan (−1.0 V → 1.0 V, step 0.0125 V, scan rate: 0.1 V s^{-1}) in the glovebox.

DATA AVAILABILITY

We have included 440 instances of training data for the device efficiency model, which can be found in train.db and test.db files under the 'data' fold in https://github.com/HongshuaiWang1/OPVGCN. This fold further houses the data pertaining to our uniquely designed molecules, including their Power Conversion Efficiencies (PCEs), denoted as C***C.db and predC***C.db, respectively. Lastly, data from the esteemed Harvard Clean Energy Project can be accessed at: https://www.matter.toronto.edu/basic-content-page/data-download.

CODE AVAILABILITY

We have generously provided unfettered access to our code and model at the following locations: https://github.com/HongshuaiWang1/OPVGCN and https://github.com/Austin6035/SLI-GNN.

Received: 28 March 2023; Accepted: 10 October 2023;
Published online: 23 October 2023

REFERENCES

1. Cheng, P., Li, G., Zhan, X. & Yang, Y. Next-generation organic photovoltaics based on non-fullerene acceptors. *Nat. Photonics* **12**, 131–142 (2018).
2. Wan, X., Li, C., Zhang, M. & Chen, Y. Acceptor–donor–acceptor type molecules for high performance organic photovoltaics–chemistry and mechanism. *Chem. Soc. Rev.* **49**, 2828–2842 (2020).
3. Kini, G. P., Jeon, S. J. & Moon, D. K. Design principles and synergistic effects of chlorination on a conjugated backbone for efficient organic photovoltaics: a critical review. *Adv. Mater.* **32**, e1906175 (2020).
4. Jain, A., Shin, Y. & Persson, K. A. Computational predictions of energy materials using density functional theory. *Nat. Rev. Mater.* **1**, 1–13 (2016).
5. Cui, Y., Zhu, P., Liao, X. & Chen, Y. Recent advances of computational chemistry in organic solar cells. *J. Mater. Chem. C* **8**, 15920-15939 (2020).
6. Mahmood, A., Irfan, A. & Wang, J.-L. Molecular level understanding of the chalcogen atom effect on chalcogen-based polymers through electrostatic potential, non-covalent interactions, excited state behaviour, and radial distribution function. *Polym. Chem.* **13**, 5993–6001 (2022).
7. Russell, S. & Norvig, P. *Artificial Intelligence: A Modern Approach*. (London, 2010).
8. Nilsson, N. J. *Principles of Artificial Intelligence*. (Morgan Kaufmann, 2014).
9. Minsky, M. Steps toward artificial intelligence. *Proc. IRE* **49**, 8–30 (1961).
10. Sun, W. et al. Artificial intelligence designer for highly-efficient organic photovoltaic materials. *J. Phys. Chem. Lett.* **12**, 8847–8854 (2021).
11. Wu, Y., Guo, J., Sun, R. & Min, J. Machine learning for accelerating the discovery of high-performance donor/acceptor pairs in non-fullerene organic solar cells. *npj Comput. Mater.* **6**, 120 (2020).
12. Wang, H., Ji, Y. & Li, Y. Simulation and design of energy materials accelerated by machine learning. *Wiley Interdiscip. Rev.* **10**, e1421 (2020).
13. Chen, C. et al. A critical review of machine learning of energy materials. *Adv. Energy Mater.* **10**, 1903242 (2020).
14. Schmidt, J., Marques, M. R., Botti, S. & Marques, M. A. Recent advances and applications of machine learning in solid-state materials science. *npj Comput. Mater.* **5**, 1–36 (2019).
15. Carrasquilla, J. & Melko, R. G. Machine learning phases of matter. *Nat. Phys.* **13**, 431–434 (2017).
16. Ma, S. & Liu, Z.-P. Machine learning for atomic simulation and activity prediction in heterogeneous catalysis: current status and future. *ACS Catal.* **10**, 13213–13226 (2020).
17. Butler, K. T. et al. Machine learning for molecular and materials science. *Nature* **559**, 547–555 (2018).
18. Mahmood, A., Irfan, A. & Wang, J.-L. 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 (2022).
19. Zhang, Q. et al. High-efficiency non-fullerene acceptors developed by machine learning and quantum chemistry. *Adv. Sci.* **9**, 2104742 (2022).
20. Mahmood, A. & Wang, J.-L. Machine learning for high performance organic solar cells: current scenario and future prospects. *Energy Environ. Sci.* **14**, 90–105 (2021).
21. Feng, J., Wang, H., Ji, Y. & Li, Y. Molecular design and performance improvement in organic solar cells guided by high-throughput screening and machine learning. *Nano Sel.* **2**, 1629–1641 (2021).
22. Saeki, A. & Kranthiraja, K. A high throughput molecular screening for organic electronics via machine learning: present status and perspective. *Jpn. J. Appl. Phys.* **59**, SD0801 (2019).

23. Mahmood, A., Sandali, Y. & Wang, J.-L. Easy and fast prediction of green solvents for small molecule donor-based organic solar cells through machine learning. *Phys. Chem. Chem. Phys.* **25**, 10417–10426 (2023).
24. Mahmood, A., Irfan, A. & Wang, J.-L. Machine learning for organic photovoltaic polymers: a minireview. *Chin. J. Polym. Sci.* **40**, 870–876 (2022).
25. Sun, W. et al. The use of deep learning to fast evaluate organic photovoltaic materials. *Adv. Theor. Simul.* **2**, 1800116 (2019).
26. Scharber, M. C. et al. Design rules for donors in bulk-heterojunction solar cells—towards 10% energy-conversion efficiency. *Adv. Mater.* **18**, 789–794 (2006).
27. Hachmann, J. et al. Lead candidates for high-performance organic photovoltaics from high-throughput quantum chemistry–the Harvard Clean Energy Project. *Energy Environ. Sci.* **7**, 698–704 (2014).
28. Hachmann, J. 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.* **2**, 2241–2251 (2011).
29. Lopez, S. A., Sanchez-Lengeling, B., de Goes Soares, J. & Aspuru-Guzik, A. Design principles and top non-fullerene acceptor candidates for organic photovoltaics. *Joule* **1**, 857–870 (2017).
30. Padula, D., Simpson, J. D. & Troisi, A. Combining electronic and structural features in machine learning models to predict organic solar cells properties. *Mater. Horiz.* **6**, 343–349 (2019).
31. Sun, W. et al. Machine learning–assisted molecular design and efficiency prediction for high-performance organic photovoltaic materials. *Sci. Adv.* **5**, eaay4275 (2019).
32. Sahu, H. & Ma, H. Unraveling correlations between molecular properties and device parameters of organic solar cells using machine learning. *J. Phys. Chem. Lett.* **10**, 7277–7284 (2019).
33. Sahu, H., Rao, W., Troisi, A. & Ma, H. Toward predicting efficiency of organic solar cells via machine learning and improved descriptors. *Adv. Energy Mater.* **8**, 1801032 (2018).
34. Dong, Z., Feng, J., Ji, Y. & Li, Y. SLI-GNN: a self-learning-input graph neural network for predicting crystal and molecular properties. *J. Phys. Chem. A* **127**, 5921–5929 (2023).
35. Lopez, S. A. et al. The Harvard organic photovoltaic dataset. *Sci. Data* **3**, 160086 (2016).
36. Zhao, Z.-W., del Cueto, M., Geng, Y. & Troisi, A. Effect of increasing the descriptor set on machine learning prediction of small molecule-based organic solar cells. *Chem. Mater.* **32**, 7777–7787 (2020).
37. Padula, D. & Troisi, A. Concurrent optimization of organic donor–acceptor pairs through machine learning. *Adv. Energy Mater.* **9**, 1902463 (2019).
38. Vehtari, A., Gelman, A. & Gabry, J. Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Stat. Comput.* **27**, 1413–1432 (2017).
39. Li, H., Bredas, J. L. & Lennartz, C. First-principles theoretical investigation of the electronic couplings in single crystals of phenanthroline-based organic semiconductors. *J. Chem. Phys.* **126**, 164704 (2007).
40. Kuzmich, A., Padula, D., Ma, H. & Troisi, A. Trends in the electronic and geometric structure of non-fullerene based acceptors for organic solar cells. *Energy Environ. Sci.* **10**, 395–401 (2017).
41. Schwarz, K. N. et al. Suppressing subnanosecond bimolecular charge recombination in a high-performance organic photovoltaic material. *J. Phys. Chem. C* **120**, 24002–24010 (2016).
42. Zhang, J., Zhu, L. & Wei, Z. Toward over 15% power conversion efficiency for organic solar cells: current status and perspectives. *Small Methods* **1**, 1700258 (2017).
43. Mühlbacher, D. et al. High photovoltaic performance of a low-bandgap polymer. *Adv. Mater.* **18**, 2884–2889 (2006).
44. Murphy, A. R. & Frechet, J. M. Organic semiconducting oligomers for use in thin film transistors. *Chem. Rev.* **107**, 1066–1096 (2007).
45. Oberhofer, H., Reuter, K. & Blumberger, J. Charge transport in molecular materials: an assessment of computational methods. *Chem. Rev.* **117**, 10319–10357 (2017).
46. Paul, A. et al. Transfer learning usingensemble neural networks for organic solar cell screening. In *2019 International Joint Conference on Neural Networks* (IJCNN). (pp. 1–8). (IEEE, 2019).
47. Pan, S. J. & Yang, Q. A survey on transfer learning. *IEEE Trans. Knowl. Data Eng.* **22**, 1345–1359 (2009).
48. Wang, X. et al. Electric dipole descriptor for machine learning prediction of catalyst surface–molecular adsorbate interactions. *J. Am. Chem. Soc.* **142**, 7737–7743 (2020).
49. Fushiki, T. Estimation of prediction error by using K-fold cross-validation. *Stat. Comput.* **21**, 137–146 (2011).
50. Hou, J., Inganas, O., Friend, R. H. & Gao, F. Organic solar cells based on non-fullerene acceptors. *Nat. Mater.* **17**, 119–128 (2018).
51. Mahmood, A., Irfan, A. & Wang, J. L. Developing efficient small molecule acceptors with sp2-hybridized nitrogen at different positions by density functional theory calculations, molecular dynamics simulations and machine learning. *Chemistry* **28**, e202103712 (2022).
52. Wang, L. et al. Non-fullerene acceptors with hetero-dihalogenated terminals induce significant difference in single crystallography and enable binary organic solar cells with 17.5% efficiency. *Energy Environ. Sci.* **15**, 320–333 (2022).
53. Zhao, X. et al. Double asymmetric core optimizes crystal packing to enable selenophene-based Acceptor with over 18 % efficiency in binary organic solar cells. *Angew. Chem. Int. Ed. Engl.* **62**, e202216340 (2023).
54. Yan, L. et al. Regioisomer-free difluoro-monochloro terminal-based hexa-halogenated acceptor with optimized crystal packing for efficient binary organic solar cells. *Angew. Chem.* **134**, e202209454 (2022).
55. Yang, C. et al. A synergistic strategy of manipulating the number of selenophene units and dissymmetric central core of small molecular acceptors enables polymer solar cells with 17.5 % efficiency. *Angew. Chem. Int. Ed. Engl.* **60**, 19241–19252 (2021).
56. Sun, Y. et al. Solution-processed small-molecule solar cells with 6.7% efficiency. *Nat. Mater.* **11**, 44–48 (2011).
57. Zhang, Q. et al. Small-molecule solar cells with efficiency over 9%. *Nat. Photonics* **9**, 35–41 (2014).
58. Guo, J. et al. 15.71% Efficiency all-small-molecule organic solar cells based on low-cost synthesized donor molecules. *Adv. Funct. Mater.* **32**, 2110159 (2021).
59. Becke, A. D. Density-functional exchange-energy approximation with correct asymptotic behavior. *Phys. Rev. A Gen. Phys.* **38**, 3098–3100 (1988).
60. Perdew, J. P. Density-functional approximation for the correlation energy of the inhomogeneous electron gas. *Phys. Rev. B* **33**, 8822 (1986).
61. Weigend, F. & Ahlrichs, R. Balanced basis sets of split valence, triple zeta valence and quadruple zeta valence quality for H to Rn: Design and assessment of accuracy. *Phys. Chem. Chem. Phys.* **7**, 3297–3305 (2005).
62. Beck, A. D. Density-functional thermochemistry. III. The role of exact exchange. *J. Chem. Phys.* **98**, 5648–5646 (1993).
63. Zhao, Y. & Truhlar, D. G. 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. Acc.* **120**, 215–241 (2008).
64. Zhao, Y. & Truhlar, D. G. Density functionals for noncovalent interaction energies of biological importance. *J. Chem. Theory Comput.* **3**, 289–300 (2007).
65. Frisch, M. et al. *Gaussian 09 Revision A. 02, 2009*. (Gaussian Inc., Wallingford CT, 2009).
66. Pedregosa, F. et al. Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
67. Cherkassky, V. & Ma, Y. Practical selection of SVM parameters and noise estimation for SVM regression. *Neural Netw.* **17**, 113–126 (2004).
68. Hearst, M. A. et al. Support vector machines. *IEEE Intell. Syst. Appl.* **13**, 18–28 (1998).
69. Myles, A. J. et al. An introduction to decision tree modeling. *J. Chemometr.* **18**, 275–285 (2004).
70. Svetnik, V. et al. Random forest: a classification and regression tool for compound classification and QSAR modeling. *J. Chem. Inf. Comput Sci.* **43**, 1947–1958 (2003).
71. Breiman, L. Random forests. *Mach. Learn.* **45**, 5–32 (2001).
72. Friedman, J. H. Stochastic gradient boosting. *Comput. Stat. Data Anal.* **38**, 367–378 (2002).
73. Ke, G. et al. Lightgbm: a highly efficient gradient boosting decision tree. *Adv. Neural Inf. Process. Syst.* **30**, 3146–3154 (2017).

ACKNOWLEDGEMENTS

This work was supported by the National Key Research Program of China (grant No. 2022YFA1503101), Science and Technology Project of Jiangsu Province (grant No. BZ2020011), National Natural Science Foundation of China (grant No. 22173067), Science and Technology Development Fund, Macau SAR (FDCT No. 0052/2021/A), Collaborative Innovation Center of Suzhou Nano Science & Technology, Priority Academic Program Development of Jiangsu Higher Education Institutions (PAPD), 111 Project, and Joint International Research Laboratory of Carbon-Based Functional Materials and Devices.

AUTHOR CONTRIBUTIONS

Y.L. and H.W. conceived the idea and initiated this project. H.W. and J.F. collected and built the datasets, H.W., J.F., Z.D., and L.J. constructed and optimized the deep learning, SVM, RF, and LightGBM models, respectively. J.Y. and M.L. designed and synthesized the OSCs materials and fabricated OSCs devices. H.S. wrote the paper. Y.L., L.J., J.F., and J.Y. contributed to the fruitful discussions and supervision of the project. All authors discussed the results and commented on the paper.

COMPETING INTERESTS

The authors declare no competing interests.

ADDITIONAL INFORMATION

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41524-023-01155-9.

Correspondence and requests for materials should be addressed to Jianyu Yuan or Youyong Li.

Reprints and permission information is available at http://www.nature.com/reprints

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