ACS OMEGA

Open Access

http://pubs.acs.org/journal/acsodf

Article

Interpretable AI and Machine Learning Classification for Identifying High-Efficiency Donor−Acceptor Pairs in Organic Solar Cells

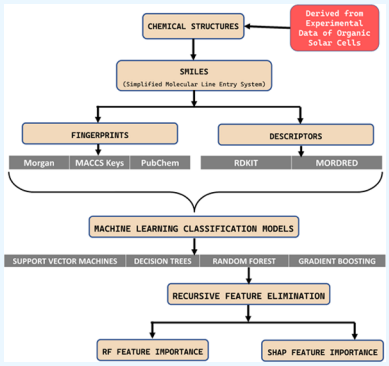
Hamza Siddiqui* and Tahsin Usmani

Cite This: *ACS Omega* 2024, 9, 34445−34455

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: To enhance the efficiency of organic solar cells, accurately predicting the efficiency of new pairs of donor and acceptor materials is crucial. Presently, most machine learning studies rely on regression models, which often struggle to establish clear rules for distinguishing between high- and low-performing donor−acceptor pairs. This study proposes a novel approach by integrating interpretable AI, specifically using Shapely values, with four supervised machine learning classification models, namely, support vector machines, decision trees, random forest, and gradient boosting. These models aim to identify high-efficiency donor−acceptor pairs based solely on chemical structures and to extract important features that establish general design principles for distinguishing between high- and low-efficiency pairs. For validation purposes, an unsupervised machine learning algorithm utilizing loading vectors obtained from the principal component analysis is employed to identify crucial features associated with high-efficiency donor−acceptor pairs. Interestingly, the features identified by the supervised machine learning approach were found to be a subset of those identified by the unsupervised method. Noteworthy features include the van der Waals surface area, partial equalization of orbital electronegativity, Moreau−Broto autocorrelation, and molecular substructures. Leveraging these features, a backward-working model can be developed, facilitating exploration across a wide array of materials used in organic solar cells. This innovative approach will help navigate the vast chemical compound space of donor and acceptor materials essential in creating high-efficiency organic solar cells.



1. INTRODUCTION

Competition and investment in next-generation solar cells, specifically organic solar cells, are driven by the dual prospects of social benefit and financial rewards in the solar energy market.[1] They offer lightweight and flexible modules, reducing the need for expensive support structures, and can be applied to various surfaces, expanding their installation versatility. They can also be produced in continuous, high-volume processes through roll-to-roll production.[2] However, the key hurdles on the path to establishing organic photovoltaics (OPVs) as the leading clean energy generation technology of the future, surpassing Si cells, lie in the imperative task of elevating their overall efficiency and operational lifespan. Additionally, there is a crucial need to enhance the manufacturing yield and scalability of OPV, ultimately paving the way for it to become the finest clean energy resource crafted by humanity.

Due to the vast and complex nature of the organic compound landscape, employing density functional theory (DFT) and time dependent-DFT (TD-DFT) methods to explore it can be arduous and impractical.[3] Similarly, experimental approaches face significant challenges and require substantial financial resources due to the large chemical compound space. Therefore, in the pursuit of identifying optimal structures for high-efficiency organic solar cells, the most viable approach is to employ large-scale data-driven machine-learning methods.[4] Machine learning (ML) methods provide a more efficient and cost-effective approach by enabling the investigation of material properties and their correlations.[5] With the increasing availability of large training data sets, advanced algorithms, and increasing processing power, ML has revolutionized materials research.[6] For organic semiconductors, ML models have been successful in predicting photovoltaic parameters, such as power conversion efficiency (PCE),[7] and facilitating tasks like quantitative structure−property relationship (QSPR) analysis,[8] design of experiments,[9] and discovery of novel materials.[10,11]

2. BACKGROUND

In 2006, Scharber et al.[12] demonstrated that the efficiency of bulk heterojunction devices utilizing PCBM as the acceptor is contingent upon the lowest-unoccupied molecular orbital (LUMO) level and the band gap of the donor, thereby establishing a correlation between energy-conversion efficiency, band gap, and the donor's LUMO level.

Received: March 5, 2024
Revised: June 8, 2024
Accepted: June 13, 2024
Published: July 31, 2024

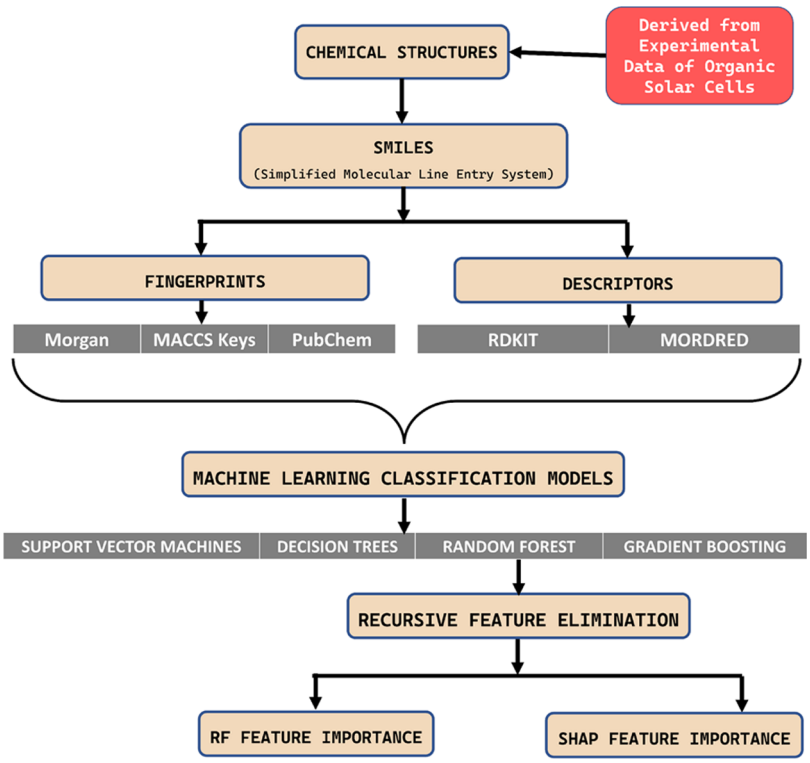


Figure 1. Workflow—machine learning classification model.

Subsequently, the Harvard Clean Energy Project (CEP)[13] utilized Scharber's model to predict the power conversion efficiencies (PCEs) of thousands of organic photoelectric molecules, with later improvements made using Gaussian process regression (GPR) in machine learning. However, the limitations of Scharber's model, such as low accuracy and time-consuming calculations, hindered its suitability for fast and accurate high-throughput screening.[14]

Apart from this, only a limited number of studies have focused on machine learning-based classification for material screening in organic solar cells.[11] Lopez et al.[15] explored the drawbacks associated with fullerene-containing OPVs and proposed an innovative approach that utilized density functional theory and Gaussian process calibration. Nagasawa et al.[16] utilized supervised learning techniques, particularly random forest (RF) screening, to facilitate the design, synthesis, and characterization of conjugated polymers. In a similar vein, Peng[17] et al. introduced the application of convolutional neural networks (CNNs) to generate and predict the properties of nonfullerene acceptors in organic solar cells. Chen's[18] research focused on the virtual screening of semiconductor polymers, specifically targeting high-performance OPV devices. Their study employed machine learning algorithms, including support vector machine (SVM) and ensemble learning, to achieve accurate predictions. Sun et al.[19] developed a comprehensive donor material database for OPVs and utilized machine learning models to establish vital structure−property relationships and screen potential new materials. In another study, Sun et al.[20] proposed a deep learning model based on a deep neural network (ResNet) that directly predicted the photovoltaic performance of diverse OPV donor materials solely based on their chemical structures. Mahmood et al.[21] harnessed the power of machine learning to predict the performance of P3HT-based organic solar cells and successfully select environmentally friendly solvents based on predicted 'Hansen solubility' parameters. Moore et al.[22] addressed the challenge of limited experimental data sets for predicting energy levels of donor molecules in OPVs by implementing transfer learning techniques in conjunction with convolutional neural networks (CNNs). Their novel approach achieved impressive accuracy, with error margins below 200 meV, and demonstrated the practical applicability of their model using commercially available donor polymers. The aforementioned studies primarily concentrated on classification tasks without delving into the underlying factors influencing the categorization of donors or acceptors into high- or low-efficiency groups. In our investigation, we extended this inquiry by incorporating descriptors from both donors and acceptors. By doing so, we aim to discern the specific features contributing to the high efficiency in both donor and acceptor entities. Notably, previous research predominantly emphasized the analysis of either donors or acceptors singularly.

3. MATERIALS AND METHODS

The objective of this research is to develop an innovative approach leveraging interpretable AI and supervised machine learning techniques, along with unsupervised methods, to accurately predict high-efficiency donor−acceptor pairs for organic solar cells and extract crucial features for general design principles in navigating the chemical compound space. This model utilizes readily available input data, specifically relevant chemical structures sourced from the literature, to classify donor−acceptor pairs. During the exhaustive literature review conducted to prepare our data set, in addition to relevant chemical structures, we have also collected important parameters such as E_g (optical band gap), highest occupied molecular orbital (HOMO), LUMO, V_{oc}, J_{sc}, FF, and device PCE.

https://doi.org/10.1021/acsomega.4c02157
ACS Omega 2024, 9, 34445−34455

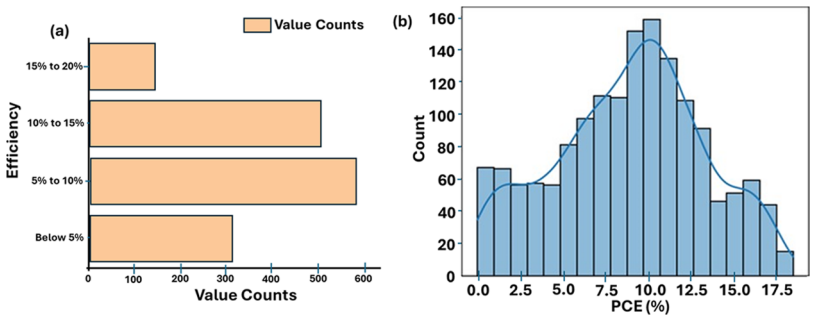


Figure 2. (a) Bar chart depicting the efficiency of donor/acceptor pairs. (b) Histogram of efficiencies of donor/acceptor pairs.

The workflow for developing the classification model is illustrated in Figure 1. In this study, we follow a multistep process. First, we create data sets that incorporate experimental data with material and photovoltaic characteristics from both fullerene and nonfullerene organic solar cells (OSCs). Second, chemical structures of all donor and acceptor materials were drawn on ChemDraw software, and their SMILES codes were generated. Using the SMILES code of distinct donor and acceptor materials, their molecular descriptors and molecular fingerprints were generated using various open-source libraries that are available freely for generating molecular descriptors and fingerprints.[23−27] We have used two types of descriptors (RDKit and Mordred) and three types of fingerprints (MACCS, PubChem, Morgan)[24] for generating data sets for both donors and acceptors using SMILES codes. This data is used for training purposes. Third, we have used four supervised ML approaches (Support Vector Machines, Decision Trees, Random Forest, and Gradient Boosting), which are accessible from the Scikit-learn Python package,[28] for the classification model based on performance metrics, particularly accuracy. Since the interpretability of the selected features is crucial in our case, we further obtain important features from the model through feature selection and interpret the selected classification machine learning models using recursive feature elimination and SHapley Additive exPlanations (SHAP). Finally, through principal component analysis, we obtain the loadings of the original features in each principal component to identify features that contribute most to the creation of each principal component. Features with higher loadings have a more substantial influence on the principal component. Therefore, these loadings are also used to infer the importance of the original features and, thus, validate our findings from SHAP analysis.

3.1. Data Preparation. The original database, obtained from a paper by Saeki et al.,[29] exhibited a right-skewed similarity distribution of PCE. In this study, the database is modified by including devices from literature reviews spanning the past 5 years encompassing various donor/acceptor pairings that show the highest achieved efficiency.[30,31] All of the extra references are mentioned in Supporting Information S7. This expanded database includes both low- and high-efficiency structures, providing a more comprehensive representation of the molecular landscape. The data set at hand comprises 1597 donor/acceptor pairs. These pairs are accompanied by additional details, such as V_{oc}, J_{sc}, FF, and PCE, as well as the HOMO and LUMO levels of both the donor and acceptor and the optical band gaps of the donors and acceptors. Upon analyzing the data set, it becomes evident from the bar charts that most of the compounds fall within the range of 10 to 15 and 5 to 10% in terms of their efficiency. Furthermore, the histogram displays a normal distribution pattern across the entire efficiency range, indicating that the distribution of the efficiency values follows a normal distribution with respect to their counts. This observation suggests that the data set is well-suited for the application of machine learning algorithms.

The distribution of % PCE is shown in Figure 2a,b. The calculated mean and median of % PCE are 8.85 and 9.2, respectively. Further details about the data set are mentioned in Supporting Information S1.

3.2. Molecular Representation and Feature Extraction. Out of the 1597 pairs containing D/A combination, 1241 are unique with 320 distinct donors and 736 distinct acceptors. Chemical structures of all donor and acceptor materials were drawn on ChemDraw software, and their SMILES codes were generated. Using the SMILES codes of distinct donor and acceptor materials, their molecular descriptors and molecular fingerprints were generated.

3.2.1. Descriptors. Molecular properties were computationally calculated by using two Python libraries: RDKit and Mordred. RDKit provided 208 descriptors, while Mordred Library[26] offered a more extensive set of 1613 descriptors. All computations were performed using open-source Python packages.[28]

3.2.2. Fingerprints. The other descriptors used in the study are molecular fingerprints. A molecular fingerprint is a binary array that represents specific structural characteristics of a chemical compound. Each bit in the fingerprint corresponds to a predefined structural feature. When the feature is present in the molecule, the bit is set to 1 (ON), and when it is absent, the bit is set to 0 (OFF). The number of bits in the fingerprint determines the amount of structural information captured.

In this study, three types of fingerprints were used:

1. Molecular ACCess System (MACCS) key (166 bits):[32] The MACCS key fingerprint consists of 166 bits and represents specific chemical substructures in the compound.
2. PubChem fingerprints (881 bits):[33] These fingerprints contain 881 bits and capture various chemical features present in the molecule, as defined by the PubChem database.
3. Morgan fingerprint (1024 bits):[34] The Morgan fingerprint, also known as circular fingerprint or extended connectivity fingerprint, is a type of structural fingerprint with 1024 bits. It encodes information about molecular substructures, including circular patterns.

Each type of fingerprint provides a unique representation of the molecular structure, and the choice of fingerprint depends on the specific analysis and applications in the study. In our

https://doi.org/10.1021/acsomega.4c02157
ACS Omega 2024, 9, 34445−34455

study, the fingerprints of all of the donors and acceptor molecules are calculated by the ChemDes Website.[24]

Hence, by utilizing molecular descriptors and fingerprints, five different data sets have been prepared, which are further used as training data sets for training the classification models and feature extraction through feature engineering.

3.3. Input Data Preparation for the ML Model. To prepare the data after descriptor calculation, missing values were replaced with 0. Subsequently, only columns with numeric data types were selected for further analysis. Additionally, numeric values exceeding the maximum representable value for the input data type were filtered out to ensure data integrity for subsequent data analysis or machine learning tasks. Furthermore, redundant features were identified by analyzing correlations among descriptors. Features with a correlation coefficient greater than 0.8 were removed from the data set as they added little unique information and could potentially introduce noise or multicollinearity during analysis.

3.4. Model Building with Machine Learning Approaches. We have used four supervised ML approaches, SVM, decision trees, random forest, and gradient boosting, which are accessible from the Scikit-learn Python package.[28] Tuning hyperparameter searches were conducted through 3-fold stratified cross-validation on training data for better model performances.

3.4.1. Support Vector Machine Classifier. In our problem, the kernel function used is the radial basis function (RBF) with $C = 1.0$. More details about all machine learning classification algorithms used are mentioned in Supporting Information S2.

3.4.2. Decision Tree Classifier. In our problem, Gini impurity or entropy is used as the criteria for splitting nodes.

3.4.3. Random Forest Classifier. For our problem, the number of estimators is set to 1400, a minimum of 10 samples is required to split an internal node, a minimum of 2 samples is required to be at a leaf node, and maximum features are determined by the square root of the total number of features, a maximum depth of 80 for each tree, and bootstrap samples are utilized during the tree-building process.

3.4.4. Gradient Boosting Classifier. In our problem, the number of boosting stages (n-estimators) was 100, the maximum depth of the individual decision trees (max-depth) was 3, the learning rate was, and the loss function was 'log_loss'.

3.5. Performance Evaluation of the ML Model. The initial phase involved employing the default parameters for all machine learning models. Given the substantial size of our data set (comprising 1558 rows), we opted to employ the hold-one-out cross-validation (HOOCV) technique to evaluate the accuracy of the classification model. Subsequently, we conducted hyperparameter tuning for the random forest model using RandomizedSearchCV within the Python programming environment. Specific details about the statistical metrics used are given in Supporting Information S3.

3.6. Interpretable Selected Feature Importance through SHAP. The importance of a descriptor is estimated by recording a reduction of mean square error for each feature when data is passed through an ensemble and averaging it over all of the ensemble.[35] By isolating these influential features through hyperparameter tuning[36] with 5-fold hold-one-out cross-validation[37] and subsequent recursive feature elimination,[38] we gain insight into the features that are the most influential for improving the accuracy of the classification model.

For the ML model explanation, we employed SHapley Additive exPlanations (SHAP). This visualization technique enabled a comprehensive and interpretable exploration of the impact each descriptor had on the model predictions. By presenting a clear depiction of how individual descriptors contribute to the overall classification process, the SHAP visualization added a layer of depth to our analysis, thereby illuminating critical insights for the advancement of high-efficiency bulk heterojunction organic solar cells.

3.7. Feature Engineering through Principal Component Analysis. The preprocessing steps entail segregating the data set based on efficiency values, generating canonical SMILES representations, imputing missing values, and standardizing data types. Subsequent feature selection involves eliminating low-variance features and highly correlated ones to enhance data set robustness. Principal component analysis (PCA) is then applied, visualizing the cumulative explained variance, determining principal component numbers, and exploring chemical space. Feature importance is evaluated through normalized principal components and the absolute sum of coefficients. Specific details are mentioned in Supporting Information S6−S8.

4. RESULTS AND DISCUSSION

4.1. Machine Learning Classification Results. *4.1.1. RDKIT Database.* The classification performance of four machine learning algorithms, support vector machines (SVM), decision trees, random forest, and gradient boosting, applied to a data set comprising 1558 samples with 101 RDKIT descriptors is evaluated. The classifiers are evaluated using hold-one-out cross-validation, and performance metrics such as accuracy, confusion matrices, Cohen's Kappa score, and Matthews correlation coefficient are reported as shown in Table 1.

Table 1. Performance Metrics for ML Classification Models on the RDKIT Database

Algorithm	Accuracy (%)	Cohen's kappa (%)	Matthews correlation (%)
Support vector machines	69.87	51.28	54.39
Decision trees	83.97	69.23	69.25
Random forest	87.82	73.08	73.13
Gradient boosting	85.90	71.79	71.89

Like the application on the RDKIT database, the classification model is applied to other data sets also. The findings revealed that certain algorithms exhibited superior predictive performance for specific data sets. For instance, gradient boosting demonstrated remarkable accuracy when employed with Mordred descriptors, while random forest excelled in predicting outcomes based on Morgan fingerprints. Moreover, the calculated performance metrics shed light on the strengths and weaknesses of each algorithm-feature combination, elucidating the potential for identifying robust design rules for high-efficiency bulk heterojunction organic solar cells. The performance evaluation matrix for a four machine learning model with six data sets in given Supporting Information S4. The summarized classification accuracy of all

https://doi.org/10.1021/acsomega.4c02157
ACS Omega 2024, 9, 34445−34455

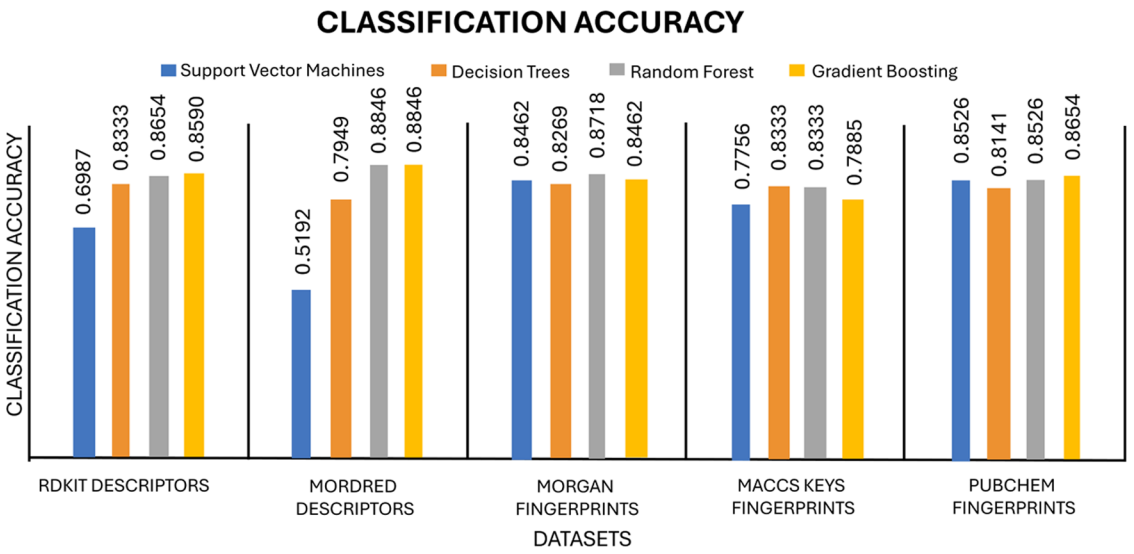


Figure 3. Summarized classification model accuracy.

four machine learning algorithms employed on five distinct data sets is shown in Figure 3.

Since the random forest classifier gave the highest accuracy, the hyperparameters representing the current state were found for the respective classification models and used for tuning.

4.2. Hyperparameter Tuning of the ML Model. For the hyperparameter tuning process, a randomized search was employed to explore a range of values for key parameters in the random forest model. The hyperparameter grid encompasses various configurations for critical aspects of the algorithm. The purpose is to systematically search through this parameter space and identify the combination that yields the optimal model performance. The hyperparameter grid consists of a number of trees (*n*_estimators): varying from 200 to 2000 in increments of 200; maximum features at the split (max_features): including 'auto' and 'sqrt'; maximum depth of trees (max_depth): ranging from 10 to 110 with intervals of 10, and including 'None' for unrestricted growth; minimum samples required to split a node (min_samples_split): examining 2, 5, and 10 samples; minimum samples required at each leaf node (min_samples_leaf): considering 1, 2, and 4 samples; and the bootstrap method (bootstrap): exploring both with and without replacement. This comprehensive exploration of hyperparameter values aims to enhance the Random Forest model's ability to generalize patterns in the data set. The randomized search is conducted to efficiently sample a diverse set of configurations, facilitating the identification of an optimal set of hyperparameters for improved predictive performance.

The hyperparameter tuning process, employing RandomizedSearchCV, resulted in the identification of optimal hyperparameters for the Random Forest model. The best configuration is as follows: the number of trees (*n*_estimators): 1400, minimum samples required to split a node (min_samples_split): 10, minimum samples required at each leaf node (min_samples_leaf): 2, maximum features at the split (max_features): 'sqrt', maximum depth of trees (max_depth): 80, and the bootstrap method (bootstrap): True. These hyperparameters are determined based on a randomized search that sampled 100 different combinations, utilizing a 3-fold cross-validation approach. The selection criterion was the negative mean absolute error (scoring = 'neg_mean_absolute_error').

The Random Forest classifier, configured with the optimized hyperparameters, demonstrated a commendable accuracy of 87.18% on the test data set. This signifies the model's ability to correctly classify instances into their respective classes, showcasing its efficacy in capturing the underlying patterns within the given data. The high accuracy obtained highlights the effectiveness of the hyperparameter tuning process in enhancing the model's predictive performance.

The above same procedure was applied across five distinct data sets, each representing diverse molecular descriptors, including Mordred descriptors, Morgan fingerprints, MACCS keys, and PubChem fingerprints. For each data set, a random forest classifier was employed, and a rigorous hyperparameter tuning process was conducted to optimize the model's performance. The optimized hyperparameters, tailored for each data set, were determined through a randomized search cross-validation approach. This involved exploring a predefined hyperparameter space to identify configurations that yield the optimal predictive accuracy. The key hyperparameters tuned included the number of trees (*n*_estimators), minimum samples required to split a node (min_samples_split), minimum samples required at each leaf node (min_samples_leaf), maximum features at the split (max_features), maximum depth of trees (max_depth), and the use of the bootstrap method (bootstrap). Upon model training and optimization, predictions were made on respective test data sets, and the accuracy of each model was computed. The achieved accuracies across the diverse data sets are reported in Table 2.

4.3. Feature Selection for Model Interpretability. In pursuit of enhancing model interpretability and identifying the most influential molecular descriptors, a rigorous feature selection process was undertaken. The random forest classifier, which was previously optimized for predictive accuracy, served as the foundation for this investigation.

Utilizing recursive feature elimination (RFE), a feature selection technique embedded in Scikit-learn, the model underwent iterative training, ranking the importance of each feature after each iteration. The optimal subset of features was determined based on the criteria of selecting the top 9 features.

https://doi.org/10.1021/acsomega.4c02157
ACS Omega 2024, 9, 34445−34455

Table 2. Achieved Accuracies Across Data Sets after Hyperparameter Tuning

	Accuracy before hyperparameter tuning for random forest classifier	Accuracy after hyperparameter tuning for random forest classifier
RDKIT descriptors	0.8590	0.8718
Mordred descriptors	0.8782	0.8782
Morgan fingerprint	0.8526	0.8590
MACCS keys	0.8205	0.8333
PubChem fingerprint	0.8462	0.8590

Subsequently, the identified features were extracted and evaluated for their significance. The selected features were determined through the get_support() method, which provided a Boolean mask indicating the chosen features. The column names corresponding to these features were then extracted for further analysis. These features serve not only as discriminative elements for the model but also as interpretable variables that hold chemical relevance.

In the pursuit of understanding the collective impact of the selected features, a crucial metric, termed "cumulative importance", was computed. The cumulative importance is computed by summing the importance scores of the selected features. This metric encapsulates the proportion of predictive power consolidated within a chosen subset of descriptors. The result unveils that the identified molecular descriptors from the RDKIT database collectively account for nearly 29% of the model's decision-making process. Understanding the maximum cumulative importance provides valuable insights into the collective information encapsulated by the chosen molecular descriptors. Similar contributions of the 9 most important features from other data sets are enlisted in Tables 3a, 3b, 3c, 3d, and 3e and Figure 4a–e. Table 4 gives the computed cumulative importance from the 9 most important features found out by recursive feature elimination.

4.4. SHAP Analysis. To pinpoint key descriptors that hold a significant influence within the random forest classification model, SHAP analysis was conducted. SHapley Additive exPlanations (SHAP) values originate from cooperative game theory and are adapted for machine learning model interpretability. The SHAP value for a specific feature quantifies the contribution of that feature to the difference between the actual prediction and the average prediction across all possible combinations of features. A more detailed theory about SHAP summary plots and SHAP waterfall plots is mentioned in Supporting Information S5.

4.4.1. Interpretation of the Summary Plot in SHAP Analysis. As shown in Figure 5a, the plot indicates that donor–acceptor pairs with high values for Acceptor_fr_bicyclic, Acceptor_S log P_VSA4, and Acceptor_VSA_Estate3 & low values for Donor_VSA_Estate5 & Acceptor_SMR_VSA7 are more likely to contribute toward high efficiency. In Figure 5b, the plot indicates that donor–acceptor pairs with high values for Acceptor_ATSC1s, Acceptor_AATSC3Z, and Acceptor_ATSC5s & low values for Donor_ATSC8dv are more likely to contribute toward high efficiency. Similarly, in Figure 5c, the SHAP dependence plot suggests that the presence of Acceptor_bit_96, Acceptor_bit_584, & Acceptor_bit_131 have a significant positive impact on the efficiency of donor–acceptor pairs, whereas the presence of Donor_bit_313, Acceptor_bit_64, & Acceptor_bit_879 have a negative impact on the efficiency of donor–acceptor pairs. As shown in Figure 5d, the SHAP dependence plot suggests that the presence of Acceptor_bit_87, Acceptor_bit_41, & Donor_bit_136 have a significant positive impact on the

Table 3a. Feature Importance Obtained by Recursive Feature Elimination for the RDKit Data Set

	Descriptor	Description
Donor	MinEStateIndex	returns the min tuple of EState indices for the molecule[39]
	VSA EState5	MOE-type descriptors using EState indices and surface area contributions
		VSA EState descriptor 5 ($5.74 \leq x < 6.00$)
	MolWt	the average molecular weight of the molecule
Acceptor	MaxEstateIndex	returns the min tuple of EState indices for the molecule[39]
	S log P VSA4	MOE-type descriptors using log P contributions and surface area contributions MOE log P VSA descriptor 4 ($0.00 \leq x < 0.10$)
	VSA EState3	MOE-type descriptors using EState indices and surface area contributions EState VSA descriptor 3 ($0.29 \leq x < 0.72$)
	fr bicyclic	no. of bicyclic rings
	allylic oxid	number of allylic oxidation sites excluding steroid dienone
	SMR VSA7	MOE-type descriptors using MR contributions and surface area contributions

Table 3b. Feature Importance Obtained by Recursive Feature Elimination for the Mordred Data Set

	Descriptor	Description
Donor	ATSC8dv	centered Moreau–Broto autocorrelation of lag 8 weighted by valence electrons
	BCUTZ-1h	the first highest eigenvalue of the Burden matrix weighted by atomic number
	ATSC6dv	centered Moreau–Broto autocorrelation of lag 6 weighted by valence electrons
	ATSC4s	centered Moreau–Broto autocorrelation of lag 4 weighted by intrinsic state
Acceptor	ATSC1s	centered Moreau–Broto autocorrelation of lag 1 weighted by intrinsic state
	ATSC5s	centered Moreau–Broto autocorrelation of lag 5 weighted by intrinsic state
	ATSC5Z	centered Moreau–Broto autocorrelation of lag 5 weighted by atomic number
	AATSC3Z	averaged and centered Moreau–Broto autocorrelation of lag 3 weighted by atomic number
	Xch-6d	6-ordered chi-chain weighted by sigma electrons

Table 3c. Feature Importance Obtained by Recursive Feature Elimination for the Morgan Fingerprints Data Set

	fingerprint (2048 bit)	description
donor	bit 40	fingerprints calculated by the algorithm in ref 34
	bit 313	
acceptor	bit 879	
	bit 79	
	bit 96	
	bit 64	
	bit 584	
	bit 31	
	bit 131	

https://doi.org/10.1021/acsomega.4c02157
ACS Omega 2024, 9, 34445–34455

Table 3d. Feature Importance Obtained by Recursive Feature Elimination for the MACCS Keys Fingerprints Data Set

	fingerprint (167 bit)	smarts[40,41,42]	description
donor	bit 136	(“[#8]=*”,1)	a double bond between an oxygen atom and any other atom
	bit 145	(“*1*****1”,1)	a chain of atoms linked by single bonds, where the atoms in the chain have the same atomic number
	bit 42	(“F”,0)	a substructure where there is a fluorine atom
	bit 127	(“*@*!@[#8]”,1)	a substructure where there is a nonring atom connected to another nonring atom by a nonring bond, and one of the nonring atoms is oxygen, and this pattern should occur exactly once
	bit 152	(“[#8][#6]([#6])[#6]”,0)	a substructure where there is a single bond between an oxygen atom and three carbon atoms, where the second carbon atom is part of a ring
	bit 109	(“*[CH_2][#8]”,0)	a substructure where there is a single bond between any atom, a methylene group, and an oxygen atom
acceptor	bit 41	(“[#6]#[#7]”,0)	a substructure where there is a triple bond between a carbon atom and a nitrogen atom
	bit 87	(“[F,Cl,Br,I]!@*@*”,0)	a substructure where there is a nonring atom, not connected to a ring, followed by any two atoms, and the nonring atom is one of F, Cl, Br, or I
	bit 158	(“[#6]-[#7]”,0)	a substructure where there is a single bond between a carbon atom and a nitrogen atom

Table 3e. Feature Importance Obtained by Recursive Feature Elimination for the PubChem Fingerprints Data Set

	fingerprint (887 bit)		description
donor	bit 36	≥8 S	no. of sulfur atoms more than 8
	bit 185	≥2 any ring size 6	no. of 6 membered rings more than 2
acceptor	bit 341	C(C)(C)(C)O	a substructure where there is a carbon atom, followed by a ring containing two additional carbon atoms and an oxygen atom
	bit 439	C(−C)(−N)(=O)	a substructure where there is a chain of three atoms: carbon, nitrogen, and oxygen, connected in the specified order
	bit 23	≥1 F	no. of fluoride atoms more than 1
	bit 174	≥5 saturated or aromatic heteroatom-containing ring size 5	no. of saturated or aromatic heteroatom-containing ring size 5 more than 5
	bit 144	≥1 saturated or aromatic carbon-only ring size 5	no. of saturated or aromatic carbon-only ring size 5 more than 1

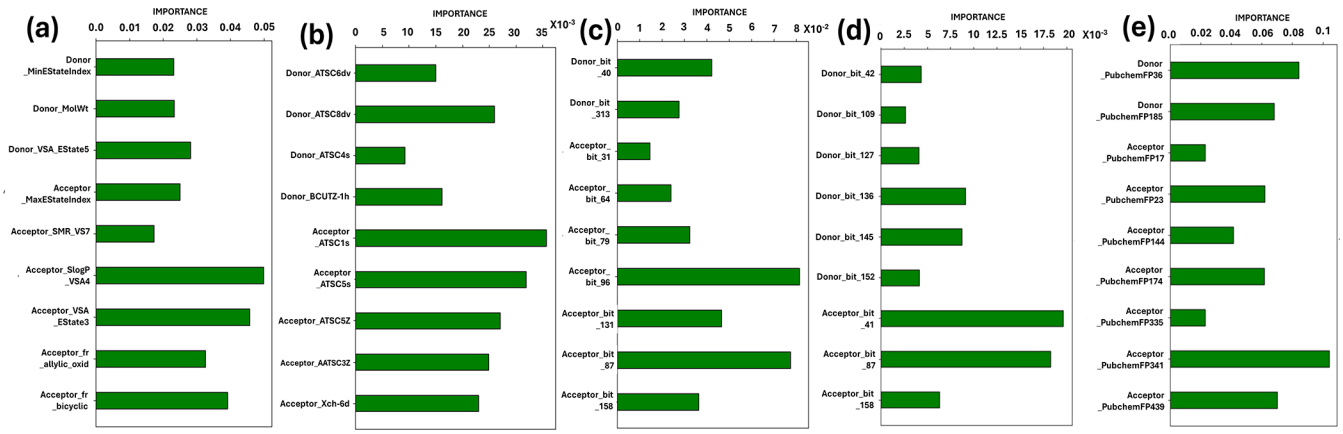


Figure 4. Feature importance obtained by recursive feature elimination for (a) RDKIT descriptor, (b) Mordred descriptor, (c) Morgan fingerprints, (d) MACCS keys fingerprints, and (e) PubChem fingerprints data sets.

Table 4. Cumulative Importance Obtained from the Nine Most Important Features

data set description	cumulative importance obtained from the nine most important features
RDKit descriptor	0.29
Mordred descriptor	0.21
Morgan fingerprints	0.38
MACCS keys fingerprints	0.77
PubChem fingerprints	0.52

efficiency of donor−acceptor pairs, whereas the presence of Donor_bit_109 has a negative impact on the efficiency of donor−acceptor pairs. In Figure 5e, the SHAP dependence plot suggests that the presence of Acceptor_PubChemFP_341 & Acceptor_PubChemFP_144 have a significant positive impact on the efficiency of donor−acceptor pairs, whereas the presence of Acceptor_PubChemFP_439 has a negative impact on the efficiency of donor−acceptor pairs.

Overall, the above SHAP Plots summarize the contribution of feature values to the efficiency of a given donor−acceptor pair and quantify the impact of respective feature values on efficiency.

4.4.2. Interpretation of Waterfall Plot in SHAP Analysis. Figure 6a−e given below shows the SHAP waterfall plots for respective data sets. In Figure 6a, the SHAP waterfall plot shows that Acceptor_fr_bicyclic is contributing significantly to the other listed feature values in the plot. Similarly, in Figure 6b, Acceptor_Xch6d and Donor_ATSC8dv are found to be

https://doi.org/10.1021/acsomega.4c02157
ACS Omega 2024, 9, 34445−34455

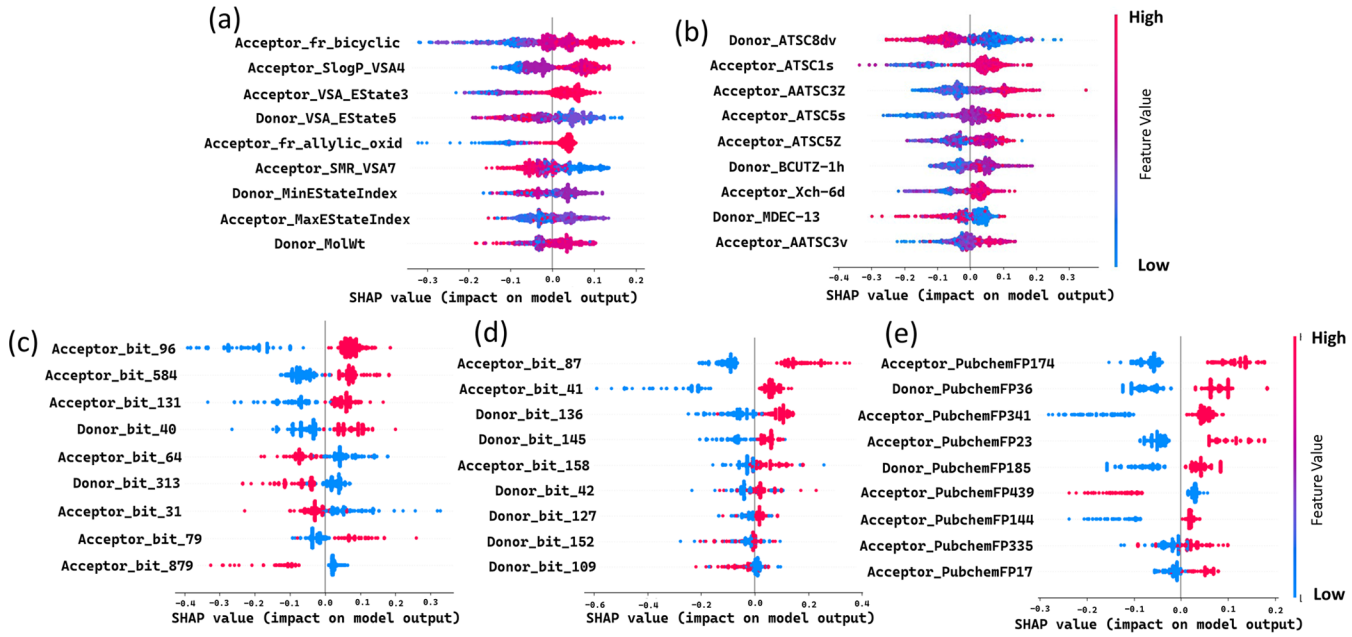


Figure 5. Feature importance summary plot obtained by SHAP analysis for the (a) RDKIT descriptor, (b) Mordred descriptor, (c) Morgan fingerprints, (d) MACCS keys fingerprints, and (e) PubChem fingerprints data sets.

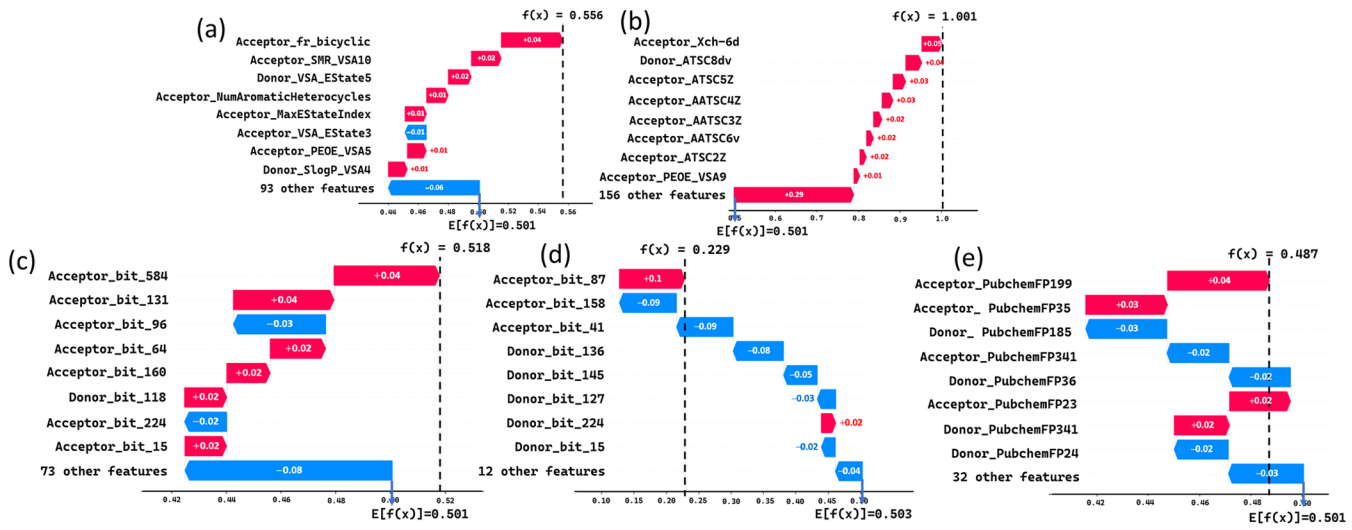


Figure 6. Feature importance waterfall plot obtained by SHAP analysis for the (a) RDKIT descriptor, (b) Mordred descriptor, (c) Morgan fingerprints, (d) MACCS keys fingerprints, and (e) PubChem fingerprints data sets.

more impactful. There is a change in the scenario in Figure 6c, where the presence of Acceptor_bit_584 and Acceptor_bit_131 has contributed cumulatively to the high-efficiency donor−acceptor pairs; in contrast, Acceptor_bit_96 corresponds to a negative contribution to the expected output. In Figure 6d, it can be observed that the presence of Acceptor_bit_87 and Donor_bit_42 outweighs all of the negative factors. In Figure 6e, Acceptor_PubChemFP_199, Acceptor_ PubChemFP_35, Acceptor_ PubChemFP_23, and Donor_ PubChemFP_341 are positive contributors toward the expected model output, $E(f(x))$.

4.5. Statistical Analysis for Validation. Principal component analysis (PCA) is a powerful technique commonly employed in data analysis to reduce the dimensionality of data sets while retaining the most significant information. In this study, we applied PCA to identify a reduced set of features that capture a substantial portion of the data set's variability.

In PCA, the importance of each feature is reflected by the magnitude of the corresponding values in the eigenvectors.[43] Often, when the data are centered and standardized, the coefficients are normalized so that the sum of the squares of the coefficients of a component is equal to the variance of the component. In this normalization, the coefficients can be interpreted as the correlation between the original variable and the principal component and are often called loadings.

For feature selection, variables are selected according to the magnitude (from largest to smallest in absolute values) of their coefficients (loadings), which are linear combinations of the original variables that make up the principal component. Absolute values near zero indicate that a variable contributes

https://doi.org/10.1021/acsomega.4c02157
ACS Omega 2024, 9, 34445−34455

Table 5. Variance from Respective Data sets after Principal Components Analysis

Data set	no. of principal components, which explain 95% variance	no. of principal components, which explain 67% variance	no. of principal components, which explain 50% variance	variance explained by the first two components (%)
RDKIT descriptors	41	13	7	22.60
Mordred descriptors	54	15	8	20.50
Morgan fingerprints	34	9	5	31.10
MACCS Keys fingerprints	12	5	3	43.10
PubChem fingerprints	18	6	3	40.70

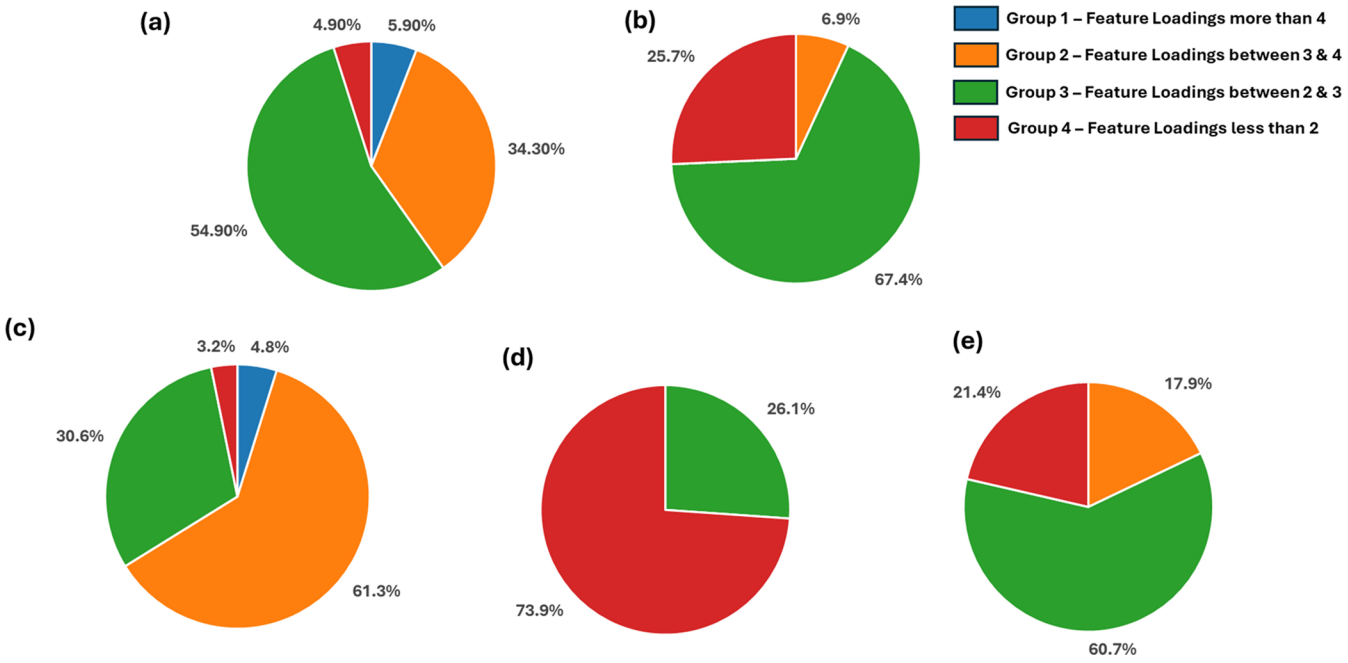


Figure 7. Distribution of features based on feature loadings or absolute sum of coefficients for the (a) RDKIT descriptor, (b) Mordred descriptor, (c) Morgan fingerprints, (d) MACCS keys fingerprints, and (e) PubChem fingerprints data sets.

little to the component, whereas larger absolute values indicate variables that contribute more to the component.[44]

For our five data sets, we have computed the absolute sum of coefficients for the number of principal components that show 95% of the explained variance, which is shown in Table 5. For example, in the RDKit database, we have taken 41 principal components as they explain 95% variance. Similarly, all of the remaining four data sets (i.e., Mordred descriptors, Morgan, MACCS keys, and PubChem fingerprint) have been analyzed. Thereafter, the absolute sum of coefficients is calculated for all of the variables that remain in the data set after preprocessing of the respective data sets, as mentioned in Methodology Section 3.7. Further specific details are mentioned in Supporting Information S6−S8.

The pie chart shown in Figure 7a−e shows the distribution of feature loadings or absolute sum of coefficients for respective data sets. Here, feature loadings are divided into four groups according to their values. Group 1 has feature loadings of more than 4, group 2 has feature loadings between 3 and 4, group 3 has feature loadings between 2 and 3, and group 4 has feature loadings less than 2.

5. CONCLUSIONS

The relationship between material properties and descriptors can be used for an inverse material design, identifying particularly promising materials based on a set of target functionalities. In our study, we have presented a machine learning-based classification model that uses the innate structure of the multidimensional property space. It tries to overcome the challenge of how to efficiently search the vast chemical design space to find materials with desired properties.

In our study, we found that the choice of feature set and model can significantly impact the predictive performance. Random forest consistently outperformed other models across multiple data sets, indicating its versatility and robustness in this context. This algorithm excelled in data sets where feature engineering was essential, such as the Mordred descriptors and PubChem fingerprint data sets. Additionally, our feature importance analysis highlighted the key molecular characteristics that influence the efficiency of organic solar cells. The identified features provide valuable insights for materials scientists and chemists working on the design of new organic materials for high-efficiency solar cells. In conclusion, this study underscores the potential of machine learning in advancing the field of organic solar cells. By leveraging diverse

https://doi.org/10.1021/acsomega.4c02157
ACS Omega 2024, 9, 34445−34455

data sets and machine learning algorithms, we have gained valuable insights into the design rules for high-efficiency organic solar cells. The identified essential features serve as a roadmap for material designers seeking to develop the next generation of organic solar cell materials. As we continue to explore the synergy between machine learning and materials science, the future holds great promise for the development of more efficient and sustainable solar energy solutions. This study represents a crucial step in the journey toward harnessing the full potential of organic solar cells, contributing to the global effort to transition to cleaner and more sustainable sources of energy.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acsomega.4c02157.

Data set description—the distribution of energy levels of the frontier orbitals of the molecules, taken as donor/acceptor pairs in the data set (S1); machine learning classification algorithms description (S2); statistical metrics for evaluating the performance of ML models (Table S3); performance evaluation of the four ML model (i.e., SVM, decision trees, random forest, gradient boosting) with five distinct data sets (RDKIT, Mordred, Morgan fingerprint, MACCS key fingerprint, and PubChem fingerprint) (Table S4); machine learning model interpretability through SHAP (S5); feature engineering through principal component analysis (S6), and list of references for preparing experimental OSC database (S7) (PDF)

■ AUTHOR INFORMATION

Corresponding Author

Hamza Siddiqui − *Organic PV Lab, Integral University, Lucknow 226026, India;* orcid.org/0009-0007-9379-248X; Email: hamsid@iul.ac.in

Author

Tahsin Usmani − *Organic PV Lab, Integral University, Lucknow 226026, India*

Complete contact information is available at: https://pubs.acs.org/10.1021/acsomega.4c02157

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

The authors would like to acknowledge Integral University for providing the necessary support (Article MCN No: MCN − IU/R&D/2023-MCN0002133).

■ REFERENCES

(1) Pastuszak, J.; Węgierek, P. Photovoltaic Cell Generations and Current Research Directions for Their Development. *Materials* **2022**, *15*, No. 5542.

(2) Reb, L. K.; Böhmer, M.; Predeschly, B.; et al. Perovskite and Organic Solar Cells on a Rocket Flight. *Joule* **2020**, *4*, 1880−1892.

(3) Huang, B.; Von Lilienfeld, O. A. Ab Initio Machine Learning in Chemical Compound Space. *Chem. Rev.* **2021**, *121*, 10001−10036.

(4) 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.* **2020**, *6*, No. 120.

(5) Liu, Y.; Zhao, T.; Ju, W.; Shi, S. Materials discovery and design using machine learning. *J. Materiomics* **2017**, *3*, 159−177.

(6) Schmidt, J.; Marques, M. R. G.; Botti, S.; Marques, M. A. L. Recent advances and applications of machine learning in solid-state materials science. *npj Comput. Mater.* **2019**, *5*, No. 83.

(7) Alwadai, N.; Ud-Din Khan, S.; Elqahtani, Z. M.; Ud-Din Khan, S. Machine Learning Assisted Prediction of Power Conversion Efficiency of All-Small Molecule Organic Solar Cells: A Data Visualization and Statistical Analysis. *Molecules* **2022**, *27*, No. 5905, DOI: 10.3390/molecules27185905.

(8) Huang, Y.; Zhang, J.; Jiang, E. S.; et al. Structure−Property Correlation Study for Organic Photovoltaic Polymer Materials Using Data Science Approach. *J. Phys. Chem. C* **2020**, *124*, 12871−12882.

(9) Cao, B.; Adutwum, L. A.; Oliynyk, A. O.; et al. How to Optimize Materials and Devices via Design of Experiments and Machine Learning: Demonstration Using Organic Photovoltaics. *ACS Nano* **2018**, *12*, 7434−7444.

(10) Zhao, Z.; Geng, Y.; Troisi, A.; Ma, H. Performance Prediction and Experimental Optimization Assisted by Machine Learning for Organic Photovoltaics. *Adv. Intell. Syst.* **2022**, *4*, No. 2100261.

(11) Malhotra, P.; Khandelwal, K.; Biswas, S.; Chen, F.-C.; Sharma, G. D. Opportunities and challenges for machine learning to select combination of donor and acceptor materials for efficient organic solar cells. *J. Mater. Chem. C* **2022**, *10*, 17781−17811.

(12) Scharber, M. C.; Mühlbacher, D.; Koppe, M.; et al. Design Rules for Donors in Bulk-Heterojunction Solar Cells—Towards 10% Energy-Conversion Efficiency. *Adv. Mater.* **2006**, *18*, 789−794.

(13) Hachmann, J.; Olivares-Amaya, R.; Atahan-Evrenk, S.; 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.* **2011**, *2*, 2241−2251.

(14) Eibeck, A.; Nurkowski, D.; Menon, A.; et al. Predicting Power Conversion Efficiency of Organic Photovoltaics: Models and Data Analysis. *ACS Omega* **2021**, *6*, 23764−23775.

(15) 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* **2017**, *1*, 857−870.

(16) Nagasawa, S.; Al-Naamani, E.; Saeki, A. Computer-Aided Screening of Conjugated Polymers for Organic Solar Cell: Classification by Random Forest. *J. Phys. Chem. Lett.* **2018**, *9*, 2639−2646.

(17) Peng, S.-P.; Zhao, Y. Convolutional Neural Networks for the Design and Analysis of Non-Fullerene Acceptors. *J. Chem. Inf. Model.* **2019**, *59*, 4993−5001.

(18) Chen, F.-C. Virtual Screening of Conjugated Polymers for Organic Photovoltaic Devices Using Support Vector Machines and Ensemble Learning. *Int. J. Polym. Sci.* **2019**, *2019*, No. 4538514.

(19) Sun, W.; Zheng, Y.; Yang, K.; et al. Machine Learning−Assisted Molecular Design and Efficiency Prediction for High-Performance Organic Photovoltaic Materials. *Sci. Adv.* **2019**, *5*, No. eaay4275.

(20) Sun, W.; Li, M.; Li, Y.; et al. The Use of Deep Learning to Fast Evaluate Organic Photovoltaic Materials. *Adv. Theory Simul.* **2019**, *2*, No. 1800116.

(21) Mahmood, A.; Wang, J.-L. A Time and Resource Efficient Machine Learning Assisted Design of Non-Fullerene Small Molecule Acceptors for P3HT-Based Organic Solar Cells and Green Solvent Selection. *J. Mater. Chem. A* **2021**, *9*, 15684−15695.

(22) Moore, G. J.; Bardagot, O.; Banerji, N. Deep Transfer Learning: A Fast and Accurate Tool to Predict the Energy Levels of Donor Molecules for Organic Photovoltaics. *Adv. Theory Simul.* **2022**, *5*, No. 2100511.

(23) Broad, J.; Binder, A. *Hacking with Kali*; Elsevier, 2014.

(24) Dong, J.; Cao, D. S.; Miao, H. Y.; et al. ChemDes: an integrated web-based platform for molecular descriptor and fingerprint computation. *J. Cheminf.* **2015**, *7*, No. 60.

(25) Hong, H.; Xie, Q.; Ge, W.; et al. Mold ^{2}, Molecular Descriptors from 2D Structures for Chemoinformatics and Toxicoinformatics. *J. Chem. Inf. Model.* **2008**, *48*, 1337−1344.
(26) Moriwaki, H.; Tian, Y.-S.; Kawashita, N.; Takagi, T. Mordred: A Molecular Descriptor Calculator. *J Cheminf.* **2018**, *10*, No. 4.
(27) Tetko, I. V.; Gasteiger, J.; Todeschini, R.; et al. Virtual Computational Chemistry Laboratory − Design and Description. *J. Comput.-Aided Mol. Des.* **2005**, *19*, 453−463.
(28) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; et al. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825−2830.
(29) Miyake, Y.; Saeki, A. Machine Learning-Assisted Development of Organic Solar Cell Materials: Issues, Analyses, and Outlooks. *J. Phys. Chem. Lett.* **2021**, *12*, 12391−12401.
(30) Zhang, G.; Lin, F. R.; Qi, F.; et al. Renewed Prospects for Organic Photovoltaics. *Chem. Rev.* **2022**, *122*, 14180−14274.
(31) Bajusz, D.; Rácz, A.; Héberger, K. Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations? *J. Cheminf.* **2015**, *7*, No. 20.
(32) Durant, J. L.; Leland, B. A.; Henry, D. R.; Nourse, J. G. Reoptimization of MDL Keys for Use in Drug Discovery. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 1273−1280.
(33) Kim, S.; Bolton, E. E.; Bryant, S. H. Similar compounds versus similar conformers: complementarity between PubChem 2-D and 3-D neighboring sets. *J. Cheminf.* **2016**, *8*, No. 62.
(34) Rogers, D.; Hahn, M. Extended-Connectivity Fingerprints. *J. Chem. Inf. Model.* **2010**, *50*, 742−754.
(35) Breiman, L.; Friedman, J. H.; Olshen, R. A.; Stone, C. J. *Classification And Regression Trees*; Routledge, 2017.
(36) Yang, L.; Shami, A. On hyperparameter optimization of machine learning algorithms: Theory and practice. *Neurocomputing* **2020**, *415*, 295−316.
(37) Vardhan, B. V. S.; Khedkar, M.; Thakre, P. In *A Comparative Analysis of Hold Out, Cross and Re-Substitution Validation in Hyper-Parameter Tuned Stochastic Short Term Load Forecasting*, 22nd National Power Systems Conference (NPSC); IEEE: New Delhi, India, 2022; pp 448−453.
(38) Darst, B. F.; Malecki, K. C.; Engelman, C. D. Using recursive feature elimination in random forest to account for correlated variables in high dimensional data. *BMC Genet.* **2018**, *19*, No. 65.
(39) Hall, L. H.; Mohney, B.; Kier, L. B. The electrotopological state: structure information at the atomic level for molecular graphs. *J. Chem. Inf. Comput. Sci.* **1991**, *31*, 76−82.
(40) SMARTS - A Language for Describing Molecular Patterns.
(41) Ehrt, C.; Krause, B.; Schmidt, R.; Ehmki, E. S. R.; Rarey, M. SMARTS.plus − A Toolbox for Chemical Pattern Design. *Mol. Inf.* **2020**, *39*, No. 2000216.
(42) MACCS Keys Fingerprint Description from Open Babel.
(43) Sarkar, S.; Boyer, K. L. Quantitative Measures of Change Based on Feature Organization: Eigenvalues and Eigenvectors. *Comput. Vis. Image Underst.* **1998**, *71*, 110−136.
(44) Jolliffe, I. T.; Trendafilov, N. T.; Uddin, M. A Modified Principal Component Technique Based on the LASSO. *J. Comput. Graph. Stat.* **2003**, *12*, 531−547.

https://doi.org/10.1021/acsomega.4c02157
ACS Omega 2024, 9, 34445−34455