Towards Predicting Efficiency of Organic Solar Cells via Machine Learning and Improved Descriptors

Harikrishna Sahu, Weining Rao, Alessandro Troisi,* and Haibo Ma*

To design efficient materials for organic photovoltaics (OPVs), it is essential to identify the largest number of parameters that control their properties and build a model using these parameters (known as descriptors) for the prediction of the power conversion efficiency (PCE). By constructing a dataset for 280 small molecule OPV systems, we found that for all high-performing devices, frontier molecular orbitals of donor molecules are nearly degenerate and in such cases, orbitals other than just HOMO and LUMO are involved in exciton formation, exciton dissociation and hole transport processes influencing the macroscopic properties of OPVs. Machine learning approaches, including random forest, gradient boosting, deep neural network are used to build models for the prediction of PCE using 13 important microscopic properties of organic materials as descriptors. Quite impressive performance of the gradient boosting model (Pearson's coefficient=0.79) indicates that it can certainly be applied to high-throughput virtual screening of promising new donor molecules for high-efficiency OPVs.

Dr. H. Sahu, W. Rao, Prof. H. Ma

Key Laboratory of Mesoscopic Chemistry of MOE

School of Chemistry and Chemical Engineering

Nanjing University, Nanjing 210023, China

E-mail: haibo@nju.edu.cn

Prof. A. Troisi

Department of Chemistry, University of Liverpool

L69 7DZ Liverpool, United Kingdom

E-mail: a.troisi@liverpool.ac.uk

1 Introduction

Organic photovoltaics (OPVs) based on conjugated molecules or polymers are regarded as a source of clean and renewable energy, with advantages such as low cost, light weight, transparency and flexibility.[1–6] The power conversion efficiency (PCE) of OPVs has already been achieved around 13%[7–9] by virtue of the recent tremendous progress in designing new organic materials and optimizing the device architecture.[10–20] However, the further improvement of PCE for OPVs necessary to compete with inorganic devices, is still highly challenging,[21,22] because the trial-and-error experimental routines suffer from stringent synthesis procedure, laborious purification steps and large time-consumption. Many challenges prevent the prediction of the PCE in OPV materials from its constituents. They include the strong electron-electron interactions and strong electron-phonon couplings as well as the complicated donor/acceptor (D/A) interface morphology, which are fundamentally different from inorganic semiconductors.[23,24] Therefore, the accurate simulating of OPVs requires high-level theoretical methods in quantum chemistry, quantum dynamics and statistical mechanics, and in recent years there have been substantial progress for the theoretical understanding of many microscopic processes such as charge transport,[25] exciton dissociation,[26,27] singlet fission,[28–30] etc. These methods are useful for few benchmark systems but cannot be used to explore the chemical space, i.e. screen a large number of candidate materials.

The predictive power of a theoretical model by high-throughput virtual screening has been explored in the recent years.[31–40] For example, in the Harvard Clean Energy Project (CEP),[41] Aspuru-Guzik and coworkers have screened $\sim$2.3 million compounds by employing the Scharber model[42,43] to find out efficient new donor molecules for OPVs. Here, the computed energies of the highest occupied molecular orbital (HOMO)/the lowest unoccupied molecular orbital (LUMO) of organic molecules are calibrated to experimentally determined values, and then employing an averaging scheme, energies of HOMO/LUMO are estimated to use them as input in the Scharber model. However, the prediction of PCE of an OPV is much more challenging compared to the energies of orbitals. Very recently, the same research team[44] noticed that the Scharber model, widely used in these virtual screening works, is not accurate enough for the prediction of PCE, and calibration of PCE by considering molecular similarity, molecular weight and band gap energy leads to improvement in correlation between experimental (49 OPVs) and calculated PCE

(Pearson's coefficient (r) is increased from 0.30 to 0.43). Further, the Scharber model is only optimized for the $PC_{61}BM$ acceptor[42] and a more general model is desired to take account of non-fullerene molecules.

In an OPV, energy conversion is accomplished by four consecutive steps: (*i*) absorption of photons and exciton formation, (*ii*) exciton diffusion to D/A interface, (*iii*) exciton dissociation and (*iv*) transport of holes/electrons to the respective electrode. Taking account of all these processes, the efficiency of a device depends on many microscopic properties of the organic material such as optical gap, charge-carrier mobility, ionization potential of donor, electron affinity of acceptor, hole-electron binding energy (E_{bind}), to name a few and it is very important to build more accurate models using all relevant and easily accessible descriptors. Here, it is important to note that we need to choose quantum chemical descriptors instead of simple topological descriptors, generally applied to drug designing, as they do not give good result when applied to database of conjugated molecules for the solar cell application.[36] Recently, Gómez-Bombarelli et al.[45] demonstrated a method to automatically generate novel chemical structures using a data-driven continuous representation of molecules in the domain of drug designing, and a similar approach can be applied for the solar cell application to generate more realistic (and therefore much less) candidate materials for the screening process, which will drastically reduce the burden of computational cost for quantum chemical descriptors. To date, designing and selection of optimal OPV molecules are primarily based on the HOMO and LUMO of organic molecules, i.e. in the Scharber model.[42] However, very recently it was demonstrated that a very low gap between LUMO and LUMO+1 in acceptor molecules is essential for high efficiency OPVs.[46–48] Since both donor and acceptor conjugated molecules have high probabilities of exhibiting energetically close HOMO and HOMO-1 (LUMO and LUMO+1),[47] it would be necessary to examine the importance of new parameters in predicting PCE for OPVs, such as the near-degeneracy of frontier molecular orbitals also in *donor* molecules.

In this work, we build a model to predict the PCE of new D/A molecules from quantum chemical calculation of properties of the isolated molecules. A number of machine learning (ML) approaches[49–52] are used to identify the connection between these properties and the PCE using experimentally available data to fit the model. Additional experimental data *not included in the fitting procedure* are used to test the predictive ability of the model. ML methods are widely adopted in drug discovery and are more recently

being adopted in the domains of heterogeneous catalyst,[53–55] reaction-mechanism study,[56–58] materials discovery,[59–66] etc. Thanks to the large number of highly efficient donor and acceptor molecules now reported and the substantial chemical differences among them, it is possibly appropriate at this point in time (i) to test different ML strategies (ii) validate hypotheses on correlation between a broad variety of molecular properties and the experimental PCE, and in general (iii) assess the reliability of data driven approaches in OPVs. In this work, we perform these tasks starting from constructing a sufficiently large database of experimental systems for which the molecular properties can be computed.

2 Materials and methods

In the present study, only small molecules are considered due to their discrete and well-defined molecular structures, which will help us to avoid errors related to uncertainty in the polymer length and conformation.[67–69] Although our current focus is mainly on the small molecules, conclusions made from this study are expected to be also applicable for the polymer OPVs. To cover a wide variety small molecule OPVs (SM-OPVs), our dataset include donor molecules such as benzodithiophene- (BDT), diketopyrrolopyrrole- (DPP), quinoxaline-, Zn-porphyrin-based systems, etc. In our dataset, there are 280 different SM-OPVs with 270 distinct donors. Available experimental data such as open-circuit voltage (V_{OC}), short-circuit current (J_{SC}), fill factor (FF) and PCE are collected for each studied system and reported in Table S1. Average PCE values are considered whenever it is available. Looking at the necessity of a wide range of PCE in the dataset, we tried to capture molecules for the entire PCE region. In the dataset, the median value is 5.245% with the upper quartile of the data ranging between 6.5 and 9.8% PCE.

We have made an attempt to build models for the prediction of PCE using 13 microscopic properties of organic materials as descriptors. All the descriptors considered in this work are properties of the isolated donor and acceptor molecules easily computable from electronic structure codes, in the spirit of allowing a rapid screening of a large number ($\sim 10^5$) of new compounds if desired. It is worth to point out that there is no rigorous method exist to choose descriptors for predicting PCE of OPVs as the number of parameters influence the optoelectronic conversion process in OPV is still unknown. Out of our selected 13 descriptors,

eleven are already known to affect the energy conversion process and other two are chosen based on our quantum chemical simulation and statistical analysis. Also, parameters which are highly correlated with other ones are not considered as descriptors. We would like to clarify that in this work $r < 0.4$, $0.4 \leq r < 0.75$ and $r \geq 0.75$ are considered as poor, moderate and high linear correlation. All descriptors considered in this study with a brief justification for choosing them are given below:

(1) *Number of unsaturated atoms in the main conjugation path of donor molecules*, $N_{\text{atom}}^{\text{D}}$. This descriptor is equivalent to the conjugation length of a molecule, and it is widely considered to be associated with photon absorption, exciton diffusion length, stability of charge carriers, etc.[5,23]

(2) *Vertical ionization potential of donor molecules*, IP(v). IP is an important parameter of a donor molecule as it is associated with the energy of HOMO which in turn correlates with the donor valence band edge energies and the V_{OC}.[70]

(3) *Polarizability of donor molecules*. A large polarizability of organic molecules is expected to reduce the exciton binding energy by stabilizing the charge separated states.

(4) *Energy of the electronic transition to a singlet excited state with the largest oscillator strength*, E_{g}. This parameter is related to photon absorption by donor molecules, therefore it is expected to correlate with the J_{SC}, and for this reason it was taken into account in an earlier virtual screening of OPV materials[44] to improve the results obtained by the Scharber model[42].

(5) *Reorganization energy for holes in donor molecules*, λ_h. This quantity is expected to correlated with the barrier for charge transport, and therefore should be smaller for higher mobility materials.[25] It has already been used in previous high-throughput virtual screening for high carrier mobility in organic semiconductors.[71]

(6) *Hole-electron binding energy in donor molecules*, E_{bind}. It is a measure of the strength of hole-electron interaction and can be estimated as the difference between the fundamental gap and optical gap.[72] The fundamental gap, i.e. the minimum energy required for the formation of a pair of separated free hole and electron, is calculated as the difference of ionization potential and electron affinity of molecules. The optical gap is the vertical excitation energy from the ground state to the first dipole-allowed excited state, which is related to the formation of Frenkel excitons in organic molecules. It strongly affects the yield of hole-electron recombination, and a small E_{bind} is always desired to increase the chance of migration of excitons to the D/A interface.

(7) *The energetic difference of HOMO of donor and LUMO of acceptor*, $E_{\mathrm{HL}}^{\mathrm{DA}}$. Energy of the charge-transfer state (E_{CT}) for the D/A interface is known to have good correlation with V_{OC} of an OPV,[21,73] and it can be approximately calculated from the $E_{\mathrm{HL}}^{\mathrm{DA}}$ [74]. We verified the relation between E_{CT} and $E_{\mathrm{HL}}^{\mathrm{DA}}$ for 14 D/A systems (Figure S1 in SI), and our results indicate a very good correlation (r=0.87) between them. Thus, $E_{\mathrm{HL}}^{\mathrm{DA}}$ can roughly estimate the driving force to dissociate excitons in the D/A interface.

(8) *Energy of the electronic transition to the lowest-lying triplet state*, $E_{\mathrm{T_1}}$. Recently, it is revealed that the charge-transfer state at D/A heterojunction can relax to the lowest-lying triplet exciton state ($\mathrm{T_1}$) of individual molecules if the process is energetically favorable, i.e., E_{CT} is greater than 0.1 eV from the energy of $\mathrm{T_1}$ state ($E_{\mathrm{T_1}}$), which may leads to recombination of holes and electrons to the ground state.[21,75] In this work $E_{\mathrm{HL}}^{\mathrm{DA}}$ is a descriptor in place of E_{CT} to avoid high computational costs (as these two parameters are linearly correlated, depicted in Figure S1). To take account the possible triplet loss channel, in addition to $E_{\mathrm{HL}}^{\mathrm{DA}}$, $E_{\mathrm{T_1}}$ is also considered as a descriptor.

(9) *The energetic difference of LUMO of donor and LUMO of acceptor*, $E_{\mathrm{LL}}^{\mathrm{DA}}$. This descriptor quantifies the degree of alignment of LUMOs of donor and acceptor molecules, which is crucial for estimating

the photoelectric conversion efficiency. If E_{LL}^{DA} is too large there is a substantial energy loss at the D/A interface; if it is not large enough the driving force for the charge separation can be insufficient, reducing the yield of charge separation.

(10) *Change in dipole moment in going from the ground state to the first excited state for donor molecules*, Δ_{ge}. The Δ_{ge} is associated with the degree of photoinduced charge transfer within a molecule, and a large value of this parameter is expected to promote the formation of a polarized exciton and reduce the geminate recombination of the hole and electron.[76] A good correlation between Δ_{ge} and the efficiency of OPVs is noted in a series of experimental works.[77,78]

(11,12) *The energetic differences of HOMO and HOMO-1*, $\Delta_H = E_{HOMO} - E_{HOMO-1}$ and *LUMO and LUMO+1*, $\Delta_L = E_{LUMO+1} - E_{LUMO}$ *of donor molecules*. The importances of Δ_H and Δ_L of organic molecules in the energy conversion process will be introduced in Section 3.

(13) *The energetic difference of LUMO and LUMO+1 of acceptors*, $\Delta_L^A = E_{LUMO+1} - E_{LUMO}$. Small Δ_L^A values of acceptors known to accelerate exciton dissociation at the D/A interface, as there is a possibility of having more than one electron accepting states in the anionic form as proposed in refs [46–48].

In the case of acceptors, we do not consider properties other than the Δ_L^A of isolated molecules as descriptors for building models because only two acceptors are taken into account in our study, and one descriptor can effectively distinguish them. The HOMO-LUMO gap and energy of HOMO of donor molecules are not considered as descriptors because in comparison to the HOMO-LUMO gap in the simple single particle approximation, the energy of major electronic transition from many-electron theory is more relevant to the photon absorption in OPVs, and energy of HOMO is linearly correlated to IP(v) (r=0.97) (See Figure S2 in the SI). Properties related to the aggregation of molecules and the effect of solvents are not considered for building models as they required significantly large computational costs. We choose

only those properties which can be quickly calculated. This will enable us to consider a large number of molecules for the screening process, and identified lead candidates may be subjected for high-level calculations with further consideration of the D/A interface morphology, role of charge-transfer and triplet states in exciton dissociation, electronic coupling between adjacent molecules, effect of side chains on solubility and aggregation of materials, etc. For example, recently Ye at al. showed that phase behaviors in OPV materials can be simulated by the atomistic molecular dynamics, and the temperature-dependent effective amorphous-amorphous interaction parameter can be derived by mapping out the phase diagram of a model amorphous polymer:fullerene material system.[79] Such studies will be highly beneficial for further screening of lead candidates before laborious synthesis of materials and device fabrication.

Ground state geometries of all the studied molecules were optimized at the density functional theory (DFT) level using hybrid meta-GGA M06-2X exchange-correlation functional in combination with the 6-31G(d) basis set. To reduce the computational cost, long alkyl side-chains are substituted by methyl groups in our calculation, as they have negligible effect on the optoelectronic properties of isolated molecules. However, it is important to point out that side chains may significantly affect solubility and morphology of materials. Harmonic vibrational frequencies were calculated to ensure that all geometries are at the minima of the potential energy hyper-surface. IP(v), λ_h and E_{bind} were calculated for each system. To obtain the vertical excitation energy and its corresponding oscillator strength, single-point time-dependent DFT (TDDFT) calculations were performed on the corresponding optimized ground state geometries using the same functional and basis set employed for the geometry optimization. It is important to note that, a few descriptors, such as Δ_{H}, Δ_{L}, E_{g} and IP(v), are also calculated using the B3LYP functional using the same ground state geometries, and values of descriptors are found to be linearly correlated with those obtained by the M06-2X functional (Figure S3 in SI), indicating reliability of the descriptors with respect to the choice of DFT functional. All quantum chemical calculations were carried out using Gaussian 09 package.[80]

To build models, we use a range of ML techniques easily accessible from Scikit-Learn.[81] In all these techniques, a set descriptors x_{i} are used as input to obtain an output y, the PCE in our case, effectively determined by a function $f(x_{\text{i}})$. The form of the function and optimization procedure of its hyperparameters vary considerably from algorithm to algorithm. In all cases, a set of descriptors and the corresponding

experimental PCE (the “training set") are used as input, and the algorithm determines the best function which is further validated using a second dataset of unknown data (“validating set"). The quality of a model is examined by the correlation between true and predicted value for a third dataset (“testing set"), which is completely separated from “training set" and “validating set", to provide an unbiased evaluation of the model fit on the training dataset. The simplest possible function is a linear combination of the descriptors, e.g. a linear regression (LR). Here we use a set of more advanced methods such as artificial neural networks (ANN), random forest (RF) and gradient boosting regression tree (GB). A brief description of these methods with additional details to repeat the present analysis are given in the SI. ANN and tree-based models are widely used and found to be successful for a number of applications.[56,65,82–84]

Technically, we split the dataset into two subsets of 250 (training and validating sets) and 30 (testing set) data points. Tanimoto and Euclidean distances of the chemical fingerprints and normalized features, respectively, between each data point of the testing set and their 30 closest points in the 250-data set and comparing them with the change in PCE for the respective systems (Figure S4) indicate that the PCE of samples are not influenced by the mere structural similarity and feature distance, and our testing set is certainly independent and an appropriate choice for the evaluation of model performance. The same figure also provides a meaningful evidence why the graph features fail to predict PCE even for a particular family of molecules.[36] Complete random division of a limited dataset may lead to over-crowded or complete absence of molecules at a certain PCE region. To avoid such inconsistency, we adopted the stratified sampling method[85–88] by dividing our dataset into 8 groups or strata, each having donor molecules with a fixed range of PCE (Table S2), and molecules are randomly divided within a group to ensure that molecules are uniformly distributed in all subsets. To find best hyperparameters of each model, 10-fold cross-validation was employed to evaluate the performance of each parameter combination. Finally, the predictive power of each model was evaluated over the testing set by training the model using 250-dataset (training and validating sets were combined for maximum use of the available data).

3 Results and discussion

Our dataset of 280 experimental systems is large enough to reveal new correlation between properties of organic molecules and macroscopic properties of devices. In order to find out potential new candidates for OPVs, we first focus on the role of orbitals in the energy conversion process, and secondly build ML-models using properties of organic molecules to predict the PCE quickly for high throughput virtual screening.

3.1 Importance of orbitals energetically close to frontier orbitals of donor molecules

In traditional theory, only HOMO and LUMO of organic molecules are taken into account to determine the quality of a material for its application in OPVs. This approximation may be valid for molecules having large Δ_H and Δ_L. Except a few extensively studied molecules such as pentacene/tetracene, most of the recently synthesized materials, especially relatively large conjugated or donor-acceptor-donor type materials, for example, molecules based on BDT,[89,90] naphtho[1,2-*c*:5,6-*c*′]bis[1,2,5]thiadiazole (NT)[14], 1,3-bis(4-(2-ethylhexyl)- thiophen-2-yl)-5,7-bis(2-ethylhexyl)benzo[1,2-*c*:4,5-*c*′]-dithiophene- 4,8-dione (BDD)[91] have considerably small Δ_H/Δ_L values ($\leq$0.48/0.07 eV). In such cases, there is a high probability for participation of orbitals other than frontier orbitals in various photophysical processes of an OPV. Therefore, it is highly desirable to revisit the importance occupied and unoccupied orbitals of organic materials for OPVs.

For all the donor molecules, Δ_H and Δ_L are calculated (see Figure S5, and Table S1 for the raw data) and the distributions of the values of them are depicted in Figure 1. More than 13% of donor molecules have Δ_H smaller than 0.2 eV and nearly 35% donors have Δ_L smaller than 0.1 eV. Thus, a large number of the studied donor molecules have considerably small Δ_H and Δ_L values, and it is important to point out that the nearly degenerate occupied/unoccupied frontier orbitals of a donor molecule may be beneficial for the hole transport/exciton dissociation process, like the near degeneracy of LUMOs in acceptors enhances the exciton dissociation rate at the donor-acceptor interface.[46–48] We verified that there is modest correlation (r=0.43) between Δ_H and Δ_L, and negligible correlation ($r \leq$0.26) between either Δ_H or Δ_L and the HOMO-LUMO gap (data in Figure S6/S7). Poor correlation between these properties indicates that they can be considered as independent descriptors. Figure S7 (c,d) shows correlation of Δ_H and Δ_L with the N^D_{atom}. Δ_H

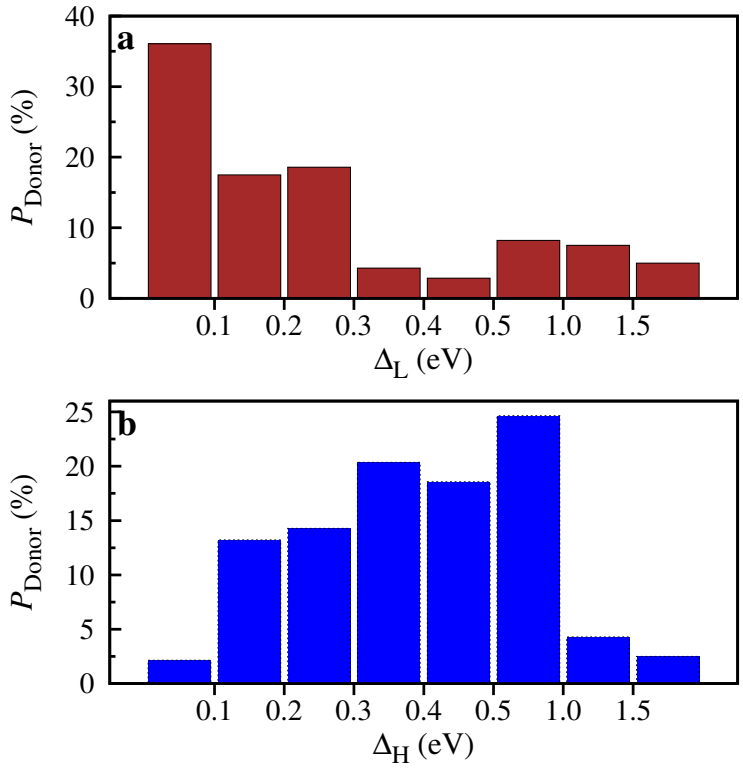


Figure 1 Percentage of donor molecules (P_{Donor}) versus Δ_L (a) and Δ_H (b).

and Δ_L decrease as expected with increasing the size of the conjugated portion of the molecule. However, the correlation with the size is fairly modest and the explicit inclusion of these parameters in the model would be desirable. We next look at the involvement of other than HOMO and LUMO orbitals in the energy conversion process of OPVs.

When light falls on donor molecules, excitons are formed by excitations of electrons from occupied to unoccupied orbitals. The main electronic transition is often considered to be dominated by the HOMO $\rightarrow$ LUMO excitation. This assumption may be true when Δ_H and Δ_L are large enough, for example, thiophene ended DPP as central core with 1,3-di-*tert*- butylbenzene as arms (CSDPP12)[92] and merocyanine dyes[93]. However, the situation may be completely different for molecules having small Δ_H and Δ_L, e.g. BDT and BDD based molecules. Figure 2 (a,b) shows contributions of orbitals other than HOMO and LUMO to the major electronic transition. In the cases of molecules having Δ_H smaller than $\sim$0.75 eV and Δ_L smaller than $\sim$0.5 eV, there is a significant contribution of orbitals other than just HOMO and LUMO

to the major electronic transition, and their contributions increase further with the decrease of Δ_H and Δ_L. Figure S8 shows the correlation between experimental absorption coefficients (ε) and theoretically calculated oscillator strengths (f_{osc}) of the main electronic transition for 82 donor molecules, and correlations of these two parameters with the Δ_H and Δ_L are depicted in Figure S9. Our results indicate that there is a modest correlation (r=0.42) between ε and f_{osc}, and correlations of these two parameters with the Δ_L are much better than their relations with the Δ_H. Populations of donor molecules with respect to contributions of orbitals other than HOMO and LUMO orbitals to the main electronic transitions are depicted in Figure 2 (c,d). As shown in the figure, the contribution is greater than 30% for nearly 50% of the studied molecules. Thus, orbitals energetically close to HOMO/LUMO also participate in the exciton formation process, and it is essential that this fact is properly accounted in any data-driven study of PCE.

Also, the dissociation of excitons into free holes and electrons, and transport of charge carriers can be influenced by Δ_H or Δ_L. The extent of electronic coupling between two interacting molecules is an important factor for both hole transport and exciton dissociation.[25] When Δ_H (Δ_L) of donor molecules is quite large, HOMO and HOMO-1 (LUMO and LUMO+1) are well separated, and only HOMO (LUMO) significantly contributes to the hole transfer (exciton dissociation) process. However, in the case of small Δ_H (Δ_L), the contribution of HOMO-1 (LUMO+1) can not be ignored in the electronic coupling calculation.[94] Four molecules having small Δ_L/Δ_H values and large PCE are randomly chosen to calculate the electronic coupling for the exciton dissociation and hole transport rates. For each system, the electronic coupling was evaluated by two different ways (a) by considering only HOMO (LUMO) orbitals, denoted as $t^{H\rightarrow H}$ ($t^{L\rightarrow L}$) and (b) by considering proper description for the initial and final states in the dimer (D/A system) and denoting $t^{eff_{HT}}$ ($t^{eff_{ED}}$), the electronic coupling between initial and final states. A detailed description of the coupling calculations is reported in the SI. To ensure reliability of the calculation with respect to the choice of DFT functional, coupling values are estimated using three different functionals, namely M06-2X, B3LYP and PBE0, and our results indicate a qualitatively similar trend for all the functionals. We found a large difference between $t^{H\rightarrow H}$ ($t^{L\rightarrow L}$) and $t^{eff_{HT}}$ ($t^{eff_{ED}}$), demonstrating that occupied orbitals other than HOMO (LUMO) significantly contribute to charge transport (exciton dissociation) process (Table S3). Thus, the hole transport and exciton dissociation rates are closely related to the Δ_H and Δ_L of donor molecules, respectively.

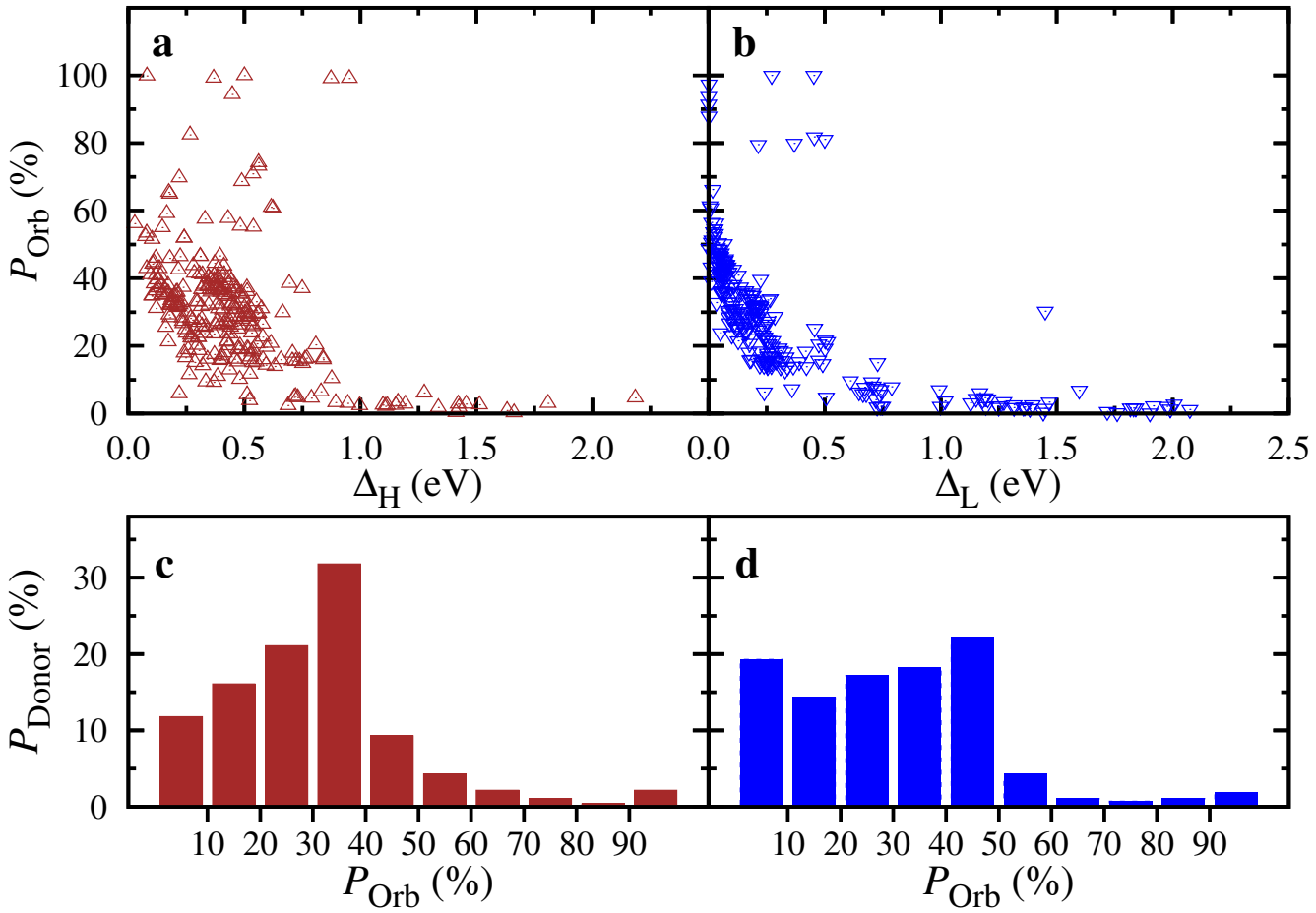


Figure 2 Contributions of orbitals other than HOMO (a) and LUMO (b) (denoted as P_{Orb}) to the most important electronic transition (largest oscillator strength) versus Δ_H and Δ_L, respectively. Percentage of donor molecules (P_{Donor}) versus P_{Orb} in cases of occupied and unoccupied orbitals are depicted in (c) and (d), respectively.

The combined computational and experimental data can be used to verify the importances of Δ_H and Δ_L in realistic devices, which is suggested by the physical models above. As shown in Figure 3 (a,b), large FF values are only found for the small energy gap between orbitals, describe the level of correlation observed reasonably good and in agreement with the physical model but of course cannot be perfect considering the many other components that may affect the efficiency. As it can be seen in Figure 3 (c,d), large PCE values are only found for small Δ_H and Δ_L values, having noticeable linear correlation with Δ_L (r=−0.41). A schematic diagram shown in Figure 4 represents involvement of occupied and unoccupied orbitals in various processes of an OPV. In case I, when Δ_H and Δ_L are large, HOMO and LUMO are mainly

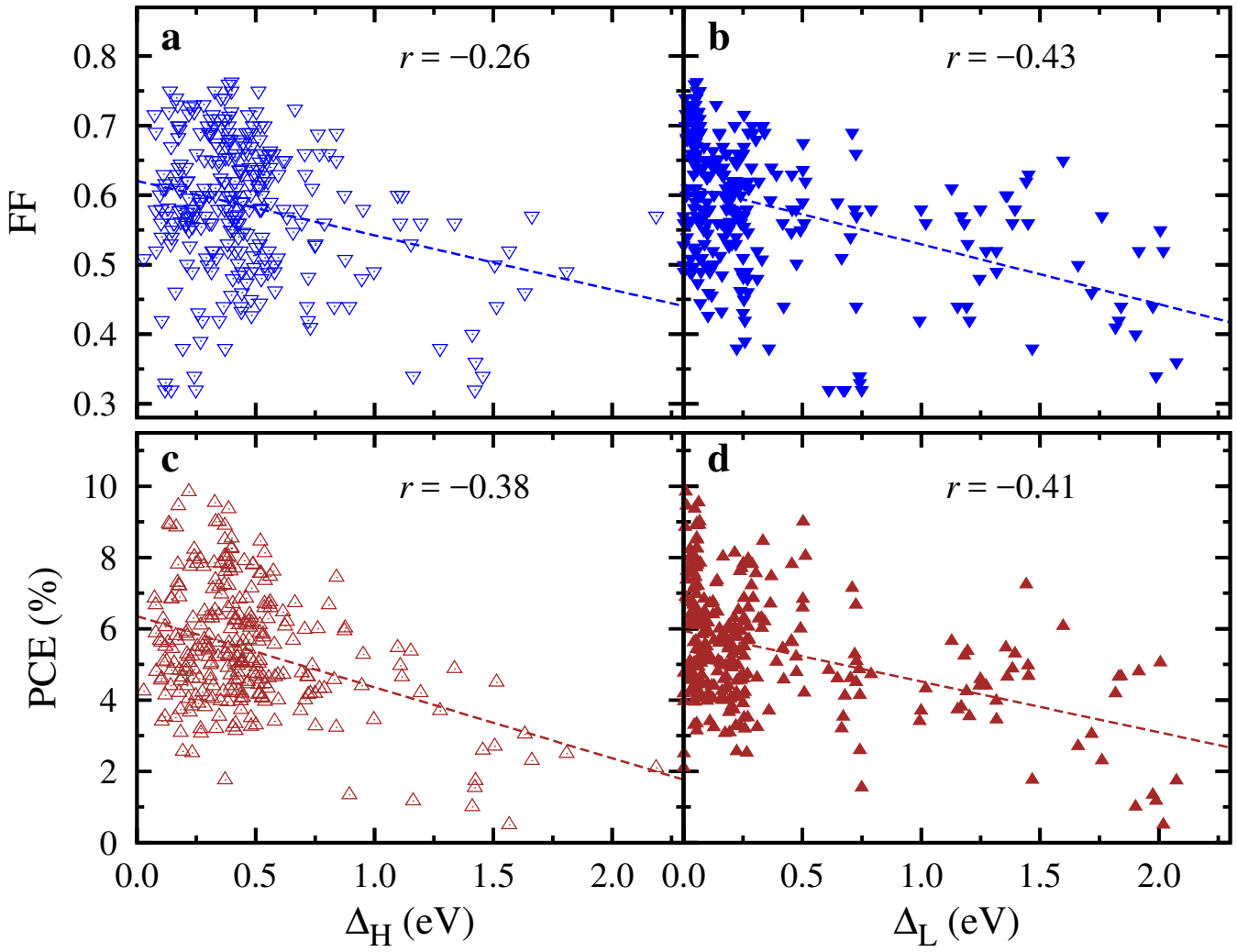


Figure 3 Correlations of FF and PCE versus Δ_H (a,c) and Δ_L (b,d). Here, Δ_H is the energetic difference between HOMO and HOMO-1 and Δ_L is the energetic differences between LUMO+1 and LUMO of donor molecules.

involved in photon harvesting, excitation transport, exciton dissociation and charge carrier transport processes. For case II, when Δ_H and Δ_L are small enough, orbitals energetically close to HOMO/LUMO also participate in these processes and sometimes they can accelerate them and enhance the PCE an OPV.

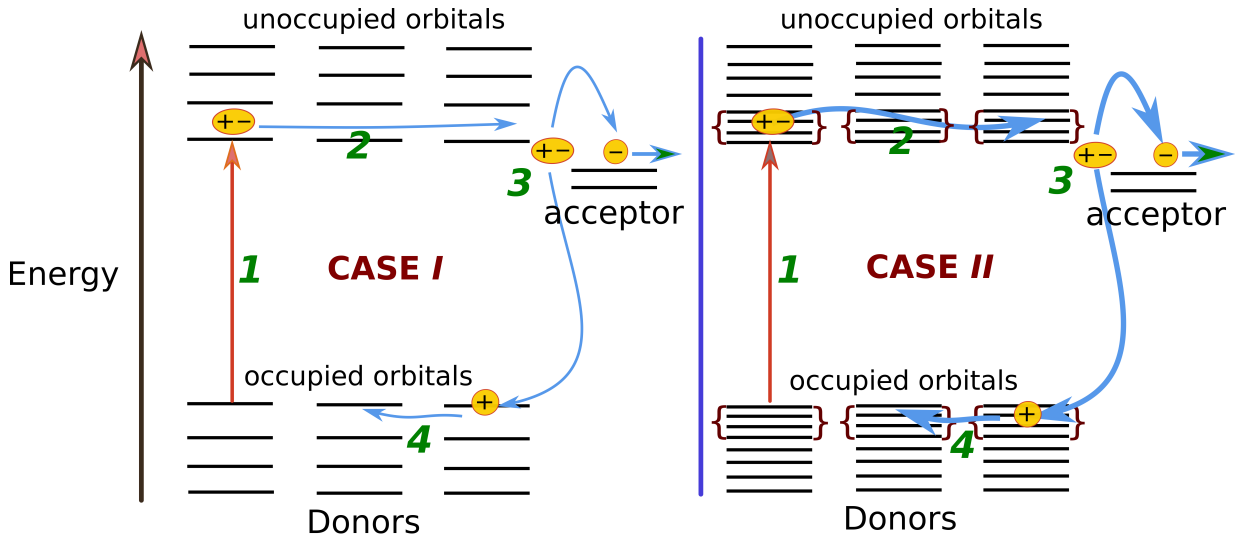


Figure 4 A schematic diagram for the involvement of orbitals energetically close to HOMO and LUMO in photophysical processes of an OPV. Cases I and II are for large and small values of Δ_H (Δ_L), respectively.

3.2 Prediction of PCE using ML algorithms

To efficiently screen a large number of candidate molecules for new high performance OPVs, theoretical models which can predict the potential PCE values accurately and very quickly are highly desired. Few attempts have already been made so far but the correlation between predicted and experimental data are very unsatisfactory ($r \sim 0.4$[44]). In this work, we build new models with improved descriptors and ML-algorithms. A range of ML approaches are applied to construct models, such as *k*-nearest neighbor, artificial neural network, random forest and gradient boosting, to assess the relative merit of these techniques.

Table 1 Prediction of the PCE by different ML algorithms. MAPE, RMSE, r and P_{out} represent mean absolute percentage error, root mean square error, Pearson's correlation coefficient and percentage of number points outside a cutoff value, respectively. Results for the testing set (30 points) and all data points obtained by models trained on 250 and 279 data points (leave-one-out cross-validation) are shown outside and inside parenthesis, respectively.

ML techniques	MAPE (%)	RMSE (%)	r	P_{out} in % ($\pm$ 1.5/$\pm$ 2.5)
LR	20.2 (23.0)	1.34 (1.25)	0.66 (0.67)	23.3/6.7 (22.1/5.4)
k-NN	18.2 (19.9)	1.21 (1.17)	0.71(0.74)	23.3/3.3 (17.9/3.9)
ANN	20.0 (20.4)	1.25 (1.16)	0.68 (0.73)	20.0/3.3 (19.6/2.9)
RF	17.2 (21.4)	1.08 (1.12)	0.78 (0.75)	13.3/3.3 (18.9/1.4)
GB	17.1 (19.9)	1.07 (1.09)	0.79 (0.76)	16.7/0.0 (18.2/1.8)

The predictive power of ML models are compared by the mean absolute percentage error (MAPE), root mean square error (RMSE) and r. Table 1 reports the predictive power measures performed over the testing set (30) and all 280 data points by training our models on 250 data points and employing the leave-one-out cross-validation technique, respectively. Results of 10-fold cross-validation over 250 data points obtained using convenience, random and stratified sampling methods are reported in Table S4. In Figure 5(a,b), predicted PCE obtained by the GB model is plotted versus experimentally estimated PCE, and results for other ML-models are shown in Figures S12-S15. The probability density of error distribution in inset of Figure 5(b) shows that the residuals between the model-predicted and experimentally estimated PCE are normally distributed, and thus there is no sampling error in the model. Similar results obtained for the testing set, 10-fold cross-validation over 250 data points, and leave-one-out cross-validation for all 280 data

points indicate that our models generalize well to new, previously unseen data. The predictive power of LR-based model is already quite remarkable but far from ideal (r=0.66). In the case of ANN-based models, a moderate improvement in results (r=0.68) is seen as nonlinear correlations are taken into account. It is interesting to find that k-NN, a simple ML algorithm, improves the prediction efficiency (r=0.71) enough for the screening process. Results of tree-based models are quite impressive, especially in the case of GB-model, where r and RMSE for the testing set are 0.79 and 1.07%. The GB model can be effectively used to predict PCE of new candidate materials, however, this model can not predict the PCE outside the 1-9.5% range due to the inherent limitation of tree-based models, i.e., they unable to extrapolate to regions, outside of what is seen in the training data. Under such circumstances, ANN model is preferable to the tree-based models. In SI, *Python scripts* with detailed description of their usage are provided to reproduce reported results and predict PCE from properties of new donor molecules.

The “importance" of a descriptor can be defined rigorously within tree-based algorithms like RF or GB. It is estimated by keeping track of the reduction of mean-square error for each descriptor when data passes through the trees and averaging it over all trees of the ensemble as proposed by Breiman et al.[95] and implemented in Scikit-learn software package[81]. The importances of descriptors for the GB and RF models are shown in Figure 5 (c) and (d), respectively. It is important to clarify that low descriptor importance does not mean the associated descriptor is irrelevant to the PCE. It may happen that more than one descriptors encode the same information and our model picked one of them. As it can be seen in the Figure 5 (c), out of 13 descriptors the importance for Δ_L^A is quite small as we have only two distinct acceptors, i.e., $PC_{61}BM$ and $PC_{71}BM$ in the dataset. The GB gives a lot of importance to descriptors other than the Δ_L^A, as they are all related to photophysical processes of an OPV. Although the descriptor importances are different for the GB and RF models, the E_{bind} is found to be the most informative descriptor for both the models with a lot of importances given to Δ_H and Δ_L. The E_{bind} estimates the strength of hole-electron interaction, and therefore it was expected to play an important role for building ML-models. The main new insight afforded by the ML analysis is the importances of Δ_H and Δ_L, which were also suggested by the visual inspection of the raw data (Figure 3).

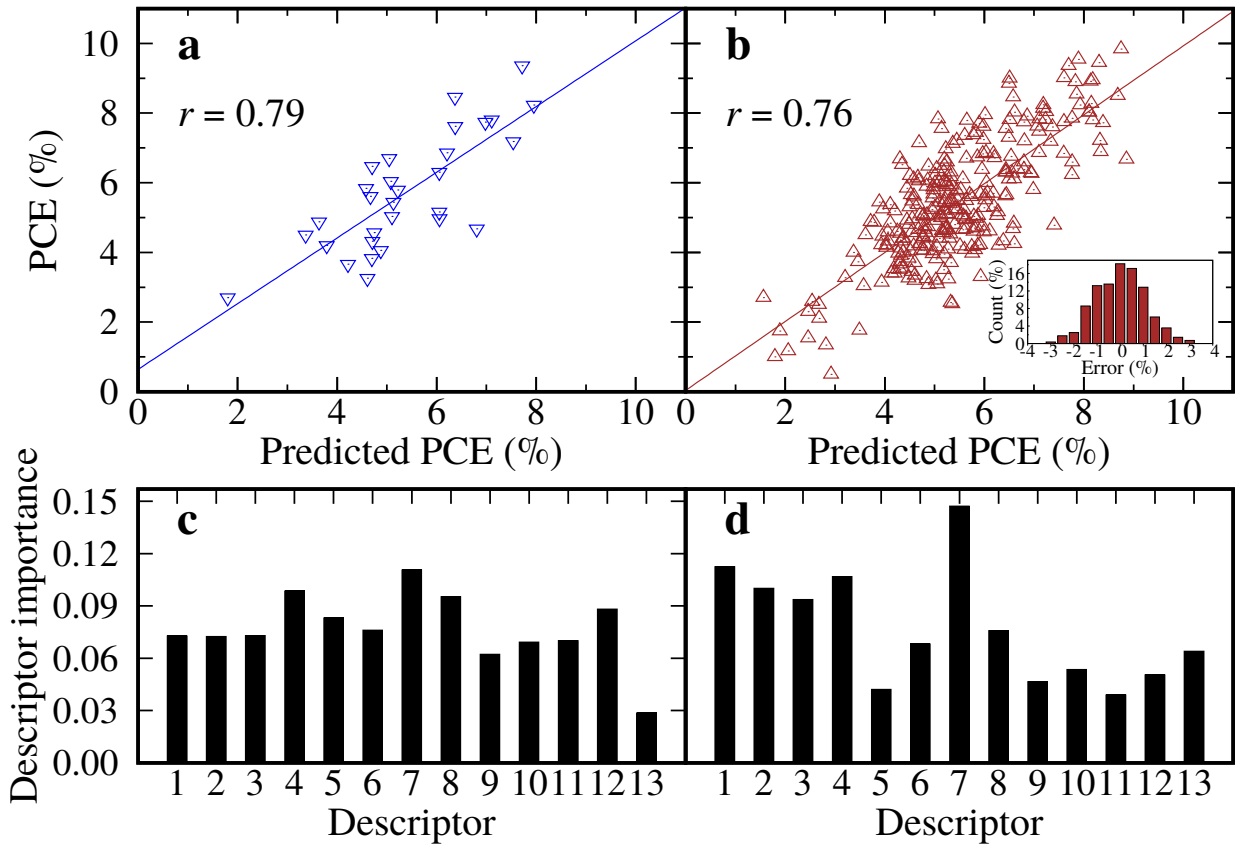


Figure 5 Theoretically predicted versus experimental PCE for the testing set (30 molecules) (a) and all data points using the leave-one-out cross-validation technique (b) for the GB model. Inset shows probability density of prediction errors. The descriptor importances for the GB (c) and RF (d) model are depicted. Descriptors are in the following order: (1) $N_{\mathrm{atom}}^{\mathrm{D}}$, (2) polarizability, (3) Δ_{L}, (4) Δ_{H}, (5) IP(v), (6) λ_h, (7) E_{bind}, (8) $E_{\mathrm{LL}}^{\mathrm{DA}}$, (9) $E_{\mathrm{HL}}^{\mathrm{DA}}$, (10) E_{g}, (11) Δ_{ge}, (12) $E_{\mathrm{T_1}}$ and (13) $\Delta_{\mathrm{L}}^{\mathrm{A}}$.

4 Conclusion and outlook

In summary, we show in this work that by using artificial intelligence ML methods we can capture the complexity of a device and build a model that can efficiently predict the efficiency (rather than simpler properties like HOMO-LUMO gap) of OPVs from its constituents with an unprecedented correlation of r=0.79. A predictive method of this quality is certainly able to make the difference in the discovery of new materials for organic solar cells. The importance of the near degeneracy of frontier molecular orbitals on various photo-physical processes in solar cell, is also demonstrated for the first time. This critical point has been overlooked by other authors, which is the basis for a better predictive model.

Before the end, we should mention that our derived ML-model does not consider the details about the morphology and many proposed microscopic mechanisms, and cannot always predict the PCE of a specified OPV device very accurately. This model is expected to examine the optimal potential of OPV materials and can be used for a preliminary virtual screening of a large number ($\sim 10^5$) of candidate molecules due to its advantage of very low computational costs. A small portion ($\sim 10^3$) of the candidates with predicted high PCE by our ML-model will be subjected for further more advanced theoretical simulations with electrodynamics and quantum dynamics models, which are much more time-consuming but can consider many effects of morphology and mechanism, such as the effect of solvents and side chains on aggregation, D/A interface morphology, exciton migration and dissociation, the coupling between electronic states and vibrational modes, singlet fission, etc. At the end, the further screened ($\sim 10^2$) molecules with more quantitative predictions by high level calculations will be suggested for experimental synthesis and device fabrication. We are quite hopeful that similar works will substantially assist development of OPVs by enhancing understanding of the operating mechanism and designing new efficient materials.

Supporting Information

Electronic Supplementary Information (ESI) available: Experimentally estimated V_{OC}, J_{SC}, FF and PCE of SM-OPVs, population of donor molecules versus PCE, correlation between various properties, details of the electronic coupling calculation and optimization of models built by various machine learning techniques are provided. A tar file "suppl_files.tar.gz" containing Cartesian coordinates of ground state geometries and SMILES for all donor/acceptor molecules, comma-separated values (CSV) files for the raw data and *Python scripts* to reproduce reported results and predict the PCE of new donor molecules is available for use.

Acknowledgement

The work was supported by the National Natural Science Foundation of China [Grant Numbers 21673109, 21722302].

References

[1] G. J. Hedley, A. Ruseckas, I. D. W. Samuel, *Chem. Rev.* **2017**, *117*, 796.

[2] S. Li, W. Liu, C.-Z. Li, M. Shi, H. Chen, *Small* **2017**, *13*, 1701120.

[3] M. Wright, R. Lin, M. J. Y. Tayebjee, G. Conibeer, *Sol. RRL* **2017**, *1*, 1700035.

[4] K. Wang, C. Liu, T. Meng, C. Yi, X. Gong, *Chem. Soc. Rev.* **2016**, *45*, 2937.

[5] C. Liu, K. Wang, X. Gong, A. J. Heeger, *Chem. Soc. Rev.* **2016**, *45*, 4825.

[6] Q. An, F. Zhang, J. Zhang, W. Tang, Z. Deng, B. Hu, *Energy Environ. Sci.* **2016**, *9*, 281.

[7] J. Hou, O. Inganäs, R. H. Friend, F. Gao, *Nature Mater.* **2018**, *17*, 119.

[8] W. Zhao, S. Li, H. Yao, S. Zhang, Y. Zhang, B. Yang, J. Hou, *J. Am. Chem. Soc.* **2017**, *139*, 7148.

[9] X. Xu, T. Yu, Z. Bi, W. Ma, Y. Li, Q. Peng, *Adv. Mater.* **2017**, 1703973.

[10] Y. Li, G. Xu, C. Cui, Y. Li, *Adv. Energy Mater.* **2017**, 1701791.

[11] M. Liu, Y. Gao, Y. Zhang, Z. Liu, L. Zhao, *Polym. Chem.* **2017**, *8*, 4613.

[12] S. D. Collins, N. A. Ran, M. C. Heiber, T.-Q. Nguyen, *Adv. Energy Mater.* **2017**, *7*, 1602242.

[13] S. Dai, F. Zhao, Q. Zhang, T.-K. Lau, T. Li, K. Liu, Q. Ling, C. Wang, X. Lu, W. You, X. Zhan, *J. Am. Chem. Soc.* **2017**, *139*, 1336.

[14] J. Wan, X. Xu, G. Zhang, Y. Li, K. Feng, Q. Peng, *Energy Environ. Sci.* **2017**, *10*, 1739.

[15] Y. Cai, L. Huo, Y. Sun, *Adv. Mater.* **2017**, *29*, 1605437.

[16] A. Tang, C. Zhan, J. Yao, E. Zhou, *Adv. Mater.* **2017**, *29*, 1600013.

[17] D. Gedefaw, M. Prosa, M. Bolognesi, M. Seri, M. R. Andersson, *Adv. Energy Mater.* **2017**, *7*, 1700575.

[18] J. Min, Y. N. Luponosov, C. Cui, B. Kan, H. Chen, X. Wan, Y. Chen, S. A. Ponomarenko, Y. Li, C. J. Brabec, *Adv. Energy Mater.* **2017**, *7*, 1700465.

[19] Y. Cai, A. Qin, B. Z. Tang, *J. Mater. Chem. C* **2017**, *5*, 7375.

[20] D. Deng, Y. Zhang, J. Zhang, Z. Wang, L. Zhu, J. Fang, B. Xia, Z. Wang, K. Lu, W. Ma, Z. Wei, *Nat. Commun.* **2016**, *7*, 13740.

[21] J. Zhang, L. Zhu, Z. Wei, *Small Methods* **2017**, *1*, 1700258.

[22] P. Cheng, X. Zhan, *Chem. Soc. Rev.* **2016**, *45*, 2544.

[23] O. Ostroverkhova, *Chem. Rev.* **2016**, *116*, 13279.

[24] A. Zhugayevych, S. Tretiak, *Annu. Rev. Phys. Chem.* **2015**, *66*, 305 .

[25] H. Oberhofer, K. Reuter, J. Blumberger, *Chem. Rev.* **2017**, *117*, 10319.

[26] K. M. Pelzer, S. B. Darling, *Mol. Syst. Des. Eng.* **2016**, *1*, 10.

[27] M. Polkehn, P. Eisenbrandt, H. Tamura, I. Burghardt, *Int. J. Quantum Chem.* **2018**, *118*, e25502.

[28] J. Xia, S. N. Sanders, W. Cheng, J. Z. Low, J. Liu, L. M. Campos, T. Sun, *Adv. Mater.* **2017**, *29*, 1601652.

[29] V. Abraham, N. J. Mayhall, *J. Phys. Chem. Lett.* **2017**, *8*, 5472.

[30] Z. Huang, Y. Fujihashi, Y. Zhao, *J. Phys. Chem. Lett.* **2017**, *8*, 3306.

[31] S. Tortorella, F. De Angelis, G. Cruciani, *J Chemom.* **2018**, *32*, e2957.

[32] C. Zanlorenzi, L. Akcelrud, *J. Polym. Sci. Part B: Polym. Phys.* **2017**, *55*, 919.

[33] K. Kuhar, A. Crovetto, M. Pandey, K. S. Thygesen, B. Seger, P. C. K. Vesborg, O. Hansen, I. Chorkendorff, K. W. Jacobsen, *Energy Environ. Sci.* **2017**, *10*, 2579.

[34] Y. Imamura, M. Tashiro, M. Katouda, M. Hada, *J. Phys. Chem. C* **2017**, *121*, 28275.

[35] V. Venkatraman, M. Foscato, V. R. Jensen, B. K. Alsberg, *J. Mater. Chem. A* **2015**, *3*, 9851.

[36] V. Venkatraman, B. K. Alsberg, *Dyes Pigm.* **2015**, *114*, 69.

[37] V. Venkatraman, P.-O. Åstrand, B. K. Alsberg, *J. Comput. Chem.* **2014**, *35*, 214.

[38] I. Y. Kanal, S. G. Owens, J. S. Bechtel, G. R. Hutchison, *J. Phys. Chem. Lett.* **2013**, *4*, 1613.

[39] V. Tripkovic, H. A. Hansen, T. Vegge, *ChemSusChem* **2018**, *11*, 629.

[40] N. M. O'Boyle, C. M. Campbell, G. R. Hutchison, *J. Phys. Chem. C* **2011**, *115*, 16200.

[41] J. Hachmann, R. Olivares-Amaya, A. Jinich, A. L. Appleton, M. A. Blood-Forsythe, L. R. Seress, C. Román-Salgado, K. Trepte, S. Atahan-Evrenk, S. Er, S. Shrestha, R. Mondal, A. Sokolov, Z. Bao, A. Aspuru-Guzik, *Energy Environ. Sci.* **2014**, *7*, 698.

[42] M. C. Scharber, D. Mühlbacher, M. Koppe, P. Denk, C. Waldauf, A. J. Heeger, C. J. Brabec, *Adv. Mater.* **2006**, *18*, 789.

[43] T. Ameri, G. Dennler, C. Lungenschmied, C. J. Brabec, *Energy Environ. Sci.* **2009**, *2*, 347.

[44] S. A. Lopez, B. Sanchez-Lengeling, J. de Goes Soares, A. Aspuru-Guzik, *Joule* **2017**, *1*, 857.

[45] R. Gómez-Bombarelli, J. N. Wei, D. Duvenaud, J. M. Hernández-Lobato, B. Sánchez-Lengeling, D. Sheberla, J. Aguilera-Iparraguirre, T. D. Hirzel, R. P. Adams, A. Aspuru-Guzik, *ACS Cent. Sci.* **2018**, *4*, 268.

[46] A. Kuzmich, D. Padula, H. Ma, A. Troisi, *Energy Environ. Sci.* **2017**, *10*, 395.

[47] H. Ma, A. Troisi, *J. Phys. Chem. C* **2014**, *118*, 27272.

[48] T. Liu, A. Troisi, *Adv. Mater.* **2013**, *25*, 1038.

[49] D. Bzdok, M. Krzywinski, N. Altman, *Nat. Methods* **2017**, *14*, 1119.

[50] J. Biamonte, P. Wittek, N. Pancotti, P. Rebentrost, N. Wiebe, S. Lloyd, *Nature* **2017**, *549*, 195.

[51] M. Krzywinski, N. Altman, *Nat. Methods* **2017**, *14*, 757.

[52] Y. LeCun, Y. Bengio, G. Hinton, *Nature* **2015**, *521*, 436.

[53] H. Li, Z. Zhang, Z. Liu, *Catalysts* **2017**, *7*, 306.

[54] R. Jinnouchi, R. Asahi, *J. Phys. Chem. Lett.* **2017**, *8*, 4279.

[55] Z. Li, S. Wang, W. S. Chin, L. E. Achenie, H. Xin, *J. Mater. Chem. A* **2017**, *5*, 24131.

[56] D. Fooshee, A. Mood, E. Gutman, M. Tavakoli, G. Urban, F. Liu, N. Huynh, D. Van Vranken, P. Baldi, *Mol. Syst. Des. Eng.* **2018**, DOI: 10.1039/C7ME00107J.

[57] C. W. Coley, R. Barzilay, T. S. Jaakkola, W. H. Green, K. F. Jensen, *ACS Cent. Sci.* **2017**, *3*, 434.

[58] P. Sadowski, D. Fooshee, N. Subrahmanya, P. Baldi, *J. Chem. Inf. Model.* **2016**, *56*, 2125.

[59] J. P. Janet, L. Chan, H. J. Kulik, *J. Phys. Chem. Lett.* **2018**, *9*, 1064.

[60] J. D. Perea, S. Langner, M. Salvador, B. Sanchez-Lengeling, N. Li, C. Zhang, G. Jarvas, J. Kontos, A. Dallos, A. Aspuru-Guzik, C. J. Brabec, *J. Phys. Chem. C* **2017**, *121*, 18153.

[61] R. Ramprasad, R. Batra, G. Pilania, A. Mannodi-Kanakkithodi, C. Kim, *Nat. Comput. Mater.* **2017**, *3*, 54.

[62] Y. T. Sun, H. Y. Bai, M. Z. Li, W. H. Wang, *J. Phys. Chem. Lett.* **2017**, *8*, 3434.

[63] J. D. Perea, S. Langner, M. Salvador, J. Kontos, G. Jarvas, F. Winkler, F. Machui, A. Görling, A. Dallos, T. Ameri, C. J. Brabec, *J. Phys. Chem. B* **2016**, *120*, 4431.

[64] D. Xue, P. V. Balachandran, J. Hogden, J. Theiler, D. Xue, T. Lookman, *Nat. Commun.* **2016**, *7*, 11241.

[65] R. Gómez-Bombarelli, J. Aguilera-Iparraguirre, T. D. Hirzel, D. Duvenaud, D. Maclaurin, M. A. Blood-Forsythe, H. S. Chae, M. Einzinger, D.-G. Ha, T. Wu, G. Markopoulos, S. Jeon, H. Kang, H. Miyazaki, M. Numata, S. Kim, W. Huang, S. I. Hong, M. Baldo, R. P. Adams, A. Aspuru-Guzik, *Nature Mater.* **2016**, *15*, 1120.

[66] L. Ward, A. Agrawal, A. Choudhary, C. Wolverton, *Nat. Comput. Mater.* **2016**, *2*, 16028.

[67] L. Nian, K. Gao, Y. Jiang, Q. Rong, X. Hu, D. Yuan, F. Liu, X. Peng, T. P. Russell, G. Zhou, *Adv. Mater.* **2017**, *29*, 1700616.

[68] J. Zhou, Y. Zuo, X. Wan, G. Long, Q. Zhang, W. Ni, Y. Liu, Z. Li, G. He, C. Li, B. Kan, M. Li, Y. Chen, *J. Am. Chem. Soc.* **2013**, *135*, 8484.

[69] Y. Lin, Y. Li, X. Zhan, *Chem. Soc. Rev.* **2012**, *41*, 4245.

[70] B. Thompson, J. Fréchet, *Angew. Chem. Int. Ed.* **2008**, *47*, 58.

[71] C. Schober, K. Reuter, H. Oberhofer, *J. Phys. Chem. Lett.* **2016**, *7*, 3973.

[72] J.-L. Brédas, *Mater. Horiz.* **2014**, *1*, 17.

[73] T. M. Burke, S. Sweetnam, K. Vandewal, M. D. McGehee, *Adv. Energy Mater.* **2015**, *5*, 1500123.

[74] D. Veldman, S. C. J. Meskers, R. A. J. Janssen, *Adv. Funct. Mater.* **2009**, *19*, 1939.

[75] K. N. Schwarz, P. B. Geraghty, D. J. Jones, T. A. Smith, K. P. Ghiggino, *J. Phys. Chem. C* **2016**, *120*, 24002.

[76] B. S. Rolczynski, J. M. Szarko, H. J. Son, Y. Liang, L. Yu, L. X. Chen, *J. Am. Chem. Soc.* **2012**, *134*, 4142.

[77] B. Carsten, J. M. Szarko, H. J. Son, W. Wang, L. Lu, F. He, B. S. Rolczynski, S. J. Lou, L. X. Chen, L. Yu, *J. Am. Chem. Soc.* **2011**, *133*, 20468.

[78] B. Carsten, J. M. Szarko, L. Lu, H. J. Son, F. He, Y. Y. Botros, L. X. Chen, L. Yu, *Macromolecules* **2012**, *45*, 6390.

[79] L. Ye, H. Hu, M. Ghasemi, T. Wang, B. A. Collins, J.-H. Kim, K. Jiang, J. H. Carpenter, H. Li, Z. Li, T. McAfee, J. Zhao, X. Chen, J. L. Y. Lai, T. Ma, J.-L. Brédas, H. Yan, H. Ade, *Nature Mater.* **2018**, *17*, 253.

[80] M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, G. Scalmani, V. Barone, B. Mennucci, G. A. Petersson, H. Nakatsuji, M. Caricato, X. Li, H. P. Hratchian, A. F. Izmaylov, J. Bloino, G. Zheng, J. L. Sonnenberg, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, T. Vreven, J. A. Montgomery, Jr., J. E. Peralta, F. Ogliaro, M. Bearpark, J. J. Heyd, E. Brothers, K. N. Kudin, V. N. Staroverov, T. Keith, R. Kobayashi, J. Normand, K. Raghavachari, A. Rendell, J. C. Burant, S. S. Iyengar, J. Tomasi, M. Cossi, N. Rega, J. M. Millam, M. Klene, J. E. Knox, J. B. Cross, V. Bakken, C. Adamo, J. Jaramillo, R. Gomperts, R. E. Stratmann, O. Yazyev, A. J. Austin, R. Cammi, C. Pomelli, J. W. Ochterski, R. L. Martin, K. Morokuma, V. G. Zakrzewski, G. A. Voth, P. Salvador, J. J. Dannenberg, S. Dapprich, A. D. Daniels, O. Farkas, J. B. Foresman, J. V. Ortiz, J. Cioslowski, D. J. Fox, Gaussian 09 Revision B.01, Gaussian, Inc., Wallingford CT, 2010.

[81] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, Édouard Duchesnay, *J. Mach. Learn Res.* **2011**, *12*, 2825.

[82] G. Skoraczyński, P. Dittwald, B. Miasojedow, S. Szymkuć, E. P. Gajewska, B. A. Grzybowski, A. Gambin, *Sci. Rep.* **2017**, *7*, 3582.

[83] R. M. Bhardwaj, A. Johnston, B. F. Johnston, A. J. Florence, *CrystEngComm* **2015**, *17*, 4272.

[84] F. Da Silva, J. Desaphy, G. Bret, D. Rognan, *J. Chem. Inf. Model.* **2015**, *55*, 2005.

[85] V. L. Parsons, *Stratified Sampling*, American Cancer Society, **2017**.

[86] T. Lumley, *Complex Surveys*, Wiley-Blackwell, **2010**.

[87] K. R. W. Brewer, *Int. Statist. Rev.* **1999**, *67*, 35.

[88] J. Neyman, *J. Royal Stat. Soc.* **1934**, *97*, 558.

[89] Z. Wang, X. Xu, Z. Li, K. Feng, K. Li, Y. Li, Q. Peng, *Adv. Electron. Mater.* **2016**, *2*, 1600061.

[90] C. Cui, X. Guo, J. Min, B. Guo, X. Cheng, M. Zhang, C. J. Brabec, Y. Li, *Adv. Mater.* **2015**, *27*, 7469.

[91] H. Zhang, Y. Liu, Y. Sun, M. Li, B. Kan, X. Ke, Q. Zhang, X. Wan, Y. Chen, *Chem. Commun.* **2017**, *53*, 451.

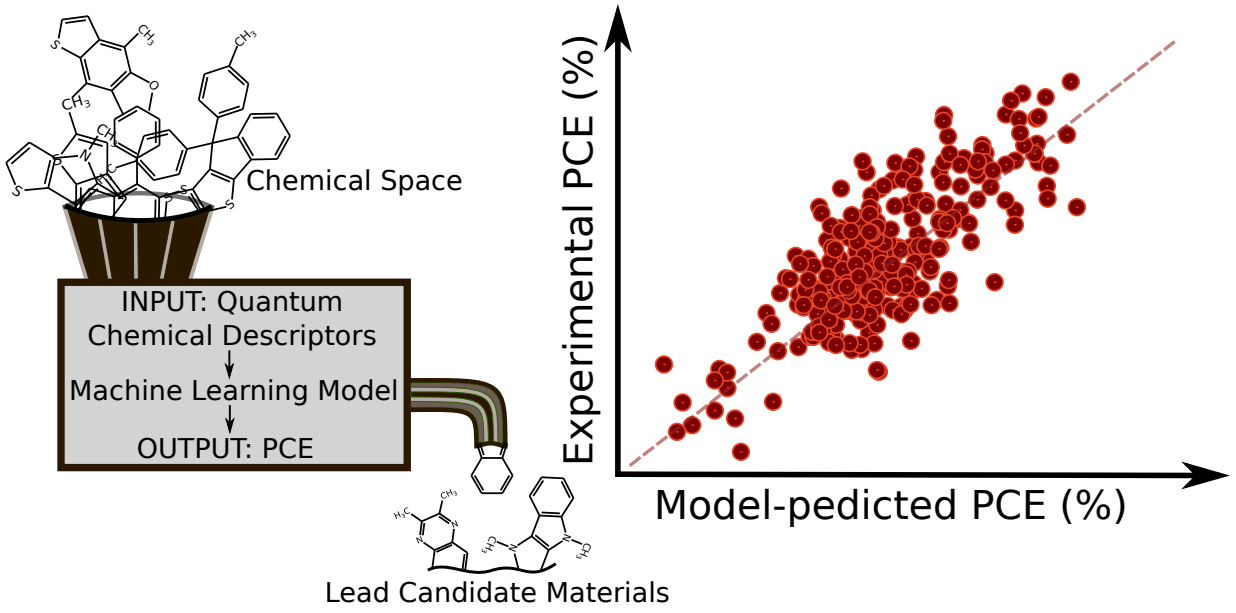
[92] C. H. P. Kumar, K. Ganesh, T. Suresh, A. Sharma, K. Bhanuprakash, G. D. Sharma, M. Chandrasekharam, *RSC Adv.* **2015**, *5*, 93579.

[93] N. M. Kronenberg, M. Deppisch, F. Wurthner, H. W. A. Lademann, K. Deing, K. Meerholz, *Chem. Commun.* **2008**, 6489–6491.

[94] H. Li, J.-L. Brédas, C. Lennartz, *J. Chem. Phys.* **2007**, *126*, 164704.

[95] L. Breiman, J. H. Friedman, R. A. Olshen, C. J. Stone, *Classification and Regression Trees*, Wadsworth and Brooks, Monterey, CA, **1984**.

Table of Contents:



Prediction of efficiency of organic solar cells by machine learning using relevant microscopic properties of donor/acceptor molecules as descriptors.