**Effect of increasing the descriptor set on machine learning prediction of small-molecule-based organic solar cells**

Zhi-Wen Zhao,a,b Marcos del Cueto,a Yun Geng,b Alessandro Troisi\*a

aDepartment of Chemistry, University of Liverpool, Liverpool, L69 3BX, UK. Email: a.troisi@liverpool.ac.uk

bInstitute of Functional Material Chemistry, Faculty of Chemistry, Northeast Normal University, Changchun, 130024, Jilin, P. R. China.

**Abstract**

In this work, we analyzed a dataset formed by 566 donor/acceptor pairs, which are part of organic solar cells recently reported. We explored the effect of different descriptors in machine learning (ML) models to predict the power conversion efficiency (PCE) of these cells. The investigated descriptors are classified into two main categories: structural (topology properties) and physical descriptors (energy-levels, molecular size, light absorption and mixing properties). In line with previous observations, ML predictions are more accurate when using both structural and physical descriptors, as opposed to only using one of them. We observed that ML predictions are also improved by using larger and more varied data sets. Importantly, the structural descriptors are the ones contributing the most to the ML models. Some physical properties are highly correlated with PCE, although they do not improve notably the ML prediction accuracy, as they carry information already encoded in the structural descriptors. Given that various descriptors have significantly different computational costs, the analysis presented here can be used as a guide to construct ML models that maximize predictive power and minimize computational costs for screening large sets of OSCs candidates.

1. **Introduction**

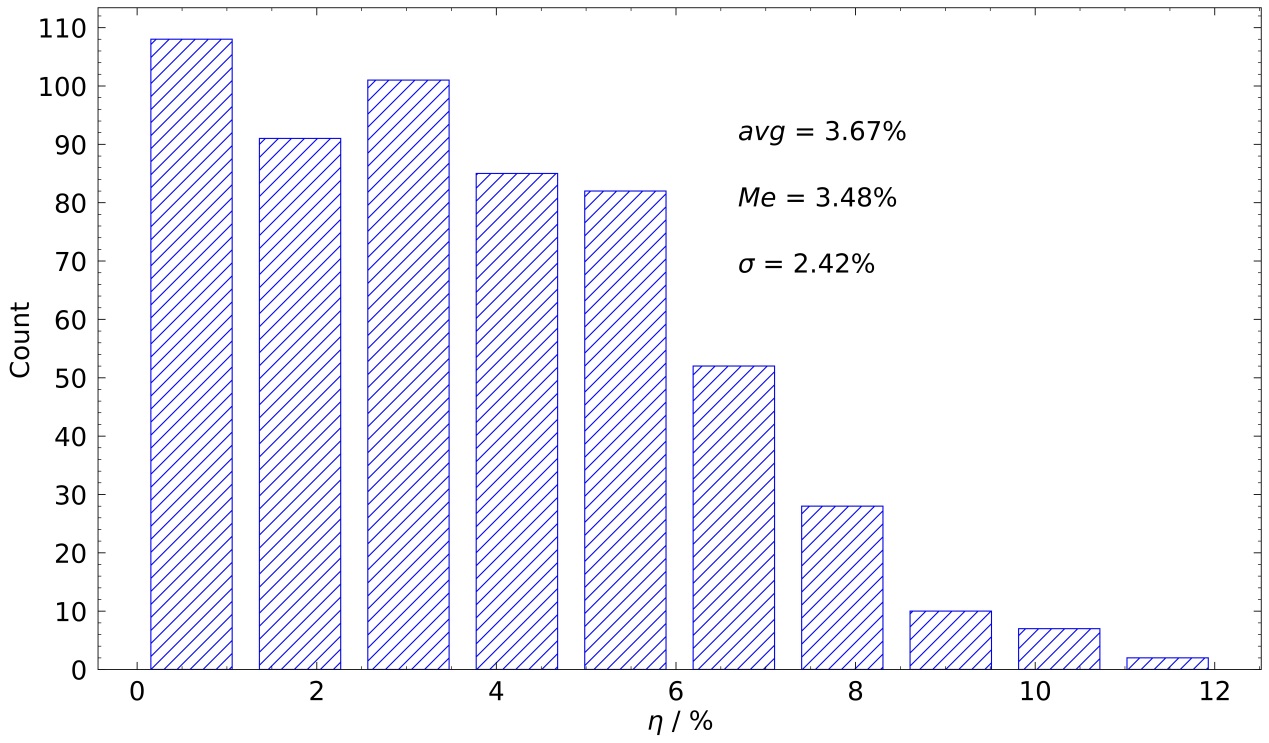
Organic solar cells (OSCs) have been greatly developed in the past decades owing to various chemical materials and diverse combinations of donor (D) and acceptor (A).1, 2 However, the findings of more efficient materials and suitable combinations of D and A largely depend on the numerous trial and error experiments conducted by a plethora of scientists.3-5 The underlying factors pertaining to performance of OSCs are complex because the understanding of the relationship between chemical structure, D/A morphology at the interface and the performance of OSC devices remains poor until now, which makes the improvements of power conversion efficiency (PCE) more challenging.6-11 Thus, a better predictivity of the performance of OSCs and a good method for data-driven molecular design is highly desired.12-14 Such predictive models would be particularly precious for the development of non-fullerene acceptors,15-17 one of the most active areas in the current development of organic photovoltaics.

As a simple model, Scharber’s model18 has often been used to estimate the performance of OSCs. However, it has some limitations since it was proposed for PC61BM while the increased prevalence of non-fullerenes calls for a more general model. In addition, the complexity of OSC devices requires additional parameters to characterize the performance, and not just the energies of frontier molecular orbitals (FMOs) of organic molecules. As a result, this model has been proved to have a very poor correlation with the recent experimental data,19, 20 and many machine learning (ML) methods have been tested to give a better predictivity of device parameters.19-25 In general, most of ML works have focused on bilayer or ternary OSCs with small molecules or polymers as donors, and fullerene and its derivatives as acceptors. Only a few ML models consider experimental non-fullerene acceptor molecules19, 26 one of which was reported by some of us, and tried to look for the best combinations of D/A pairs in OSCs. Meanwhile, the hidden design principles in most of ML works only focused on electronic properties.21, 27, 28 For example, Lee reported the relationship between device parameters of ternary OSCs and the electronic properties with different kinds of ML learning algorithms,22 in which only FMOs properties were considered. Therefore, some previous works ignore important factors such as photophysical properties and did not consider complex relationships between different properties and device parameters. Fortunately, in recent works,24, 29 the correlation between different kinds of physical parameters and optimal performance of OSCs has been enhanced through ML analysis. For instance, in recent work19 published by some of us, the descriptors of structural similarity metrics and five electronic parameters for donor and acceptor molecules were adopted and the predictivity of the ML model has been improved when considering both chemical topology and electronic structure. Also, ML work using Random Forest (RF) and Gradient Boosting (GB) on OSCs reported by Ma et al., has revealed the correlation between molecular properties and device parameters of small molecule-based OSCs (SM-OSCs) considering 13 descriptors chosen from 28 properties, which include excited-state information.29 Although the computational cost in this work is large, properties related to miscibility were not considered. ML models built in these works to predict the device parameters of OSCs are diverse and the optimal performance of OSCs could be enhanced through ML analysis. However, it is known that the morphology of D/A interface, which is largely affected by the electronic and morphological properties of molecular aggregates, is related to exciton dissociation.7, 8, 30 The characterization of the morphology is challenging because of the high-level calculations and complex simulations required, so the properties related to morphology are not taken into account. Thereby, it is worthwhile to find a trade-off method to characterize the morphology to try to describe the D/A packing patterns and interactions.

The goal of this work is to build on the previous successes on ML prediction of organic photovoltaic devices19 and to consider a few new aspects. The dataset is expanded to follow the recent addition to the literature, particularly valuable in the area of non-fullerene acceptors where the explored chemical space has recently become more diverse, and the predictive ability of ML is consequently expected to improve. Furthermore, in this work, we also expand the number of descriptors, with properties like excited-state properties, morphology parameters and other relevant parameters. Some of the additional features considered in this work include (i) information on the morphology of the D/A interface and (ii) excited-state properties of the component materials. Rather than considering just a larger fixed set of descriptors, we analyze in detail the correlation between descriptors and devise a range of strategies to select them. In this way, we can identify the best concise representation of the D/A pair in terms of accuracy and also computational cost. Through the detailed work presented here, we could explain why there seems to be a limit to the predictive ability when descriptors highly correlated with the PCE are also included.

1. **Data gathering**

The dataset obtained searching the literature includes 566 organic donor-acceptor (D/A) pairs, with 513 distinct donors and 33 different acceptors (such as C60, PC61BM, PC71BM, ITIC, IDTBR, IDIC, PDIs etc.). One can find 1080 distinct entries in the Web of Science31 for the query “(solar cell OR heterojunction OR photovoltaic) AND small molecule” for the time period [01/01/2015 - 11/07/2019]. Any articles related to perovskite, ternary or additive, polymer or copolymer and theoretical articles were ignored, which means only data on heterojunction with small molecule donors and acceptors have been considered. There are no fundamental limitations in using ML methods for predicting properties of polymeric materials, and several advances have been made in this direction. However, there are some challenges not yet solved that are not present in the prediction of molecular materials properties (difficulties in fingerprinting,32 univocally computing the electronic properties and computing polymer processing information), which make it not advisable to combine information of the polymers and small molecules in the same model. The available device parameters (open-circuit voltage (*V*OC), short-circuit current (*J*SC), fill factor (*FF*) and power conversion efficiency (PCE)) are collected for each D/A system in journals with sufficiently broad diffusion which, for the present data we define as those having a 2019 impact factor greater than 2.5. Average PCE value is considered when the device has multiple reported PCE values and any duplicate data was removed from the database. Keeping molecules that differ only in their alkyl side chains may result in artificially good performance due to having very similar entries, so molecules differing in their alkyl side chains are considered only once in the dataset. The distribution of PCE in the dataset is shown in Figure 1. There is, as expected, a prevalence of low efficiency cell and a comparatively smaller number of entries with efficiency larger than 9%. The average value of PCE is 3.67% with median PCE value of 3.48% and standard deviation (σ) of 2.42%.



**Figure** **1**. Distribution of reported efficiencies for D/A pairs considered in the dataset.

1. **Methods**

**3.1. Descriptors**

After defining the data set, the next step for the construction of our ML model is the definition of a set of properties (also referred indistinctively as descriptors or features) to characterize all systems studied. In this work we consider a set of properties larger than the ones considered so far in literature.19, 23, 24, 29 In the preliminary phase of this study, we initially exclude descriptors that are highly correlated with each other, e.g. descriptor HOMO-LUMO gap was removed because it is too highly correlated with the HOMO energy level. After studying the correlation between each of them and other OPV parameters, we consider various strategies to select the best descriptors for our ML model.

The descriptors have been divided into five groups:

(1) ***Molecular topology***. Connectivity between atoms defines a chemical species uniquely and it can be used to provide a quantitative measure of the “chemical distance” between two species. The chemical structure is encoded in the SMILES representation of a molecule and, from this, the Morgan fingerprint33 (with radius 2 and 2048 bits) was used as a descriptor of each species, the choice is based on discussions present also in other works.19, 23, 34

(2) ***Properties related to molecular size***. The molecular weight *W*, and the molecular polarizability *P* (often related to size) are included as they influence the miscibility of the materials and the nature of the D/A interface.35, 36 The reorganization energy *λ* of holes (for donors) and electrons (for acceptors) related to the size of the conjugated portion of the molecule37 may affect transport and change separation dynamics.38

(3) ***Properties related to molecular energy-levels***. Orbital energies of HOMO (*EHOMO*), LUMO (*ELUMO*) are obviously essential, and the separation between nearest frontier orbitals (*EH=EHOMO**EHOMO*-1 and *EL=ELUMO*+1*ELUMO*) was shown to improve ML models.19, 24, 39, 40 The orbital energies are obtained from quantum chemical calculations, considering that they must be homogeneous “descriptors” of such properties that do not depend on the availability of data or the specific way they have been reported. Besides, the fundamental gap *Efund* for donor molecule, which is defined as the difference between vertical ionization potential (*IP*(*v*)) and electron affinity, is related to electron-hole separation.29 As we know, hole-electron binding energyis calculated by the energy difference between *Efund*and optical gap. A smaller value of hole-electron binding energy is favourable for the migration of charge into D/A interface.29, 41 Moreover, excitation energy of the lowest-lying triplet state of the donor (*ET1*) was considered because the triplet loss channel could influence *V*OC and increase charge-recombination rates. If the energy difference between *ET1* and the energy of charge-transfer state is larger than 0.1 eV, the recombination process will be enhanced.24, 42

(4) ***Absorption-properties***. We consider the excited energy of the lowest-excited state(*ES1*), thesinglet state with the largest oscillator strength (*Efmax*) and the sum of oscillator strength (Σ*f*) of the excited states with absorption wavelength longer than 300 nm of donor and acceptor. These parameters are expected to have a relation with *J*SC in the solar cell, as it can characterize the photo-absorption ability of molecules.24, 29, 40, 43 Σ*f* might also be relevant because it takes the absorption-properties of non-fullerenes into consideration.

(5) ***Mixing Properties***. The morphology of the bulk heterojunction is known to influence the PCE and depends on the mixing properties of donors and acceptors.9, 44-48 The dipole moment (*μ)* for donor and acceptor, obtained from electronic structure optimization calculations, is a first indicator of the electrostatic interaction between molecules, which is related to miscibility.49 A range of additional descriptors related to miscibility have been selected from the literature of drug discovery, where the solubility of small molecules in different environments plays a key role.50, 51 For example, the number of H-bond acceptors (H-bond acc.) for donor and acceptor should be globally associated with solubility and permeability,52 and can be readily extracted from the chemical structure. The number of rotatable bonds (Rot. Bonds) for donor and acceptor is the number of bonds in a molecule allowing free rotation around themselves, which is defined as any single bond, not in a ring, bound to a nonterminal heavy atom and they are known to determine the flexibility of the molecule and its solubility.53 Finally, as the partition coefficient between n-octanol and water (Po/w) is a classical description of lipophilicity, we include among the descriptors a computed estimate of Log(Po/w) for donors and acceptors, indicated as XLOGP3.54

Full computational details for the properties above are given in the part of computational methods and Table S1 in the Supporting Information (SI). Electronic structure calculations were computed at the B3LYP/6-31G\* level, with the exclusion of ** where the smaller basis set 3-21G\* was used which have been proved in other papers.55, 56 The energy levels appear to be very marginally affected by introducing dispersion correction (with B3LYP-D3) in the optimization (RMSE of HOMO and LUMO of 0.029 and 0.016 eV on a subset of 10 molecules). This is because the energy levels are determined by the rigid conjugated portions of the molecule, which is not much affected by the dispersion correction. The excited state properties are computed at the M06-2X/6-31G\* level to ensure a better description of electronic excited states with CT character.57 Fingerprints have been computed with the RDKit package58 and the miscibility properties have been generated with the code SwissADME50 (with the exception of several molecules computed with XLOGP359).

* 1. **Distance metrics**

The different algorithms discussed in the next section have, as a common feature, the need of defining a “distance” between systems described by a set of descriptors. Extending previous work,19, 23 we can calculate the distance between two donor-acceptor pairs, **p***ij* and **p***mn*, as a linear combination of the distance based on physical properties (*Dph*), and the distance based on the chemical structure (i.e. fingerprint) separately for donor (*Dfpd*) and acceptor (*Dfpa*):

 (1)

where *γ*1, *γ*2 and *γ*3 are three hyper parameters. In this work, different hyper parameters are defined to tune the relative importance of physical and fingerprint distance, and we have considered three distinct cases: i) use only physical descriptors (*γ*2 = *γ*3 =0), ii) use only structural descriptors (*γ*1 = 0) and iii) use both physical and structural descriptors. The values of (*γ*1, *γ*2, *γ*3) were optimized by minimizing the RMSE of the test set with a differential evolution algorithm.60

The physical distances were calculated as the Euclidean distance between the vectors containing physical properties:

 (2)

The fingerprint distances61 were calculated from the Tanimoto similarity index (T)62 between the vectors containing the Morgan fingerprints of the corresponding donor and acceptor (*dfp* and *afp*) :

 (3)

 (4)

 (5)

**3.3. Machine Learning (ML) Algorithms**

We used the ML algorithms (i.e. *k*-NN, KRR, SVR), as implemented in scikit-learn python package.63 For each of the these algorithms, we have divided our data set into a training set to train the specific model, and a test set to validate the model’s predictions. This division of training and data sets was performed with the leave-one-out (LOO) method. The data were scaled to have a null average and standard deviation equal to unity. Comparing the predicted and actual value of each point in the test set for each LOO iteration, we can obtain a RMSE value. We chose the optimum hyper parameters of each algorithm as those that minimize the RMSE. In the case of *k*-NN regression, we chose the optimum number of neighbors, *k*, from a list of possibilities (*k*=3-9). The KRR and SVR hyper parameters were optimized via a differential evolution algorithm,60 as implemented in SciPy.64

(i) *k*-Nearest Neighbours (*k*-NN)

Within the *k*-NN regression algorithm,65 the value of a configuration is calculated as the weighted average of its nearest *k* neighbours, where the weights are inversely proportional to the distance between the neighbour configurations. Distance metrics, as described above, are able to measure how ‘close’ two points are to each other. Different values of *k* were tested, as shown in Figure S1, to reduce the errors while maintaining the predictivity. In our case, the RMSE was minimized for *k*=3.

(ii) Kernel Ridge Regression (KRR):

Kernel ridge regression (KRR) uses the kernel trick to allow for a non-linear version of regularised least squares. KRR can predict the target values ****of a new configuration **p***mn* as:

 (6)

where α is a regularization hyperparameter, *I* is the identity matrix, *K* and ** can be obtained from

 (7)

The distance of physical and structural properties described above are used to compute *f* mapping vectors into a scalar which can be described as

 (8)

where *f* is the kernel function that maps the coordinate vectors to a scalar. The particular form also determines the non-linearity introduced in the procedure. A kernel matrix can be interpreted as matrix of distances between couples of configurations in input space.

(iii) Support Vector Regression (SVR):

The SVR algorithm66 uses the same kernel that we have described above for KRR. The SVR algorithm is very similar to KRR and the main difference is that SVR uses the epsilon-insentive loss, instead of the squared error loss. The optimized hyper parameters in SVR include the regularization parameters *C* and *ε*, which specifies the region within which no penalty is associated in the training loss function.

1. **Results**

**4.1. Correlation analysis**

An initial insight on the data can be obtained considering the correlation between the computed descriptors, and between them and the experimental PCE, *V*OC, *J*SC and *FF* values. The correlation matrices are shown in Figure 2 and Figure 3, separated for easier representation for the descriptors of donors and acceptors. Separating the correlation matrixes is useful also because there are 513 distinct donors but only 33 distinct acceptors, so the statistics for the latter will not be as good as that of the donors. The correlation between descriptors and the solar cell parameters is an excellent indicator of their importance. The data shows that features related to molecular size, optical absorption and mixing properties have higher correlations with performance of OSCs than ground-state properties, which suggests that their inclusion in the model can be useful. Some mixing properties of the donor, like Rot. Bonds-D and XLOGP3-D have relatively high correlation (0.31 and 0.32) with *FF*, expected because the *FF* reflects the quality of the heterojunction morphology (the correlation with the same property of the acceptor is lower).

Some of the descriptors are strongly correlated with each other, e.g. both molecular mass *W* and polarizability *P* are positively correlated with the miscibility parameter Rot. Bonds and XLOGP3, and is therefore positively correlated with *FF* and PCE. Analogously, the fundamental gap *Efund* is negatively correlated with the miscibility parameter and therefore with *FF* and *PCE* too. Such correlations cannot be used to establish a relation of causality between properties and device characteristics but they can be convenient to exclude, in future studies, parameters that are expensive to compute (e.g. *P*) but are highly correlated with parameters easy to compute (e.g. *W*). Notably, *FF* and *J*SC have a much larger correlation coefficient with PCE than *V*OC, as shown in Figure 2, probably because there is a smaller variability in *V*OC. This observation suggests that it is more important to develop predictors than to capture the physics determining *FF* and *J*SC.

The correlation among descriptors can give additional insight and can be used to remove descriptors that contain very similar information,[67-68]. For example, the correlation coefficient between *P*-D and *W*-D (in the same group) is 0.97, and *P*-D is preferable because of its greater correlation with PCE. Similarly, the correlation coefficient between *E*S1-D and *E*fmax-D (in the group of absorption) is 0.63 and *E*S1-D is preferable. Firstly,the strong correlation between *ES1*-D and *ET1*-D is mostly because higher singlet implies higher triplet energy. *ES1*-D also has a large correlation coefficient with *Efmax*-D, which implies that most of the investigated molecules have strongest absorption at the first excited state. Another group of descriptors (*W*-D, *P*-D, Σ*f*–D, Rot. Bonds-D and XLOGP3-D) are highly correlated for physical reasons. Long side chains present in many molecules increase polarizability, molecular weight, rotatable bonds, and solubility in organic phase.



**Figure 2.** Pearson’s correlation coefficient for all combinations of investigated descriptors of donor molecules or device parameters. Blocks corresponding to groups of parameters are outlined.



**Figure 3.** Pearson’s correlation coefficient for all combinations of investigated descriptors of acceptor molecules or device parameters.

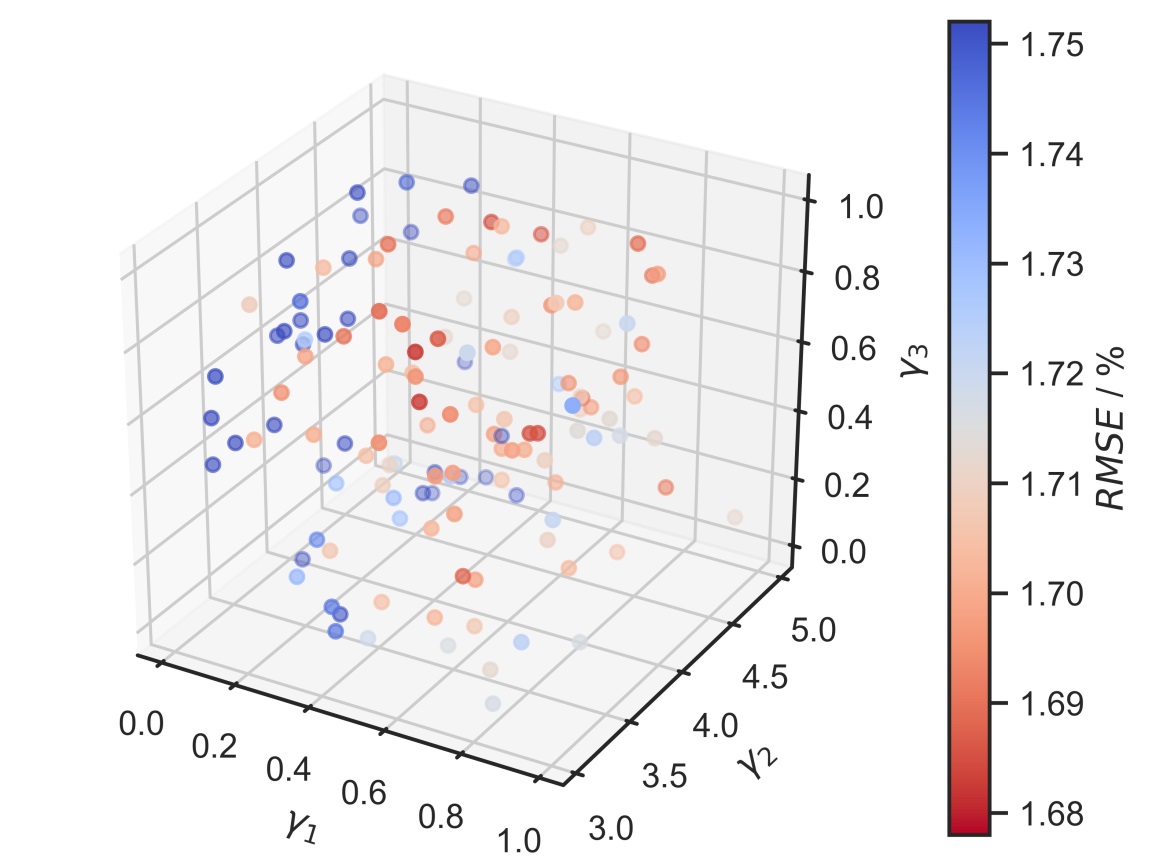
**4.2. ML models and discussions**

Due to the large number of descriptors used in this work, it would be extremely time consuming to make a systematic analysis of all possible combinations. To explore the effect of new descriptors introduced in this work, we started by using the same descriptors adopted in a previous work.19 Then we explored the role of adding further descriptors in different ways and we tried to find descriptors that are as cheap as possible to obtain and give as much information as possible.

We start by considering just five physical descriptors (*HOMO*-D, *LUMO*-D, *LUMO*-A, *λh*,*λe*) and the structural descriptors (fingerprint), combined and by themselves, to compare with the results presented in previous works19, 23 with a smaller data set (therefore considering the effect of the inclusion of more recent data) and two cross validation approaches. Figure 4 shows the effect of considering physical, structural and combined physical and structural descriptors in a *k*-NN model with LOO and 5-fold cross validation and the detailed values (also for 10-fold cross validation) are shown in Table S2-S3. The similarity of the results with the two cross-validation approaches allows one to consider the simplest (LOO) when considering a variety of ML models as we will do below. Another important observation, in agreement with previous work,19 is that ML based just on physical descriptors is not useful for practical purposes, while the majority of the “learning” is done via the structural descriptor alone. Figure 4 also shows the performance of the three ML algorithms considered here. When we just use structural descriptors the accuracy is very similar with all three methods; for the combined structural plus physical properties, the performance of the *k*-NN algorithm is slightly better. Here, we obtained with *k*-NN a Pearson’s correlation coefficient (*r*) of 0.72, much larger than the one presented in previous work with the same method (*r*=0.61).19 The results obtained with more complex methods (i.e. with more hyper parameters) like KRR and SVR are nearly equivalent to the results of *k*-NN seen in the in Table S4-S5, which suggests that one might make good predictions with simple ML methods if the dataset is complete enough. An alternative visualization of the relative importance of the different descriptors is given in Figure 5, where the RMSE for different values of the hyper parameters are given, including their optimized value for *k*-NN at γ1=0.65, γ2=4.08 and γ3=0.97. Looking at Equation 1, one can see that these hyper parameter values indicate quantitatively that the fingerprints of donors give the largest contribution to prediction accuracy, with the contribution of the fingerprint of acceptors and the five physical descriptors having a reduced, but still meaningful, contribution. Note that this larger contribution of donor fingerprints might be partially caused by the greater number of unique donors in the field of organic photovoltaics (see further analysis in the SI, Table S6). Further developments in the field, with a larger number of unique acceptors, might allow for more detailed analyses that result in slightly different relative weights of donor’s and acceptor’s fingerprints.



**Figure 4.** *k*-NN analysis with leave-one-out (LOO) and 5-fold cross validation and KRR and SVR leave-one-out (LOO) predictions of PCE obtained with basic five physical descriptors and two structural descriptors.

****

**Figure 5.** Hyper parameters for photovoltaic efficiency prediction by *k*-NN algorithm with leave-one-out (LOO) cross validation using basic five physical descriptors and fingerprints.

As described in the above section, some descriptors have considerable correlations with parameters of OSCs. Besides structural descriptors, we tried to select physical descriptors according to the correlations and the group of descriptors to have good prediction of PCE. Table 1 contains the key results (*r* and RMSE) of a range of ML models obtained with different descriptors. In Table 1, we show the ML performance using five physical descriptors in Set 0, according to our previous work.19 For Set 1, donors’ physical descriptors (Δ*EL*-D, *Efund*-D, *P*-D, Σ*f*–D and XLOGP3-D) are chosen from the correlation matrix based on introducing descriptors from different groups with the largest correlation with the PCE, added to the fingerprint descriptors and labelled as “5-D-Dph+2Dfp”. Then we explore the effect of increasing the number of donors’ physical descriptors to seven (Δ*EL*-D, *Efund*-D, *P*-D, Σ*f*–D, XLOGP3-D, *LUMO*-D and *ES1*-D) and sixteen (*HOMO*-D, *LUMO*-D, Δ*EH*-D, Δ*EL*-D, *Efund*-D, *λh*, *P*-D, *W*-D, *ES1*-D, *ET1*-D, *Efmax*-D, Σ*f*–D, *μ*-D, Rot. Bond*s*-D, H-bond acc*.*-D and XLOGP3-D), which are labelled as “7-D-Dph+2Dfp” and “16-D-Dph+2Dfp”, respectively, also in Set 1. The correlation *r* and RMSE do not change much for these new descriptors compared to the basic five descriptors. In Table S3, we show that there is little improvement also on the prediction of the *FF*. However, the choice is not completely indifferent as some models like “7-D-Dph+2Dfp” are computationally much cheaper than “5Dph+2Dfp”, which require the expensive calculation for reorganization energy.

We explored separately the effect of physical descriptors of acceptors in Set 2 and select five descriptors (Δ*EH*-A, *P*-A, Σ*f*-A, H-bond acc.-A and XLOGP3-A) according to correlation matrix and find that the result becomes less accurate. When we add all of fifteen acceptors’ descriptors, the result almost does not change compared to the former five descriptors. To probe the effect of different combinations of donors and acceptors descriptors, we tried to include descriptors of both donors and acceptors chosen from correlation matrix and labelled them as “10Dph(5D+5A)+2Dfp”, “12Dph(7D+5A)+2Dfp” and “16-D-15-A-Dph+2Dfp” in Set 3, with no benefit to the results. Thus, we choose the five basic descriptors plus extra physical descriptors from each Set according to the correlation coefficient, shown in Set 4 in Table 1. Similarly, there is no appreciable improvement if physical descriptors from each group according to the correlation coefficient are included (method “7Dph(5basic+1D+1A)+2Dfp” in Table 1).

The results obtained with the descriptors labelled as “9Dph(5basic+2D+2A)+2Dfp” (nine descriptors consisting of the five basic ones, two additional absorption descriptors (Σ*f*–D and Σ*f*–A), two additional miscibility descriptors XLOGP3-D and XLOGP3-A) and “11Dph(5basic+3D+3A)+2Dfp” (same as “9Dph(5basic+2D+2A)+2Dfp”, plus two descriptors in the molecular size group, *P*-D and *P*-A) are slightly better, with equivalent *r* values, and slightly smaller RMSE values. However, further increasing the numbers of descriptors (see for examples the descriptors labelled as “13Dph(5basic+4D+4A)+2Dfp” expanded with *Efund*-D and H-bond acc. A, and “15Dph(5basic+5D+5A)+2Dfp” expanded with Δ*EL*-D andΔ*EH*-A) are not favourable. Although the prediction results are improved when considering physical descriptors in addition to the structural ones, the additional descriptors only change slightly the accuracy of our model, as presented in more detail in Table S7. We have performed this analysis to find the most convenient combination of descriptors labelled as “9Dph(5basic+2D+2A)+2Dfp” with *k*-NN. We show in Table S8 that KRR and SVR have a similar prediction capacity when using these descriptors. We also show a detailed description of the selected descriptors and the accuracy change when considering only structural or physical descriptors in the SI.

**Table 1.** The correlation *r* and RMSE for structural descriptors plus different combinations of physical descriptors. (Set 0: five basic physical descriptors; Set 1: donors’ physical descriptors; Set 2: acceptors’ physical descriptors; Set 3: physical descriptors of donors and acceptors; Set 4: five basic descriptors plus descriptors selected from different groups according to the correlation matrix.) Results correspond to PCE predictions obtained with *k*-NN (*k*=3) regression and LOO cross validation.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Number of descriptors | Descriptors | *r* | RMSE (%) |
| Set 0 | 7 | 5Dph+2Dfp | 0.7248 | 1.6827 |
| Set 1 | 7 | 5-D-Dph+2Dfp | 0.7104 | 1.7229 |
| 9 | 7-D-Dph+2Dfp | 0.7171 | 1.7005 |
| 18 | 16-D-Dph+2Dfp | 0.7187 | 1.6921 |
| Set 2 | 7 | 5-A-Dph+2Dfp | 0.6889 | 1.7859 |
| 17 | 15-A-Dph+2Dfp | 0.6880 | 1.7879 |
| Set 3 | 12 | 10Dph(5D+5A)+2Dfp | 0.7029 | 1.7416 |
| 14 | 12Dph(7D+5A)+2Dfp | 0.7088 | 1.7251 |
| 33 | 16-D-15-A-Dph+2Dfp | 0.7200 | 1.6895 |
| Set 4 | 9 | 7Dph(5basic+1D+1A)+2Dfp | 0.7251 | 1.6772 |
| 11 | 9Dph(5basic+2D+2A)+2Dfp | 0.7286 | 1.6700 |
| 13 | 11Dph(5basic+3D+3A)+2Dfp | 0.7260 | 1.6754 |
| 15 | 13Dph(5basic+4D+4A)+2Dfp | 0.7225 | 1.6845 |
| 17 | 15Dph(5basic+5D+5A)+2Dfp | 0.7150 | 1.7064 |

Taken computational cost and the prediction accuracy in terms of *r* and RMSE into account, some of the best results are obtained when using seven descriptors chosen from correlation matrix and the five basic descriptors plus *Σf* and XLOGP3 of donor and acceptor (“7-D-Dph+2Dfp” and “9Dph(5basic+2D+2A)+2Dfp” in Table 1). In Figure 6, we show the PCE value predicted by the five basic descriptors (“5Dph+2Dfp” in Table 1) seven descriptors (“7-D-Dph+2Dfp”) and descriptors labeled “9Dph(5basic+2D+2A)+2Dfp”, with *k*-NN prediction, as a function of the actual value for these three good-performing sets of descriptors. As suggested by previous works69, 70, the learning curves for the best descriptors combination are reported in Figure S2 to confirm that the models are not affected by fit biases. Note that one could also use learning curves to guide hyper-parameters optimizations, or perform further optimization of the hyperpameter in the region of greatest interest as proposed for example in ref.71. Considering our best results in Table 1, we have improved the correlation *r* with respect to our previous work19 and the accuracy of our results are comparable to alternative ML methods proposed by others.72-74

**C:\Users\wuy\Desktop\revise-data\data-redo-3-5\pce-real-pred-knn-loo-different-descriptors\knn-3set-result.tif**

**Figure 6.** *k*-NN leave-one-out (LOO) predictions of PCE. (Left: using the “5Dph+2Dfp” descriptors; Middle: using the “7-D-Dph+2Dfp” descriptors; Right: using the “9Dph(5basic+2D+2A)+2Dfp” descriptors).

However, the slightly improvement of new descriptors should be analyzed deeply. We added more hyper-parameters as more heterogeneous properties included in the ML process to see the effect of different weight of descriptors. In this condition, we can define the distance *D* between neighbor points *i* and *j*, as a linear combination of donor distance (*Dfpd*), acceptor distance (*Dfpa*) and physical distances (*Dph*):



Where *Dphk* is the Euclidean distance between the vectors containing an arbitrary number of descriptors. This allows more flexibility as one can use as many weights and descriptors as one wants. This distance is calculated as:. However, the results generated by optimizing more hyper parameters have not improved the RMSE by more than 0.02%, as seen in the Table S9. Moreover, we have tested alternative choices of the descriptors by feature elimination in Table S10-S11 and Figure S375 and discuss this as well in the SI, although they do not change the above conclusions. This information is complementary to the results in Table 1 since the improvement of the quality of the prediction when adding descriptors provide a different indication of the descriptor’s importance. (further methods are given in Ref. 68) However, since fingerprints are the major contributors to the prediction, the ranking of the other physical descriptors is not too informative in this case.

The results presented in this work suggest that it is possible to improve the predictive ability of ML method by introducing new parameters, but the improvement is much less than one would expect considering that some of the novel parameters have a very strong correlation with PCE or *FF*. We believe that this is due to the fact that fingerprints are able to encode most of the information that other physical parameters include. As an example, we show in Figure S4 that one can predict XLOGP3, from just the fingerprint using a *k*-NN algorithm, resulting in correlation coefficients of 0.87 for donors and 0.92 for acceptors. Also, it is possible to predict a measure of absorption intensity (Σ*f*) from the fingerprints with correlation coefficient 0.83 for donors and 0.99 for acceptors.

A simple application of the method that does not require any additional tool is based on the (manual) formulation of chemically plausible hypotheses of novel materials by combining known fragments. Some recent works on generative models deal with the problem of synthetic cost or ease of synthesis.76, 77 Alternatively, one can screen libraries of existing compounds78 or compounds that are easy to make and based on commercially available precursors.79 In this work, as it is important that the test is unbiased to avoid cherry-picking we considered all reports from WoS on the 4th July 2020 for a period of four weeks prior to the search. Nine small molecule OPV systems were reported in this period and have been used as an additional test set which is reported in Table 2. The results appear encouraging. The nine predicted PCEs are divided into two groups, one group with predicted efficiency lower than 7.7% (four elements) and the other with efficiency larger than 9.4% (five elements). The experimental efficiency is also clustered in a low efficiency set (in the range 9-8.7% and a well separate high efficiency set with efficiency >12.4%). The predictions place each system correctly in the high and low efficiency set with one exception. The rank correlation coefficient between prediction and experiment is 0.72 - in line with the expectation based on cross validation.

**Table 2**. Power conversion efficiency (PCE) predictions of D/A pairs of small molecule OSCs for latest four weeks before 4th July 2020. Predictions obtained with *k*‑NN algorithm (*k*=3) using the “9Dph(5basic+2D+2A)+2Dfp” descriptors.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Index | Donor ID | Acceptor ID | Ref. | Expt. PCE (%) | Predicted PCE (%) |
| 1 | BSCl-C2 | IDIC-4Cl | [80] | 12.40 | 7.03 |
| 2 | 2FDC5T | IDIC-4F | [81] | 8.70 | 7.22 |
| 3 | DC5T | IDIC-4F | [81] | 6.81 | 7.32 |
| 4 | BTRCl | PC71BM | [82] | 9.00 | 7.66 |
| 5 | SM1-S | Y6 | [83] | 12.94 | 9.47 |
| 6 | BTTzR | Y6 | [84] | 13.50 | 9.51 |
| 7 | BTRCl | Y6 | [82] | 13.30 | 9.52 |
| 8 | SM1-F | Y6 | [83] | 14.07 | 9.54 |
| 9 | SM1 | Y6 | [83] | 12.72 | 10.61 |

1. **Conclusions**

In summary, we have constructed a large database of small molecule OPV containing their experimental characteristics and a broad set of computed properties. We have used the database to study the correlation between descriptors and experimental efficiency. We have also analyzed the effect of increasing the number of descriptors on ML prediction efficiency. The results reveal that the larger and more recent dataset could improve the ML prediction accuracy. We should keep in mind that not all ML models have the same computational cost, and some low-cost method, like those including simple physical descriptors and structural features, already have *r* ≈ 0.7. The choice of method depends on the number of hypotheses to be tested. For few test case manually generated one can use the most predictive and expensive method that would take few CPU hours per molecule. For very large databases of hypothetical molecules to be tested, one can perform a first screening based on structural fingerprints (at negligible computational cost) followed by a screening of intermediate accuracy limited to the most promising candidates.

Although we have included excited-state and miscibility descriptors in the ML analysis, the ML prediction accuracy has not improved as one would naively expect. These descriptors encode information that is already present in the structural fingerprints, so their contribution is minimal or even redundant. On the other hand, good predictive power based on chemical fingerprint may prevent the discovery of new molecule with a completely different chemistry making the ML more interpolative than predictive. Retaining a good selection of physical descriptors is therefore necessary to maintain a good predictive ability when the ML model is fed with completely new chemical compounds.

Finally, it should not be forgotten that, while the language of ML is relatively new in the area of materials discovery, the attempt to find correlations between physical properties of the same system has always been at the heart of scientific investigation. For expensive investigations like OPV, many correlations have been observed on the basis of a small number of observations and a study like the one proposed here, because of the large number of descriptors considered, provides the means to test the validity of such correlations on a broad set of observations.

**Supporting information and Data available**

Computational methods, ML results in detail, combinations of descriptors and the results about predictions of excited state properties and miscibility. The raw data and scripts used to reproduce the data in the figures of the main manuscript are available. {https://github.com/marcosdelcueto/MachineLearning\_AcceptorDonor.git}

**Acknowledgements**

Z-W.Z. acknowledges the support of the China Scholarship Council. Z-W.Z. also thanks Dr. Hai-Bin Li for useful suggestions. A.T. acknowledges the support of EPSRC. We thank Mr. Rex Manurung, Mr. Ömer H. Omar and Dr. Daniele Padula for fruitful discussions.

**References:**

1. Lee, C.; Lee, S.; Kim, G. U.; Lee, W.; Kim, B. J., Recent Advances, Design Guidelines, and Prospects of All-Polymer Solar Cells. *Chem. Rev.* **2019,** *119* (13), 8028-8086.

2. Yao, H.; Ye, L.; Zhang, H.; Li, S.; Zhang, S.; Hou, J., Molecular Design of Benzodithiophene-Based Organic Photovoltaic Materials. *Chem. Rev.* **2016,** *116* (12), 7397-457.

3. Zhang, G.; Zhao, J.; Chow, P. C. Y.; Jiang, K.; Zhang, J.; Zhu, Z.; Zhang, J.; Huang, F.; Yan, H., Nonfullerene Acceptor Molecules for Bulk Heterojunction Organic Solar Cells. *Chem. Rev.* **2018,** *118*, 3447-3507.

4. Fu, H.; Wang, Z.; Sun, Y., Polymer Donors for High-Performance Non-Fullerene Organic Solar Cells. *Angew. Chem. Int. Ed. Engl.* **2019,** *58*, 4442-4453.

5. Yao, H.; Wang, J.; Xu, Y.; Zhang, S.; Hou, J., Recent Progress in Chlorinated Organic Photovoltaic Materials. *Acc. Chem. Res.* **2020,** *53*, 822-832.

6. Hu, H.; Li, Y.; Zhang, J.; Peng, Z.; Ma, L.-k.; Xin, J.; Huang, J.; Ma, T.; Jiang, K.; Zhang, G.; Ma, W.; Ade, H.; Yan, H., Effect of Ring-Fusion on Miscibility and Domain Purity: Key Factors Determining the Performance of PDI-Based Nonfullerene Organic Solar Cells. *Adv. Energy Mater.* **2018,** *8* (26), 1800234.

7. Ma, W.; Tumbleston, J. R.; Wang, M.; Gann, E.; Huang, F.; Ade, H., Domain Purity, Miscibility, and Molecular Orientation at Donor/Acceptor Interfaces in High Performance Organic Solar Cells: Paths to Further Improvement. *Adv. Energy Mater.* **2013,** *3* (7), 864-872.

8. Min, J.; Luponosov, Y. N.; Gasparini, N.; Richter, M.; Bakirov, A. V.; Shcherbina, M. A.; Chvalun, S. N.; Grodd, L.; Grigorian, S.; Ameri, T.; Ponomarenko, S. A.; Brabec, C. J., Effects of Alkyl Terminal Chains on Morphology, Charge Generation, Transport, and Recombination Mechanisms in Solution-Processed Small Molecule Bulk Heterojunction Solar Cells. *Adv. Energy Mater.* **2015,** *5* (17), 1500386.

9. Huang, Y.; Kramer, E. J.; Heeger, A. J.; Bazan, G. C., Bulk Heterojunction Solar Cells: Morphology and Performance Relationships. *Chem. Rev.* **2014,** *114* (14), 7006-7043.

10. Ye, L.; Hu, H.; Ghasemi, M.; Wang, T.; Collins, B. A.; Kim, J. H.; Jiang, K.; Carpenter, J. H.; Li, H.; Li, Z.; McAfee, T.; Zhao, J.; Chen, X.; Lai, J. L. Y.; Ma, T.; Brédas, J.-L.; Yan, H.; Ade, H., Quantitative relations between interaction parameter, miscibility and function in organic solar cells. *Nat. Mater.* **2018,** *17* (3), 253-260.

11. Manley, E. F.; Harschneck, T.; Eastham, N. D.; Leonardi, M. J.; Zhou, N.; Strzalka, J.; Chang, R. P. H.; Chen, L. X.; Marks, T. J., Side Chain and Solvent Direction of Film Morphology in Small-Molecule Organic Solar Materials. *Chem. Mater.* **2019**, *31* (20), 8308-8319.

12. Choudhary, K.; Bercx, M.; Jiang, J.; Pachter, R.; Lamoen, D.; Tavazza, F., Accelerated Discovery of Efficient Solar-cell Materials using Quantum and Machine-learning Methods. *Chem. Mater.* **2019**, *31* (15), 5900–5908.

13. Ye, L.; Li, W.; Guo, X.; Zhang, M.; Ade, H., Polymer Side-Chain Variation Induces Microstructural Disparity in Nonfullerene Solar Cells. *Chem. Mater.* **2019**, *31* (17), 6568-6577.

14. Oliynyk, A. O.; Buriak, J. M., Virtual Issue on Machine-Learning Discoveries in Materials Science. *Chem. Mater.* **2019**, *31* (20), 8243-8247.

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. Liu, X.; Li, Y.; Ding, K.; Forrest, S., Energy Loss in Organic Photovoltaics: Nonfullerene Versus Fullerene Acceptors. *Phys. Rev. Appl.* **2019,** *11*, 024060.

17. Zhao, C.; Guo, Y.; Zhang, Y.; Yan, N.; You, S.; Li, W., Diketopyrrolopyrrole-based conjugated materials for non-fullerene organic solar cells. *J. Mater. Chem. A* **2019,** *7*, 10174-10199.

18. Scharber, M. C.; Mühlbacher, D.; Koppe, M.; Denk, P.; Waldauf, C.; Heeger, A. J.; Brabec, C. J., Design Rules for Donors in Bulk-Heterojunction Solar Cells—Towards 10 % Energy-Conversion Efficiency. *Adv. Mater.* **2006,** *18* (6), 789-794.

19. Padula, D.; Simpson, J. D.; Troisi, A., Combining electronic and structural features in machine learning models to predict organic solar cells properties. *Mater. Horiz.* **2019,** *6* (2), 343-349.

20. Sahu, H.; Yang, F.; Ye, X.; Ma, J.; Fang, W.; Ma, H., Designing promising molecules for organic solar cells via machine learning assisted virtual screening. *J. Mater. Chem. A* **2019,** *7* (29), 17480-17488.

21. 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* (10), 2639-2646.

22. Lee, M. H., Insights from Machine Learning Techniques for Predicting the Efficiency of Fullerene Derivatives‐Based Ternary Organic Solar Cells at Ternary Blend Design. *Adv. Energy Mater.* **2019**, 1900891.

23. Padula, D.; Troisi, A., Concurrent Optimization of Organic Donor–Acceptor Pairs through Machine Learning. *Adv. Energy Mater.* **2019,** *9* (40), 1902463.

24. Sahu, H.; Rao, W.; Troisi, A.; Ma, H., Toward Predicting Efficiency of Organic Solar Cells via Machine Learning and Improved Descriptors. *Adv. Energy Mater.* **2018,** *8* (24), 1801032.

25. Cao, B.; Adutwum, L. A.; Oliynyk, A. O.; Luber, E. J.; Olsen, B. C.; Mar, A.; Buriak, J. M., How To Optimize Materials and Devices via Design of Experiments and Machine Learning: Demonstration Using Organic Photovoltaics. *ACS nano* **2018,** *12*, 7434-7444.

26. Lee, M.-H., Robust random forest based non-fullerene organic solar cells efficiency prediction. *Org. Electron.* **2020,** *76*, 105465.

27. 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.

28. Chen, C.; Zuo, Y.; Ye, W.; Li, X.; Deng, Z.; Ong, S. P., A Critical Review of Machine Learning of Energy Materials. *Adv. Energy Mater.* **2020,** *10*, 1903242

29. Sahu, H.; Ma, H., Unraveling Correlations between Molecular Properties and Device Parameters of Organic Solar Cells Using Machine Learning. *J. Phys. Chem. Lett.* **2019,** *10* (22), 7277-7284.

30. Perea, J. D.; Langner, S.; Salvador, M.; Kontos, J.; Jarvas, G.; Winkler, F.; Machui, F.; Gorling, A.; Dallos, A.; Ameri, T.; Brabec, C. J., Combined Computational Approach Based on Density Functional Theory and Artificial Neural Networks for Predicting The Solubility Parameters of Fullerenes. *J. Phys. Chem. B* **2016,** *120* (19), 4431-8.

31. Web of Science, http://www.webofknowledge.com (accessed 11 July 2019).

32. Lin, T.-S.; Coley, C. W.; Mochigase, H.; Beech, H. K.; Wang, W.; Wang, Z.; Woods, E.; Craig, S. L.; Johnson, J. A.; Kalow, J. A.; Jensen, K. F.; Olsen, B. D., BigSMILES: A Structurally-Based Line Notation for Describing Macromolecules. *ACS Cent. Sci.* **2019,** *5* (9), 1523-1531.

33. Rogers, D.; Hahn, M., Extended-Connectivity Fingerprints. *J. Chem. Inf. Model* **2010,** *50* (5), 742-754.

34. Lopez, S. A.; Pyzer-Knapp, E. O.; Simm, G. N.; Lutzow, T.; Li, K.; Seress, L. R.; Hachmann, J.; Aspuru-Guzik, A., The Harvard organic photovoltaic dataset. *Sci. Data* **2016,** *3* (1).

35. Perea, J. D.; Langner, S.; Salvador, M.; Kontos, J.; Jarvas, G.; Winkler, F.; Machui, F.; Gorling, A.; Dallos, A.; Ameri, T.; Brabec, C. J., Combined Computational Approach Based on Density Functional Theory and Artificial Neural Networks for Predicting The Solubility Parameters of Fullerenes. *J. Phys. Chem. B* **2016,** *120* (19), 4431-8.

36. Mothy, S.; Guillaume, M.; Ide, J.; Castet, F.; Ducasse, L.; Cornil, J.; Beljonne, D., Tuning the Interfacial Electronic Structure at Organic Heterojunctions by Chemical Design. *J. Phys. Chem. Lett.* **2012,** *3* (17), 2374-8.

37. Devos, A.; Lannoo, M., Electron-phonon coupling for aromatic molecular crystals: Possible consequences for their superconductivity. *Phys. Rev. B* **1998,** *58* (13), 8236-8239.

38. Oberhofer, H.; Reuter, K.; Blumberger, J., Charge Transport in Molecular Materials: An Assessment of Computational Methods. *Chem. Rev.* **2017,** *117* (15), 10319-10357.

39. Bakulin, A. A.; Dimitrov, S. D.; Rao, A.; Chow, P. C.; Nielsen, C. B.; Schroeder, B. C.; McCulloch, I.; Bakker, H. J.; Durrant, J. R.; Friend, R. H., Charge-Transfer State Dynamics Following Hole and Electron Transfer in Organic Photovoltaic Devices. *J. Phys. Chem. Lett.* **2013,** *4* (1), 209-15.

40. Kuzmich, A.; Padula, D.; Ma, H.; Troisi, A., Trends in the electronic and geometric structure of non-fullerene based acceptors for organic solar cells. *Energy Environ. Sci.* **2017,** *10* (2), 395-401.

41. Brédas, J.-L., Mind the gap! *Mater. Horiz.* **2014,** *1* (1), 17-19.

42. Zhang, J.; Zhu, L.; Wei, Z., Toward Over 15% Power Conversion Efficiency for Organic Solar Cells: Current Status and Perspectives. *Small Methods* **2017,** *1* (12), 1700258.

43. Ip, C. M.; Eleuteri, A.; Troisi, A., Predicting with confidence the efficiency of new dyes in dye sensitized solar cells. *Phys. Chem. Chem. Phys.* **2014,** *16* (36), 19106-19110.

44. Yi, X.; Peng, Z.; Xu, B.; Seyitliyev, D.; Ho, C. H. Y.; Danilov, E. O.; Kim, T.; Reynolds, J. R.; Amassian, A.; Gundogdu, K.; Ade, H.; So, F., Critical Role of Polymer Aggregation and Miscibility in Nonfullerene-Based Organic Photovoltaics. *Adv. Energy Mater.* **2020,** *10*, 1902430.

45. Zhou, Z.; Liu, W.; Zhou, G.; Zhang, M.; Qian, D.; Zhang, J.; Chen, S.; Xu, S.; Yang, C.; Gao, F.; Zhu, H.; Liu, F.; Zhu, X., Subtle Molecular Tailoring Induces Significant Morphology Optimization Enabling over 16% Efficiency Organic Solar Cells with Efficient Charge Generation. *Adv. Mater.* **2020,** *32*, 1906324.

46. Yi, N.; Ai, Q.; Zhou, W.; Huang, L.; Zhang, L.; Xing, Z.; Li, X.; Zeng, J.; Chen, Y., Miscibility Matching and Bimolecular Crystallization Affording High-Performance Ternary Nonfullerene Solar Cells. *Chem. Mater.* **2019,** *31*, 10211-10224.

47. Zhou, R.; Jiang, Z.; Yang, C.; Yu, J.; Feng, J.; Adil, M. A.; Deng, D.; Zou, W.; Zhang, J.; Lu, K.; Ma, W.; Gao, F.; Wei, Z., All-small-molecule organic solar cells with over 14% efficiency by optimizing hierarchical morphologies. *Nat. Commun.* **2019,** *10*, 5393.

48. Zhu, L.; Zhang, M.; Zhou, G.; Hao, T.; Xu, J.; Wang, J.; Qiu, C.; Prine, N.; Ali, J.; Feng, W.; Gu, X.; Ma, Z.; Tang, Z.; Zhu, H.; Ying, L.; Zhang, Y.; Liu, F., Efficient Organic Solar Cell with 16.88% Efficiency Enabled by Refined Acceptor Crystallization and Morphology with Improved Charge Transfer and Transport Properties. *Adv. Energy Mater.* **2020,** 1904234.

49. Petrov, J. G.; Andreeva, T. D.; Moehwald, H., Dipolar Interactions and Miscibility in Binary Langmuir Monolayers with Opposite Dipole Moments of the Hydrophilic Heads. *Langmuir* **2009,** *25* (6), 3659-3666.

50. Daina, A.; Michielin, O.; Zoete, V., SwissADME: a free web tool to evaluate pharmacokinetics, drug-likeness and medicinal chemistry friendliness of small molecules. *Sci. Rep.* **2017,** *7* (1).

51. Daina, A.; Michielin, O.; Zoete, V., iLOGP: a simple, robust, and efficient description of n-octanol/water partition coefficient for drug design using the GB/SA approach. *J. Chem. Inf. Model.* **2014,** *54* (12), 3284-301.

52. Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J., Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Deliv. Rev.* **1997,** *23* (1), 3-25.

53. Veber, D. F.; Johnson, S. R.; Cheng, H.-Y.; Smith, B. R.; Ward, K. W.; Kopple, K. D., Molecular Properties That Influence the Oral Bioavailability of Drug Candidates. *J. Med. Chem.* **2002,** *45* (12), 2615-2623.

54. Cheng, T.; Zhao, Y.; Li, X.; Lin, F.; Xu, Y.; Zhang, X.; Li, Y.; Wang, R.; Lai, L., Computation of Octanol−Water Partition Coefficients by Guiding an Additive Model with Knowledge. *J. Chem. Inf. Model* **2007,** *47* (6), 2140-2148.

55. Brédas, J.-L.; Beljonne, D.; Coropceanu, V.; Cornil, J., Charge-Transfer and Energy-Transfer Processes in π-Conjugated Oligomers and Polymers:  A Molecular Picture. *Chem. Rev.* **2004,** *104* (11), 4971-5004.

56. Nematiaram, T.; Padula, D.; Landi, A.; Troisi, A., On the Largest Possible Mobility of Molecular Semiconductors and How to Achieve It. *Adv. Funct. Mater.* **2020,** *30*, 2001906.

57. Zhao, Y.; Truhlar, D. G., The M06 suite of density functionals for main group thermochemistry, thermochemical kinetics, noncovalent interactions, excited states, and transition elements: two new functionals and systematic testing of four M06-class functionals and 12 other functionals. *Theor. Chem. Acc.* **2008,** *120*, 215-241.

58. Landrum, G., RDKit: Open-source cheminformatics.

59. Cheng, T.; Zhao, Y.; Li, X.; Lin, F.; Xu, Y.; Zhang, X.; Li, Y.; Wang, R.; Lai, L., Computation of Octanol−Water Partition Coefficients by Guiding an Additive Model with Knowledge. *J. Chem. Inf. Model* **2007,** *47* (6), 2140-2148.

60. Storn, R.; Price, K., Differential Evolution – A Simple and Efficient Heuristic for global Optimization over Continuous Spaces. *J. Glob. Optim.* **1997,** *11* (4), 341-359.

61. Pyzer-Knapp, E. O.; Simm, G. N.; Aspuru Guzik, A., A Bayesian approach to calibrating high-throughput virtual screening results and application to organic photovoltaic materials. *Mater. Horiz.* **2016,** *3* (3), 226-233.

62. Bajusz, D.; Racz, A.; Heberger, K., Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations? *J. Cheminformatics* **2015,** *7*, 20.

63. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; Duchesnay, E.; Louppe, G., Scikit-learn: Machine Learning in Python. *J. Mach. Learn Res.* **2012,** *12*.

64. E. Jones, T. O., P. Peterson et al., SciPy: Open source scientic tools for Python. **2001**, <http://www.scipy.org/> (accessed 4 July 2020).

65. Altman, N. S., An Introduction to Kernel and Nearest-Neighbor Nonparametric Regression. *Am. Stat.* **1992,** *46* (3), 175-185.

66. Smola, A. J.; Schölkopf, B., A tutorial on support vector regression. *Stat Comput.* **2004,** *14* (3), 199-222.

67. Yu, Y.; Tan, X.; Ning, S.; Wu, Y., Machine Learning for Understanding Compatibility of Organic–Inorganic Hybrid Perovskites with Post-Treatment Amines. *ACS Energy Lett.* **2019,** *4* (2), 397-404.

68. Holleis, L.; Shivaram, B. S.; Balachandran, P. V., Machine learning guided design of single-molecule magnets for magnetocaloric applications. *Appl. Phys. Lett.* **2019,** *114* (22), 222404.

69.Elton, D. C.; Boukouvalas, Z.; Butrico, M. S.; Fuge, M. D.; Chung, P. W., Applying machine learning techniques to predict the properties of energetic materials. *Sci. Rep.* **2018,** *8* (1), 9059.

70. Ma, X.-Y.; Lewis, J. P.; Yan, Q.-B.; Su, G., Accelerated Discovery of Two-Dimensional Optoelectronic Octahedral Oxyhalides via High-Throughput Ab Initio Calculations and Machine Learning. *J. Phys. Chem. Lett.* **2019,** *10* (21), 6734-6740.

71. Wei, L.; Xu, X.; Gurudayal; Bullock, J.; Ager, J. W., Machine Learning Optimization of p-Type Transparent Conducting Films. *Chem. Mater.* **2019,** *31*, 7340-7350.

72. David, T. W.; Anizelli, H.; Tyagi, P.; Gray, C.; Teahan, W.; Kettle, J., Using Large Datasets of Organic Photovoltaic Performance Data to Elucidate Trends in Reliability Between 2009 and 2019. *IEEE J. Photovolt.* **2019,** 9, 1768-1773.

73. Lee, M.-H., Performance and Matching Band Structure Analysis of Tandem Organic Solar Cells Using Machine Learning Approaches. *Energy Technol.* **2019,** *8*. 1900974.

74. Paul, A.; Furmanchuk, A.; Liao, W. K.; Choudhary, A.; Agrawal, A., Property Prediction of Organic Donor Molecules for Photovoltaic Applications Using Extremely Randomized Trees. *Mol. Inform.* **2019,** *38*, e1900038.

75. Guyon, I.; Weston, J.; Barnhill, S.; Vapnik, V., Gene Selection for Cancer Classification using Support Vector Machines. *Mach. Learn*. **2002,** 46, 389-422.

76. Benjamin, S.-L.; Carlos, O.; Gabriel L., G.; Alan, A.-G., Optimizing distributions over molecular space. An Objective-Reinforced Generative Adversarial Network for Inverse-design Chemistry (ORGANIC). **2017,** *ChemRxiv*, [https://chemrxiv.org/articles/ORGANIC\_ 1\_pdf/5309668](https://chemrxiv.org/articles/ORGANIC_%201_pdf/5309668)(accessed 4 July 2020).

77. Peng, S. P.; Zhao, Y., Convolutional Neural Networks for the Design and Analysis of Non-Fullerene Acceptors. J. Chem. Inf. Model. **2019,** *59* (12), 4993-5001.

78. Padula, D.; Omar, Ö. H.; Nematiaram, T.; Troisi, A., Singlet fission molecules among known compounds: finding a few needles in a haystack. *Energy Environ. Sci.* **2019,** *12*, 2412-2416.

79. Yuan, Q.; Santana-Bonilla, A.; Zwijnenburg, M. A.; Jelfs, K. E., Molecular generation targeting desired electronic properties via deep generative models. *Nanoscale* **2020,** *12*, 6744-6758.

80. Wu, Q.; Deng, D.; Zhou, R.; Zhang, J.; Zou, W.; Liu, L.; Wu, S.; Lu, K.; Wei, Z., Modulation of Donor Alkyl Terminal Chains with the Shifting Branching Point Leads to the Optimized Morphology and Efficient All-Small-Molecule Organic Solar Cells. *ACS Appl. Mater. Interfaces* **2020,** *12*, 25100-25107.

81. Xu, T.; Chang, Y.; Yan, C.; Yang, Q.; Kan, Z.; Singh, R.; Kumar, M.; Li, G.; Lu, S.; Duan, T., Fluorinated oligothiophene donors for high-performance nonfullerene small-molecule organic solar cells. *Sustain. Energy Fuels* **2020,** *4*, 2680-2685.

82. Tang, H.; Chen, H.; Yan, C.; Huang, J.; Fong, P. W. K.; Lv, J.; Hu, D.; Singh, R.; Kumar, M.; Xiao, Z.; Kan, Z.; Lu, S.; Li, G., Delicate Morphology Control Triggers 14.7% Efficiency All‐Small‐Molecule Organic Solar Cells. *Adv. Energy Mater.* **2020,** *10*, 2001076.

83. Qiu, B.; Chen, Z.; Qin, S.; Yao, J.; Huang, W.; Meng, L.; Zhu, H.; Yang, Y. M.; Zhang, Z. G.; Li, Y., Highly Efficient All-Small-Molecule Organic Solar Cells with Appropriate Active Layer Morphology by Side Chain Engineering of Donor Molecules and Thermal Annealing. *Adv. Mater.* **2020,** *32*, e1908373.

84. Wang, Y.; Wang, Y.; Zhu, L.; Liu, H.; Fang, J.; Guo, X.; Liu, F.; Tang, Z.; Zhang, M.; Li, Y., A novel wide-bandgap small molecule donor for high efficiency all-small-molecule organic solar cells with small non-radiative energy losses. *Energy Environ. Sci.* **2020,** *13*, 1309-1317.

**TOC Graphics – Actual size**

