

## Supplementary Materials for

### Machine learning–assisted molecular design and efficiency prediction for high-performance organic photovoltaic materials

Wenbo Sun, Yujie Zheng, Ke Yang, Qi Zhang, Akeel A. Shah, Zhou Wu, Yuyang Sun, Liang Feng, Dongyang Chen, Zeyun Xiao\*, Shirong Lu\*, Yong Li, Kuan Sun\*

\*Corresponding author. Email: kuan.sun@cqu.edu.cn (K.S.); lushirong@cigit.ac.cn (S.L.); xiao.z@cigit.ac.cn (Z.X.)

Published 8 November 2019, *Sci. Adv.* **5**, eaay4275 (2019)  
DOI: 10.1126/sciadv.aay4275

#### This PDF file includes:

Section S1. ML methods and machine language expressions of molecule

Section S2. Process of experiment and proof of the reliability of the ML model

Fig. S1. Introduction to different ML algorithms.

Fig. S2. Chemical structures of the 10 new donor materials.

Fig. S3. Prediction results versus experimental data for the 10 new donor materials.

Table S1. Details of PaDEL descriptors.

Table S2. Details of RDKit descriptors.

Table S3. Complete MACCS fingerprint of P3HT and PTB7.

Table S4. Photovoltaic parameters of OPV devices fabricated with different donor materials.

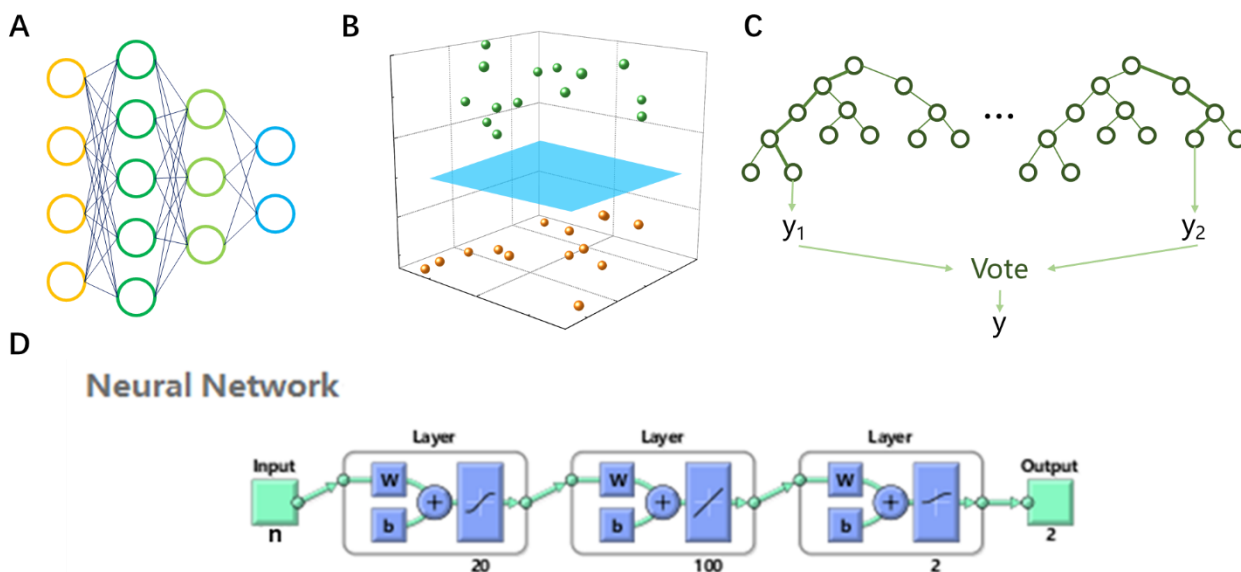
Table S5. Prediction results from DNN, RF, and SVM using Hybridization and FP2 fingerprints as inputs, as well as DNN and RF using Daylight fingerprints.

Table S6. Prediction results from BPNN using Daylight fingerprints when classification threshold is 10%.

References (48, 49)

# Supplementary Materials

## Section S1. ML methods and machine language expressions of molecule



**Fig. S1. Introduction to different ML algorithms.** Schematic diagram of the working principles for (A) artificial neural network, (B) support vector machine and (C) random forest. (D) The structure of the BPNN used in this work.

The basic structure of a multi-layered ANN is shown in fig. S1A. Nodes in one layer are connected with the adjacent layers, and data in ANN flow from one layer to the next (in feedforward approaches). The input of a node is the weighted sum of the output of nodes in the preceding layer. After the operation of an activation function, the node will produce an output. The algorithm of back propagation is applied in ANN to optimize the weights, so that the relationship between the input and output can be learned. One difference between BPNN and DNN is the activation function. Activation functions used in BPNN typically consist of the *logsig* function  $f(n) = \frac{1}{1+e^{-n}}$ , *tansig* function  $f(n) = \frac{2}{1+e^{-2n}} - 1$  or linear *purelin* function  $f(n) = n$ . However, all these functions suffer from the problem that gradient may vanish, which will cause an error during training. Meanwhile, this matter limits the model

going deeper. DNN use the ReLU function  $f(x) = \begin{cases} 0 & (x \leq 0) \\ x & (x > 0) \end{cases}$  as their activation functions

due to its capacity for producing quick convergence. Thus, the number of layers of a DNN can be large, rendering the model a powerful data-processing tool. Notably, a special structure of convolutional neural networks is used in deep learning (25), which is specialized in image input.

SVM (26) can analyze data for classification based on statistical learning theory. In most cases, real-world data is nonlinear, which can cause problems for separating data by hyperplanes. In kernel SVM (as shown in fig. S1B) data can be mapped onto a higher dimensional space called a feature space, so that they can be divided by a hyperplane to achieve the correct classification. A kernel function, which expresses the inner product between any two data points (known as feature vector) in the feature space, is the key in SVM, because it implicitly defines the map between low and high dimensional space.

Calculation of kernel function is based on data in the low dimensional space, but the final result is displayed in the high dimensional space. Thus, complex calculations directly in high-dimensional space can be avoided and nonlinear data can be processed by kernel SVM.

RF (28) is a machine learning method based on decision trees. fig. S1C shows the working principle of RF. Bootstrap aggregating (46) is the key idea in RF. When training a model, each tree randomly chooses multiple samples from training set to form a new subset, and then randomly chooses multiple features from the input to make a decision. Through voting, an output is produced from hundreds of decision trees to provide the best answer.

In this work, BPNN, DNN, SVM and RF were implemented in *Matlab* (27, 29), while deep learning was implemented in *caffe* (48). Our best-performing BPNN contain four layers in total (two hidden layers). The number of neurons in the four layers are n, 20, 100 and 2 respectively, where n is the length of the input. The corresponding activation functions of four

layers are *purelin*, *tansig*, *purelin* and *logsig* (as shown in fig. S1D). We obtain a DNN with two hidden layers through structural design and performance optimization. They have 50 and 1 neurons, respectively. The activation functions in the hidden layers are ReLU. A fine-tuned GoogLeNet (19, 49) is included to establish the deep learning model. We optimized the number of decision trees and other parameters to find the most accurate RF and SVM approaches.

**Table S1. Details of PaDEL descriptors.**

Descriptor	Dimension of descriptors	Type and amount of descriptors
PaDEL	1D	Constitutional descriptors (120)
	2D	Autocorrelation descriptors (346)
		Basak descriptors (42)
		BCUT descriptors (6)
		Burden descriptors (96)
		Connectivity descriptors (56)
		E-state descriptors (489)
		Kappa descriptors (3)
		Molecular property descriptors (15)
		Quantum chemical descriptors (6)
		Topological descriptors (265)
	3D	CPSA descriptors (29)
		RDF descriptors (210)
		Geometrical descriptors (21)
		WHIM descriptors (91)
		3D Autocorrelation descriptors (80)

Available from: [http://www.scbdd.com/padel\\_desc/descriptors/](http://www.scbdd.com/padel_desc/descriptors/)

**Table S2. Details of RDKit descriptors.**

Descriptor	Dimension of descriptors	Type and amount of descriptors
RDKit	1D	Constitutional descriptors (106)
	2D	Connectivity descriptors (12)
		MOE-type descriptors (58)
		Molecular property descriptors (5)
		Topological descriptors (15)

Available from: [http://www.scbdd.com/rdk\\_desc/descriptors/](http://www.scbdd.com/rdk_desc/descriptors/)

**Table S3. Complete MACCS fingerprint of P3HT and PTB7.**

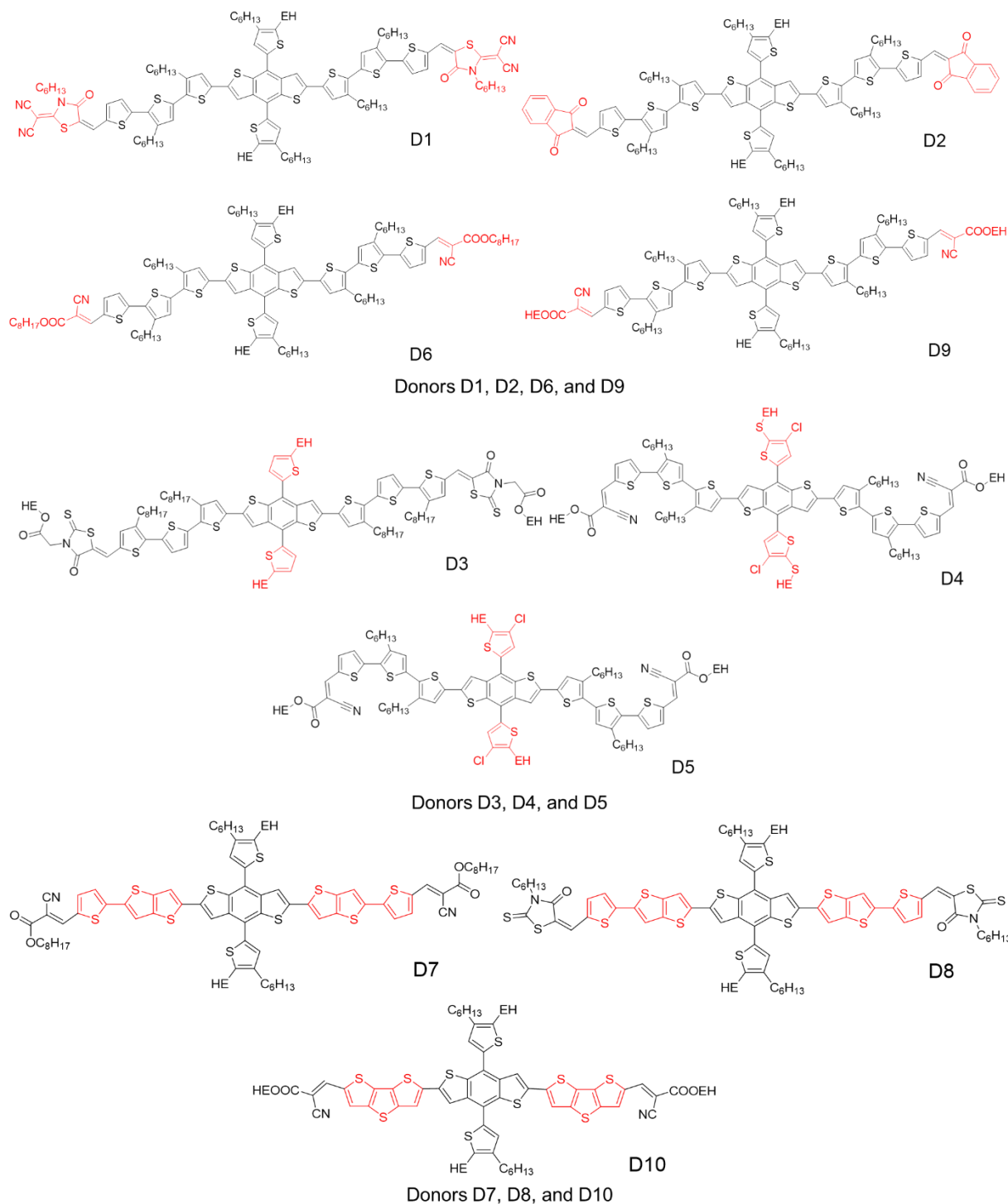
Donor material	MACCS fingerprint														
P3HT	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	0	0	0	0	0	1	0	1	0	0	0	0	1	0	0
	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
	0	0	1	0	0	0	0	0	1	1	1	0	1	0	0
	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0
	0	1	0	0	0	0	0	0	1	0	0	1	0	0	1
0	0	0	0	1	0	0	0	0	1	0	1	0	0	1	
PTB7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0

	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
	0	0	0	0	0	1	0	1	0	0	0	1	1	0	0
	0	0	0	0	0	1	0	0	0	0	1	0	0	0	1
	0	1	1	1	0	0	0	1	1	1	1	0	1	0	1
	0	0	1	0	0	1	1	1	1	0	0	0	0	1	0
	0	1	1	0	1	1	0	1	1	0	1	1	0	1	1
	0	1	1	1	1	0	1	0	1	1	0	1	1	1	1

## Section S2. Process of experiment and proof of the reliability of the ML model

### 2.1 New molecules

All reagents or materials are were used as received from commercial sources, unless otherwise stated. The new donor materials D1-D10 are synthesized using standard synthetic procedures and the materials were fully characterized with  $^1\text{H}$  NMR,  $^{13}\text{C}$  NMR, and high-resolution mass spectrometry. An example synthesis route can be found in our recently published paper (22). The detailed synthesis of the other 9 materials will be described in following papers. Generally, the 10 new materials are divided into three groups with changes or modifications on the end groups (A),  $\pi$  links, core (D), and alkyl chains. In each of the groups, the characteristics making the donors different from BTR or published structures are highlighted in red.



**Fig. S2. Chemical structures of the 10 new donor materials.**

## 2.2 Fabrication of OPV devices

OPVs with standard structures were manufactured on patterned indium tin oxide (ITO)-coated glass substrates (15  $\Omega$ /sq, AE Tech.). The typical procedure is as follows: first, these

substrates were cleaned sequentially with detergent, de-ionized water, acetone, and isopropyl alcohol for 20 min under sonication. They were then dried by nitrogen flow and treated with UV ozone for 30 min. Subsequently, 50 uL PEDOT:PSS aqueous solution was spin coated on an ITO substrate at 6000 rpm for 40 s, followed by a thermal annealing on a hot plate at 120 °C for 20 min to form the hole transport layer (HTL). The substrates were then transferred into a glovebox filled with nitrogen ( $O_2 < 10$  ppm;  $H_2O < 1$  ppm). 40 uL mixed solution of donor and acceptor dissolved in chloroform was dripped on the PEDOT: PSS layer, before these substrates were spin coated at 2000 rpm for 30 s. To form a better molecular packing, solvent vapor treatment by tetrahydrofuran (THF) was introduced. Subsequently, DPO with a concentration of 0.5 mg/mL dissolved in isopropanol (IPA) was deposited on the top of active layer by spin coating at 2000 rpm for 30 s. Finally, these semi-finished devices were moved to a thermal evaporation chamber with a base pressure of approximately  $2 \times 10^{-4}$  Pa, where 100 nm Ag was deposited through a shadow mask with an active area of  $0.11 \text{ cm}^2$ .

**Table S4. Photovoltaic parameters of OPV devices fabricated with different donor materials.**

Device Structure	$V_{oc}$ (V)	$J_{sc}$ ( $\text{mA}/\text{cm}^2$ )	FF (%)	PCE (%)
ITO/PEDOT:PSS/1:PC <sub>71</sub> BM/DPO/Ag	0.69	3.59	39.20	0.97
ITO/PEDOT:PSS/2:PC <sub>71</sub> BM/DPO/Ag	0.89	13.42	71.04	8.46
ITO/PEDOT:PSS/3:IDIC/DPO/Ag	0.88	15.37	66.56	8.85
ITO/PEDOT:PSS/4:PC <sub>71</sub> BM/DPO/Ag	0.98	8.91	60.69	5.30
ITO/PEDOT:PSS/5:PC <sub>71</sub> BM/DPO/Ag	1.02	6.70	58.20	3.97
ITO/PEDOT:PSS/6:PC <sub>71</sub> BM/DPO/Ag	0.96	10.71	69.17	7.10



ITO/PEDOT:PSS/7:Y6/DPO/Ag	0.77	11.80	38.40	3.50
ITO/PEDOT:PSS/8:PC <sub>71</sub> BM/DPO/Ag	0.90	4.28	54.70	2.10
ITO/PEDOT:PSS/9:PC <sub>71</sub> BM/DPO/Ag	1.00	12.61	69.24	8.72
ITO/PEDOT:PSS/10:PC <sub>71</sub> BM/DPO/Ag	1.00	3.78	40.01	1.51

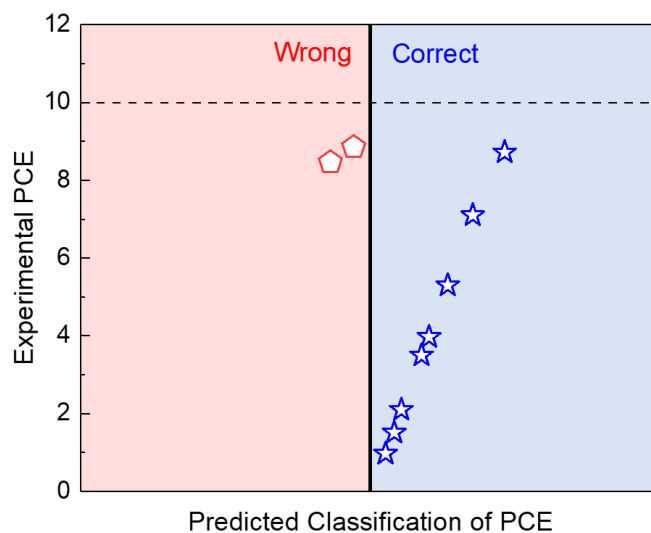
### 2.3 Predictions of the ML models

**Table S5. Prediction results from DNN, RF, and SVM using Hybridization and FP2 fingerprints as inputs, as well as DNN and RF using Daylight fingerprints.**

Donor (Real PCE)	PCE 0~2.99%	PCE > 3.00%
1 (0.97%)	✓	
2 (8.46%)		✓
3 (8.85%)		✓
4 (5.30%)		✓
5 (3.97%)		✓
6 (7.10%)		✓
7 (3.50%)		✓
8 (2.10%)		✗
9 (8.72%)		✓
10 (1.51%)		✗

**Table S6. Prediction results from BPNN using Daylight fingerprints when classification threshold is 10%.**

Donor (Measured PCE)	Predicted PCE 0~9.99%	Predicted PCE > 10.00%
1 (0.97%)	✓	
2 (8.46%)		✗
3 (8.85%)		✗
4 (5.30%)	✓	
5 (3.97%)	✓	
6 (7.10%)	✓	
7 (3.50%)	✓	
8 (2.10%)	✓	
9 (8.72%)	✓	
10 (1.51%)	✓	



**Fig. S3. Prediction results versus experimental data for the 10 new donor materials.** The model is based on the BPNN and Daylight fingerprints as input, and the classification threshold is 10%.