Electronic Supplementary Material (ESI) for Digital Discovery.

Supporting Information

Limitations of machine learning models when predicting compounds with completely new chemistries: possible improvements applied to the discovery of new non-fullerene acceptors

Zhi-Wen Zhao,[a,b][+] Marcos del Cueto, *[a][+] Alessandro Troisi[a]

[a]Department of Chemistry, University of Liverpool, Liverpool, L69 3BX, UK.

[b]Institute of Functional Material Chemistry, Faculty of Chemistry, Northeast Normal University, Changchun, 130024, Jilin, P. R. China.

[+]These authors contributed equally to this work.

* m.del-cueto@liverpool.ac.uk

Contents:

1. Data and code availability

The 566 donor/acceptor pairs with their detailed information and codes used to discuss the ML results in this work are available from a public GitHub repository at github.com/marcosdelcueto/NonFullereneAcceptorPrediction.

2. Computational methods

2.1 Generating properties

Physical properties related to ground-state were calculated at the B3LYP/6-31G* level, with the exclusion of where the smaller basis set 3-21G* was used for reorganization energy calculations.[1] And the excited state properties computed at the M06-2X/6-31G* level[2] and the triplet state energies were computed by ΔSCF procedure.[3] Fingerprints have been computed with the RDKit package[4] and the miscibility properties have been generated with the code SwissADME[5] (with the exception of several molecules computed with XLOGP3[6]).

2.2 Distance Metrics and chemical similarity

The distance between two donor-acceptor pairs, $\mathbf{p}_{ij}$ and $\mathbf{p}_{mn}$, is calculated as a linear combination of the distance based on physical properties (D_{ph}), and the distance based on the fingerprint separately for donors (D_{fpd}) and acceptors (D_{fpa}) can be calculated as:

$$D = \gamma_{ph} D_{ph}(p_{ij}, p_{mn}) + \gamma_d D_{fpd}(p_{ij}, p_{mn}) + \gamma_a D_{fpa}(p_{ij}, p_{mn}) \quad (1)$$

Different hyperparameters (γ_{ph}, γ_a and γ_d) are defined to tune the relative importance of physical and fingerprint distance, and we have considered three distinct cases: i) use only physical descriptors ($\gamma_d = \gamma_a = 0$), ii) use only fingerprints ($\gamma_{ph} = 0$) and iii) use both physical descriptors and fingerprints.

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

$$D_{ph}\left(\mathbf{p}_{ij}, \mathbf{p}_{mn}\right) = \|\mathbf{p}_{ij}^{ph} - \mathbf{p}_{mn}^{ph}\|_2 \quad (2)$$

The chemical similarity of two molecules is a routine task in cheminformatics. It has been reported that Tanimoto similarity was identified as one of the best similarity metrics.[7] The fingerprint distances[8] were calculated from the Tanimoto similarity index $T(\mathbf{r}, \mathbf{s})$ between the vectors containing the Morgan fingerprints of the corresponding donor and acceptor (d^{fp} and a^{fp}) :

$$D_{fpd}\left(\mathbf{p}_{ij}, \mathbf{p}_{mn}\right) = 1 - T\left(d_i^{fp}, d_m^{fp}\right) \quad (3)$$

$$D_{fpa}\left(\mathbf{p}_{ij},\mathbf{p}_{mn}\right) = 1 - T\left(a_j^{fp}, a_n^{fp}\right) \quad (4)$$

$$T\left(\mathbf{r},\mathbf{s}\right) = \frac{\mathbf{r}^T\mathbf{s}}{\mathbf{r}^T\mathbf{r}+\mathbf{s}^T\mathbf{s}-\mathbf{r}^T\mathbf{s}} \quad (5)$$

Since distances are bit vectors, similarity will have a value between 0 and 1.

2.3. Machine learning methods

2.3.1 *k*-nearest neighbors (*k*-NN)

The *k*-NN regression algorithm[9] with weights and proximity determined by the distances expressed in Eq. 1 adopted here to predict values of the properties as a weighted average for the nearest *k* neighbors. The predictions were computed using a leave-one-out (LOO) procedure and 10-fold cross-validation scanning for various values of *k*, which allowed us optimizing *k* to give the best results.

2.3.2 Kernel Ridge Regression (KRR)

KRR as a modified version of regularized least squares could obtain predictions depending on the vicinity with previous observations.[10] The target values $\hat{\boldsymbol{y}}'$ of new donor/acceptor pair $\mathbf{p}_{mn}$ is defined as $\hat{\boldsymbol{y}}' = y^T\left(K+\alpha I\right)^{-1}\kappa'$, where α is a regularization hyperparameter, *I* is the identity matrix, *K* and κ' can be obtained from $K_{ij,mn} = f\left(\mathbf{p}_{ij},\mathbf{p}_{mn}\right)$ and $\kappa'_{ij,mn} = f\left(\mathbf{p}_{ij},\mathbf{p}'_{mn}\right)$. And the distance of physical and structural properties are used to compute f by mapping vectors into a scalar as follows: $f(p_{ij},p_{mn}) = e^{-(\gamma_{ph}D_{ph}^2(p_{ij},p_{mn})+\gamma_d D_{fpd}^2(p_{ij},p_{mn})+\gamma_a D_{fpa}^2(p_{ij},p_{mn}))}$ which can take into account electronic and/or structural properties.

2.3.3 Support Vector Regression (SVR)

SVR[11] uses the same kernel as described above for KRR. Both algorithms are similar, and the main difference is the use of ε-sensitive loss with SVR instead of the squared error loss. The optimization parameters in SVR are the regularization parameter C, and ε, which defines the region within which there is no penalty in the training loss function.

2.3.4 Hyperparameters optimization

Comparing the predicted and actual value of each point in the test set for each iteration, we can obtain an RMSE value. The iteration could be LOO or *leave-one-group-out* (LOGO). For each set of hyperparameters, each of the points of

the test set were predicted by training a model on the remaining N points. In the case of *k*-NN regression, we obtain the optimum number of neighbors, *k*, from a list of possibilities (*k*=1-20). We are using a different number of physical descriptors, plus the fingerprints of the donor and acceptor with a recombination rate of 0.7, mutation of 0.5-1, popsize of 15, and $0<\gamma<6$, $0<\alpha,C,\varepsilon<10$ for stochastic optimization in KRR and SVR. The data have been scaled so that the average of each descriptor is zero and the standard deviation is one.

3. Grouping process

As shown in Table S1, the 33 acceptors were encoded by different fragments. In practice, each entry in the data set of acceptors is labelled as 1, 2, 3, etc. Those labels are used to create training and testing sets. We then grouped the acceptor when they contained similar fragments. The encoded number is 1 if they contain the specific fragment and with 0 if they do not contain any fragment identified. For each investigated acceptor, the corresponding chemical building blocks were considered as fragments, and the number of new molecules for each fragment entry was counted. Fullerenes were not counted by group while all PDIs and other non-fullerenes were considered when grouping molecules to be consistent with our main aim in this work. In the beginning, we paid attention to the group with a number of new molecules larger than 3 and got 8 groups. For molecule like acceptor 7, which is not obvious what fragment to consider, we considered LUMO contributions as a criterion. As seen from Table S2, molecule 7 will be removed from Group 6 (IDT contributes 18.9% to LUMO) and kept in Group 3 (BT contributes 58.3% to LUMO). As seen in Table S2, we can re-assign molecules in different groups by fragment contribution to LUMO larger than 50%: (i) remove molecule 7 in Group 6 and keep it in Group 3; (ii) keep Group 2, Group 3, Group 4 and Group 5, and remove Group 6, Group 7 and Group 8.

Table S1. The encode of fragments for each acceptor.

molecule	C_{60}	C_{70}	$PC_{61}BM$	$PC_{71}BM$	ICBA	bis-PCBM	PDI(1)	PDI(2)	PDI(3+)	DPP	BT	IID-T	IC	IDT	NI	fluorene	BDTP	IDTT	6T	oth1	oth2	oth3	oth4	oth5	oth6
1	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
2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0
3	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1
4	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0
5	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1
6	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0
7	0	0	0	0	0	0	0	0	0	0	1	0	0	1	0	0	0	0	0	0	1	0	0	0	0
8	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
9	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
10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1	0	0
11	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0
12	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
13	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
14	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
15	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
16	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
17	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0
18	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
19	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
20	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
21	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
22	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0
23	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
24	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0
25	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
26	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0
27	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0
28	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
29	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	1	0	0	0	0	0	0	0	0	0
30	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	1	0	0	0	0	0	0	0	0	0
31	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
32	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
33	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

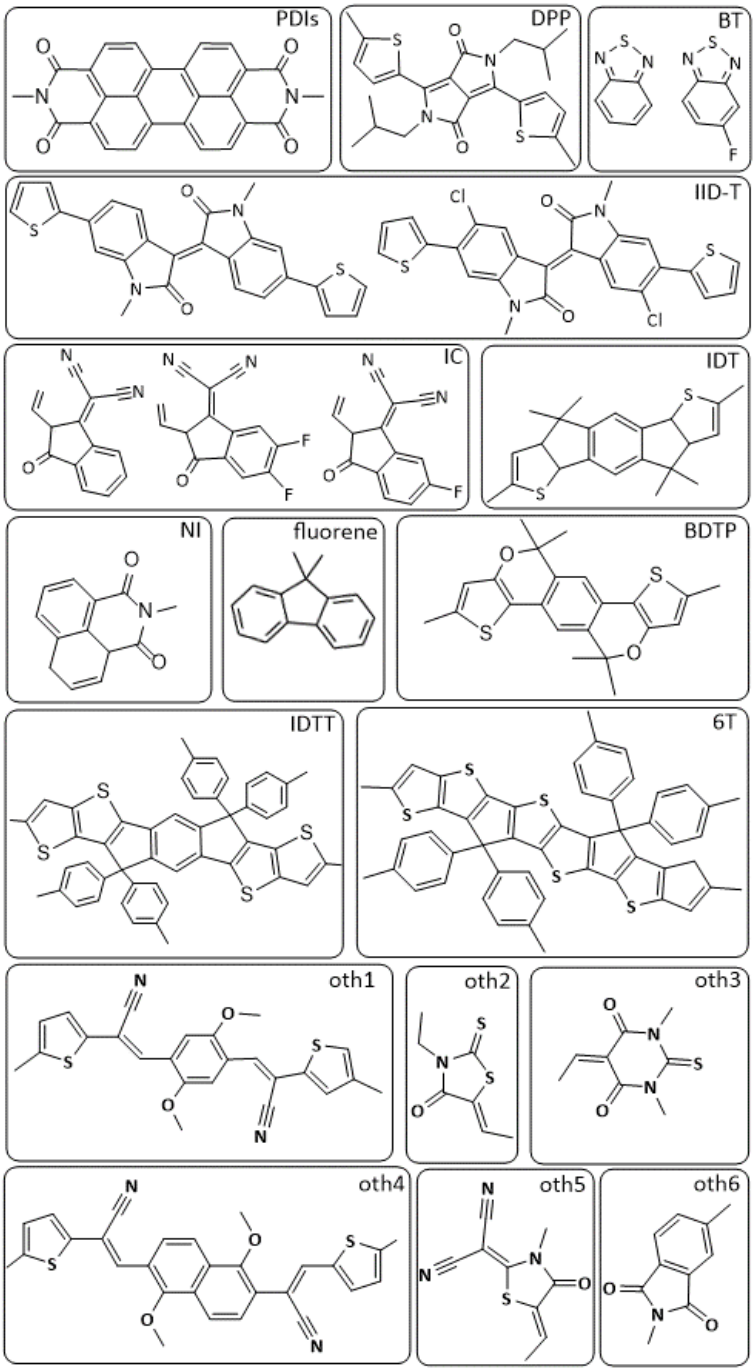


Figure S1. The chemical structures of considered fragments.

Table S2. Fragments contribution to LUMO (Mulliken partition).

Group 2 (DPP)	Group 3 (BT)	Group 4 (IID-T)	Group 5 (IC)	Group 6 (IDT)	Group 7 (NI)	Group 8 (fluorene)
23(52%)	7(58%)	3(92%)	17(62%)	7(19%)	2(23%)	14(17%)
31(64%)	29(57%)	4(92%)	22(60%)	17(39%)	4(4%)	29(6%)
32(61%)	30(59%)	5(97%)	24(60%)	26(38%)	6(2%)	30(8%)
		6(97%)	25(55%)		10(24%)	
			26(62%)		11(22%)	
			27(64%)		29(29%)	
					30(26%)	

4. Machine learning results

Table S3. Optimized hyperparameters and the resulting RMSE/r for k-NN predictions of PCE with different cross-validation strategies.

	Features	Group	k	γ_{ph}	γ_d	γ_a	RMSE (%)	r
LOO-interp.	**Fp.**	-		0.0	5.870	2.878	1.783	0.690
	Phys.	-		1.0	0.0	0.0	2.155	0.513
LOO-extrap.	**Fp.**	**G1**	4	0.0	4.873	3.514	2.960	0.323
		G2	4	0.0	5.031	2.904		
		G3	4	0.0	5.031	2.904		
		G4	4	0.0	5.680	2.883		
		G5	4	0.0	4.500	2.146		
	Phys.	**G1**	15	1.0	0.0	0.0	3.209	-0.077
		G2	15	1.0	0.0	0.0		
		G3	15	1.0	0.0	0.0		
		G4	15	1.0	0.0	0.0		
		G5	15	1.0	0.0	0.0		
LOGO-extrap.	**Fp.**	**G1**	2	0.0	3.404	4.873	3.564	-0.011
		G2	5	0.0	3.053	5.432		
		G3	5	0.0	2.629	0.529		
		G4	20	0.0	0.704	3.443		
		G5	2	0.0	4.500	2.146		

	Phys.	G1	4	1.0	0.0	0.0	3.197	0.018
		G2	5	1.0	0.0	0.0		
		G3	4	1.0	0.0	0.0		
		G4	7	1.0	0.0	0.0		
		G5	6	1.0	0.0	0.0		

Table S4. Optimized hyperparameters and the resulting RMSE/r for KRR predictions of PCE with different cross-validation strategies.

	Features	Group	α	γ_{ph}	γ_d	γ_a	RMSE (%)	r
LOO-interp.	Fp.	-	0.349	0.0	2.766	0.627	1.751	0.691
	Phys.	-	0.139	0.110	0.0	0.0	2.009	0.563
LOO-extrap.	Fp.	G1	0.366	0.0	3.054	0.736	3.522	0.078
		G2	0.383	0.0	3.184	0.913		
		G3	0.360	0.0	2.980	0.884		
		G4	0.389	0.0	3.111	0.652		
		G5	0.347	0.0	3.632	0.306		
	Phys.	G1	0.105	0.111	0.0	0.0	4.108	0.167
		G2	0.155	0.105	0.0	0.0		
		G3	0.119	0.115	0.0	0.0		
		G4	0.139	0.110	0.0	0.0		
		G5	0.097	0.108	0.0	0.0		
LOGO-extrap.	Fp.	G1	0.015	0.0	0.037	0.350	3.766	0.067
		G2	0.241	0.0	0.761	2.59×10^{-5}		
		G3	0.371	0.0	2.508	6.32×10^{-4}		
		G4	0.065	0.0	0.012	0.024		
		G5	0.180	0.0	1.500	0.100		
	Phys.	G1	7.93×10^{-3}	1.86×10^{-4}	0.0	0.0	2.845	0.309
		G2	2.57×10^{-3}	9.60×10^{-5}	0.0	0.0		
		G3	0.196	5.94×10^{-3}	0.0	0.0		
		G4	0.310	3.53×10^{-3}	0.0	0.0		
		G5	0.659	4.30×10^{-3}	0.0	0.0		

Table S5. Optimized hyperparameters and the resulting RMSE/r for SVR predictions of PCE with different cross-validation strategies.

	Features	Group	(C, ε)	γ_{ph}	γ_d	γ_a	RMSE (%)	r
LOO-interp.	**Fp.**	-	3.979, 0.017	0.0	3.977	0.818	1.667	0.726
	Phys.	-	7.216, 1.157	0.157	0.0	0.0	1.961	0.587
LOO-extrap.	**Fp.**	**G1**	4.138, 0.030	0.0	5.411	1.098	3.268	0.169
		G2	3.694, 0.0112	0.0	4.877	1.203		
		G3	3.762, 0.097	0.0	4.442	0.818		
		G4	2.982, 0.021	0.0	4.951	0.664		
		G5	3.297, 0.023	0.0	5.256	0.692		
	Phys.	**G1**	9.161, 1.199	0.115	0.0	0.0	3.246	0.101
		G2	5.326, 1.237	0.172	0.0	0.0		
		G3	4.884, 1.083	0.162	0.0	0.0		
		G4	7.952, 1.162	0.151	0.0	0.0		
		G5	9.651, 1.157	0.132	0.0	0.0		
LOGO-extrap.	**Fp.**	**G1**	10.000, 1.141	0.0	0.178	0.664	3.323	-0.055
		G2	0.110, 4.902	0.0	5.331	2.192		
		G3	0.130, 4.946	0.0	5.178	4.699		
		G4	9.959, 2.618	0.0	0.077	0.508		
		G5	9.604, 7.19×10^{-3}	0.0	3.497	0.115		
	Phys.	**G1**	9.933, 2.720	$6.33x10^{-3}$	0.0	0.0	2.833	0.232
		G2	0.152, 4.926	2.829	0.0	0.0		
		G3	0.180, 4.932	1.633	0.0	0.0		
		G4	7.207, 2.996	2.86×10^{-3}	0.0	0.0		
		G5	9.999, 3.357	7.59×10^{-3}	0.0	0.0		

5. LOGO convergence to LOO

To better understand the relation between the proposed LOGO-extrapolation (five chemically different groups) and LOO-interpolation (566 groups with one donor/acceptor pair each), we considered three intermediate cases:

(i) LOGO - 5 groups with non-fullerene pairs. Our database consists of 49 non-fullerene donor/acceptor pairs, corresponding to 23 unique non-fullerene acceptor molecules. As a first step to increase the heterogeneity of our initial groups, we have randomly split these 23 non-fullerene acceptor molecules into five groups, in a way that each group has a number of pairs per group similar to the LOGO-extrapolation case (~10).

(ii) In total, our database has 33 unique acceptor molecules, so we have considered 33 groups, as a next step to increase group heterogeneity,

where each of them contains all pairs with a given acceptor molecule.

(iii) To increase the number of groups further, we have split the 33 groups with the most pairs into new groups, so each group has approximately 40 pairs per group at most. We end up with 55 groups, where some of them contain pairs with the same acceptor molecules as other groups, increasing the heterogeneity between groups further, approaching the LOO extreme.

6. LOGO-extrapolation vs LOO-extrapolation

6.1. Predicted vs Experimental PCE

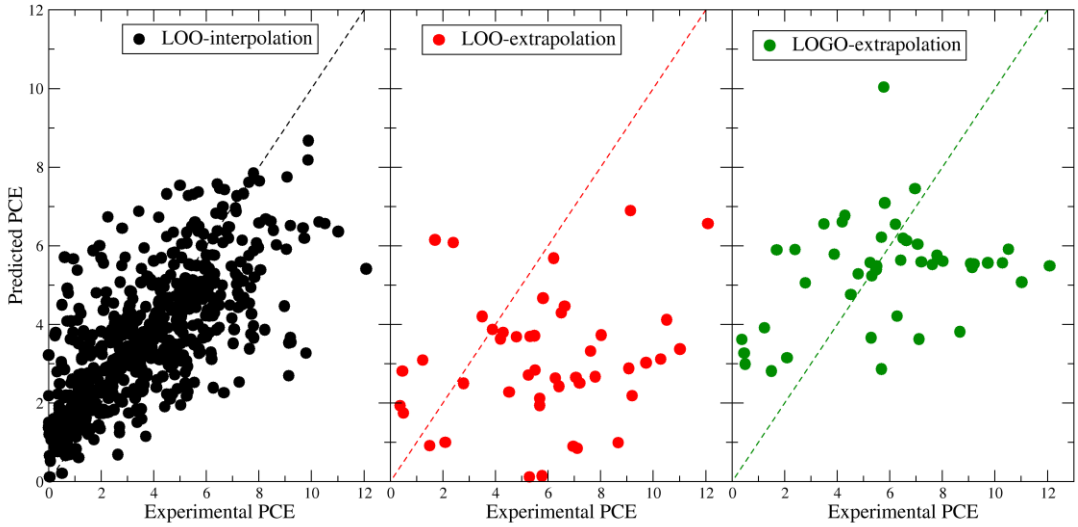


Figure S2. Predicted vs experimental PCE values when using LOO-interpolation (left panel) LOO-extrapolation (middle panel) and LOGO-extrapolation (right panel) when using KRR and physical descriptors.

6.2. Different Grouping

The benefit of using LOGO-extrapolation with respect to LOO-extrapolation can be seen in Table S6. When we group our unique 23 non-fullerene acceptors in five random groups, LOO-extrapolation and LOGO-extrapolation return virtually the same result. However, when we group the acceptors in five chemically distinct groups, LOGO-extrapolation results in a significantly lower RMSE.

Table S6. RMSE resulting of using LOO-extrapolation and LOGO-extrapolation when using different grouping. Results were obtained with KRR, using fingerprints and physical descriptors.

	LOO-extrapolation	**LOGO-extrapolation**
Five random groups	2.58 %	2.57 %
Five chemically distinct groups	3.36 %	2.84 %

7. Screening new candidates

To see how our model would perform when predicting the PCE of molecules from new chemical groups, we can use KRR to optimize the hyperparameters using the LOGO-extrapolation approach with physical descriptors, using all our database during training. This results in the optimized values: $\alpha = 0.07950,\ \gamma_{ph} = 0.001596.$

As a first step to identify the ranges for desired properties of new molecules, we identified that the four acceptors with the largest PCE in our database have $E_{LUMO} \approx -3.5\ eV,\ \lambda \approx 0.2\ eV,\ \sum f \approx 2.9$ and $XLOGP3 \approx 23.6$. When we take the acceptor with the largest PCE in our database[12], and we use it as a test point when training the model with all our database, it results in a predicted PCE value of 6.46%. We can change slightly each descriptor to identify which ranges result in a relative decrease of ~25% ($PCE \geq 4.8\%$). Therefore, we would expect that acceptors with $E_{LUMO} < -2.85\ eV,\ \lambda > 0.12\ eV,\ \sum f > 0.8$ and $XLOGP3 > 2$ would result in a large PCE.

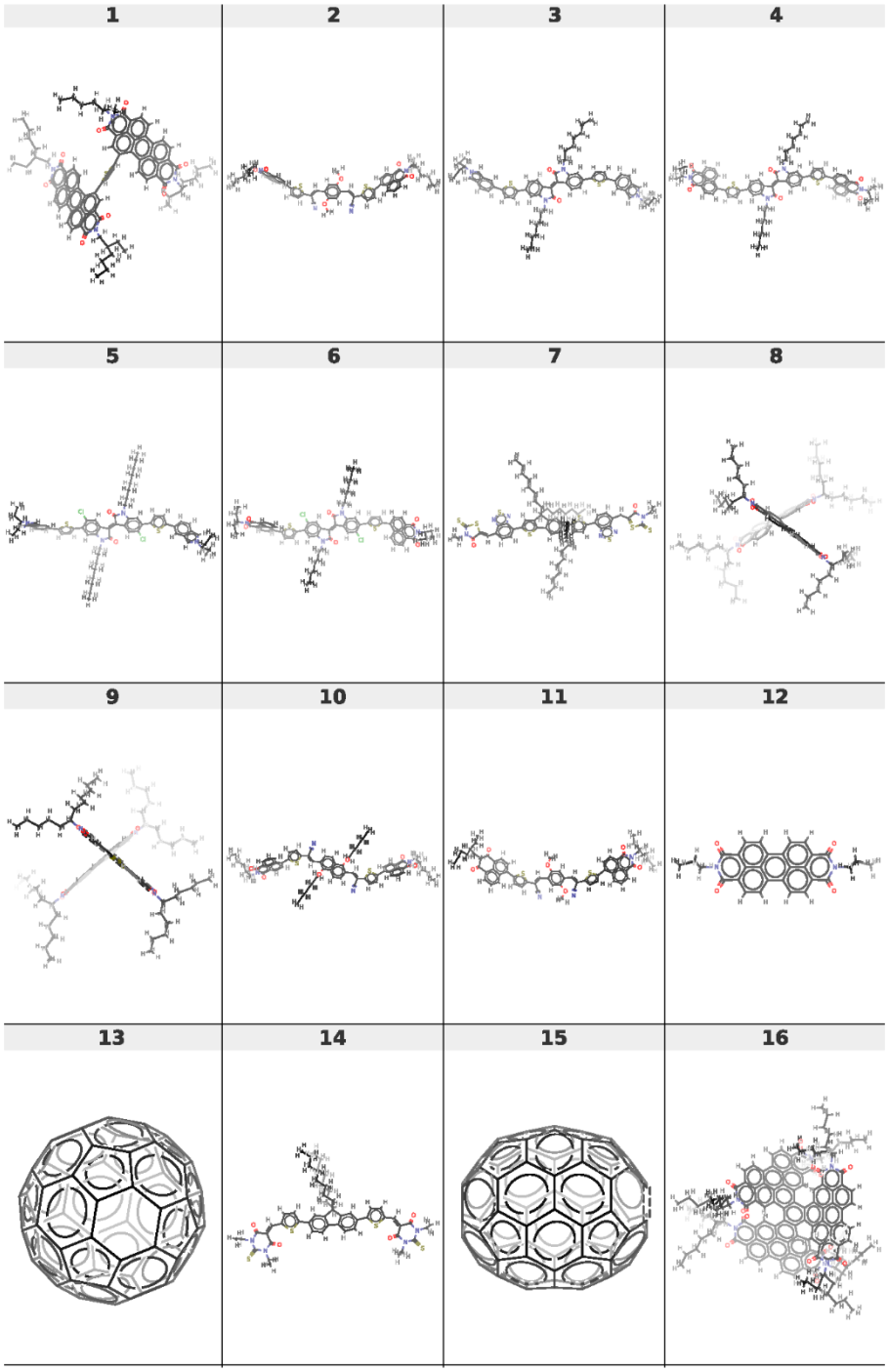
The database from ref. [13] contains multiple organic molecules, including their LUMO energy and oscillator strength for the three first excited states at a TDDFT/M06-2X/def2-SVP level of theory. We screened molecules that were within the $E_{LUMO} < -2.85\ eV,\ \sum f > 0.8$ range, had less than 70 heavy atoms, were not part of any of the groups in our database and were not too similar to each other. After this, we have used the SwissADME web tool[5] to calculate their XLOGP3 value, and added some side chains to those with a XLOGP3 value of approximately 2 or less. This procedure resulted in nine candidates, whose physical descriptors were calculated with the same level of theory specified in Section S2 of this SI. We show in Table S7 the computed descriptors of these molecules and the resulting predicted PCE.

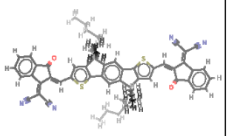
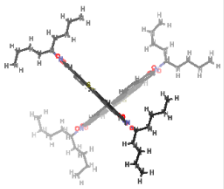
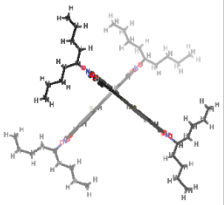
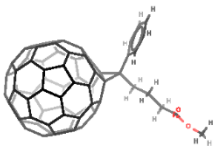
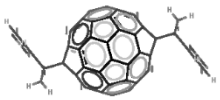
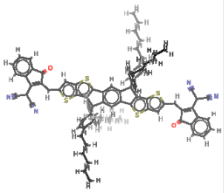
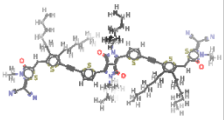
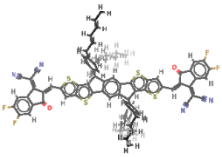
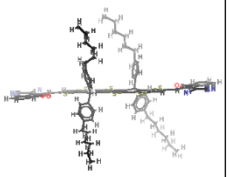
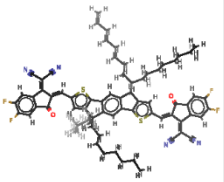
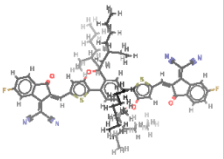
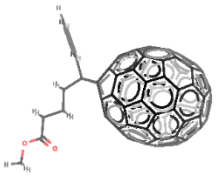
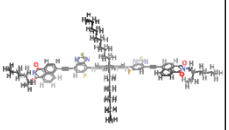
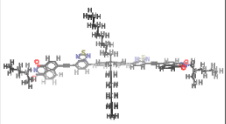
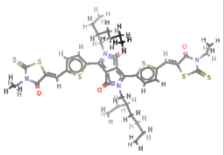
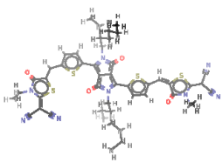
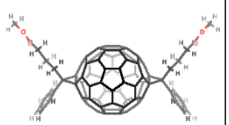
Table S7. Predicted PCE for nine molecules from families not present in the training dataset. Results were obtained using KRR with physical descriptors.

Molecule no.	LUMO (eV)	λ (eV)	$\sum f$	XLOGP3	Predicted PCE (%)
1	-4.7266	0.1848	0.9699	8.31	7.32
2	-3.8994	0.1363	0.7528	9.31	6.52
3	-3.5786	0.1274	1.5634	9.18	6.36
4	-4.2161	0.2366	0.7953	8.42	5.58
5	-3.4055	0.2003	1.3320	16.41	4.81
6	-4.2452	0.2800	0.5608	8.57	4.73
7	-3.5152	0.1985	1.2209	7.93	4.58
8	-3.5372	0.2352	0.0000	9.16	3.16
9	-3.9097	0.3763	0.6825	9.51	2.04

8. Acceptors structures

Table S8. The distinct 33 acceptors in the database of experimental photovoltaic cells.



17	18	19	20
			
21	**22**	**23**	**24**
			
25	**26**	**27**	**28**
			
29	**30**	**31**	**32**
			
33			
			

References:

1 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.* **104**, 4971-5004, (2004).

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

3 Grotjahn, R., Maier, T. M., Michl, J. & Kaupp, M. Development of a TDDFT-Based Protocol with Local Hybrid Functionals for the Screening of Potential Singlet Fission Chromophores. *J. Chem. Theory Comput.* **13**, 4984-4996, (2017).

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

5 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.* **7**, 42717, (2017).

6 Cheng, T. *et al.* Computation of Octanol−Water Partition Coefficients by Guiding an Additive Model with Knowledge. *J. Chem. Inf. Model.* **47**, 2140-2148, (2007).

7 Bajusz, D., Racz, A. & Heberger, K. Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations? *J. Cheminformatics* **7**, 1-13, (2015).

8 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. Horizons* **3**, 226-233, (2016).

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

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

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

12 Gao, K. *et al.* Over 12% Efficiency Nonfullerene All-Small-Molecule Organic Solar Cells with sequentially Evolved Multilength Scale Morphologies. *Adv. Mater.* **31**, 1807842, (2019)

13 Padula, D., Omar, O. H., Nematiaram, T., Troisi, A. Singlet fission molecules among known compounds: finding a few needles in a haystack. *Energy Environ. Sci.* **12**, 2412-2416 (2019). Dataset available at: https://datacat.liverpool.ac.uk/id/eprint/1472