Digital Discovery

ROYAL SOCIETY OF CHEMISTRY

PAPER

View Article Online
View Journal | View Issue

Check for updates

Cite this: *Digital Discovery*, 2022, **1**, 266

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,‡[ab] Marcos del Cueto ‡*[a] and Alessandro Troisi [a]

We try to determine if machine learning (ML) methods, applied to the discovery of new materials on the basis of existing data sets, have the power to predict completely new classes of compounds (extrapolating) or perform well only when interpolating between known materials. We introduce the leave-one-group-out cross-validation, in which the ML model is trained to explicitly perform extrapolations of unseen chemical families. This approach can be used across materials science and chemistry problems to improve the added value of ML predictions, instead of using extrapolative ML models that were trained with a regular cross-validation. We consider as a case study the problem of the discovery of non-fullerene acceptors because novel classes of acceptors are naturally classified into distinct chemical families. We show that conventional ML methods are not useful in practice when attempting to predict the efficiency of a completely novel class of materials. The approach proposed in this work increases the accuracy of the predictions to enable at least the categorization of materials with a performance above and below the median value.

Received 17th February 2022
Accepted 23rd March 2022

DOI: 10.1039/d2dd00004k

rsc.li/digitaldiscovery

Open Access Article. Published on 25 March 2022. Downloaded on 6/22/2024 3:41:12 AM.
This article is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported Licence.

I. Introduction

One of the most exciting recent developments of materials discovery is the adoption of machine learning (ML) to guide the exploration of chemical and materials space.[1–4] In a typical application, existing datasets of experimental characterizations are often combined with computed features of the same materials and used to predict the property of interest of novel materials. The field is very frequently reviewed,[5–9] and some examples include alloys,[10] polymers,[11] perovskites[12] and other inorganic solids.[13–18] Although the results are impressive and have prompted a widespread adoption of such methods across all areas of materials discovery, there is still some uncertainty over the ability of ML to explore completely new chemical/material spaces.[19] In general, the predictive ability of ML is computed *via* cross-validation, *i.e.*, predicting the performance of materials in a testing set that has not been included in the training set to optimize the ML algorithm, where training and testing sets are randomly generated from all data available in various ways. It is known that such algorithms perform better with larger data sets and with training data as close as possible to the ones to be predicted. Cross-validation generally gives the same weight to all predictions, whether they are for entries very close to existing ones in the training set (producing, in essence, an interpolation) or they are entirely novel (producing a much more challenging extrapolation). Moreover, materials are generally clustered by scientists into families based on their related chemical structures. Any discovery of a new family of compounds is regarded as a breakthrough, while discovering a novel member of an existing family is considered a more incremental advance. Therefore, accurate predictions within the families of known compounds and outside such families have a completely different value to the community. It should be noted that there are other non-random cross-validation methods, like scaffold-splits, time-splits and stratified-splits, which offer an alternative way to evaluate models. The general impact of such methods on the evaluation of the model can be found in ref. 20 and 21. Stratified sampling has also been used to train models in organic solar cells, although is has a minimal impact.[22] In our preliminary work with this dataset, time-splits do not perform well, as the validation families are developed at similar points in time. Scaffold splits would use certain structural properties to split the data into groups, although here we opted for a combination of structural and electronic characteristics to categorize in groups, as explained in Section 3 of the ESI.†

The goal of this work is to assess the ability of ML to predict the efficiency of interesting energy materials from completely

[a]*Department of Chemistry, University of Liverpool, Liverpool, L69 3BX, UK. E-mail: m.del-cueto@liverpool.ac.uk*

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

† Electronic supplementary information (ESI) available. See DOI: 10.1039/d2dd00004k

‡ These authors contributed equally to this work.

© 2022 The Author(s). Published by the Royal Society of Chemistry

Open Access Article. Published on 25 March 2022. Downloaded on 6/22/2024 3:41:12 AM.
This article is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported Licence.

new families and offer a new method to do so. In this context, by completely new, we mean materials that belong to a chemical family that is not present when training the model, and can be generated using chemical intuition, database searching or generative models.[23] Then, one can use our methodology to screen these candidates and decide which ones will have a larger performance, reducing the number of candidates and accelerating the production of materials from new families. The methodology, in the most general terms, consists of constructing an ML model that is trained without any information on a new family of materials and assessing its quality in predicting the property of known elements of such family. Note that here we refer to training as the process of finding the optimum hyperparameters through a specific validation method. A practical problem is that the definition of "new" is not mathematically accurate, and the novelty of a material is related to (a combination of) electronic, geometric or synthetic features that cannot be captured by an algorithm, while they will appear as self-evident to any expert in the relevant scientific domain. The problem of predicting the target properties of data outside of the training domain is also often tackled with transfer learning, where a previously trained method is used a starting point when predicting data in a new domain. We chose to study the ability of ML to explore new chemical space in the context of predicting novel non-fullerene acceptors for organic solar cells (OSCs).[24–26] The topic is of significant contemporary interest as the identification of non-fullerene acceptors is considered essential to develop a competitive OSC technology and recent improvements have seen an almost three-fold increase in efficiency in five years since the report of non-fullerene electron acceptor (ITIC).[27,28] For this scientific problem, there are well-defined families of acceptors recognized by the community and used to categorize the recent advances in the field. We can, therefore, ask whether new families of non-fullerene acceptors could have been predicted without any information on any member of that family. In this work, we discuss how a conventional cross-validation results in an over-optimistic evaluation of models when they are eventually used to predict new classes of compounds. We aim to draw conclusions on the specific field of computer-aided discovery of OSCs acceptors but also, more generally, on a practical approach to assess the usefulness of ML methods for more exploratory research. There have been other recent studies that evaluate ML models with out-of-sample tests in materials discovery.[29,30] There has been similar work to predict out-of-sample reaction yields,[31,32] and work in risk minimization applied to organic molecules to improve domain generalization.[33] We introduce in this work a modification providing a simple framework to train models to perform extrapolations. This change is shown to improve significantly the accuracy of the model when predicting out-of-sample materials, with respect to models trained with a usual cross-validation.

The growing interest on non-fullerene acceptor devices[34] has produced a large amount of valuable data. While there is no general standardization of the processing condition to measure the power conversion efficiency (PCE) and some of the experimental details are not always available, the experimental datasets appear to be sufficiently accurate to enable data science analysis with a range of works reporting good predictive abilities. For instance, Haibo Ma's team collected 300 experimental data of small-molecule OSCs, and trained an ML model using 10-fold cross-validation with a PCE prediction accuracy RMSE of approximately 1.2%.[35] They also trained another model with a different database by leave-one-out (LOO) cross-validation, achieving a good prediction accuracy.[22] Arindam Paul *et al.* used extremely randomized tree models to predict the HOMO energies of the HOPV dataset,[36] to accelerate the screening process of OSCs.[37] Similarly, Salvy P. Russo's group focused on the screening uses of ML by training ML models with DFT data to approximate properties of organic photovoltaics (OPV) materials.[38] Jie Min *et al.* adopted five common algorithms for polymer/non-fullerene OSC devices by 10-fold cross-validation with the best results achieved with the Random Forest method.[39] The approach has been recently extended to the study of perovskite solar cells.[40] Jeff Kettle's group used ML to analyze a dataset with 1850 OSC devices, and were able to identify which material properties play a major role in the device stability and degradation.[41]

A word of caution when assessing the accuracy of ML models is that experimental datasets will likely have distribution biases that will affect the reliability of predictive ML models trained with those datasets.[42] There have been recent approaches to correct these biases by, for example, re-introducing data of failed experiments,[43] adopting unbiased design of experiments[44] and using new frameworks and metrics.[19,45] Even though advances in this area will undoubtedly be beneficial, here we are interested in the accuracy of ML models when predicting new classes of materials, even in the best-case scenario when the dataset is balanced and representative of the field.

The data set used in this work contains experimentally investigated small-molecule organic photovoltaics whose chemical structures and PCE values are collated from the literature, with detailed information in our previous report[46] and a public repository.[47] In this work, we have used two distinct types of features that have previously proven useful in predicting different properties of donor–acceptor pairs:[46,48] (i) fingerprints (also referred to as chemical descriptors in other works) consisting of the Morgan fingerprints[49] of donors and acceptors, and (ii) physical descriptors consisting of:

HOMO and LUMO energies of the donor, the LUMO energy of the acceptor. The energies of the frontier orbitals are expected to affect the PCE of the solar cell, and they have been shown to improve the PCE prediction of ML models in organic solar cells.[22]

Reorganization energy of hole (for donor) and electron (for acceptor) transport. The reorganization energy is expected to correlate with the charge transport properties of the system.[50]

Sum of the oscillator strengths of the states below UV range (3.54 eV) for donor and acceptor. This parameter measures the optical absorption of the molecule and a high value is beneficial for photovoltaic activity.[51]

Measure of miscibility evaluated for both the donor and acceptor. We have approximated the miscibility of each molecule as the logarithm of the octanol water partition function, as a mixture of compounds with different hydrophobicity is more

Open Access Article. Published on 25 March 2022. Downloaded on 6/22/2024 3:41:12 AM.
This article is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported Licence.

likely to segregate. The logarithm of the octanol water partition function was calculated by the XLOGP3 (ref. 52) method, which is commonly used for organic molecules,[53] using the SwissADME[54] web tool.

More details on these descriptors, why they were selected for this dataset, and how they compare against other descriptors, can be found in ref. 46. We detail the level of theory used to calculate each descriptor in the ESI.† Our final database consists of 566 donor/acceptor pairs, 49 of which contain non-fullerene acceptors and it is available as a stand-alone dataset.[47] In these 566 pairs, we have a total of 33 distinct acceptor molecules shown in the ESI,† which also include the computational details used to obtain all features.

In this work, we have used the kernel ridge regression (KRR)[55] algorithm, which is commonly applied to organic molecules datasets by several authors[6,56] and was used for the same dataset in ref. 46. We have also used *k*-nearest neighbors regression (*k*-NN)[57] and support vector regression (SVR).[58] The results in the manuscript were obtained with KRR, and we show in the ESI† how *k*-NN and SVR produce relatively similar results. These algorithms use the "distance" with the training data to predict the PCE of unknown data, and we show a more in-depth explanation in the ESI.† We chose these algorithms as they are easy to implement and are commonly used for this type of application in materials science.[59] All these algorithms struggle when predicting the PCE of new families of compounds, which is why we have proposed a new training framework in this work, which can be applied to any algorithm, and improves their extrapolation capabilities. Novel molecules are expected to be more distant in the parameter space, so we want to explore the ability of the algorithm and features to predict properties without nearby known structures. As described in ref. 48, the physical descriptors distances are calculated as the Euclidean distance between physical descriptor values. The chemical similarity of materials in the database is evaluated *via* the Tanimoto index,[60] which is obtained from the Morgan fingerprints to characterize how similar each molecule is to others. The hyperparameters were optimized using a differential evolution algorithm,[61] as implemented in SciPy.[62] When training the model, the hyperparameters (including feature weights) are optimized, and their values for each case are shown in Tables S3–S5 in the ESI.† We show in Fig. 1 the workflow of this work, in which one first preprocess the data to generate a suitable database (described in more detail in ref. 46), then one trains the model by selecting a specific validation method to optimize the hyperparameters of the model (discussed in Sections II–IV), and finally one can deploy the model to screen candidate materials by their predicted PCE values (example shown in Section V).

II. Leave one group out cross-validation

The full data set, D, is formed by a set of vectors with features $\{x_i\}$, and target property values $\{y_i\}$, $D = \{(x_i, y_i)\}$. In a model like KRR, the target property is a function of the features, x, the hyperparameters, h, and the data set, D:

$$y = f(x; h, D) \quad (1)$$

The hyperparameters, h, are normally found by cross-validation. Subsets A_1, A_2,… of D are selected. Indicating with $D - A_k$ the set obtained by removing the subset A_k from D (with $-$ indicating the exclusion operator) the hyperparameters are chosen to minimize the total square error:

$$\Delta(h, D)^2 = \sum_k \sum_{x_i \in A_k} [f(x_i;\ h,\ D - A_k) - y_i]^2 \quad (2)$$

It is common to construct the sets $\{A_k\}$ as a random partition of the data in n subsets of equal size and the resulting method is known as n-fold cross-validation. Another common approach is to partition D in as many subsets $\{A_k\}$ as the elements of D, with each subset containing just one element. This approach is known as leave-one-out (LOO) and corresponds to optimizing the ability of the function to predict a particular data point without any information on it.

If the data set is made of different families of related materials, the cross-validation methods above would favour the process of interpolation between data points. The subset $D - A_k$ will always contain many elements of the same family and the optimization of the error in eqn (2) does not really reflect the ability of the function f to predict properties of a completely new set of materials. To emphasize this aspect, we refer to the results using this type of cross-validation as LOO-interpolation.

A naïve approach to deal with this issue would be to exclude all elements of a particular family A_n from the cross-validation, perform any form of cross-validation with the remaining data $D - A_n$ and evaluate the predictive ability of the resulting method on the elements of A_n. This approach provides a measure of how well a standard ML approach predicts the properties of new families of compounds if no element of that family was used in its training. For example, one could perform a LOO cross-validation excluding in turn families of molecules, and we refer to this elementary approach as LOO-extrapolation. We show an example dataset with four distinct families in Fig. 2A, and we show in Fig. 2B how the data would be split with LOO-extrapolation to predict the values of one of these families.

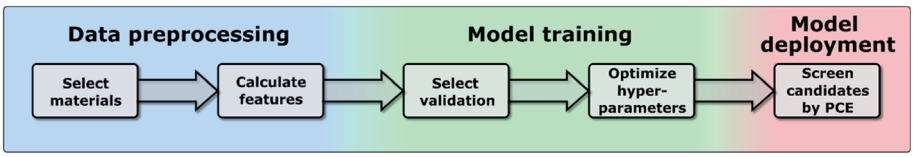

Fig. 1 Workflow of the proposed framework.

© 2022 The Author(s). Published by the Royal Society of Chemistry

Open Access Article. Published on 25 March 2022. Downloaded on 6/22/2024 3:41:12 AM.
This article is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported Licence.

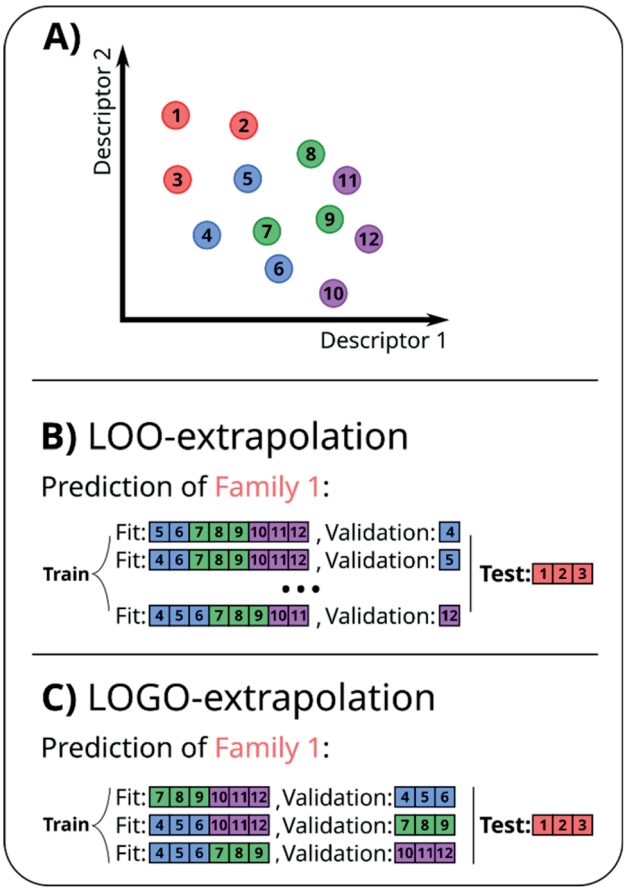

Fig. 2 (A) Example of a small 2D database with 12 points split among four families, indicated by their color. (B) Representation of the LOO-extrapolation to predict the values of 'Family 1', where the algorithm is trained with a LOO cross-validation. (C) Representation of our proposed LOGO-extrapolation to predict the values of 'Family 1', where the algorithm is trained to extrapolate to other families and is finally tested to extrapolate the values of that new family.

There have been similar approaches suggested recently, like the leave-one-cluster-out cross-validation,[29] where the dataset is split in clusters and one tries to predict the values for each cluster, which has been left out of the training set. Another approach is the *k*-fold-*m*-step forward cross-validation,[30] which ranks data by their target property value and evaluates how well the model can predict target values outside of the training domain, *i.e.* the model is trained to perform extrapolations of candidates in a different range of the target property, not necessarily to extrapolate to candidates from a different domain in the feature space. For example, a function tuned for prediction within the set $D - A_n$ will be exploiting the existence of elements of the same family within $D - A_n$ and it may not result in the best function to predict properties when no information on similar materials is available. The best chance to build a model able to perform accurate predictions on new families of materials is to train the model to do so.

We partition D in subgroups containing chemically distinct families $\{A_n\}$. For each subgroup A_m, we find the parameters h that minimize the following error:

$$\Delta_m(h, D - A_m) = \sum_{n \neq m} \sum_{x_i \in A_n} [\, f(x_i;\ h,\ D - A_n - A_m) - y_i]^2 \qquad (3)$$

In essence, leaving out group A_m, we consider in turn all the other groups A_n and compute the error in predicting elements in A_n without using the data in A_n (and neither A_m). The scheme is illustrated in Fig. 2C. Minimizing eqn (3) with respect to h is equivalent to finding the best function at performing extrapolations. This approach simulates a situation where no element of the family A_m has been discovered, and the remaining elements can be divided into distinct families. We call this method leave-one-group-out (LOGO) cross-validation. The procedure can be repeated for every A_m, where each one results in an optimal set of hyperparameters $h^{m,D}_{\min,\mathrm{LOGO}}$. The RMSE error can then be defined as:

$$\mathrm{RMSE}_{\mathrm{LOGO}} = \sqrt{\frac{1}{\sum_m \mathrm{size}(A_m)} \sum_m \sum_{x_i \in A_m} \left[f\left(x_i; h^{m,D}_{\min,\mathrm{LOGO}}, D - A_m\right) - y_i \right]^2} \qquad (4)$$

We have referred to this use of a LOGO cross-validation to train the ML model to perform extrapolations as LOGO-extrapolation (see Fig. 2C). Note that this training can be applied to any ML algorithm, and it is ultimately based on a cross-validation approach that mimics the discovery of novel material classes to overcome the inherent 'leakage' of information that is present in other cross-validation methods. We are effectively optimizing the model's parameters so that they are good at generalizing predictions of novel groups. Similar goals to optimize the training process for a specific task can be found in the meta-learning of neural networks.[63,64]

III. Chemical groups definition

For non-fullerene acceptors, the grouping is based on the available literature with some additional considerations. Some recent reviews[25,65–68] have made attempts to classify the main acceptors by different chemical fragments such as perylene diimide (PDI),[65,67] isoindigo-based molecule (IID),[69] 1,8-naphthalimide (NI),[70] 2-(3-oxo-2,3-dihydroinden-1-ylidene (IC),[71] diketopyrrolopyrrole (DPP),[72] fluorene,[73] indacenodithiophene (IDT),[24,74] benzothiadiazole (BT),[24] benzodi(thienopyran) (BDTP),[75,76] (indacenodithieno[3,2-*b*]thiophene (IDTT),[24] terthieno[3,2-*b*]thiophene (6T)[77] and so forth. These conjugated fragments are responsible for the energy values of the acceptor's LUMO and donor's HOMO.[78] However, such classification does not immediately yield a partitioning of all non-fullerene acceptors into distinct non-overlapping groups. If the classification is too coarse-grained (*e.g.*, molecules containing thiophenes), too many elements of different families will be grouped together; and if the classification is too fine-grained, there could be just one element in each group. We started by considering the type of fragments present in each acceptor (*e.g.*, PDI, DPP, *etc.*). We show in Table S1† how we fingerprinted the

Open Access Article. Published on 25 March 2022. Downloaded on 6/22/2024 3:41:12 AM.
This article is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported Licence.

33 distinct acceptors in the database, by considering which of these fragments they contained. When it was ambiguous in which group to classify a particular acceptor, priority was given to the fragment where the LUMO – which gives the acceptor character – is localized. Any acceptor fragment containing more than 50% weight of the LUMO orbital density can be considered as a valid fragment, as shown in more detail in the ESI.† Note that this grouping process is relevant to the organic solar cells field, where groups of acceptor molecules with unique fragments have emerged over time, but specific knowledge of a field would be necessary to group data points into novel groups, as there is no unique standardized definition of novelty. At the end of this process, we have identified five groups of non-fullerene acceptors (G1-5): PDIs (with one or more monomers), DPP, BT, IID-T (IID unit connected with thiophene) and IC, which are illustrated in Fig. 3. In total, we classify 45 donor/acceptor pairs with non-fullerene acceptors in our five groups. Each of these groups contains between four and 25 donor/acceptor pairs from our database that broadly match the classification proposed in the literature.[69]

IV. LOO *vs.* LOGO

(a). LOGO extrapolation capabilities

In Table 1, we show the resulting RMSE in the ML prediction of several OPV molecules comparing a regular cross-validation (LOO) with a LOO-extrapolation and a LOGO extrapolation performed on the five groups identified in Fig. 3. When the extrapolation RMSE is much larger than the LOO RMSE, the ML method fails to describe the new family on the basis of existing knowledge. It has been studied before that fingerprints yield a similar or even better accuracy than physical descriptors when performing interpolation tasks.[46] However, this trend is reversed when trying to predict the PCE of new groups by training the model to extrapolate, where using physical descriptors is significantly better. This is an important point, as it shows that when one is trying to make predictions of molecules similar to those known by the model, one can rely solely on fingerprints to make accurate predictions, but one needs to use physical descriptors (and therefore a minimal understanding of the physics of the material) when predicting molecules from a new chemical family. Note that we are interested in describing extrapolations in structure space: predict the PCE of molecules that contain chemical groups and motifs unknown to the model. However, such an extrapolation does not necessarily mean an extrapolation of the physical descriptors, which are continuous.

A preliminary assessment of the data is offered by comparing the RMSE, as shown in Table 1. In this table, we can see how the correlation coefficient is very low and RMSE is large when using the fingerprints with either LOO-extrapolation or LOGO-extrapolation, indicating that extrapolating data based on this information alone is more challenging. We can also see how the LOGO extrapolation presents a clear improvement over the LOO-extrapolation. The best performance obtained with LOO-extrapolation results in a RMSE of 3.52%, and LOGO-extrapolation improves the RMSE to 2.84% (a relative improvement of 19%). Similarly, the best correlation obtained with LOGO-extrapolation ($r = 0.31$) is also significantly larger than the correlation obtained with LOO-extrapolation (0.08–0.17). To see this improvement when using LOGO-extrapolation, using a grouping with chemically distinct families, we present in Section S6.2 of the ESI† a comparison of LOO-extrapolation and LOGO-extrapolation when using another

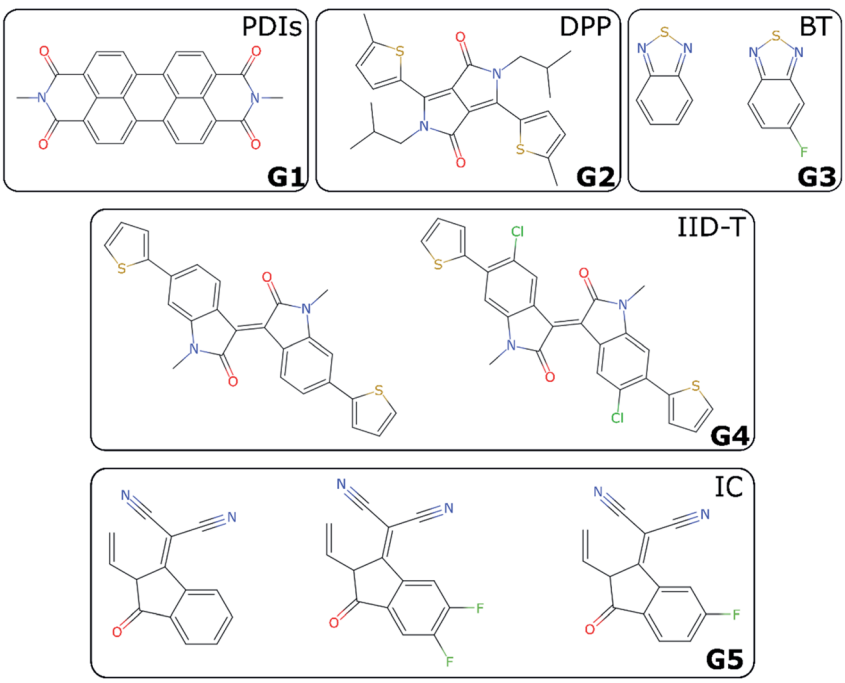

Fig. 3 Illustration of investigated groups containing different fragments.

© 2022 The Author(s). Published by the Royal Society of Chemistry

Open Access Article. Published on 25 March 2022. Downloaded on 6/22/2024 3:41:12 AM.
This article is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported Licence.

Table 1 RMSE (%) and Pearson correlation coefficient, r, results of PCE prediction using different cross-validation methods, with KRR. We have tried different features: fingerprints or physical descriptors

	Features	RMSE (%)	r
LOO-interpolation	Fingerprints	1.75	0.69
	Physical descriptors	2.01	0.56
LOO-extrapolation	Fingerprints	3.52	0.08
	Physical descriptors	4.11	0.17
LOGO-extrapolation	Fingerprints	3.77	0.07
	Physical descriptors	2.84	0.31

grouping. The best RMSE achieved with LOGO is still far from the one obtained with a regular LOO cross-validation (interpolating known data), and it results in a coefficient of determination of $R^2 = 0.04$, which makes it not suitable to perform quantitative predictions. However, researchers are not necessarily interested in a model that can make accurate quantitative predictions, and a model that can do a binary classification to separate candidate materials into well-performing and bad-performing can be equally helpful.

We show in Fig. 4 the PCE distribution of the complete dataset and the distribution for the 45 points corresponding to the non-fullerene materials classified into one of our five chemical groups, which we will try to predict. In Fig. 5, we show our PCE prediction of these 45 points when using LOO-extrapolation and LOGO-extrapolation, as well as the experimental PCE. One can clearly see how the LOGO-extrapolation distribution is much closer than the LOO-extrapolation one to the experimental PCE distribution (closer median and lower/upper quartile values), although both models struggle to predict high PCE values.

We have already mentioned the overall smaller RMSE and larger correlation (see Table 1) when adopting LOGO-extrapolation. However, we can go one step further and

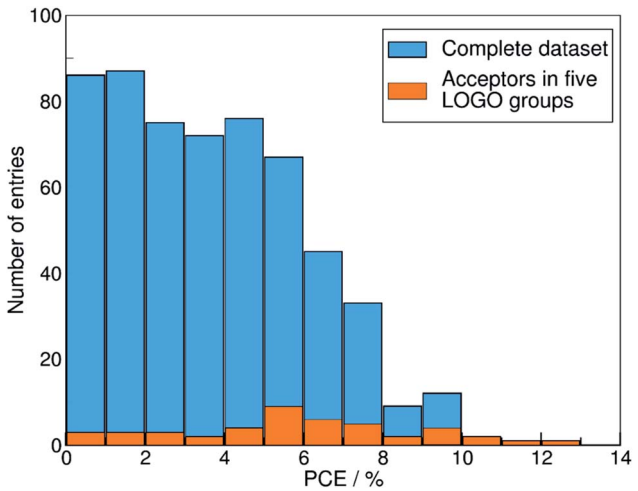

Fig. 4 Histogram with the PCE distribution of our complete dataset (in blue), and the PCE distribution of the 45 pairs with non-fullerene acceptors classified in the five groups used in the LOGO-extrapolation (in orange).

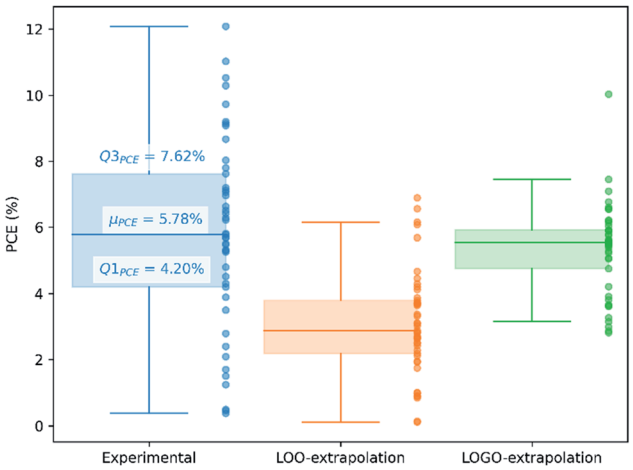

Fig. 5 PCE distribution of non-fullerene acceptors in our database, indicating the lower quartile ($Q1_{PCE}$), median (μ_{PCE}) and upper quartile ($Q3_{PCE}$). We also show the predicted PCE distributions with LOO-extrapolation and LOGO-extrapolation.

quantify how advantageous LOGO-extrapolation would be with respect to LOO-extrapolation when trying to identify materials over a certain threshold. We have used three different thresholds. We have chosen the lower quartile ($Q1_{PCE}$), median value (μ_{PCE}) and upper quartile ($Q3_{PCE}$) as statistically significant values to judge the ability of the model to do qualitative classifications (the threshold values are shown in Fig. 5). These thresholds allow to judge how the model performs for different classifications with an increasing difficulty:

(i) Identify materials with PCE > $Q1_{PCE}$. This simple threshold allows us to evaluate how well the model would do in identifying the worse performing materials.

(ii) Identify materials with PCE > μ_{PCE}. With this threshold, we have the best data distribution and we can classify candidate materials as well-performing and bad-performing, reducing possible candidates by half.

(iii) Identify materials with PCE > $Q3_{PCE}$. This threshold is more challenging, and allow us to quantify how many materials in the top 25% of our dataset are correctly predicted within that range.

Each predicted PCE can be classified as true positive (TP), false positive (FP), true negative (TN) or false negative (FN), as shown in Table 2.

We can directly measure the accuracy of our model by calculating the probability of making a correct prediction:

$$A = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}}. \quad (5)$$

Additionally, we also calculate the precision, P, as the fraction of predicted well-performing materials that are actually well-performing,

$$P = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \quad (6)$$

Open Access Article. Published on 25 March 2022. Downloaded on 6/22/2024 3:41:12 AM.
This article is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported Licence.

Table 2 Confusion matrix when classifying PCE above a threshold

	Predicted PCE > threshold	Predicted PCE < threshold
Actual PCE > threshold	True positive (TP)	False negative (FN)
Actual PCE < threshold	False positive (FP)	True negative (TN)

Table 3 Metrics obtained with a LOO-extrapolation and LOGO-extrapolation, with KRR, when classifying materials with PCE > $Q1_{PCE}$ and PCE > μ_{PCE}

	PCE > $Q1_{PCE}$		PCE > μ_{PCE}	
Metric	LOO-extrapolation	LOGO-extrapolation	LOO-extrapolation	LOGO-extrapolation
TP	6	30	2	8
FP	3	5	2	7
TN	9	6	21	15
FN	27	4	20	15
Accuracy	0.33	0.80	0.51	0.51
Precision	0.67	0.86	0.50	0.53
Recall	0.18	0.88	0.09	0.35
F_1-Score	0.29	0.87	0.15	0.42

and the recall, R, as the fraction of actual well-performing materials that are predicted to be well-performing

$$R = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}. \quad (7)$$

These metrics are common for these types of binary classifications,[79,80] and they are often averaged in a single metric, the F_1-score, which we use as another indicator of the classification accuracy:

$$F_1 = 2 \times \frac{P \times R}{P + R}. \quad (8)$$

We show the results for the two thresholds PCE > $Q1_{PCE}$ and PCE > μ_{PCE} in Table 3. For the lowest threshold of PCE > $Q1_{PCE}$, it is worth noting how a LOO-extrapolation results in a mediocre accuracy and F_1-score ($A = 0.33$ and $F_1 = 0.29$). The precision is relatively high ($P = 0.67$), but the recall is quite low ($R = 0.18$), which means that we are missing most of the actual well-performing materials. However, when we use a LOGO cross-validation, both the accuracy and F_1-score are significantly larger ($A = 0.80$ and $F_1 = 0.87$), and the number of false negatives is reduced further, resulting in a larger recall ($R = 0.88$). In other words, if no special care is taken, an ML algorithm trained with a regular cross-validation performs very poorly when trying to predict new classes of molecules. However, a suitable cross-validation, like the one proposed in the LOGO-extrapolation, can significantly improve the predictive power of the ML model. When we use a higher threshold, PCE > μ_{PCE}, we observe a similar trend, although now both LOO-extrapolation and LOGO-extrapolation result in the same accuracy ($A = 0.51$) and a similar precision ($P = 0.50$ and $P = 0.53$, respectively). However, LOGO-extrapolation still results in a much larger recall and F_1-score ($R = 0.35$ and $F_1 = 0.42$) when compared to LOO-extrapolation ($R = 0.09$ and $F_1 = 0.15$). When we use the most challenging threshold of PCE > $Q3_{PCE}$, both LOO-extrapolation and LOGO-extrapolation struggle and they are not able to correctly predict any of the points in that interval. Therefore, it seems that this approach is advantageous to identify low-performing materials, but is unable to correctly identify high-performing materials. We show in Fig. S2 in the ESI† all real and predicted values when using LOO-extrapolation and LOGO-extrapolation.

(b). LOGO convergence to LOO

The relative ability to predict completely new chemistries is intuitively related to the novelty of the molecules to be considered, with respect to the training set. To better understand the

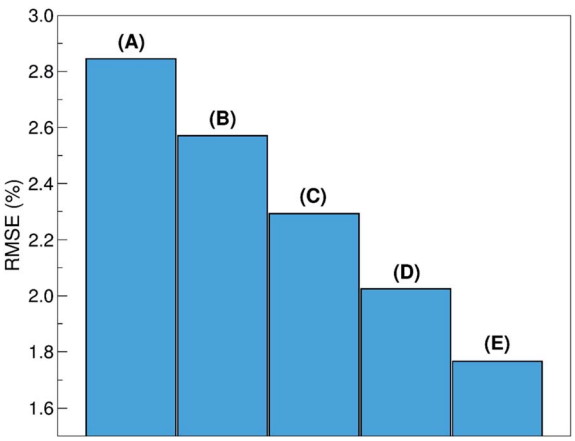

Fig. 6 Prediction RMSE, using LOGO-extrapolation with different groups: (A) non-fullerene acceptors split in five chemically distinct families; (B) non-fullerene acceptors split in five random groups; (C) all acceptors split in 33 random groups; (D) all acceptors split in 55 random groups. (E) Prediction RMSE using LOO-interpolation.

© 2022 The Author(s). Published by the Royal Society of Chemistry

Open Access Article. Published on 25 March 2022. Downloaded on 6/22/2024 3:41:12 AM.
This article is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported Licence.

relation between the proposed LOGO-extrapolation and the conventional LOO-interpolation, it is instructive to see how the prediction RMSE is affected by changing the definition of groups, which we exemplify in Fig. 6 through the use of five models. The prediction RMSE is the largest when we classify the donor/acceptor pairs with non-fullerene acceptors in the five groups defined on the basis of chemical similarity, as in Section III (model A). The RMSE improves as the pairs with non-fullerene acceptor molecules are grouped in five random groups (model B). The RMSE is reduced further when we consider all donor/acceptor pairs, where each group contains just the pairs with one of the 33 acceptor molecules (model C), and even more if we increase the number of groups to 55 (model D), where some acceptor molecules are present in more than one group. The best RMSE value is found when the LOO cross-validation is used (model E). This progression of cases exemplifies the gradual change from a model performing pure extrapolation (less accurate but more useful) toward a model performing pure interpolation (more accurate but less useful). We show in the ESI† more details on how these groups were selected.

V. Example of model deployment

As any other ML methodology, this approach can be used to evaluate the efficiency of molecules not present in the data set, which can be obtained from experimental intuition, a database search or a more complicated generative model.[23] To illustrate the possible applications, we identified nine molecules from a database of computed electronic properties of known molecules developed in ref. 81. We considered molecules with physical descriptors in the same range of high performing non-fullerene acceptors: LUMO energy lower than -2.85 eV and $\sum f$ larger than 0.8, and we have added alkyl side chains to match more closely the solubility parameter of the best performing examples. See Section S7 of the ESI† for more details of how

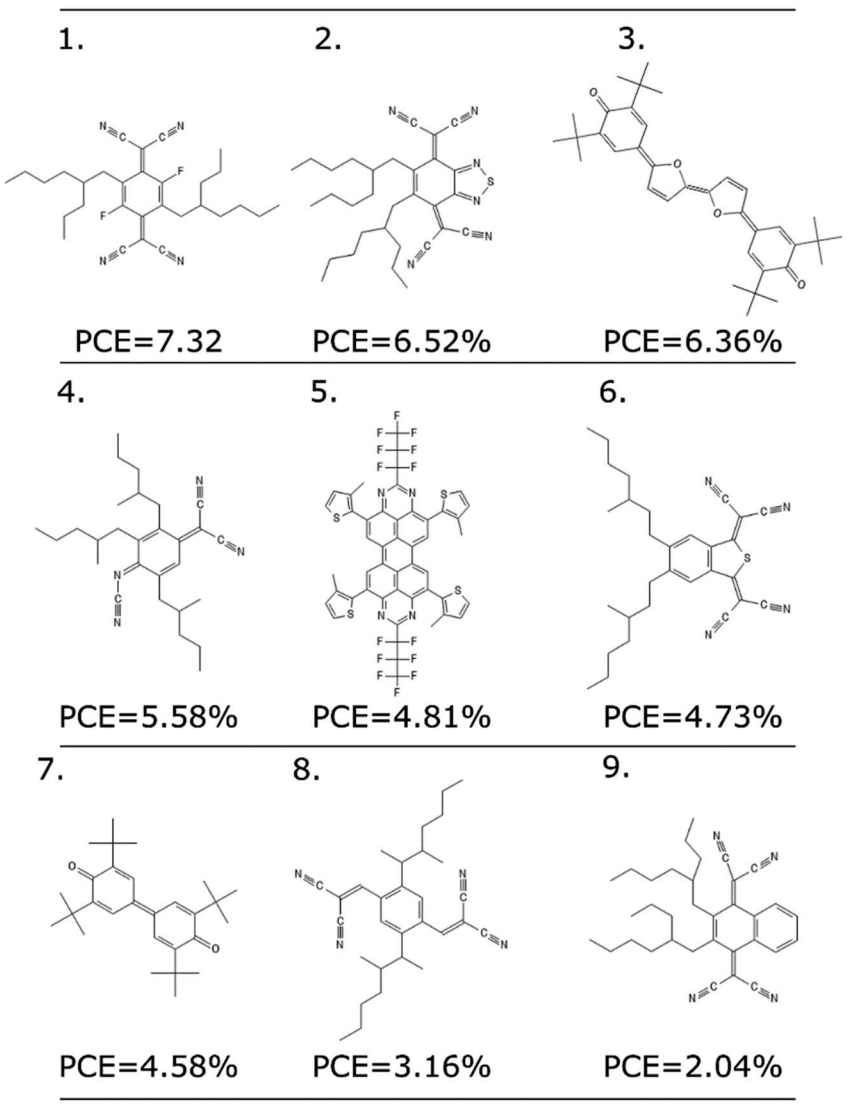

Fig. 7 Nine molecules unknown to the model, and their suggested PCE using the LOGO-extrapolation framework, with KRR and physical descriptors.

Open Access Article. Published on 25 March 2022. Downloaded on 6/22/2024 3:41:12 AM.
This article is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported Licence.

these molecules were selected and the optimized hyperparameters of the model.

We excluded molecules that belonged to any of the known classes of non-fullerene acceptors in our dataset. Fig. 7 reports nine molecules, along with their predicted PCE, when combined with the best performing donor in our database (ZnP-TBO[82]). We can see how the PCE range of these molecules is similar to the range predicted for the other non-fullerene acceptor groups (see Fig. 5), and three of them have a predicted PCE above the median value for all other NFA in the database, which suggests them as high interest candidates. Additional considerations like cost and ease of synthesis from precursor could be considered, and domain knowledge would be particularly critical in the design phase.

VI. Conclusions

In summary, we proposed a training method to improve the extrapolation capabilities of ML models to materials from completely new families, and we applied it to the prediction of the performance of non-fullerene acceptors. Our results show that the quantitative prediction error in these extrapolative tasks is larger than the one achieved by predicting molecules within known chemical families. To address this shortcoming, we propose a method based on a leave-one-group-out (LOGO) cross-validation, in which we train an ML model to accurately perform extrapolations on unseen families of compounds. We have shown this choice results in an RMSE improvement and a significant success rate in classifying unseen materials above and below the median efficiency, when compared to a method trained with an ordinary cross-validation. We have observed that when extrapolating to new chemical families, physical descriptors are needed to make accurate predictions, and cheap fingerprints are not enough, unlike when one is predicting molecules within known chemical families. Finally, we have seen how the LOGO cross-validation accuracy converges towards the accuracy obtained with a regular cross-validation when one ignores the existence of chemical families. These results are promising and suggest that, in fields where data becomes structured into recognizable families, using training methods like the proposed LOGO cross-validation can accelerate the discovery of completely new families.

Data availability

The dataset used during the current study, as well as the code used to perform the analysis and scripts to reproduce the main results are available in a public GitHub repository at www.github.com/marcosdelcueto/NonFullereneAcceptorPrediction.

Author contributions

A. T. designed the project. M. d. C and A. T. developed the theoretical formalism. Z.-W. Z. created the database. M. d. C. performed the ML analysis. All authors contributed to the analysis of the results and writing of the manuscript.

Conflicts of interest

There are no conflicts to declare.

Acknowledgements

Z.-W. Z. acknowledges the support of the China Scholarship Council. A. T. acknowledges the support of EPSRC and ERC. We thank Mr Rex Manurung, Mr Ömer H. Omar and Dr Daniele Padula for fruitful discussions.

References

1 P. S. Gromski, A. B. Henson, J. M. Granda and L. Cronin, *Nat. Rev. Chem.*, 2019, **3**, 119–128.
2 M. Awale, R. Visini, D. Probst, J. Arus-Pous and J. L. Reymond, *Chimia*, 2017, **71**, 661–666.
3 C. W. Coley, N. S. Eyke and J. F. Jensen, *Angew. Chem., Int. Ed.*, 2019, **59**, 23414–23436.
4 A. Mahmood and J.-L. Wang, *Energy Environ. Sci.*, 2021, **14**, 90–105.
5 A. O. Oliynyk and J. M. Buriak, *Chem. Mater.*, 2019, **31**, 8243–8247.
6 C. Chen, Y. Zuo, W. Ye, X. Li, Z. Deng and S. P. Ong, *Adv. Energy Mater.*, 2020, **10**, 1903242.
7 J. Schmidt, M. R. G. Marques, S. Botti and M. A. L. Marques, *npj Comput. Mater.*, 2019, **5**, 83.
8 X. Rodríguez-Martínez, E. Pascual-San-José and M. Campoy-Quiles, *Energy Environ. Sci.*, 2021, **14**, 3301–3322.
9 Y. Liu, O. C. Esan, Z. Pan and L. An, *Energy and AI*, 2021, **3**, 100049.
10 Z. Zhou, Y. Zhou, Q. He, Z. Ding, F. Li and Y. Yang, *npj Comput. Mater.*, 2019, **5**, 128.
11 S. Wu, Y. Kondo, M. Kakimoto, B. Yang, H. Tamada, I. Kuwajima, G. Lambard, K. Hongo, Y. Xu, J. Shiomi, C. Schick, J. Morikawa and R. Yoshida, *npj Comput. Mater.*, 2019, **5**, 66.
12 P. V. Balachandran, B. Kowalski, A. Sehirlioglu and T. Lookman, *Nat. Commun.*, 2018, **9**, 1668.
13 C. C. Fischer, K. J. Tibbetts, D. Morgan and G. Ceder, *Nat. Mater.*, 2006, **5**, 641–646.
14 G. Hautier, C. C. Fischer, A. Jain, T. Mueller and G. Ceder, *Chem. Mater.*, 2010, **22**, 3762–3767.
15 G. Hautier, C. Fischer, V. Ehrlacher, A. Jain and G. Ceder, *Inorg. Chem.*, 2011, **50**, 656–663.
16 P. Dey, J. Bible, D. Datta, S. Broderick, J. Jasinski, M. Sunkara, M. Menon and K. Rajan, *Comput. Mater. Sci.*, 2014, **83**, 185–195.
17 A. O. Oliynyk, L. A. Adutwum, J. J. Harynuk and A. Mar, *Chem. Mater.*, 2016, **28**, 6672–6681.
18 K. Ryan, J. Lengyel and M. Shatruk, *J. Am. Chem. Soc.*, 2018, **140**, 10158–10168.
19 B. Kailkhura, B. Gallagher, S. Kim, A. Hiszpanski and T. Y.-J. Han, *npj Comput. Mater.*, 2019, **5**, 108.
20 J. B. O. Mitchell, *Wiley Interdiscip. Rev.: Comput. Mol. Sci.*, 2014, **4**, 468–481.

© 2022 The Author(s). Published by the Royal Society of Chemistry

Open Access Article. Published on 25 March 2022. Downloaded on 6/22/2024 3:41:12 AM. This article is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported Licence.

21 D. Krstajic, L. J. Buturovic, D. E. Leahy and S. Thomas, *J. Cheminf.*, 2014, **6**, 10.
22 H. Sahu, W. Rao, A. Troisi and H. Ma., *Adv. Energy Mater.*, 2018, **8**, 1801032.
23 S.-P. Peng and Y. Zhao, *J. Chem. Inf. Model.*, 2019, **59**, 4993–5001.
24 A. Wadsworth, M. Moser, A. Marks, M. S. Little, N. Gasparini, C. J. Brabec, D. Baran and I. McCulloch, *Chem. Soc. Rev.*, 2019, **48**, 1596–1625.
25 J. Hou, O. Inganäs, R. H. Friend and F. Gao, *Nat. Mater.*, 2018, **17**, 119–128.
26 C. Yan, S. Barlow, Z. Wang, H. Yan, A. K.-Y. Jen, S. R. Marder and X. Zhan, *Nat. Rev. Mater.*, 2018, **3**, 18003.
27 Y. Lin, J. Wang, Z.-G. Zhang, H. Bai, Y. Li, D. Zhu and X. Zhan, *Adv. Mater.*, 2015, **27**, 1170–1174.
28 Y. Cui, H. Yao, J. Zhang, K. Xian, T. Zhang, L. Hong, Y. Wang, Y. Xu, K. Ma, C. An, C. He, Z. Wei, G. Gao and J. Hou, *Adv. Mater.*, 2020, **32**, 1908205.
29 B. Meredig, E. Antono, C. Church, M. Hutchinson, J. Ling, S. Paradiso, B. Blaiszik, I. Foster, B. Gibbons, J. Hattrick-Simpers, A. Mehta and L. Ward, *Mol. Syst. Des. Eng.*, 2018, **3**, 819–825.
30 Z. Xiong, Y. Cui, Z. Liu, Y. Zhao, M. Hu and J. Hu, *Comput. Mater. Sci.*, 2020, **171**, 109203.
31 D. T. Ahneman, J. G. Estrada, S. Lin, S. D. Dreher and A. G. Doyle, *Science*, 2018, **360**, 6385.
32 J. G. Estrada, D. T. Ahneman, R. P. Sheridan, S. D. Dreher and A. G. Doyle, *Science*, 2018, **362**, 6416.
33 W. Jin, R. Barzilay and T. Jaakkola, arXiv:2006,03908, 2020.
34 P. Cheng, G. Li, X. Zhan and Y. Yang, *Nat. Photonics*, 2018, **12**, 131–142.
35 H. Sahu, F. Yang, X. Ye, J. Ma, W. Fang and H. Ma, *J. Mater. Chem. A*, 2019, **7**, 17480–17488.
36 S. A. Lopez, E. O. Pyzer-Knapp, G. N. Simm, T. Lutzow, K. Li, L. R. Seress, J. Hachmann and A. Aspuru-Guzik, *Sci. Data*, 2016, **3**, 160086.
37 A. Paul, A. Furmanchuk, W. K. Liao, A. Choudhary and A. Agrawal, *Mol. Inf.*, 2019, **38**, e1900038.
38 N. Meftahi, M. Klymenko, A. J. Christofferson, U. Bach, D. A. Winkler and S. P. Russo, *npj Comput. Mater.*, 2020, **6**, 166.
39 Y. Wu, J. Guo, R. Sun and J. Min, *npj Comput. Mater.*, 2020, **6**, 120.
40 J. Li, B. Pradhan, S. Gaur and J. Thomas, *Adv. Energy Mater.*, 2019, **9**, 1901891.
41 T. W. David, H. Anizelli, T. J. Jacobsson, C. Gray, W. Teahan and J. Kettle, *Nano Energy*, 2020, **78**, 105342.
42 B. Krawczyk, *Prog. Artif. Intell.*, 2016, **5**, 221–232.
43 P. Raccuglia, K. C. Elbert, P. D. F. Adler, C. Falk, M. B. Wenny, A. Mollo, M. Zeller, S. A. Friedler, J. Schrier and A. J. Norquist, *Nature*, 2016, **533**, 73–76.
44 B. Cao, L. A. Adutwum, A. O. Oliynyk, E. J. Luber, B. C. Olsen, A. Mar and J. M Buriak, *ACS Nano*, 2018, **12**, 7434–7444.
45 M. del Cueto and A. Troisi, *Phys. Chem. Chem. Phys.*, 2021, **23**, 14156–14163.
46 Z.-W. Zhao, M. del Cueto, Y. Geng and A. Troisi, *Chem. Mater.*, 2020, **32**, 7777–7787.
47 M. del Cueto, Non-Fullerene Acceptor Prediction, github.com/marcosdelcueto/NonFullereneAcceptorPrediction, 2022.
48 D. Padula and A. Troisi, *Adv. Energy Mater.*, 2019, **9**, 1902463.
49 D. Rogers and M. Hahn, *J. Chem. Inf. Model.*, 2010, **50**, 742–754.
50 C. Schober, K. Reuter and H. Oberhofer, *J. Phys. Chem. Lett.*, 2016, **7**, 3973–3977.
51 S. A. Lopez, B. Sanchez-Lengeling, J. G. Soares and A. Aspuru-Guzik, *Joule*, 2017, **1**, 857.
52 T. Cheng, Y. Zhao, X. Li, F. Lin, Y. Xu, X. Zhang, Y. Li, R. Wang and L. Lai, *J. Chem. Inf. Model.*, 2007, **47**, 2140–2148.
53 R. Mannhold, G. I. Poda, C. Ostermann and I. V. Tetko, *J. Pharm. Sci.*, 2009, **98**, 861–893.
54 A. Daina, O. Michielin and V. Zoete, *Sci. Rep.*, 2017, **7**, 42717.
55 D. Padula, J. D. Simpson and A. Troisi, *Mater. Horiz.*, 2019, **6**, 343–349.
56 H. Jung, S. Stocker, C. Kunkel, H. Oberhofer, B. Han, K. Reuter and J. T. Margraf, *ChemSystemsChem*, 2020, **2**, e1900052.
57 N. S. Altman, *Am. Stat.*, 1992, **46**, 175–185.
58 A. J. Smola and B. A. Schölkopf, *Stat. Comput.*, 2004, **14**, 199–222.
59 C. Chen, Y. Zuo, W. Ye, X. Li, Z. Deng and S. P. Ong, *Adv. Energy Mater.*, 2020, **10**, 1903242.
60 D. Bajusz, A. Racz and K. Heberger, *J. Cheminf.*, 2015, **7**, 1–13.
61 R. Storn and K. Price, *J. Glob. Optim.*, 1997, **11**, 341–359.
62 P. Virtanen, *et al.*, *Nat. Methods*, 2020, **17**, 261–272.
63 C. Finn, P. Abbeel and S. Levine, *Proceedings of the 38th International Conference on Machine Learning*, 2017, vol. 70, pp. 1126–1135.
64 Y. Bai, M. Chen, P. Zhou, T. Zhao, J. Lee, S. Kakade, H. Wang and C. Xiong, *Proceedings of the 38th International Conference on Machine Learning*, 2021, vol. 139, pp. 543–553.
65 G. Zhang, J. Zheo, P. C. Y. Chow, K. Jiang, J. Zhang, Z. Zhu, J. Zhang, F. Huang and H. Yan, *Chem. Rev.*, 2018, **118**, 3447–3507.
66 J. Zhang, H. S. Tan, X. Guo, A. Facchetti and H. Yan, *Nat. Energy*, 2018, **3**, 720–731.
67 C. B. Nielsen, S. Holliday, H. Y. Chen, S. J. Cryer and I. McCulloch, *Acc. Chem. Res.*, 2015, **48**, 2803–2812.
68 S. Li, W. Liu, C.-Z. Li, M. Shi and H. Chen, *Small*, 2017, **13**, 1701120.
69 B. Zhang, N. An, H. Wu, Y. Geng, Y. Sun, Z. Ma, W. Li, Q. Guo and E. Zhou, *Sci. China: Chem.*, 2020, **63**, 1262–1271.
70 J. Zhang, X. Zhang, H. Xiao, G. Li, Y. Liu, C. Li, H. Huang, X. Chen and Z. Bo, *ACS Appl. Mater. Interfaces*, 2016, **8**, 5475–5483.
71 Suman and S. P. Singh, *J. Mater. Chem. A*, 2019, **7**, 22701–22729.
72 J. C. Bijleveld, V. S. Gevaerts, D. Di Nuzzo, M. Turbiez, S. G. J. Mathijssen, D. de Leeuw, M. M. Wienk and R. A. J. Janssen, *Adv. Mater.*, 2010, **22**, E242–E246.
73 Suman, A. Bagui, R. Datt, V. Gupta and S. P. Singh, *Chem. Commun.*, 2017, **53**, 12790–12793.
74 F. Wu, L. Zhong, H. Hiu, Y. Li, Z. Zhang, Y. Li, Z.-G. Zhang, H. Ade, Z.-Q. Jiang and L.-S. Liao, *J. Mater. Chem. A*, 2019, **7**, 4063–4071.

Open Access Article. Published on 25 March 2022. Downloaded on 6/22/2024 3:41:12 AM.
This article is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported Licence.

75 H. Wu, H. Fan, S. Xu, C. Zhang, S. Chen, C. Yang, D. Chen, F. Liu and X. Zhu, *Sol. RRL*, 2017, **1**, 1700165.
76 H. Wu, H. Fan, S. Xu, L. Ye, Y. Guo, Y. Yi, H. Ade and X. Zhu, *Small*, 2019, **15**, 1804271.
77 X. Shi, J. Chen, K. Gao, L. Zuo, Z. Yao, F. Liu, J. Tang and A. K.-Y. Jen, *Adv. Energy Mater.*, 2018, **8**, 1702831.
78 A. Kuzmich, D. Padula, H. Ma and A. Troisi, *Energy Environ. Sci.*, 2017, **10**, 395–401.
79 W. Li, R. Jacobs and D. Morgan, *Comput. Mater. Sci.*, 2018, **150**, 454–463.
80 L. Weston and C. Stampfl, *Phys. Rev. Mater.*, 2018, **2**, 085407.
81 D. Padula, O. H. Omar, T. Nematiaram and A. Troisi, *Energy Environ. Sci.*, 2019, **12**, 2412–2416.
82 K. Gao, S. B. Jo, X. Shi, L. Nian, M. Zhang, Y. Kan, F. Lin, B. Kan, B. Xu, Q. Rong, L. Shui, F. Liu, X. Peng, G. Zhou, Y. Cao and A. K.-Y. Jen, *Adv. Mater.*, 2019, **31**, 1807842.

© 2022 The Author(s). Published by the Royal Society of Chemistry