npj Computational Materials

www.nature.com/npjcompumats

ARTICLE OPEN

Check for updates

Machine learning property prediction for organic photovoltaic devices

Nastaran Meftahi[1]✉, Mykhailo Klymenko[1], Andrew J. Christofferson[2], Udo Bach[3], David A. Winkler[4,5,6,7] and Salvy P. Russo[1]✉

Organic photovoltaic (OPV) materials are promising candidates for cheap, printable solar cells. However, there are a very large number of potential donors and acceptors, making selection of the best materials difficult. Here, we show that machine-learning approaches can leverage computationally expensive DFT calculations to estimate important OPV materials properties quickly and accurately. We generate quantitative relationships between simple and interpretable chemical signature and one-hot descriptors and OPV power conversion efficiency (PCE), open circuit potential (V_{oc}), short circuit density (J_{sc}), highest occupied molecular orbital (HOMO) energy, lowest unoccupied molecular orbital (LUMO) energy, and the HOMO–LUMO gap. The most robust and predictive models could predict PCE (computed by DFT) with a standard error of ±0.5 for percentage PCE for both the training and test set. This model is useful for pre-screening potential donor and acceptor materials for OPV applications, accelerating design of these devices for green energy applications.

npj Computational Materials (2020)6:166 ; https://doi.org/10.1038/s41524-020-00429-w

INTRODUCTION

Worrying increases in anthropomorphic greenhouse gas emissions have driven a strong expansion of research into discovery, design, and optimization of materials for energy applications. Organic photovoltaic (OPV) materials are of great interest because of their potential to generate cheap, printable semiconductor devices that convert light into electrical energy. They promise sustainable sources of clean energy if their efficiencies and stabilities can be improved. Organic solar cell manufacture is intrinsically a simple and low-cost process, and devices can be lightweight and flexible[1,2]. However, the relationships between materials properties, device configuration, and performance are complex and often poorly understood. Given the potentially vast number of materials and device configurations possible exhaustive experimentation, even using high throughput methods, cannot guarantee finding the highest performing materials.

The power conversion efficiency (PCE, % incident light energy converted to electricity) is one of the most crucial properties for OPV solar cells. Density functional theory (DFT) can calculate several important properties of photovoltaic materials that affect PCE:[3] V_{oc} (open circuit potential), J_{sc} (short circuit density), energy of the donor highest occupied molecular orbital (HOMO), energy of the acceptor lowest unoccupied molecular orbital (LUMO), and the HOMO–LUMO gap, but this requires extensive computational resources and time. Many physiochemical phenomena relating to light absorption in solar cells, such as the exciton formation[4] and migration[5] process, charge transport[6] and recombination, need to be considered[7,8].

Machine learning (ML) can potentially model the complex relationships between materials, device properties, and OPV performance, given sufficient data, allowing efficient leveraging of expensive and time-consuming experiments and quantum chemical calculations. Carefully chosen, a relatively small number of DFT calculations, validated by experiments, can train ML models that predict relevant OPV properties for materials not yet synthesized. Apart from the availability of sufficient and reliable training data, the most important element of ML models is the choice of descriptors, mathematical representations of the structural and physicochemical properties of the donors and acceptors used in the OPV devices. Clearly, device construction parameters are also relevant and can be included in the models if they are available. Different ML algorithms often give similar quality models for a given set of descriptors, whereas a given ML algorithm trained on different types of descriptors can generate models of highly variable quality[9]. Many types of molecular descriptors are available, including topological, electronic, geometrical, molecular fragment, and quantum chemical, among others.

ML approaches have been a popular choice for predicting photovoltaic properties[10]. For example, Padula and co-workers modeled the photovoltaic properties of 249 organic donor–acceptors pairs to using k-NN (k-nearest neighbor)[11] regression and kernel ridge regression[12,13] methods trained on a combination of electronic and structural parameters[14]. Sahu et al. used random forest (RF)[15,16], gradient boosting (GB)[17], and deep neural networks (DNN) to model PCE for 280 small OPV molecules using 13 microscopic properties of as descriptors to train the models. The models predicted PCE for 30 molecules in a test set with modest R^2 values of 0.44, 0.50, 0.46, 0.61, and 0.62 for linear regression, k-NN, neural network, RF, GB models, respectively[18]. Root-mean-square error (RMSE) values, a robust estimate of model quality[19], ranged from 1.07% PCE for GB to 1.34% PCE for linear regression. Note: as PCE is the percentage conversion of light energy, in this paper when we refer to standard error or RMSE values we mean the uncertainty in this property (e.g. 10.6 ± 0.5%),

[1]ARC Centre of Excellence in Exciton Science, School of Science, RMIT University, Melbourne, VIC 3001, Australia. [2]School of Science, College of Science, Engineering and Health, RMIT University, Melbourne, VIC 3001, Australia. [3]ARC Centre of Excellence in Exciton Science, Department of Chemical Engineering, Monash University, Wellington Road, Clayton, VIC 3800, Australia. [4]La Trobe Institute for Molecular Science, La Trobe University, Kingsbury Drive, Bundoora, VIC 3086, Australia. [5]Monash Institute of Pharmaceutical Sciences, Monash University, Parkville, VIC 3052, Australia. [6]School of Pharmacy, University of Nottingham, Nottingham NG7 2QL, UK. [7]CSIRO Data61, Pullenvale, QLD 4069, Australia. ✉email: nastaran.meftahi@rmit.edu.au; salvy.russo@rmit.edu.au

Published in partnership with the Shanghai Institute of Ceramics of the Chinese Academy of Sciences

npj nature partner journals

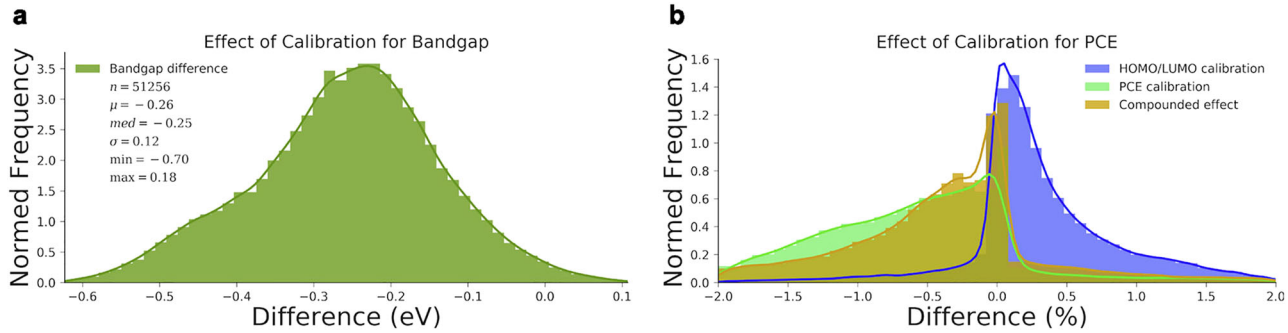

Fig. 1 Effects of calibration on HOMO–LUMO gap and PCE. The distributions for **a** HOMO–LUMO gap and **b** PCE are presented as a histogram and a kernel density estimate from the difference between a baseline and a calibration. The band gap also shows the general statistics of the distribution. Reprinted with permission from ref. [26]. Copyright 2017, with permission from Elsevier.

not the percentage error in this property (e.g. $100 \times 0.5/10.6$). Pereira and co-workers[20] also generated ML models for the energy of the HOMO and LUMO of OPV materials. They used a dataset of 111,725 molecules, fingerprint and modified distance descriptors, and RF[15,16], support vector machine[21], and a standard feedforward neural network to perform feature selection and property modeling. They found that the RF algorithm trained on modified distance descriptors generated the best predictions for HOMO and LUMO for an external test set of 9989 compounds, with R^2 values of 0.89 and 0.93 and RMSE of 0.21 and 0.23 eV for the HOMO and LUMO energies, respectively. The HOMO–LUMO gap could be predicted with an R^2 of 0.91 and RMSE of 0.30 eV.

Although ML methods can model OPV properties well, one of the main problems is that the models are opaque, and the descriptors used to train them arcane. It is hard to extract information from the models that is useful for designing improved OPV materials. Here we show how efficient and chemically interpretable descriptors that can be computed quickly and do not require additional experimental measurements or resource-intensive DFT calculations can predict important OPV properties with good accuracy.

RESULTS AND DISCUSSION

Model development

Here, we used the Harvard Photovoltaic Dataset (HOPV15) dataset[22] that includes data from quantum chemical calculations and that calculated by the Scharber model plus experimental properties collected from literature[23]. The Scharber model uses a single parameter, the computed HOMO–LUMO gap, in which V_{oc} is assumed to be the HOMO–LUMO gap minus a band offset, J_{sc} is assumed to be 0.65 of the current resulting from absorbing all incident photons above the HOMO–LUMO gap, and FF is set 0.65 (ref. [24]).

Our aim is to demonstrate that simple, interpretable molecular descriptors and ML methods can model and predict important OPV properties. While it is clearly ideal to model experimentally measured properties directly, there are many variables that can affect the OPV performance metrics, for example, the device design; processing conditions; dopants, dyes, solvents, and other additives; and others. Thus, measured OPV properties can vary from experiment to experiment and between labs. Data points from different sources can be inconsistent and affect reproducibility, constituting a relatively large source of error. Our goal in this work was to show that ML methods in general, and signatures specifically, are well suited to modeling and predicting a wide range of OPV properties. Therefore, we used the reproducible large dataset of photovoltaic properties calculated by a range of DFT methods in HOPV15. Clearly, these methods can be usefully applied to experimental data collected under conditions where all relevant information and device characteristics are carefully controlled, once the results are available in sufficient quantity to train ML models. Generally, there is not a good correlation between experimental and raw DFT calculated values[25]. DFT calculations occasionally calculate physically unrealistic negative values for PCE, and these uncorrected computed values PCE do not correlate well with the experimental values. However, Lopez et al.[26] showed that calibration of PCE values in the HOPV15 dataset can significantly improve the correlation between experimental and Scharber PCEs (Fig. 1).

As our goal was to generate models with good predictive performance that are chemically interpretable, we employed molecular signature descriptors. These represent chemical fragments in the donor and acceptor molecules. The "Methods" section describes the signature descriptors fully. We generated models for PCE, V_{oc}, J_{sc}, HOMO energy, LUMO energy, and the HOMO–LUMO gap for the 344 compounds in the dataset. We initially generated multiple linear regression models using an expectation maximization algorithm with a Laplacian prior[27] to select a sparse subset of descriptors. These linear models generally exhibited low test set predictivities, with R^2 values ≤0.2. Consequently, we modeled these properties using the well-proven nonlinear BRANNLP (Bayesian Regularized Artificial Neural Network with Laplacian prior)[28,29] method.

Clearly, OPV devices comprise donor and acceptor materials, and ML models must encode the properties of both. We employed three modeling strategies (Supplementary Methods) with increasing complexity in how the acceptor material was encoded. The first and simplest strategy generated separate OPV properties models for each type of acceptor (Supplementary Table 2). The second strategy accounted for different acceptors using a simple "1-hot" indicator variable. Here the different acceptors were encoded in the model as 1 if present and 0 if absent (Appendix B, Supplementary Information). This captures essential differences between the acceptors, the most relevant being the acceptor LUMO energies. Thirdly, we generated models for the six OPV properties in which donor and acceptors were encoded using signature descriptors (Supplementary Table 3). We aimed to make the best predictions for materials and, if possible, to interpret the models in terms of molecular functionality in the donor and acceptor materials structures.

All three modeling strategies predicted PCE, V_{oc}, J_{sc}, HOMO, LUMO, and HOMO–LUMO gap with moderate to good efficacy. The PCE models were always robust and predictive, with $R^2 > 0.64$ for training set and >0.58 for test set prediction. Exceptions were the HOMO energy and V_{oc} prediction for the PC61BM acceptor subset, the J_{sc} prediction for the TiO_2 subset, and the V_{oc} prediction for the total dataset with acceptors encoded by signature descriptors (Supplementary Tables 2 and 3). In the latter case, it is likely that the V_{oc} model is overfitted given that that it employs 90 effective parameters in the neural network.

Table 1. Performance of BRANNLP models in predicting OPV properties.

Property	N_{desc}	N_{eff}	Training set		Test set	
			R^2	SEE	R^2	SEP
PCE (%)	59	61	0.72	0.50	0.78	0.48
V_{oc} (V)	26	28	0.65	0.16	0.58	0.16
J_{sc} (mA cm^{-2})	39	40	0.57	18	0.60	22
HOMO (eV)	49	55	0.87	0.004	0.49	0.007
LUMO (eV)	90	101	0.94	0.003	0.67	0.008
Gap (eV)	37	40	0.83	0.007	0.65	0.010

N_{eff} is the number of effective parameters (weights) and N_{desc} is the final number of signature descriptors in the models. SEE is the standard error of estimation and SEP is the standard error of prediction. The neural network contained two hidden layer neurons.

The best models were generated by the second strategy, trained on 344 donor–acceptors pairs, with donors encoded by signature descriptors and acceptors captured by 1-hot binary vectors. We summarize the results of this study below, and the results of modeling OPV properties using strategies 1 and 3 in the Supplementary Material. For strategy 2 models, the dataset was divided into a training set of 276 donor–acceptor pairs and a test set of 68 donor–acceptor pairs by k-means clustering. The BRANNLP nonlinear modeling and variable selection method was used to generate the QSPR models. Table 1 summarizes the performance of these models.

Figure 2 illustrates the performance of the BRANNLP models for the six OPV properties. The majority of models predicting V_{oc}, J_{sc}, HOMO, LUMO, and HOMO–LUMO gap resulted in R^2 for the training and test sets greater than 0.5, which indicates that all these models are sufficiently predictable to provide useful estimates for these properties.

Model validation

It is essential to validate models to determine their predictive power, robustness, and reliability. We assessed this in three ways: predicting properties of a test set partitioned from the dataset and never used in training; randomly scrambling the property values and rebuilding the models (*y*-scrambling); de novo prediction of OPV properties of materials from the literature not used in the modeling study. Model predictivity was assessed by the R^2 statistic and the standard error of estimation or prediction for training and test set[30,31]. The ability of the ML models to recapitulate the properties of materials in the test set partitioned from the dataset is summarized in Table 1.

In *y*-scrambling, we randomly distributed the property values and generated ML models using this randomized data[31]. Low R^2 values for the training set and test set compared to the initial model shows that these models are not chance correlations nor overfitted[32]. We conducted three *y*-scrambling tests for each model as shown in Table 2. The R^2 values were near zero for the ML models trained on these data, showing that the primary models, whose predictions are presented in Table 1, are robust, reliable, and predictive.

We also used another external test set to assess model predictivity[32]. After generating the OPV property models and validating using the test sets partitioned from the data and by *y*-scrambling, we returned to the literature to find additional donor and acceptor materials whose properties could be predicted by the ML models. It is preferable that the DFT method that calculate the properties in external validation set is the same as that used for the dataset used to train the models. Our external validation dataset comprised eight donors and one acceptor (PC61BM) whose OPV properties were reported in the literature[33]. The HOMO, LUMO, and HOMO–LUMO gap energies were calculated by the same B3LYP/def2-SVP functional and basis set employed in the HOPV15 dataset. Table 3 shows the statistics for the line of best fit (trend) between the reported and predicted frontier orbital properties.

Table 4 shows the predicted absolute values for the HOMO, LUMO, and HOMO–LUMO gap energies compared to the reported values. The RMSE values for these predictions for the external test set were 0.19, 0.43, and 0.41 eV respectively. These results show that the models have useful abilities to predict at least the frontier orbital energies of materials, provided that are within or close to the domains of applicability of the models used to predict these properties. This proof of concept test of the ability of this type of descriptor and machine-learning method suggests that when larger training sets are available, it will be possible to predict important OPV properties of a larger range of materials.

Descriptor analysis

Machine-learning models, including artificial neural networks, can be hard to interpret in terms of the chemistries needed to improve the OPV properties[34,35]. Often the problem is due to the use of arcane descriptors rather than the modeling algorithm per se. This was the motivation to assess the ability of chemically interpretable signature descriptors to model OPV properties.

To this end, we performed analysis on the most relevant descriptors used to build the ML models based on strategy 2. Supplementary Table 4 shows the molecular structures of the most relevant descriptors selected by the BRANNLP method for each of the six OPV properties. In general, for the six properties modeled, there needs to be a balance between electron-withdrawing and donating functional groups, hydrophilicity and hydrophobicity, and conjugation length.

The OPV properties are not completely independent, some are significantly correlated, as reflected in the ML models. PCE is related to V_{oc}, J_{sc}, fill factor (FF), and the input power (P_{in}) by Eq. (1).

$$\mathrm{PCE} = 100 \times \frac{V_{oc} \times \mathrm{FF} \times J_{sc}}{P_{in}}. \quad (1)$$

The device efficiency PCE in the Scharber and related models is related to V_{oc} (Eq. 1), which is a function of the HOMO–LUMO gap. Frontier orbital energies are generally raised by electron-donating substituents and lowered by electron-withdrawing substituents[36,37]. The gap is influenced by the presence of strong electron-withdrawing substituents, such as nitro, trifluoromethyl, sulfone, nitrile, and methylene malononitrile (one of the strongest electron-withdrawing functional groups) moieties which lower the energy of the HOMO and LUMO and by electron-donating functional groups such as amine that raise the energy of the frontier orbitals. Thus, HOMO–LUMO gap can be raised or lowered by these substituents and is usually less influenced by these substituent effects. The key molecular fragments identified by sparse feature selection for the V_{oc}, HOMO, LUMO, and gap models are largely consistent with this theory and experimental observations. HOMO energies were also modulated by F and S substitution in the polyene chain. Hydrophilic groups such as carboxylic acid and N–O–N also modulated the PCE, as did fragments with extended conjugated double bonds such as polyenes, especially those with heteroatoms (O, N) embedded functionality. The most important functional groups for J_{sc} were hydrophilic moieties such as carboxylic acid and amines, and polyenes with sulfur substitution. Figure 3 shows an example of effective fragments on PCE. Additional examples to illustrate the rest of the relevant descriptors for each OPV properties that we described above are shown in Supplementary Fig. 1.

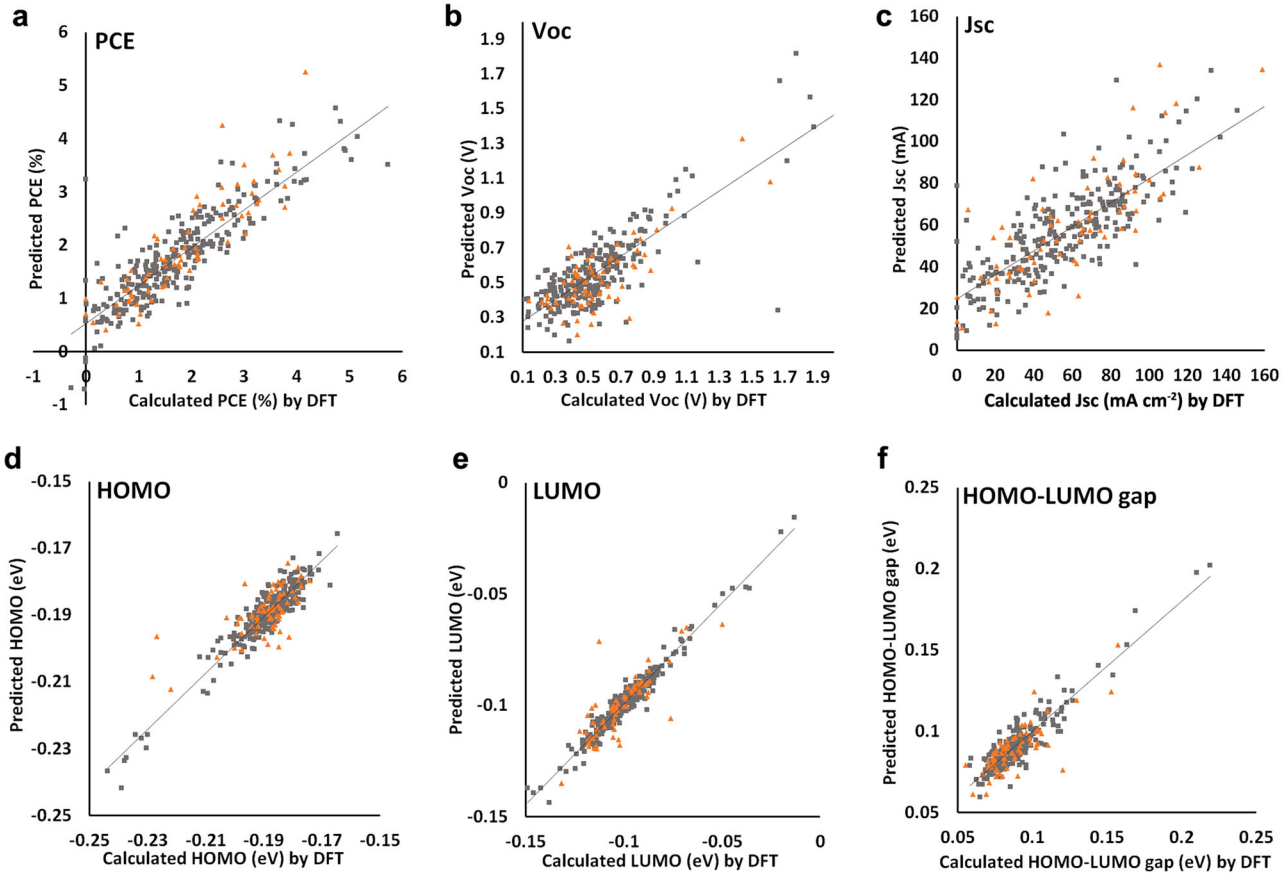

Fig. 2 Prediction of six OPV properties for the training set and test set using the BRANNLP method and signature descriptors for the donors and 1-hot descriptors for the acceptors. a PCE, **b** V_{oc}, **c** J_{sc}, **d** HOMO energy, **e** LUMO energy, **f** HOMO–LUMO gap. Training set predictions are in black and test set predictions in orange. Data on the performance of these models are shown in Table 1.

Table 2. Statistics of three rounds of *Y*-scrambling for the models presented in Table 1.

	R^2 training set	R^2 test set
PCE (%)	0.081 ± 0.004	0.020 ± 0.018
V_{oc} (V)	0.055 ± 0.023	0.008 ± 0.004
J_{sc} (mA cm^{-2})	0.118 ± 0.038	0.043 ± 0.034
HOMO (eV)	0.118 ± 0.025	0.004 ± 0.002
LUMO (eV)	0.082 ± 0.016	0.013 ± 0.003
Gap (eV)	0.079 ± 0.027	0.042 ± 0.054

Table 3. Performance of the models trained using HOPV15 data in predicting the HOMO, LUMO, and HOMO–LUMO gap values of the external validation set (trend line fit).

Property	R^2	SEP (eV)
HOMO	0.66	0.05
LUMO	0.77	0.10
HOMO–LUMO gap	0.53	0.08

In summary, we have shown that chemically interpretable chemical fragment-based descriptors can be used to train ML models that predict six key properties of OPV devices. Our approach leverages resource-intensive DFT calculations into larger regions of materials space, allowing fast and accurate estimates of these important photovoltaic properties for a relatively large number of donor and acceptor materials that may not yet be synthesized. Our study used a synergistic combination of efficient and chemically interpretable descriptors, sparse feature selection, and self-optimizing Bayesian regularized neural networks. The most relevant descriptors for each model provide guidance for materials chemists as to how to synthesize materials with improved OPV properties, or to mine them from larger databases of real or virtual materials. The ML models predicted PCE, V_{oc}, J_{sc}, HOMO, LUMO, and HOMO–LUMO gap using simple signature descriptors that encode the molecular properties of the molecules with good efficacy. Although some individual models had relatively low prediction accuracies for the test set, a consensus of all models for each property could identify a statistically robust model for each of the six OPV properties. This work demonstrated the importance of using nonlinear ML methods to map molecular descriptors to important OPV properties, and the value of signature descriptors in building robust, chemically interpretable, and predictive models of these properties. When using these models to screen large libraries of candidate OPV materials prior to synthesis, care must be taken to ensure such libraries lie near the domain of applicability of the models. Clearly, the quality of predictions in this study is dependent of the size, diversity, and accuracy of the underlying dataset. The ML algorithms we have developed in this study can be applied to any dataset of OPV structures and in future work we intend to extend the scope of this work to include large and more accurate computed and experimental datasets.

Table 4. Reported, predicted, and error values of HOMO, LUMO, and HOMO–LUMO gap energies of the external test set.

Donor	HOMO			LUMO			Gap		
	Reported	Predicted	Error	Reported	Predicted	Error	Reported	Predicted	Error
1a	−5.75	−6.05	0.30	−3.21	−2.68	−0.53	2.54	2.85	−0.31
1b	−6.07	−6.15	0.08	−3.84	−3.57	−0.27	2.23	2.90	−0.67
2a	−5.79	−6.14	0.35	−3.29	−2.80	−0.48	2.50	2.80	−0.30
2b	−6.10	−6.24	0.14	−3.89	−3.67	−0.22	2.21	2.85	−0.64
3a	−6.02	−6.13	0.11	−2.78	−2.33	−0.45	3.24	2.79	0.45
3b	−6.25	−6.23	−0.02	−3.39	−3.45	0.06	2.86	2.85	0.01
4a	−6.19	−6.23	0.04	−3.15	−2.79	−0.36	3.04	2.77	0.27
4b	−6.44	−6.23	−0.21	−3.74	−3.05	−0.69	2.70	2.81	−0.11

METHODS

Dataset

Being data-driven methods, machine-learning methods are critically dependent on the amount and quality of training data. As QSPR models are only valid within their domain of applicability, a large, diverse dataset with a wide range of properties can generate models with a better generalization ability[38]. Clearly, all objects in the dataset must be measured under same condition and be reproducible and accurate. The closer the distribution of training data is to a normal distribution, the more accurate the models generated from it are likely to be[30]. In this study, we employed the Harvard Photovoltaic Dataset (HOPV15, https://www.nature.com/articles/sdata201686#Tab1)[22], one of the largest and most diverse datasets available in literature for OPV properties, to train QSPR models. This dataset was compiled from 350 small molecule and polymer electron donors and acceptors, and includes experimental properties collected from literature plus data from quantum chemical calculations and the Scharber model[23]. The calculated properties include the values of open circuit potential (V_{oc}), short circuit density (J_{sc}), and PCE, which were derived from the model given by Scharber, whereas HOMO energy, LUMO energy of donor, and the HOMO–LUMO gap where calculating using ab initio (Hybrid DFT) methods. Lopez et al.[22] generated all possible conformers of each donor materials in the HOPV15 dataset and used various DFT functionals combined with def2-SVP, and the Scharber model to calculate the OPV properties. In this dataset, four different DFT functionals (B3LYP, BP86, M06-2X, and PBE0) were used in the quantum chemical calculations. The property values were averaged over all conformers as there was negligible dependence of properties on conformation. We compared the properties calculated by these methods and found that B3LYP, BP86, and PBE0 generated very similar values, while M06-2X dramatically overestimated the HOMO–LUMO gap and consequently predicted PCEs of ~0. Our QSPR models used the values of PCE, V_{oc}, and J_{sc} calculated using the Scharber model and HOMO, LUMO, and HOMO–LUMO gap using the B3LYP[39,40] DFT functional combined with the def2-SVP[41] basis set. While it is well known that the B3LYP functional overestimates electron delocalization, reasonable reproduction of experimental HOMO–LUMO gaps for conjugated systems is still possible[42]. Moreover, the errors tend to be systematic, and trends based on relative values can still be meaningful. We chose B3LYP in order to be consistent with previous studies, but calibration to experimental results has been shown to remove the dependence on the specific functional chosen for DFT calculations[25,26]. The advantage of using calculated properties over the experimental ones is that we are confident that these properties are measured with the same method, while the experimental data could have been measured under different conditions. The chemical names of electron acceptors, their SMILES (simplified molecular input line entry system) strings for donors, and values of OPV properties for each pair of donor and acceptors are listed in Appendix A of the Supplementary Information. We removed three donors due to the lack of information about the acceptors used in the PV device (compounds number 18, 82, and 273 in Appendix A) and another three donors because of duplications (compounds number 73, 204, and 334 in Appendix A). Table 5 presents the range of reported properties. A k-means clustering algorithm was used to divide the datasets for each property into a training set (80% of the dataset) used to train the model and a test set (20% of the dataset) used to evaluate the prediction accuracies of the model. This was done to ensure the test set lay within the domain of applicability of the model, and to allow others to reproduce the results we report here.

Molecular descriptors

An important aim of this project was to use molecular descriptors to describe the donors and acceptors that are both efficient and chemically interpretable. In models used for virtual screening of potentially unsynthesized materials, the use of experimentally determined electronic properties or those derived from computationally expensive DFT calculations as descriptors would be costly and time-consuming, as the descriptors for each new molecule of interest would have to be determined individually. On the other hand, signature descriptors can be generated for thousands of candidate materials in a matter of minutes, using freely available software. We used signature descriptors in this study to generate interpretable and predictive models of OPV properties because they are better able to guide synthesis towards improved materials than the arcane descriptors commonly used. As well as generating good models, they are also easier to visualize and provide better guidance as to what functionality in the molecules contributes to, or degrades, properties. Signature descriptors, shown to be effective in other areas of property prediction, are based on the connection path of atoms in the molecule. They provide a systematic calculation system that can describe the "neighborhood" of atoms in a molecule. To understand the concept of signatures, it is important to define the molecular graphs which represent a molecule based on atoms and bonds. Atoms are defined by a set of atom types, which could be provided either from the periodic table or a molecular force field. In molecular graphs, every atom will be assigned an atom type by a function, where atom type considers the possible covalent bonds of each atom. The signature of an atom is a subgraph of molecular graph in a shape of a tree that contains all atoms and all bonds within a specified distance. That is, the signature descriptor of a given atom is the connected path of atoms of a specified length, *l*. This effectively dissects molecules into an ensemble of fragments, creating a fingerprint whose elements indicate the number of each type of fragment exists in each molecule that is characteristic of the material[43–45]. Figure 4 demonstrates how signature descriptors can be computed from chemical structures. In this project, we applied the MolSig program[46] to compute the signature descriptors. We used the Open Babel package[47] to convert the SMILES strings, which encode the 2D structure of molecules as a string of characters, to Cartesian coordinates of atomic positions in.mol file format. The signature descriptors were generated for path lengths 0–4. These descriptors were then collected, sorted based on size, and any with less than two examples were removed. A total pool of 695 acceptors and donor materials descriptors was generated. We then used the variable selection method to choose the most relevant molecular features for each OPV properties. By mapping back the most relevant signature descriptors onto prototype molecules in the dataset, we can provide important guidance for material scientists to synthesize new materials or improve existing ones. The most relevant signature descriptors for each property are provided in Supplementary Table 4.

Feature selection

The signature descriptor method can generate fingerprints containing a very large number of fragment elements. While large pools of descriptors

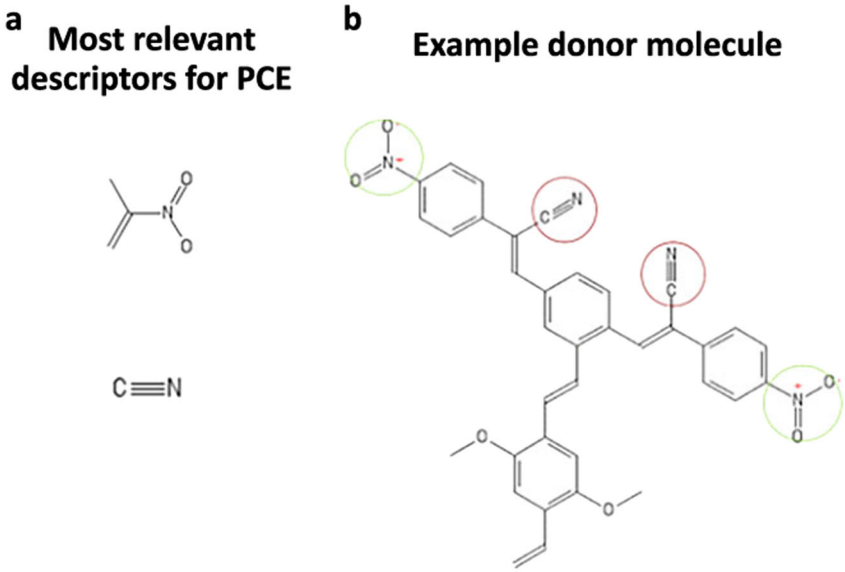

Fig. 3 Examples of most relevant descriptors for PCE. a Examples of most relevant descriptors for PCE. **b** An example donor molecule with the descriptors circled. Additional examples of effective descriptors for PCE and other OPV properties are provided in Supplementary Fig. 1.

Table 5. The ranges of the OPV properties calculated at the B3LYP/def2-SVP level of theory.

	Minimum	Maximum	Range
PCE (%)	−0.285	5.722	6.007
V_{oc} (V)	−0.119	2.033	2.152
J_{sc} (mA cm^{-2})	0	186.4	186.4
HOMO (eV)	−0.224	−0.165	0.059
LUMO (eV)	−0.149	−0.013	0.136
Gap (eV)	0.056	0.215	0.281

for materials are very useful, great care must be taken to choose a subset of the most relevant descriptors to avoid overfitting models. Overfitted models predict the training set very well but have little predictive power for new data. To aid interpreting models and to avoid overfitting them, it is essential to reduce the dimensionality of the descriptors. Careless selection of subsets of descriptors from a larger pool can also lead to chance correlations[48]. We employed very sparse feature selection methods based on L1 regression and Bayesian regularized neural networks with sparse (Laplacian) prior to achieve very efficient selection of the most relevant descriptors for each OPV property modeled. These methods have been shown in many studies to yield parsimonious subsets of description in a context-dependent way that provide models derived from them with excellent predictive power.

Nonlinear property modeling

Neural networks are universal approximators[49] that can model any linear or nonlinear and continuous relationships given sufficient training data. However, they have some drawbacks such as overfitting (too many adjustable parameters relative to the number of training data) and overtraining (memorizing training data better but generalizing worse) that generate models with low predictability. ANN (Artificial Neural Network) models are also said to be difficult to interpret, although this is as much to do with interpretable descriptors as the ML method used[50]. Burden and Winkler[51] showed that Bayesian regularization of standard backpropagation neural networks can overcome many of the disadvantages of ANN used to model molecules or materials. The BRANNGP method[50,52] (Bayesian Regularized Artificial Neural Network with Gaussian Prior) was shown to generate robust models of diverse ranges of molecules and properties. BRANNGP can effectively prune less relevant weights from networks (making the models effectively invariant to the number of hidden layer nodes), providing instead an estimation of number of effective parameters[28]. When the Gaussian prior is replaced by a sparsity-inducing Laplacian prior Bayesian the resulting neural network (BRANNLP) can also prune less relevant descriptors and well as less relevant weights. Thus, a Bayesian regularized neural network with a Laplacian prior is a feedforward, fully connected neural network that uses Bayesian regularization to optimize the sparsity of models, that is, to find the right balance between model complexity (variance) and simplicity (bias)[28]. This sparse feature selection method is based on L1 regression, similar to the LASSO method[53,54]. This neural network method has been shown to generate robust and optimally sparse models of diverse materials properties[55–58]. Here we have used BRANNLP implemented in the CSIRO-Biomodeller package[59–61] to predict the properties photovoltaic and electronic properties of compounds. These networks generate predictions of the training and test set properties that are relatively insensitive to the number of nodes in the hidden layer, with the effective parameters in the models being relatively constant as the number of hidden layer nodes increase above a minimum. These networks rarely need more than 2–3 nodes in the hidden layer to model most materials properties. Two hidden layer nodes were used in this study. Our shallow neural networks consist of input (descriptors), hidden (computation), and output layers and are fully connected[9]. The input and output nodes use linear transfer functions, and the hidden layer node sigmoidal functions. The data are mean centered and normalized prior to modeling. A Levenberg–Marquardt algorithm is used for backpropagation. In contrast, deep learning methods use a very large number of hidden layer nodes and non-differentiable transfer functions (e.g. ReLU) and weight dropout to avoid overfitting[62,63]. The universal approximation theorem states that shallow neural networks (like the BRANNLP network here) generate models of similar predictive accuracies to DNN given the same training data. Comparative modeling of large dataset of drugs by shallow and DNN algorithms showed that the predictability of models is indeed similar[9].

To evaluate the performance of the models, the R^2 statistic and the standard error of estimation (SEE) and standard error of prediction (SEP) for training set and test set, respectively, were calculated. R^2 is the square correlation coefficient between the predicted and measured values of data points in training set and test set. SEE and SEP represent the root-mean-square error between the predicted and measured values of data points, adjusted for degrees of freedom, in the training and test set, respectively[27,28]. SEE and SEP are more robust measures of model quality than R^2 values[19].

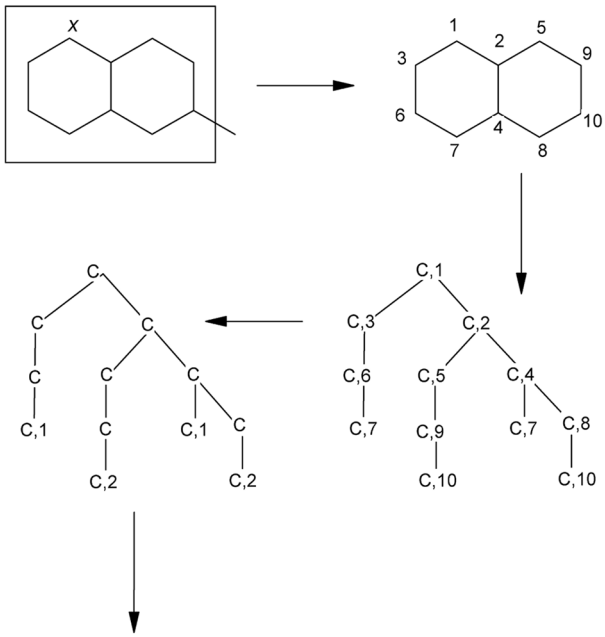

Fig. 4 Illustration of the process of generating signature descriptors. C1 and C2 define a complete ring with the length of 0 to 4. Reprinted with permission from ref. [45]. Copyright 2003 American Chemical Society.

DATA AVAILABILITY

Additional data not found in the text and Supplementary Information is available from the corresponding authors upon reasonable request.

CODE AVAILABILITY

Bayesian regularized neural network and LASSO[53,54] (a similar sparse feature selection method to the one we used) have become available in R, MATLAB, and other statistical packages. For example, in 2016, Okut published a book chapter explaining the Bayesian regularized neural networks with a MATLAB code for application of this algorithm[64]. Also, recently Rodriguez and Gianola released the latest version of BRNN package, based on R program and CRAN repository. In this package Bayesian regularization for feedforward neural network is implemented to build machine-learning models[65]. Our in-house ML package that implements the BRANNLP algorithm is useful for generating sparse models that generalize well and are hard to overtrain or overfit. To ensure accessibility, we repeated the PCE model prediction study using a public-domain conventional ANN. We used TensorFlow (Google) to build this machine-learning algorithm. We reproduced the PCE model using a two-layer perceptron feedforward artificial neural network (code and the input files are available on GitHub: https://github.com/Nas796/Machine-learning-for-photovoltaic-material-property-prediction). This ANN model predicted the properties of the training and test set data with R^2 values of 0.83 and 0.72, respectively. The statistics of the same model generated by the BRANNLP methods had R^2 of 0.72 and 0.78 for prediction of the training and test set properties respectively. The results are quite similar for both modeling methods, again illustrating that the choice of descriptors is more important than the type of modeling algorithm.

Received: 21 May 2020; Accepted: 29 September 2020;
Published online: 06 November 2020

REFERENCES

1. Abdulrazzaq, O. A., Saini, V., Bourdo, S., Dervishi, E. & Biris, A. S. Organic solar cells: a review of materials, limitations, and possibilities for improvement. *Part. Sci. Technol.* **31**, 427–442 (2013).
2. Cui, Y. et al. Over 16% efficiency organic photovoltaic cells enabled by a chlorinated acceptor with increased open-circuit voltages. *Nat. Commun.* **10**, 2515 (2019).
3. Mosconi, E., Amat, A., Nazeeruddin, M. K., Grätzel, M. & De Angelis, F. First-principles modeling of mixed halide organometal perovskites for photovoltaic applications. *J. Phys. Chem. C* **117**, 13902–13913 (2013).
4. Janković, V. & Vukmirović, N. Dynamics of exciton formation and relaxation in photoexcited semiconductors. *Phys. Rev. B* **92**, 235208 (2015).
5. Mikhnenko, O. V., Blom, P. W. & Nguyen, T.-Q. Exciton diffusion in organic semiconductors. *Energy Environ. Sci.* **8**, 1867–1888 (2015).
6. Coropceanu, V. et al. Charge transport in organic semiconductors. *Chem. Rev.* **107**, 926–952 (2007).
7. Proctor, C. M., Kuik, M. & Nguyen, T.-Q. Charge carrier recombination in organic solar cells. *Prog. Polym. Sci.* **38**, 1941–1960 (2013).
8. Ran, N. A. et al. Charge generation and recombination in an organic solar cell with low energetic offsets. *Adv. Energy Mater.* **8**, 1701073 (2018).
9. Winkler, D. A. & Le, T. C. Performance of deep and shallow neural networks, the universal approximation theorem, activity cliffs, and QSAR. *Mol. Inform.* **36**, 1600118 (2017).
10. Mesta, M., Chang, J. H., Shil, S., Thygesen, K. S. & García-Lastra, J. M. A protocol for fast prediction of electronic and optical properties of donor-acceptor polymers using density functional theory and tight-binding method. *J. Phys. Chem. A.* **123**, 4980–4989 (2019).
11. Altman, N. S. An introduction to kernel and nearest-neighbor nonparametric regression. *Am. Stat.* **46**, 175–185 (1992).
12. Rupp, M. Machine learning for quantum mechanics in a nutshell. *Int. J. Quant. Chem.* **115**, 1058–1073 (2015).
13. Vu, K. et al. Understanding kernel ridge regression: common behaviors from simple functions to density functionals. *Int. J. Quant. Chem.* **115**, 1115–1128 (2015).
14. Padula, D., Simpson, J. D. & Troisi, A. Combining electronic and structural features in machine learning models to predict organic solar cells properties. *Mater. Horiz.* **6**, 343–349 (2019).
15. Breiman, L. Random forests. *Mach. Learn.* **45**, 5–32 (2001).
16. Svetnik, V. et al. Random forest: a classification and regression tool for compound classification and QSAR modeling. *J. Chem. Inf. Comput. Sci.* **43**, 1947–1958 (2003).
17. Guelman, L. Gradient boosting trees for auto insurance loss cost modeling and prediction. *Expert Syst. Appl.* **39**, 3659–3667 (2012).
18. Sahu, H., Rao, W., Troisi, A. & Ma, H. Toward predicting efficiency of organic solar cells via machine learning and improved descriptors. *Adv. Energy Mater.* **8**, 1801032 (2018).
19. Alexander, D. L., Tropsha, A. & Winkler, D. A. Beware of R 2: simple, unambiguous assessment of the prediction accuracy of QSAR and QSPR models. *J. Chem. Inf. Model* **55**, 1316–1322 (2015).
20. Pereira, F. et al. Machine learning methods to predict density functional theory B3LYP energies of HOMO and LUMO orbitals. *J. Chem. Inf. Model* **57**, 11–21 (2016).
21. Cortes, C. & Vapnik, V. Support-vector networks. *Mach. Learn.* **20**, 273–297 (1995).
22. Lopez, S. A. et al. The Harvard organic photovoltaic dataset. *Sci. Data* **3**, 160086 (2016).
23. Scharber, M. C. et al. Design rules for donors in bulk-heterojunction solar cells—towards 10% energy-conversion efficiency. *Adv. Mater.* **18**, 789–794 (2006).
24. Iharbi, F. et al. An efficient descriptor model for designing materials for solar cells. *npj Comput. Mater.* **1**, 15003 (2015).
25. Pyzer-Knapp, E. O., Simm, G. N. & Guzik, A. A Bayesian approach to calibrating high-throughput virtual screening results and application to organic photovoltaic materials. *Mater. Horiz.* **3**, 226–233 (2016).
26. 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* **1**, 857–870 (2017).
27. Burden, F. & Winkler, D. Optimal sparse descriptor selection for QSAR using Bayesian methods. *QSAR Comb. Sci.* **28**, 645–653 (2009).
28. Burden, F. R. & Winkler, D. A. An optimal self-pruning neural network and non-linear descriptor selection in QSAR. *QSAR Comb. Sci.* **28**, 1092–1097 (2009).
29. Winkler, D. A. & Burden, F. R. Bayesian neural nets for modeling in drug discovery. *Drug Discov. Today. BIOSILICO* **2**, 104–111 (2004).
30. Katritzky, A. R. et al. Quantitative correlation of physical and chemical properties with chemical structure: utility for prediction. *Chem. Rev.* **110**, 5714–5789 (2010).
31. Wold, S., Eriksson, L. & Clementi, S. in *Chemometric Methods in Molecular Design* (ed. van de Waterbeemd, H.) 309–338 (Wiley, Weinheim, 1995).

32. Tropsha, A., Gramatica, P. & Gombar, V. K. The importance of being earnest: validation is the absolute essential for successful application and interpretation of QSPR models. *QSAR Comb. Sci.* **22**, 69–77 (2003).
33. Nowak-Król, A. et al. Modulation of band gap and p-versus n-semiconductor character of ADA dyes by core and acceptor group variation. *Org. Chem. Front.* **3**, 545–555 (2016).
34. Fujita, T. & Winkler, D. A. Understanding the roles of the "two QSARs". *J. Chem. Inf. Model.* **56**, 269–274 (2016).
35. Johansson, U., Sönströd, C., Norinder, U. & Boström, H. Trade-off between accuracy and interpretability for predictive in silico modeling. *Future Med. Chem.* **3**, 647–663 (2011).
36. Salzner, U. & Kiziltepe, T. Theoretical analysis of substituent effects on building blocks of conducting polymers: 3,4′-substituted bithiophenes. *J. Org. Chem.* **64**, 764–769 (1999).
37. Luponosov, Y. N. et al. Effects of electron-withdrawing group and electron-donating core combinations on physical properties and photovoltaic performance in D-π-A star-shaped small molecules. *Org. Electron.* **32**, 157–168 (2016).
38. Golbraikh, A. Molecular dataset diversity indices and their applications to comparison of chemical databases and QSAR analysis. *J. Chem. Inf. Comput. Sci.* **40**, 414–425 (2000).
39. Becke, A. Density-functional thermochemistry: the role of extract exchange. *J. Chem. Phys.* **98**, 648–645 (1993).
40. Becke, A. D. Density-functional exchange-energy approximation with correct asymptotic behavior. *Phys. Rev. A* **38**, 3098–3100 (1988).
41. Weigend, F. & Ahlrichs, R. Balanced basis sets of split valence, triple zeta valence and quadruple zeta valence quality for H to Rn: design and assessment of accuracy. *Phys. Chem. Chem. Phys.* **7**, 3297–3305 (2005).
42. Ma, J., Li, S. & Jiang, Y. A time-dependent DFT study on band gaps and effective conjugation lengths of polyacetylene, polyphenylene, polypentafulvene, polycyclopentadiene, polypyrrole, polyfuran, polysilole, polyphosphole, and polythiophene. *Macromolecules* **35**, 1109–1115 (2002).
43. Churchwell, C. J. et al. The signature molecular descriptor: 3. Inverse-quantitative structure–activity relationship of ICAM-1 inhibitory peptides. *J. Mol. Graph. Model.* **22**, 263–273 (2004).
44. Faulon, J.-L., Churchwell, C. J. & Visco, D. P. The signature molecular descriptor. 2. Enumerating molecules from their extended valence sequences. *J. Chem. Inf. Comput. Sci.* **43**, 721–734 (2003).
45. Faulon, J.-L., Visco, D. P. & Pophale, R. S. The signature molecular descriptor. 1. Using extended valence sequences in QSAR and QSPR studies. *J. Chem. Inf. Comput. Sci.* **43**, 707–720 (2003).
46. Carbonell, P., Carlsson, L. & Faulon, J.-L. Stereo signature molecular descriptor. *J. Chem. Inf. Model.* **53**, 887–897 (2013).
47. O'Boyle, N. M. et al. Open Babel: an open chemical toolbox. *J. Cheminformatics* **3**, 33 (2011).
48. Topliss, J. G. & Costello, R. J. Chance correlations in structure-activity studies using multiple regression analysis. *J. Med. Chem.* **15**, 1066–1068 (1972).
49. MacKay, D. Bayesian framework for backpropagation networks. *Neural Comput.* **4**, 448–472 (1992).
50. Lucic, B., Amic, D. & Trinajstic, N. Nonlinear multivariate regression outperforms several concisely designed neural networks on three QSPR data sets. *J. Chem. Inf. Comput. Sci.* **40**, 403–413 (2000).
51. Burden, F. & Winkler, D. in *Artificial Neural Networks: Methods and Applicatio*ns (ed. Livingstone, D. J.) 23–42 (Humana Press, 2009).
52. Neal, R. M. in *Bayesian Learning for Neural Networks*. 29–53 (Springer, New York, 1996).
53. Gauraha, N. Introduction to the LASSO. *Resonance* **23**, 439–464 (2018).
54. Tibshirani, R. Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. B* **58**, 267–288 (1996).
55. Feiler, C. et al. In silico screening of modulators of magnesium dissolution. *Corros. Sci.* **163**, 108245 (2019).
56. Manallack, D. T., Burden, F. R. & Winkler, D. A. Modelling inhalational anaesthetics using Bayesian feature selection and QSAR modelling methods. *ChemMedChem* **5**, 1318–1323 (2010).
57. Mikulskis, P., Alexander, M. R. & Winkler, D. A. Towards Interpretable machine learning models for materials discovery. *Adv. Intell. Syst.* **1**, 1900045 (2019).
58. Rasi Ghaemi, S. et al. High-throughput assessment and modeling of a polymer library regulating human dental pulp-derived stem cell behavior. *ACS Appl. Mater. Interfaces* **10**, 38739–38748 (2018).
59. Burden, F. R. & Winkler, D. A. Robust QSAR models using Bayesian regularized neural networks. *J. Med. Chem.* **42**, 3183–3187 (1999).
60. Burden, F. R. & Winkler, D. A. New QSAR methods applied to structure−activity mapping and combinatorial chemistry. *J. Chem. Inf. Comput. Sci.* **39**, 236–242 (1999).
61. Winkler, D. A. & Burden, F. R. Robust QSAR models from novel descriptors and Bayesian regularised neural networks. *Mol. Simul.* **24**, 243–258 (2000).
62. LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521**, 436–444 (2015).
63. Glorot, X., Bordes, A. & Bengio, Y. in *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*. 315–323 (Fort Lauderdale, FL, USA, 2011).
64. Okut, H. in *Artificial Neural Networks—Models and Applications* (ed. Rosa, J. L. G.) 27–48 (IntechOPen, London, 2016).
65. Perez-Rodriguez, P. & Gianola, D. *brnn: Bayesian Regularization for Feed-Forward Neural Networks.* https://CRAN.R-project.org/package=brnn (2020).

ACKNOWLEDGEMENTS

This work was supported by the Australian Government through the Australian Research Council (ARC) under the Centre of Excellence scheme (project number CE170100026). This work was also supported by computational resources provided by the Australian Government through the National Computational Infrastructure National Facility and the Pawsey Supercomputer Centre. Professor Frank Burden is gratefully acknowledged for generating the CSIRO-Biomodeller code and Professor Chris F. McConville for helpful discussions.

AUTHOR CONTRIBUTIONS

N.M., U.B., D.A.W., and S.P.R. conceived the study. N.M. and A.J.C. trained the M.L. models. M.K. wrote the publicly available code to reproduce the models. All authors contributed to the discussion and interpretation of results and writing of the paper.

COMPETING INTERESTS

The authors declare no competing interests.

ADDITIONAL INFORMATION

Supplementary information is available for this paper at https://doi.org/10.1038/s41524-020-00429-w.

Correspondence and requests for materials should be addressed to N.M. or S.P.R.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.