Property Prediction of Organic Donor Molecules for Photovoltaic Applications using Extremely Randomized Trees

Arindam Paul[†,*], Alona Furmanchuk[‡], Wei-keng Liao[†], Alok Choudhary[†], and Ankit Agrawal[†]

[†]Department of Electrical Engineering and Computer Science, Northwestern University
[‡] Institute for Public Health and Medicine, Feinberg School of Medicine, Center for Health Information Partnerships, Northwestern University
[*]Corresponding author: arindam.paul@eecs.northwestern.edu

Received: date / Accepted: date

Abstract Organic solar cells are an inexpensive, flexible alternative to traditional silicon-based solar cells but disadvantaged by low power conversion efficiency due to empirical design and complex manufacturing processes. This process can be accelerated by generating a comprehensive set of potential candidates. However, this would require a laborious trial and error method of modeling all possible polymer configurations. A machine learning model has the potential to accelerate the process of screening potential donor candidates by associating structural features of the compound using molecular fingerprints with their highest occupied molecular orbital energies. In this paper, extremely randomized tree learning models are employed for the prediction of HOMO values for donor compounds, and a web application is developed[1]. The proposed models outperform neural networks trained on molecular fingerprints as well as SMILES, as well as other state-of-the-art architectures such as Chemception and Molecular Graph Convolution on two datasets of varying sizes.

Keywords Solar Cells · Data Mining · Organic Photovoltaics · Machine Learning · Materials Informatics

1 Introduction

Solar energy is a vital source of clean, versatile renewable energy and an important component in solving the worldwide energy problem [1, 2]. Organic Photovoltaic cells (OPVs) [3–6] are lightweight, flexible, inexpensive and more customizable compared to traditional silicon-based photovoltaics [7]. However, there are challenges impeding the usage of OPVs in a commercial environment. The major issue surrounding OPVs is low power conversion efficiency of fabricated cells. Maximum cell efficiency observed in organic solar cells is currently 13.2% [8], and commercial devices usually achieve around 5-8% [9], which is much lower than silicon-based photovoltaics. The primary bottleneck in the improvement of OPV device design is complex manufacturing processes that lead to the reduction of active layer performance [10]. Traditionally, the design of a potential OPV material is dependent on conjectures from experiments, and expertise of materials scientists, followed by a laborious process of synthesis, characterization, and optimization of a prototype device.

The screening of OPV materials could be semi-automated through utilization of various modeling techniques (finite element [11, 12] to ab initio [13, 14] and molecular modeling [15]). Yosipof et al. [16] establishes the importance of data reduction and visualization using Principle Component Analysis and Self Organizing Maps, wherein two metal oxide solar cell libraries are analyzed. Jorgensen et al. [17] describes deep generative models for predicting molecular properties, and in particular, delineates screening of OPV using molecule generation via context-free grammar VAE. Kaspi et al [18] introduces a machine learning/data mining-based decision support system PVAnalyzer for identification of interesting trends not easily observable using simple bi-parametric correlations, and provides scope of finding new insights into factors affecting solar cells performances. The task of screening is complicated due to the difficulty in capturing complex effects culminating from multiple local minimum configurations a polymer could

---

[1] http://info.eecs.northwestern.edu/OPVPredictor

adopt during the manufacturing of the active layer [19–22].

Machine learning applied to available experimental observations and theoretical simulations could potentially generate many comprehensive models with advanced predictive capabilities. This approach has been successfully applied in several materials and molecular designs [23–36] across application areas.

In this paper, machine learning models using extremely randomized trees (ERTs) [37] were developed to advance the organic monomer screening process for photovoltaic applications [38, 39]. The results of ab initio simulations were combined with the cataloged description of the structural details of the monomers. The variance of structural morphology in the actual device was approximated with sets of local conformers that possibly could be created during manufacturing. Models developed in this paper predict highest occupied molecular orbital (HOMO) energy of the donor monomers in the active layer of the device that is averaged across multiple configurations using Boltzmann averaging. The predicted value paired with the complementary lowest unoccupied molecular orbital of the acceptor molecule could be used in speeding up the screening process. The proposed models outperform neural networks trained on molecular fingerprints as well as SMILES [40–42], as well as other state-of-the-art architectures such as Chemception and Molecular Graph Convolutions on both the smaller Harvard Organic Photovoltaic (HOPV) dataset as well as on a subset of the Clean Energy Project (CEP) dataset. For end-user convenience, the machine learning models were implemented as a web application at `http://info.eecs.northwestern.edu/OPVPredictor`.

2 Method

2.1 Extremely Randomized Trees

ERTs use an ensemble of decision trees [37] in which a node split is selected completely randomly with respect to both variable index and variable splitting value. The principle behind ERTs is using several small decision trees that are individually weak learners but when aggregated in an ensemble leads to a very robust learner. ERTs are similar to other tree based ensemble algorithms such as random forests (RFs) but unlike RFs, the same training set is used for training all the trees. Further, ERTs split a node based on both variable index and variable splitting value while random forests only splits by variable value. This makes ERTs both more computationally efficient than RFs and generalizable. Figure 1 illustrates the working of ERTs by aggregating results from several smaller trees.

2.2 Scharber Model

For a solar cell, the most important property is power conversion efficiency (PCE) or the amount of electricity which can be generated due to the interaction of electron donors and acceptors. The Scharber model [43] provides a relation between the voltage V_{oc} and the energies of the HOMO and the lowest unoccupied molecular orbital (LUMO) level of the donor and acceptor molecules respectively, which in turn can be related to the power conversion efficiency (PCE), the maximum efficiency of solar cells. In the following equation, J_{sc} is the short-circuit current density, FF is electrical fill factor and P_{in} is incident-light intensity. E^{Donor}_{HOMO} and $E^{Acceptor}_{LUMO}$ indicate the HOMO and LUMO energy levels of the donor and acceptor molecules respectively.

$$V_{oc} = 1/e(|E^{Donor|}_{HOMO} - E^{|Acceptor|})_{LUMO} - 0.3V$$

$$PCE = 100 * (V_{oc} * FF * J_{sc})/P_{in}$$

2.3 Datasets

The HOPV dataset [44] used in this work is a collection of photovoltaic measurements for a diverse set of 350 organic donor compounds generated by extensively searching the literature. In our experiments, the dataset was reduced to 344 molecules after removing redundant isomeric samples [45]. The dataset provides density functional theory (DFT) calculations of HOMO energy values for four functionals B3LYP, BP86, PBE and M06 using the basis set def2-SVP [46]. We get the expected values for HOMO values across all conformers by calculating the boltzmann average. Each molecule in the HOPV dataset is represented by a subset of 3-18 conformers obtained at kT, where k is the Boltzmann constant and T is the temperature of the OPV device. The global minimum (T = 0K) structures used for prediction of HOMO energies are far from the donor molecule structures in real OPV devices, after various manufacturing steps. We observe from Figure 2 that the PCE of the OPV device and the HOMO energy values are correlated with each other. We abstained from building models on the experimental values as HOMO values were missing for many molecules, and manufacturing information was not provided.

The band gap of the processed organic layer (made up of donors, acceptors, and other additives) would be altered from their global minimum value due to the shift

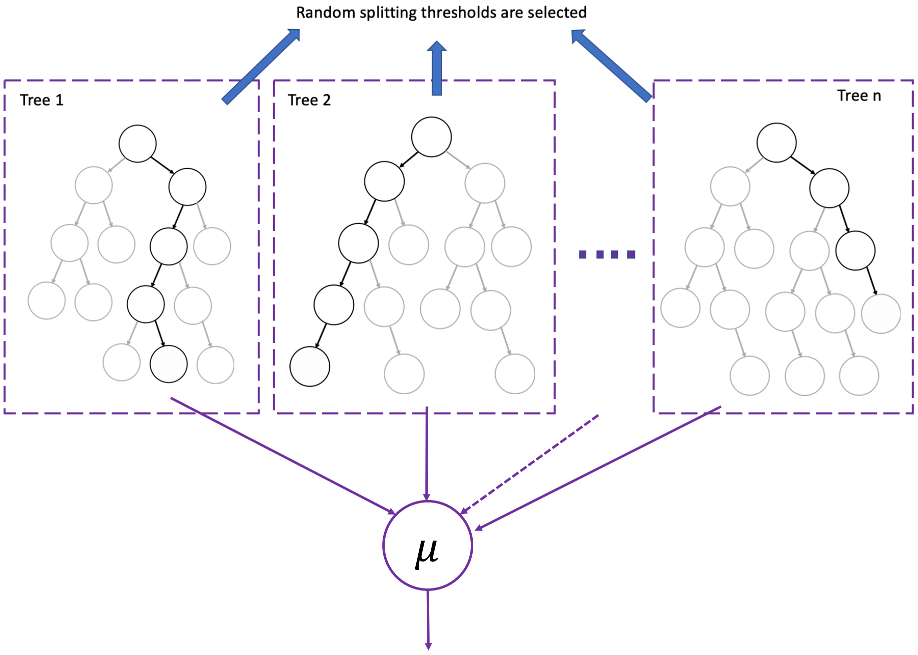


Fig. 1: Extremely randomized trees (ERT) architecture : ERTs are a forest of decision trees where node split is selected randomly with respect to both variable index as well as variable splitting value. Results from several small trees (indicated in dashed boxes) are aggregated in ERTs. The black paths represent the decision tree path for a given data point, and the gray paths represent the decision tree paths that are not selected. The output of each individual tree is aggregated and the final predicted value is the arithmetic mean (indicated by μ).

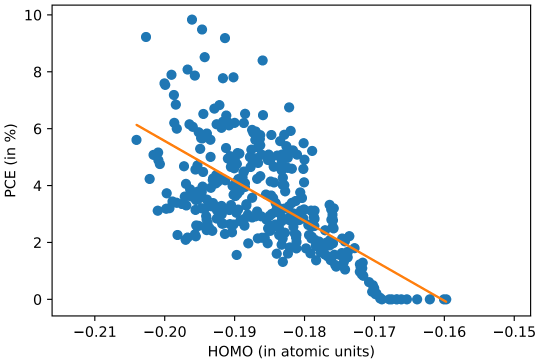
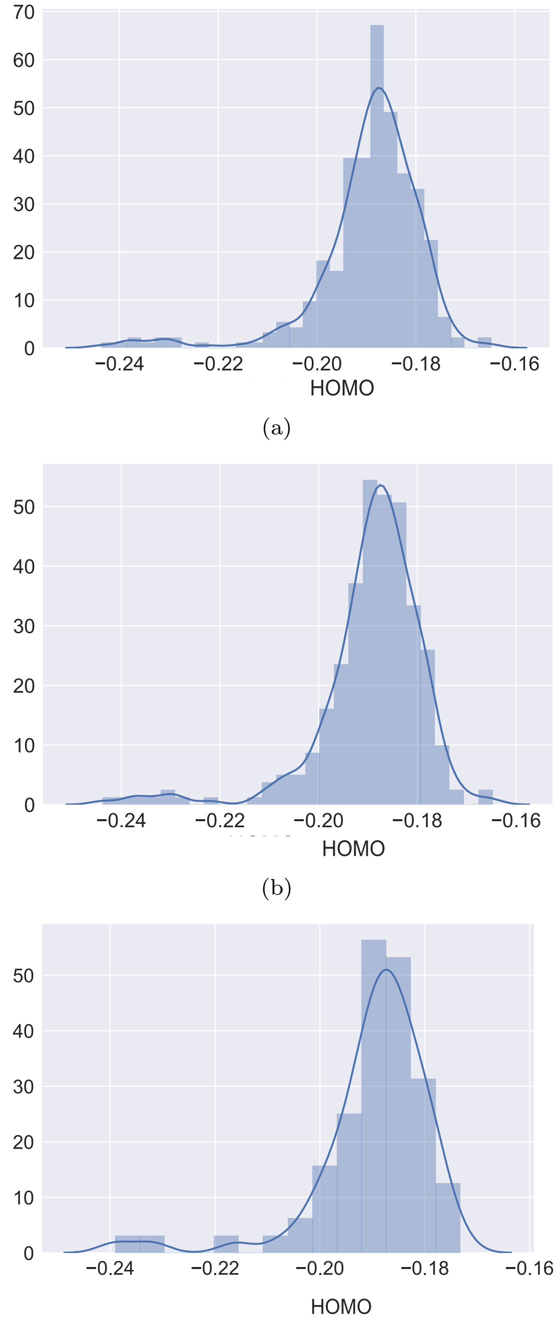


Fig. 2: The scatter-plot (with line of best fit) demonstrates the linear relationship between PCE of the device and HOMO values of the donor compound. The boltzmann average of the HOMO values for each conformer is used to determine the HOMO for a given donor.

of molecules from their ideal configuration. The degree of alteration would depend on the exact routine used in manufacturing, and is hard to predict. The boltzmann averaging is an attempt to account for the effect of structural variation in the experimental device. This is because different conformers of the same molecule occur in real OPV devices, and hence HOMO energies averaged over all conformers into the predictive model is expected to improve the relevance of the predicted HOMO values to the performance of the actual device.

To evaluate the validity of ERTs to scale to other datasets, we experimented on a subset of the Harvard CEP Dataset [47] which contains DFT-calculated molecular structures and properties for many candidate donor structures for organic photovoltaic cells. The CEP is a virtual high-throughput discovery and design effort for the next generation of plastic solar cell materials. It studies many candidate structures to identify suitable compounds for the harvesting of renewable energy from the sun and for other organic electronic applications. To establish the generalization of the models for larger datasets, we scraped a portion of the CEP database available. For scraping, we used the python libraries selenium [48] and beautiful soup [49]. This dataset is made available in the supplementary material. We restricted our extraction to 22,179 data points as the online CEP database had restrictions in place preventing automatic web-extraction of the entire database.



(a)

(b)

(c)

Fig. 3: Distribution of the datasets : (a) entire HOPV dataset, (b) training set, and (c) held-out test set. All the HOMO values are in atomic units (a.u.). 1 atomic unit is equal to 27.21 eV.

2.4 Data Preparation

For both the datasets, the original data was divided into training and test subsets. Figure 3 illustrates the distribution of the HOMO values across the complete HOPV dataset, the training and test sets. The dataset is split into training and test subsets with 80% and 20% of the data points respectively. We use stratified shuffle splitting to ensure similar distribution across the training and test set. The HOPV dataset provided DFT calculations for 4 functionals : PBE, B3LYP, BP86 and M06. In this paper, we restricted ourselves to PBE calculations. Further, we found that all the other functionals can be expressed as a linear transformation of the PBE functional values.

Two fingerprint representations - MACCS and Atom Pair were used for generating features [50–55]. For Atom Pair fingerprints, we initially calculated the original unhashed count vector of length 4 million for all the molecules using RDKit. After that, features that are invariant across the entire dataset were removed. This led to the reduction of the length of the unfolded fingerprint from 4 million to 2696 . The uncompressed MACCS fingerprint was only 166 bits long, and hence no feature reduction or transformation was performed. We did not use 1024 bit compressed fingerprint representation for Atom Pair as the original meaning of the fingerprint would be lost.

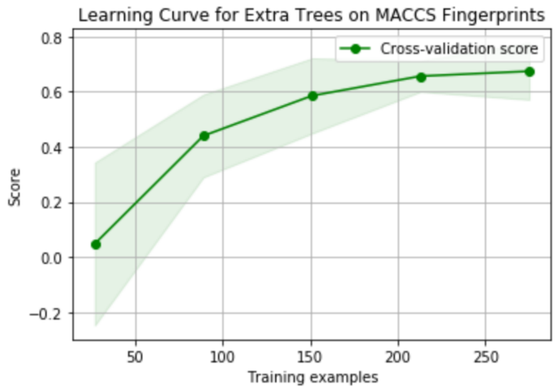
The fingerprints were prepared from their simplified molecular-input line-entry system (SMILES) [56] formulae using RDKit Python Library [57]. SMILES is a form of line notation for the chemical structure of molecules, and considered a versatile system. Molecule editors can generate 2D and 3D models from the line notation. The HOPV dataset provides canonical Standard SMILES implementations both in standard and shortened format.

Extensive grid search was performed across hyperparameters to discover the model architecture with the least mean absolute error for 5-fold cross-validation. This model was chosen and trained on the entire training dataset.

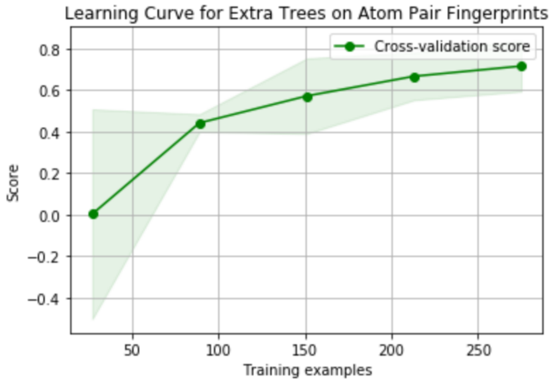
3 Results & Discussion

3.1 Experimental Results

In this work, we provide a framework for reducing the design space by screening new donor candidates using machine learning models developed on the HOPV dataset. Although both donors and acceptors are essential for an OPV application, the current work is restricted to donors as there are only a small number



(a)



(b)

Fig. 4: Learning curves for the cross-validated ERT models across different set of training examples for the MACCS and Atom Pair Fingerprints. The goodness of prediction (Q^2) is used as the score.

of known acceptors [58, 59] compared to hundreds of thousands of potential donor molecules. Therefore, developing a machine learning-based screening solution for donor molecules would lead to the identification of OPV devices with high PCE.

Figure 4 demonstrates the learning curve of the cross-validated ERT models across different set of training examples. The learning curves help demonstrate the increase of the learning capacity of the model as the dataset is increased. Further, the variance of the cross-validated models (indicated by the shaded green band surrounding the corresponding curve) decreases as the number of training examples increase.

To compare their performance, we also trained other state-of-the-art architectures for all datasets used. This includes a fully connected (FC) network trained on the fingerprint representations. Further, we also compare against 1-D CNN, RNN and CNN-RNN architectures trained on SMILES as recent papers have demonstrated their superiority over FC methods [40–42]. Lastly, we compare against other state of the art neural networks used in molecular informatics such as ConvGraph and Chemception. While the ConvGraph architecture uses the molecular structure encoded as graphs as input and then performs graph convolutions, Chemception architecture [60], based on the Inception architecture for image classification [61], directly develops a very deep neural network model by training directly on images of molecules. Bagging, RandomForest, ERTs and AdaBoost algorithms were implemented using Scikit-Learn Python Library [62]. The XGBoost package [63] was utilized for creating the xgboost model. The FC, CNN, RNN, CNN-RNN and Chemception models were implemented using Keras [64] with Tensorflow [65] backend. The ConvGraph was implemented using DeepChem library [66].

In Table 1, we present the results of the experiments across all the models for the HOPV dataset. We present the % Mean Absolute Error (MAE), Root Mean Squarer Error (RMSE) and goodness of prediction (Q^2). We can observe the superiority of ERTs for both the MACCS and Atom Pair fingerprints over the other models. ERTs trained on MACCS and Atom Pair had a mean absolute percentage error (% MAE) of 1.91% and 1.97%. The RNN, CNN and CNN-RNNs trained on the SMILES had % MAE between 2.62% and 3.25%. Convolutional Graphs had % MAE of 2.82 % and all other methods based on deep neural networks had even higher % MAE. Two ensemble tree based algorithms XGBoost and Random Forest outperform all other methods except ERTs. Even other ensemble tree-based algorithms such as AdaBoost and Bagging perform relatively well and at par with the best neural network based methods (RNNs and CNN-RNNs). It must be noted that although ERT models outperform RF models based on % MAE (lower %MAE) and Q^2 (much higher Q^2), RF models have slightly lower RMSE.

In Table 2, the results of the randomization tests such as y-Randomization and pseudo-Descriptor tests are delineated. y-Randomization (also known as y-scrambling or response randomization) is a form of a permutation test, where the values of the response variable are randomly ascribed to different compounds, while the descriptors values are left intact. In the pseudo-descriptors test, the descriptors are replaced by random numbers that are also subsequently used to train the models. In our case as the features in fingerprints are bit vectors, we generate random bit strings for features. A comparison across the performance metrics such as % MAE, RMSE and Q^2 of the ERT models between the original dataset (in Table 1) and the randomization tests (in Table 2) demonstrates that our proposed models perform much better than models based on ran-

Table 1: Comparison of performance of ERT models with other algorithms for the HOPV dataset

Algorithm	**Feature**	**% MAE**	**RMSE**	Q^2
AdaBoost	Molecular Fingerprint (MACCS)	2.6443	0.0061	0.1670
AdaBoost	Molecular Fingerprint (AtomPair)	2.5395	0.0058	0.2269
XGBoost	Molecular Fingerprint (MACCS)	2.0472	0.0057	0.7277
XGBoost	Molecular Fingerprint (AtomPair)	2.0141	0.0057	0.7263
Bagging	Molecular Fingerprint (MACCS)	2.6162	0.0063	0.1098
Bagging	Molecular Fingerprint (AtomPair)	2.4500	0.0058	0.2503
Random Forest	Molecular Fingerprint (MACCS)	2.0977	0.0054	0.4982
Random Forest	Molecular Fingerprint (AtomPair)	2.0589	0.0053	0.5169
ERTs	**Molecular Fingerprint (MACCS)**	**1.9703**	**0.0057**	**0.7390**
ERTs	**Molecular Fingerprint (AtomPair)**	**1.9100**	**0.0056**	**0.7427**
FC	Molecular Fingerprint (MACCS)	3.6850	0.0084	-0.5906
FC	Molecular Fingerprint (AtomPair)	3.5135	0.0078	-0.3975
CNN	SMILES	3.2536	0.0072	-0.1885
RNN	SMILES	2.6240	0.0062	0.1200
CNN-RNN	SMILES	2.6443	0.0061	0.1670
ConvGraph	Molecular Graphs	2.8170	0.0079	0.1082
Chemception	Molecule Image	3.2738	0.0079	-0.4089

Table 2: Performance metrics of the randomization tests performed using the MACCS and AtomPair fingerprints as features

Features	**Model**	**% MAE**	**RMSE**	Q^2
MACCS	y-Randomization	4.6036	0.0117	-2.1476
	Pseudo-Descriptors	6.5617	0.0167	-5.3666
Atom Pair	y-Randomization	3.3981	0.0083	-0.5600
	Pseudo-Descriptors	5.5822	0.0147	-3.9450

dom input features (pseudo-Descriptors) or labels (y-Randomization).

3.2 Correlation of fingerprint features

We wanted to explore the correlation between the most important features for our model for understanding their impact on the HOMO value. Figures 5 and 6 depict the correlation matrices for top 5 features important for MACCS and Atom Pair Fingerprints, as they perform best across all the fingerprints. We restricted to top 5 features as the contribution of other features was very close to 0. The length of MACCS fingerprints is 166, which is much shorter compared to other fingerprints, and is least affected by the curse of dimensionality. The correlation plots demonstrate that presence of any ring (Feature 0), presence of a C=C double bond (Feature 3) and presence of an aromatic ring (Feature 4) is positively correlated with HOMO value, whereas a C≡N triple bond (Feature 1) and a N=O double bond (Feature 2) is negatively correlated with HOMO value. Further, the correlation plot illustrates that presence of any ring, the presence of C=C bond and presence of an aromatic ring are strongly positively correlated with each other and hence we can conclude that these features often co-occur together in compounds with high HOMO value. Similarly, C≡N triple bond and N=O double bond have a weak positive correlation with each other, and their co-occurrence together leads to a compound with low HOMO value.

Figure 7 depicts two compounds with the highest HOMO value, and the abundance of rings including aromatic rings correspond to our observation from the correlation plots. Figure 8 illustrates two compounds from the HOPV dataset with the lowest HOMO value, and the presence and abundance of C≡N triple bond and N=O double bond are per our expectation based on correlation values. Although all compounds in the HOPV dataset had aromatic rings as the fingerprints are count vectors and not bit vectors, it demonstrates that the number of rings positively correlate to higher HOMO value rather than the presence or absence of rings.

Figure 9 depicts the best-predicted structures from the dataset with respect to predictions based on both atom pair and MACCS fingerprints. All the compounds that are predicted well have many aromatic rings, in agreement to our models as the number of rings and the number of aromatic rings are essential features. On the contrary in Figure 10, the compounds have fewer aromatic rings, and also have many features that are not part of the important features in the extremely randomized tree model. This makes it difficult to ac-

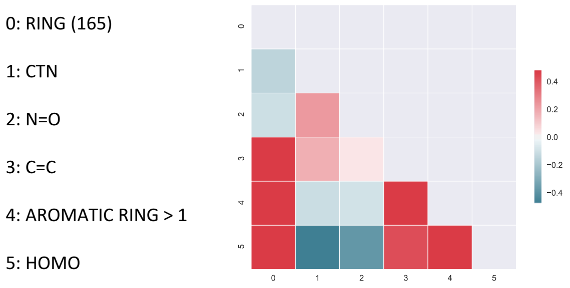


Fig. 5: Correlation for MACCS Fingerprints across the top 5 features and HOMO

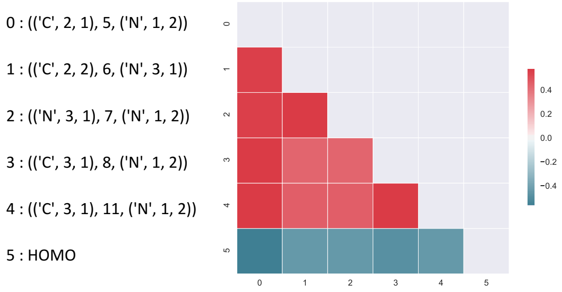


Fig. 6: Correlation for Atom Pair Fingerprints across the top 5 features and HOMO

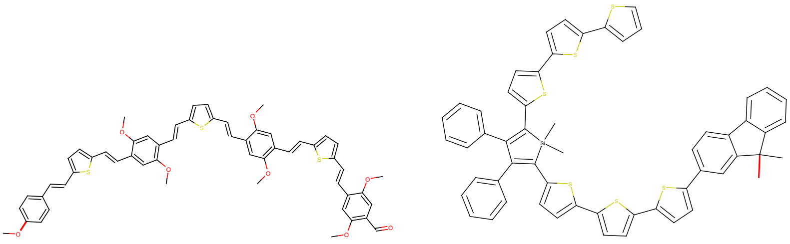

(a) (b)

Fig. 7: Specimen donor molecules with the highest HOMO

curately predict the HOMO value. Although in this paper, the predicted feature is HOMO and not PCE, the demonstrated dependence of HOMO and PCE (via the Scharber model as well as illustrated in Figure 2 implies that PCE values are correlated directly to HOMO.

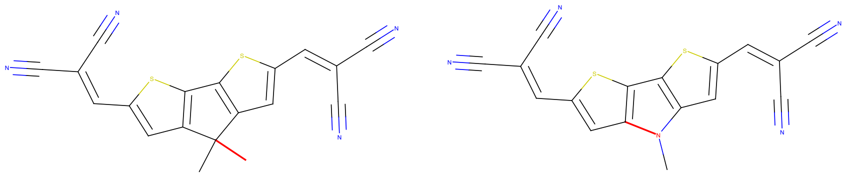
3.3 Generalization on Larger Dataset

We explored ERTs on the larger dataset of 22,179 molecules extracted from the Harvard CEP Database. We present the distribution of the HOMO values of the larger dataset in Figure 11. The reported HOMO values in the CEP dataset are an aggregate across several functionals. Table 3 compares the performance of the ERT models with other algorithms. As this dataset is much larger compared to the 350 molecule HOPV dataset, some deep neural methods such as convolutional graphs expectedly perform comparable to the ERTs, and SMILES-based models slightly outperform the ERT models. As the dataset is larger, we increased the number of trees in our model to 200.

4 Web Application

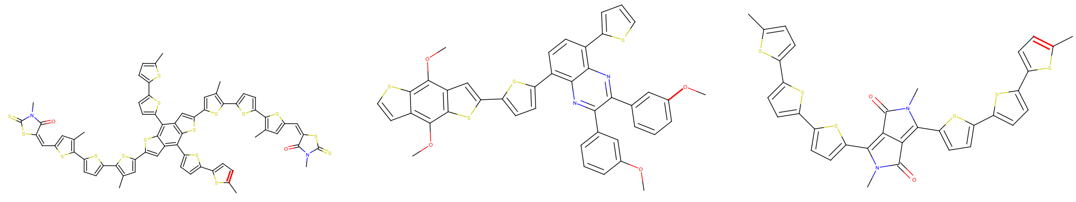
A web application is developed for the convenience of end users. The application accepts a single donor molecule in canonical SMILES format, converts it to the corresponding fingerprint, performs feature reduction, and the machine learning model is run to predict its HOMO value. Further, the results for other functionals are also calculated using the linear correlation between B3LYP and these functionals.

The RDKit library is used for generating the fingerprint from the SMILES notation as well as generat-



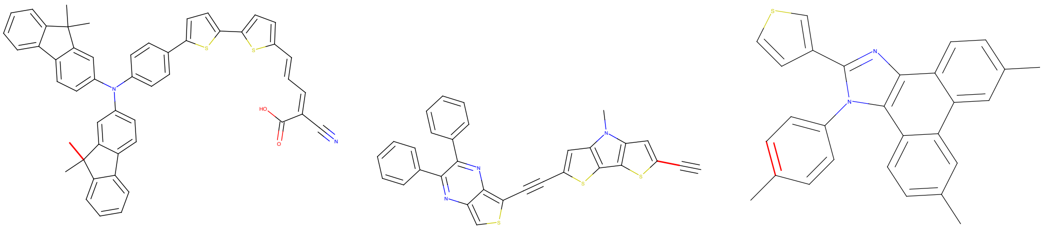
(a) (b)

Fig. 8: Specimen donor molecules with the lowest HOMO



(a) (b) (c)

Fig. 9: Best predicted structures based on prediction by both MACCS and Atom Pair Fingerprints



(a) (b) (c)

Fig. 10: Worst Predicted Structures based on prediction by both MACCS and Atom Pair Fingerprints

Table 3: Comparison of extremely randomized tree models with other algorithms for the 22,179 molecule CEP dataset

Algorithm	**Feature**	**% MAE**	**RMSE**	Q^2
AdaBoost	MACCS	2.0349	0.1284	0.7210
AdaBoost	AtomPair	2.0170	0.1272	0.7261
XGBoost	MACCS	0.9430	0.0611	0.9558
XGBoost	AtomPair	0.9378	0.0622	0.9523
Bagging	MACCS	1.6434	0.107	0.8065
Bagging	AtomPair	1.6418	0.1076	0.8551
Random Forest	MACCS	1.4331	0.0946	0.8864
Random Forest	AtomPair	1.4654	0.0967	0.8819
ERTs	**MACCS**	**0.8991**	**0.0598**	**0.9572**
ERTs	**AtomPair**	**0.8696**	**0.0584**	**0.9604**
FC	MACCS	1.6444	0.1070	0.8065
FC	AtomPair	1.6226	0.1058	0.8107
CNN	SMILES	0.7804	0.0521	0.9673
RNN	SMILES	0.7815	0.0527	0.9663
CNN-RNN	SMILES	0.7786	0.0529	0.9667
ConvGraph	Molecular Graphs	0.9104	0.0519	0.9619
Chemception	Molecule Image	1.4681	0.0974	0.8762

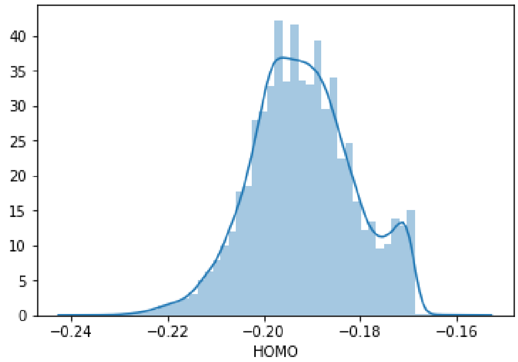


Fig. 11: Distribution of the CEP subset. All the HOMO values are in atomic units (a.u.). 1 atomic unit is equal to 27.21 eV.

ing the molecular structure representation. The scikit learn library is used for loading and running the trained model on the input molecule. The libraries Tornado and Flask are used for generating and displaying the output from the machine learning model on the website.

It must be noted that prediction of properties based on the Scharber model indicates the highest possible power efficiency [47], and the actual efficiency after accounting for the morphology of the final photovoltaic device is usually lower. Figure 12 illustrates the screenshot of the homepage of the web application where a user can input the SMILES of a potential donor compound.

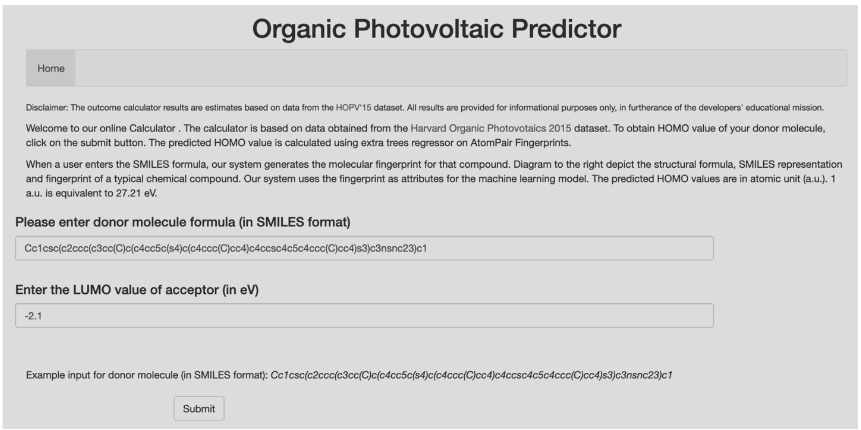
Figure 13 depicts the screenshot corresponding to a response for the predicted HOMO values across 4 functionals: PBE, B3LYP, BP86 and M06 is displayed in both a.u. as well as eV alongside the molecular structure of the compound. Further, we also calculate the open circuit voltage(V_{oc}) for the corresponding donor-acceptor combination when the user provides LUMO value of the acceptor. It must be noted that our models initially predict PBE and then a linear transformation is used to calculate values for other functionals based on the PBE prediction.

Although the Scharber model is simplistic to account for all the complex physics of an OPV explicitly, it nonetheless provides a valuable indication of the inherent promise of a candidate compound. Further, as the HOPV dataset was small, the web application must be used with caution. Due to the low mean absolute percentage error (% MAE), it will have high precision for compounds that are similar to those in the HOPV dataset. For instance, the HOPV dataset has only 3 compounds that have Selenium in the donor molecule.

5 Conclusions

A methodology for predicting properties using fingerprints of donor molecules is presented. The elegance of an ensemble based regression technique such as ERTs lies in the fact that it minimizes the need for feature reduction or normalization. In particular, ERTs are generalizable and less prone to overfitting which is essential while learning from a small dataset. Further, ERTs are easily interpretable - a desired trait for further understanding of which features are most important for the predicted property of a given monomer. One of the goals of machine learning models is reusability. In the proposed work, although the models were trained using the PBE functional values, we ascertained that HOMO values of other functionals namely B3LYP, BP86, and M06 could be expressed as a linear transformation of their corresponding values for B3LYP functionals. Hence, the models developed for PBE can be extended to predict for other functionals. For the smaller OPV dataset, ERT models achieve better performance than other methods -both tree-based as well as those based on neural network. Further, we evaluated ERTs on the larger dataset and it performed almost at par with CNN or RNN-based neural networks trained on SMILES. We also provide a web application where users can receive the predicted HOMO values for the chemical compound of the donor as well as V_{oc} of the donor-acceptor combination for a given acceptor.

This work reveals the potential of integration of feature manipulation combined with extensive grid search on a small experiment-theory calibrated dataset of organic photovoltaic donors. Our system allows researchers to get an estimate of the HOMO energy values of donor compounds used in OPV applications, and motivate the development of an inexpensive photovoltaic solution. Directed efforts are needed to standardize the collection and representation of experimental manufacturing and processing data for effective use with machine learning techniques. Leveraging machine learning with computational and experimental chemistry could play an essential role in the expedition of systematic design of high-efficiency OPV materials, and holds significant promise as a potential solution to future energy needs. The success of using machine learning models on a small but well-curated calibrated dataset exposes an exciting area in materials discovery, and in particular for solar cell technology. This, in turn, can provide a path towards solving the world energy problem in a clean and environmentally friendly way.



Organic Photovoltaic Predictor

Disclaimer: The outcome calculator results are estimates based on data from the HOPV'15 dataset. All results are provided for informational purposes only, in furtherance of the developers' educational mission.

Welcome to our online Calculator . The calculator is based on data obtained from the Harvard Organic Photovotaics 2015 dataset. To obtain HOMO value of your donor molecule, click on the submit button. The predicted HOMO value is calculated using extra trees regressor on AtomPair Fingerprints.

When a user enters the SMILES formula, our system generates the molecular fingerprint for that compound. Diagram to the right depict the structural formula, SMILES representation and fingerprint of a typical chemical compound. Our system uses the fingerprint as attributes for the machine learning model. The predicted HOMO values are in atomic unit (a.u.). 1 a.u. is equivalent to 27.21 eV.

Please enter donor molecule formula (in SMILES format)

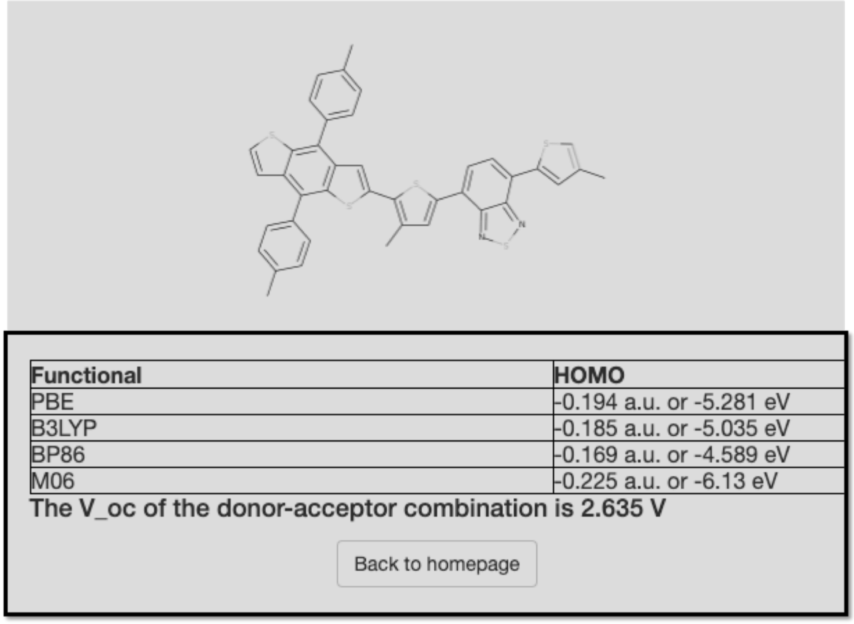
Cc1csc(c2ccc(c3cc(C)c(c4cc5c(s4)c(c4ccc(C)cc4)c4ccsc4c5c4ccc(C)cc4)s3)c3nsnc23)c1

Enter the LUMO value of acceptor (in eV)

-2.1

Example input for donor molecule (in SMILES format): *Cc1csc(c2ccc(c3cc(C)c(c4cc5c(s4)c(c4ccc(C)cc4)c4ccsc4c5c4ccc(C)cc4)s3)c3nsnc23)c1*

Fig. 12: Screenshot of the the web application with request page for a given donor compound, and LUMO value of a potential acceptor



Functional	HOMO
PBE	-0.194 a.u. or -5.281 eV
B3LYP	-0.185 a.u. or -5.035 eV
BP86	-0.169 a.u. or -4.589 eV
M06	-0.225 a.u. or -6.13 eV

The V_oc of the donor-acceptor combination is 2.635 V

Fig. 13: Screenshot of the the web application with response illustrating the molecular compound of the SMILES input with the predicted HOMO values across 4 functionals: PBE, B3LYP, BP86 and M06

Acknowledgements This work was performed under the following financial assistance award 70NANB19H005 from U.S. Department of Commerce, National Institute of Standards and Technology as part of the Center for Hierarchical Materials Design (CHiMaD). Partial support is also acknowledged from DOE awards DE-SC0014330, DE-SC0019358.

References

1. Rosaria Ciriminna, Francesco Meneguzzo, Mario Pecoraino, and Mario Pagliaro. Rethinking solar energy education on the dawn of the solar economy. *Renewable and Sustainable Energy Reviews*, 63:13–18, 2016.
2. Oecd global science forum - oecd, 2016.
3. Wichien Sang-aroon, Seksan Laopha, Phrompak Chaiamornnugool, Sarawut Tontapha, Samarn Saekow, and Vittaya Amornkitbamrung. Dft and tddft study on the electronic structure and photoelectrochemical properties of dyes derived from cochineal and lac insects as photosensitizer for dye-sensitized solar cells. *Journal of molecular modeling*, 19(3):1407–1415, 2013.
4. Zhong Hu, Vedbar S Khadka, Wei Wang, David W Galipeau, and Xingzhong Yan. Theoretical study of two-photon absorption properties and up-conversion efficiency of new symmetric organic π-conjugated molecules for photovoltaic devices. *Journal of molecular modeling*, 18(8):3657–3667, 2012.
5. Mazmira Mohamad, Rashid Ahmed, Amirudin Shaari, and Souraya Goumri-Said. First principles investigations of vinazene molecule and molecular crystal: a prospective candidate for organic photovoltaic applications. *Journal of molecular modeling*, 21(2):27, 2015.
6. Natalia Inostroza, Fernando Mendizabal, Ramiro Arratia-Pérez, Carlos Orellana, and Cristian Linares-Flores. Improvement of photovoltaic performance by substituent effect of donor and acceptor structure of tpa-based dye-sensitized solar cells. *Journal of molecular modeling*, 22(1):25, 2016.
7. Claudia N Hoth, Roland Steim, Pavel Schilinsky, Stelios A Choulis, Sandro F Tedde, Oliver Hayden, and Christoph J Brabec. Topographical and morphological aspects of spray coated organic photovoltaics. *Organic Electronics*, 10(4):587–593, 2009.
8. MC Scharber and NS Sarciftci. Bulk heterojunction organic solar cells: Working principles and power conversion efficiencies. *Nanostructured Materials for Type III Photovoltaics*, 45:33, 2017.
9. Jiangeng Xue, Soichi Uchida, Barry P Rand, and Stephen R Forrest. Asymmetric tandem organic photovoltaic cells with hybrid planar-mixed molecular heterojunctions. *Applied Physics Letters*, 85(23):5757–5759, 2004.
10. Jianhui Hou and Xia Guo. Active layer materials for organic solar cells. In *Organic Solar Cells*, pages 17–42. Springer, 2013.
11. P Granero, VS Balderrama, J Ferré-Borrull, J Pallarès, and LF Marsal. Two-dimensional finite-element modeling of periodical interdigitated full organic solar cells. *Journal of Applied Physics*, 113(4):043107, 2013.
12. Yongjeong Lee, Kyungnam Kang, Sanghwa Lee, Hyeong Pil Kim, Jin Jang, and Jungho Kim. Integrated optoelectronic model for organic solar cells

based on the finite element method including the effect of oblique sunlight incidence and a non-ohmic electrode contact. *Japanese Journal of Applied Physics*, 55(10):102301, 2016.

13. Warren J. Hehre. *Ab initio molecular orbital theory.* Wiley-Interscience, 1986.
14. Jeremy Taylor, Hong Guo, and Jian Wang. Ab initio modeling of quantum transport properties of molecular electronic devices. *Physical Review B*, 63(24):245407, 2001.
15. Jean-Luc Brédas, Joseph E Norton, Jérôme Cornil, and Veaceslav Coropceanu. Molecular understanding of organic solar cells: the challenges. *Accounts of chemical research*, 42(11):1691–1699, 2009.
16. Abraham Yosipof, Omer Kaspi, Koushik Majhi, and Hanoch Senderowitz. Visualization based data mining for comparison between two solar cell libraries. *Molecular informatics*, 35(11-12):622–628, 2016.
17. Peter B Jørgensen, Mikkel N Schmidt, and Ole Winther. Deep generative models for molecular science. *Molecular informatics*, 37(1-2):1700133, 2018.
18. Omer Kaspi, Abraham Yosipof, and Hanoch Senderowitz. Pv analyzer: A decision support system for photovoltaic solar cells libraries. *Molecular informatics*, 37(9-10):1800067, 2018.
19. Jean Roncali, Philippe Leriche, and Philippe Blanchard. Molecular materials for organic photovoltaics: small is beautiful. *Advanced Materials*, 26(23):3821–3838, 2014.
20. Sam-Shajing Sun and Niyazi Serdar Sariciftci. *Organic photovoltaics: mechanisms, materials, and devices.* CRC press, 2017.
21. Sarah Holliday, Yilin Li, and Christine Luscombe. Recent advances in high performance donor-acceptor polymers for organic photovoltaics. *Progress in Polymer Science*, 2017.
22. Johannes Hachmann, Roberto Olivares-Amaya, Sule Atahan-Evrenk, Carlos Amador-Bedolla, Roel S Sánchez-Carrera, Aryeh Gold-Parker, Leslie Vogt, Anna M Brockway, and Alán Aspuru-Guzik. The harvard clean energy project: large-scale computational screening and design of organic photovoltaics on the world community grid. *The Journal of Physical Chemistry Letters*, 2(17):2241–2251, 2011.
23. Harikrishna Sahu, Weining Rao, Alessandro Troisi, and Haibo Ma. Toward predicting efficiency of organic solar cells via machine learning and improved descriptors. *Advanced Energy Materials*, 8(24):1801032, 2018.
24. Dipendra Jha, Logan Ward, Arindam Paul, Wei-keng Liao, Alok Choudhary, Chris Wolverton, and Ankit Agrawal. Elemnet: Deep learning the chemistry of materials from only elemental composition. *Scientific reports*, 8(1):17593, 2018.
25. Keith T Butler, Daniel W Davies, Hugh Cartwright, Olexandr Isayev, and Aron Walsh. Machine learning for molecular and materials science. *Nature*, 559(7715):547, 2018.
26. Arindam Paul, Pinar Acar, Ruoqian Liu, Wei-Keng Liao, Alok Choudhary, Veera Sundararaghavan, and Ankit Agrawal. Data sampling schemes for microstructure design with vibrational tuning constraints. *AIAA Journal*, 56(3):1239–1250, 2018.
27. Peter Bjørn Jørgensen, Murat Mesta, Suranjan Shil, Juan Maria García Lastra, Karsten Wedel Jacobsen, Kristian Sommer Thygesen, and Mikkel N Schmidt. Machine learning-based screening of complex molecules for polymer solar cells. *The Journal of chemical physics*, 148(24):241735, 2018.
28. Arindam Paul, Pinar Acar, Wei-keng Liao, Alok Choudhary, Veera Sundararaghavan, and Ankit Agrawal. Microstructure optimization with constrained design objectives using machine learning-based feedback-aware data-generation. *Computational Materials Science*, 160:334–351, 2019.
29. Bing Cao, Lawrence A Adutwum, Anton O Oliynyk, Erik J Luber, Brian C Olsen, Arthur Mar, and Jillian M Buriak. How to optimize materials and devices via design of experiments and machine learning: Demonstration using organic photovoltaics. *ACS nano*, 12(8):7434–7444, 2018.
30. Dipendra Jha, Logan Ward, Zijiang Yang, Christopher Wolverton, Ian Foster, Wei-keng Liao, Alok Choudhary, and Ankit Agrawal. Irnet: A general purpose deep residual regression framework for materials discovery. 2019.
31. Dipendra Jha, Aaron Gilad Kusne, Nam Nguyen, Wei-keng Liao, Alok Choudhary, and Ankit Agrawal. Peak area detection network for directly learning phase regions from raw x-ray diffraction patterns. In *2019 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2019.
32. Zijiang Yang, Yuksel C Yabansu, Dipendra Jha, Wei-keng Liao, Alok N Choudhary, Surya R Kalidindi, and Ankit Agrawal. Establishing structure-property localization linkages for elastic deformation of three-dimensional high contrast composites using deep learning approaches. *Acta Materialia*, 166:335–345, 2019.
33. Ankit Agrawal and Alok Choudhary. Deep materials informatics: Applications of deep learning in materials science. *MRS Communications*, pages 1–14, 2019.

34. Juan-Pablo Correa-Baena, Kedar Hippalgaonkar, Jeroen van Duren, Shaffiq Jaffer, Vijay R Chandrasekhar, Vladan Stevanovic, Cyrus Wadia, Supratik Guha, and Tonio Buonassisi. Accelerating materials development via automation, machine learning, and high-performance computing. *Joule*, 2(8):1410–1420, 2018.
35. Jan Ulaczyk, Krzysztof Morawiec, Paweł Zabierowski, Tomasz Drobiazg, and Nicolas Barreau. Finding relevant parameters for the thin-film photovoltaic cells production process with the application of data mining methods. *Molecular informatics*, 36(9), 2017.
36. Adam C Mater and Michelle L Coote. Deep learning in chemistry. *Journal of chemical information and modeling*, 2019.
37. Pierre Geurts, Damien Ernst, and Louis Wehenkel. Extremely randomized trees. *Machine learning*, 63(1):3–42, 2006.
38. Omer Kaspi, Abraham Yosipof, and Hanoch Senderowitz. Random sample consensus (ransac) algorithm for material-informatics: application to photovoltaic solar cells. *Journal of cheminformatics*, 9(1):34, 2017.
39. Johannes Hachmann, Roberto Olivares-Amaya, Adrian Jinich, Anthony L Appleton, Martin A Blood-Forsythe, Laszlo R Seress, Carolina Roman-Salgado, Kai Trepte, Sule Atahan-Evrenk, Süleyman Er, et al. Lead candidates for high-performance organic photovoltaics from high-throughput quantum chemistry–the harvard clean energy project. *Energy & Environmental Science*, 7(2):698–704, 2014.
40. Garrett B Goh, Nathan O Hodas, Charles Siegel, and Abhinav Vishnu. Smiles2vec: An interpretable general-purpose deep neural network for predicting chemical properties. *arXiv preprint arXiv:1712.02034*, 2017.
41. Arindam Paul, Dipendra Jha, Reda Al-Bahrani, Wei-keng Liao, Alok Choudhary, and Ankit Agrawal. Chemixnet: Mixed dnn architectures for predicting chemical properties using multiple molecular representations. In *NeurIPS Workshop on Molecules and Materials*, 2018.
42. Arindam Paul, Dipendra Jha, Reda Al-Bahrani, Wei-keng Liao, Alok Choudhary, and Ankit Agrawal. Transfer learning using ensemble neural networks for organic solar cell screening. In *2019 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2019.
43. Markus C Scharber, David Mühlbacher, Markus Koppe, Patrick Denk, Christoph Waldauf, Alan J Heeger, and Christoph J Brabec. Design rules for donors in bulk-heterojunction solar cells—towards 10% energy-conversion efficiency. *Advanced materials*, 18(6):789–794, 2006.
44. The harvard organic photovoltaics 2015 (hopv) dataset: An experiment-theory calibration resource. `https://figshare.com/articles/HOPV15_Dataset/1610063`, 2016.
45. O Anatole von Lilienfeld, Raghunathan Ramakrishnan, Matthias Rupp, and Aaron Knoll. Fourier series of atomic radial distribution functions: A molecular fingerprint for machine learning models of quantum chemical properties. *International Journal of Quantum Chemistry*, 115(16):1084–1093, 2015.
46. Florian Weigend and Reinhart Ahlrichs. Balanced basis sets of split valence, triple zeta valence and quadruple zeta valence quality for h to rn: Design and assessment of accuracy. *Physical Chemistry Chemical Physics*, 7(18):3297–3305, 2005.
47. Edward O Pyzer-Knapp, Kewei Li, and Alan Aspuru-Guzik. Learning from the harvard clean energy project: The use of neural networks to accelerate materials discovery. *Advanced Functional Materials*, 25(41):6495–6502, 2015.
48. Satya Avasarala. *Selenium WebDriver practical guide*. Packt Publishing Ltd, 2014.
49. Leonard Richardson. Beautiful soup documentation. *April*, 2007.
50. Iiris Kahn, Andre Lomaka, and Mati Karelson. Topological fingerprints as an aid in finding structural patterns for lrrk2 inhibition. *Molecular informatics*, 33(4):269–275, 2014.
51. Adrià Cereto-Massagué, María José Ojeda, Cristina Valls, Miquel Mulero, Santiago Garcia-Vallvé, and Gerard Pujadas. Molecular fingerprint similarity search in virtual screening. *Methods*, 71:58–63, 2015.
52. Sereina Riniker and Gregory A Landrum. Open-source platform to benchmark fingerprints for ligand-based virtual screening. *Journal of cheminformatics*, 5(1):1, 2013.
53. Andreas Bender and Robert C Glen. Molecular similarity: a key technique in molecular informatics. *Organic & biomolecular chemistry*, 2(22):3204–3218, 2004.
54. Michael Reutlinger, Christian P Koch, Daniel Reker, Nickolay Todoroff, Petra Schneider, Tiago Rodrigues, and Gisbert Schneider. Chemically advanced template search (cats) for scaffold-hopping and prospective target prediction for orphan molecules. *Molecular informatics*, 32(2):133–138, 2013.

55. Yasuo Tabei and Koji Tsuda. Sketchsort: Fast all pairs similarity search for large databases of molecular fingerprints. *Molecular informatics*, 30(9):801–807, 2011.
56. David Weininger. Smiles, a chemical language and information system. 1. introduction to methodology and encoding rules. *Journal of chemical information and computer sciences*, 28(1):31–36, 1988.
57. G Landrum. Rdkit: open-source cheminformatics software, 2016.
58. Navpreet Kaur, Mandeep Singh, Dinesh Pathak, Tomas Wagner, and JM Nunzi. Organic materials for photovoltaic applications: Review and mechanism. *Synthetic Metals*, 190:20–26, 2014.
59. Yuze Lin and Xiaowei Zhan. Non-fullerene acceptors for organic photovoltaics: an emerging horizon. *Materials Horizons*, 1(5):470–488, 2014.
60. Garrett B Goh, Charles Siegel, Abhinav Vishnu, Nathan O Hodas, and Nathan Baker. Chemception: A deep neural network with minimal chemistry knowledge matches the performance of expert-developed qsar/qspr models. *arXiv preprint arXiv:1706.06689*, 2017.
61. Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015.
62. F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.
63. Tianqi Chen, Tong He, Michael Benesty, Vadim Khotilovich, and Yuan Tang. Xgboost: extreme gradient boosting. *R package version 0.4-2*, pages 1–4, 2015.
64. François Chollet et al. Keras, 2015.
65. Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, et al. Tensorflow: A system for large-scale machine learning. In *12th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 16)*, pages 265–283, 2016.
66. deepchem package — deepchem master documentation. `https://deepchem.io/docs/deepchem.html`. (Accessed on 06/23/2019).