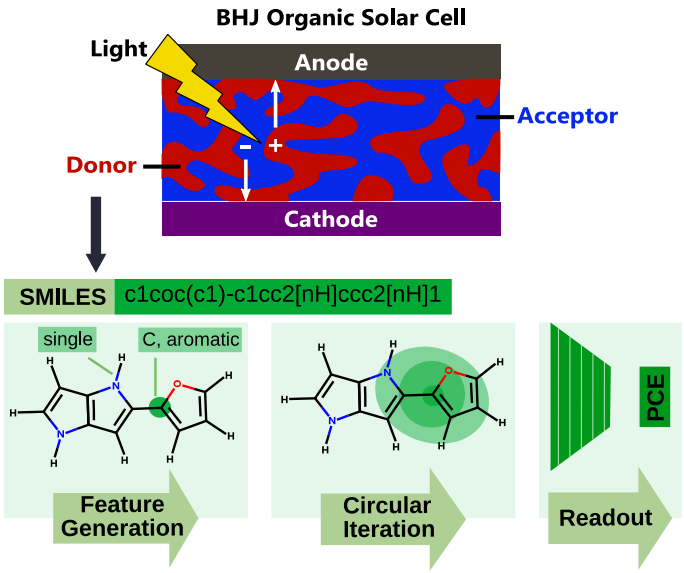
User Story

Machine Learning techniques to evaluate power conversion efficiency of organic photovoltaics



THE CHALLENGE

Test the ability of Machine Learning (ML) models to predict Power Conversion Efficiency (PCE) of organic photovoltaics (OPVs) based on SMILES-derived structural information of the donor candidates, as well as assess impact and implications of the choice of training data.

THE SOLUTION

- Three neural and three baseline models implemented, trained and assessed for predicting PCEs.
- Computational (CEPDB) and experimental (HOPV15) datasets considered

THE RESULTS

Performance of the tested models in predicting PCEs:

- The CEPDB fitted well by all models, but predicted PCEs disagree with experiments.
- The HOPV15 fitted poorly, with the baseline models being better than the neural
- Graph-based neural networks were found to be the best performing models

OVERVIEW

With a strong global push towards clean energy generation, more resources are being invested in researching and developing photovoltaic devices. One particular type of such devices are organic photovoltaics (OPVs) which, recently, have been gaining increasing interest on account of their unique properties such as a low weight, flexibility or easy of manufacture. In order to design a well performing OPV material, a lot of experiments is required, which is challenging both time and resource-wise. Therefore, computational methods are frequently employed to enable rapid screening of candidate materials looking for the materials that would give the highest power conversion efficiency (PCE) of the final OPV product.

This use-case focuses on building and testing the performance of computational models for predicting power conversion efficiency of OPVs based on the SMILES-derived structural information of the donor candidates as well as assessing impact and implications of the choice of training data: large but synthetic vs small but experimental datasets.

CASE DESCRIPTION

- Three neural and three baseline models were built, trained and assessed on two common datasets: computational (CEPDB [1]) and experimental (HOPV15 [2])
- The neural models comprised: (i) Bi-directional Long Short-Term Memory (gFSI/BiLSTM), Attentive Fingerprints (Attentive FP), and Simple Graph neural network (Simple GNN) models
- The baseline models comprised: (i) Support Vector Regression (SVR), (ii) Random Forests, and (iii) High Dimensional Model Representation (HDMR) models
- All the models were trained using structural information derived from the donor molecule SMILES strings
- Single and nested cross-validation techniques were used to minimise data partitioning impact on the model final assessment score.

1. J. Hanchmann, R. Olivares-Amaya et al., *Chemistry Letters.*, *2241-2251, 2011*
2. S. A. Lopez, E. O. Pyzer-Knapp et al., *Scientific Data, 1-7, 2016*

User Story

Table 1: Performance of models trained on CEPDB dataset in predicting PCE of organic photovoltaics

Model / Dataset	CEPDB				
	set	MSE	MAE	R^2	r
g-FSI/BiLSTM	train	0.249	0.369	0.957	0.978
	val	0.544	0.520	0.906	0.952
	test	0.554	0.520	0.904	0.951
Simple GNN	train	0.064	0.192	0.989	0.994
	val	0.154	0.287	0.973	0.987
	test	0.164	0.288	0.972	0.986
Attentive FP	train	0.024	0.118	0.996	0.998
	val	0.062	0.176	0.989	0.995
	test	0.071	0.180	0.988	0.994
SVR	train	0.009	0.010	0.999	0.999
	test	0.345	0.418	0.940	0.970
RF	train	0.009	0.062	0.998	0.999
	test	0.620	0.556	0.893	0.945
HDMR	train	0.413	0.475	0.929	0.964
	test	0.530	0.536	0.909	0.953

Table 2: Performance of models trained on HOPV15 datasets in predicting PCE of organic photovoltaics

Model / Dataset	HOPV15				
	set	$\overline{\text{MSE}}$	$\overline{\text{MAE}}$	$\overline{\text{R}^2}$	$\bar{\text{r}}$
g-FSI/BiLSTM	train	2.519 ± 0.446	1.282 ± 0.130	0.505 ± 0.055	0.737 ± 0.040
	val	3.208 ± 1.015	1.427 ± 0.200	0.309 ± 0.101	0.569 ± 0.076
	test	3.935 ± 0.493	1.612 ± 0.104	0.188 ± 0.171	0.467 ± 0.156
Simple GNN	train	1.279 ± 0.580	0.863 ± 0.210	0.746 ± 0.114	0.877 ± 0.058
	val	2.410 ± 0.551	1.220 ± 0.111	0.470 ± 0.074	0.698 ± 0.041
	test	4.059 ± 0.524	1.567 ± 0.134	0.176 ± 0.109	0.521 ± 0.053
Attentive FP	train	2.936 ± 1.101	1.377 ± 0.377	0.420 ± 0.211	0.648 ± 0.138
	val	3.020 ± 0.964	1.397 ± 0.227	0.355 ± 0.081	0.597 ± 0.070
	test	4.417 ± 1.503	1.672 ± 0.223	0.127 ± 0.193	0.455 ± 0.113
SVR	train	0.253 ± 0.444	0.187 ± 0.301	0.950 ± 0.088	0.976 ± 0.043
	test	2.846 ± 0.459	1.360 ± 0.102	0.423 ± 0.086	0.668 ± 0.076
RF	train	0.696 ± 0.616	0.573 ± 0.348	0.859 ± 0.124	0.934 ± 0.060
	test	2.902 ± 0.379	1.329 ± 0.052	0.409 ± 0.081	0.652 ± 0.055
HDMR	train	0.724 ± 0.171	0.673 ± 0.070	0.855 ± 0.035	0.927 ± 0.019
	test	3.185 ± 0.540	1.411 ± 0.080	0.350 ± 0.135	0.623 ± 0.078

[a] Provided model accuracy metrics are given as a mean across all m-folds and the error bars are given as σ, which for a normal distribution would correspond to a confidence level of 68%

RESULTS

Main model assessment results on CEPDB dataset are provided in **Table 1**. It can be seen that all of the models were able to derive high correlation coefficients between the learned and test PCE values, with the Attentive FP method reaching state of the art performance having the test set Mean Squared Error of 0.07.

Main model assessment results on HOPV15 dataset are provided in **Table 2**. In this case, the nested cross-validation technique was additionally used to reduce the impact of data partitioning into the training, validation and test sets. Therefore, the final performance metrics represent a mean over obtained metrics across all outer cross validation folds (5 in total) including standard deviation across these folds as the error bars. It can be seen that, overall, all models performed rather poorly. Attentive FP now performs the worst, with Simple GNN and g-FSI/BiLSTM also presenting very large errors. Contrary to the CEPDB case, the baseline models now outperform the neural methods, which could be due to the fact that the neural methods need to train the weights and have insufficient data to do so. Still, the performance of the machine learning models was not observed to be good.

The poor models performance in the HOPV15 case can be likely due to the nature of the HOPV15 dataset, which is smaller and also much less homogeneous than that in the CEPDB case. This is due to expected differences in experimental set ups, larger chemical complexity of the species in the dataset, and possibly a larger number of variables influencing real-world organic solar cell PCEs that may not be strongly correlated with the structural information of the donor molecules encoded in the SMILES strings, such as bulk solar cell properties.

APPLICATION AREAS

- Organic Solar Cells

PRODUCTS/SERVICES USED

- *Technical Services team*
- *MoDS*™