Deep learning based PCE Prediction and Classification of high performance OPV molecules from SMILES.

Dipika Boro

Abstract

With the increase in cell efficiency and the increasing drive to move towards renewable resources that have less environmental effect, there is an increased attention in the study and research of designing OPVs in the recent years. This study aims to predict PCE values and efficiency labels as ‘high’ or ‘low’ for Organic Photo Voltaic materials as an initial step towards designing OPVs. SMILES representation is used for this study, because it is useful to feed them to ML models. Since SMILES is a string-based representation, LSTM is used as the main model in this study because they capture patterns in sequences.

Introduction

With the global push towards using renewable energy resources for power generation and storage, a significant portion of resources are being invested in researching to better harness solar energy. Researches are being done for developing photovoltaic devices. Currently silicon based solar cells are the most prominent solar cells followed by perovskite-based ones. But with improving cell efficiency, organic photovoltaic materials are a rapidly emerging photovoltaic technology. In organic photovoltaic cells the absorbing layer is based on organic semiconductors (OSC). They are advantageous over other solar cells because of their low weight, flexibility, low environment impact and ease of manufacturing.

A typical OPV device consists of one or several photoactive materials sandwiched between two electrodes. Figure 1 shows the Bulk Heterojunction cell where the donor and acceptor are well mixed at the nanoscale level – allowing interfaces at an appropriate diffusion distance to be dispersed across the active layer whilst maintaining the necessary thickness for absorption.

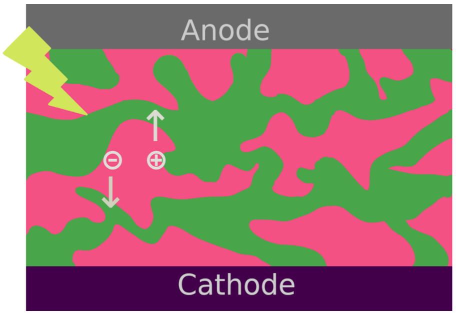


Figure 1: Bulk Heterojunction cell.

Although organic solar cells have much lower PCEs than its silicon-based or perovskite-based counterparts, recent studies have shown that organic solar cells also have potential to reach PCEs as high as 20% or even 30% in some cases. Predicting Power Conversion Efficiency of organic photovoltaics is an important initial first step towards designing an OPV with high PCE.

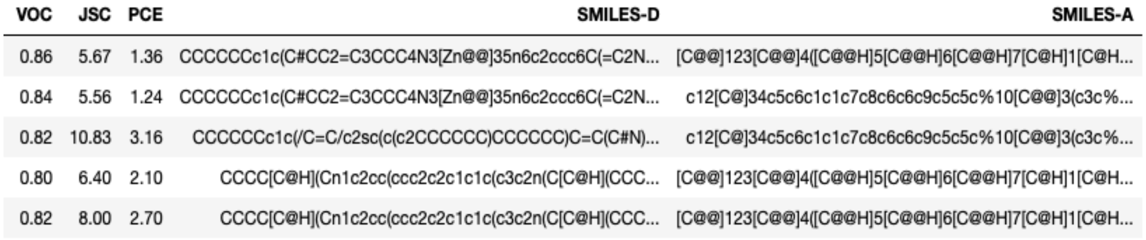
Conducting experiments can be extremely tedious and challenging. So many machine learning algorithms are being used to derive quantitative structure property relationship. Deep Learning has shown extraordinary prediction capabilities in fields such as Computer Vision and Natural Language processing. This study aims to use deep learning to predict the PCE value, and also to classify a given material into ‘high performance’ or ‘low performance’ OPV material. More details about how this was achieved is discussed in Methods and Experiments sections.

Related work

A wide variety of machine learning algorithms have been applied to predict the performance of organic photovoltaics by using different target data sets. The Harvard Clean Energy Project Database (CEPDB) [1], is one such target data set for ML models that contains computationally determined PCE values for 2.3 million organic photovoltaic candidates. In [2], the authors study and analyze 5 ML and DL based models for predicting PCE values. Sun et.al.[3], use 5 ML based algorithms for classifying OPVs as high or low performance. They use SMILES and fingerprint representations for making the efficiency prediction. Some researched use SMILES representation to detect chemical motifs. Hirohara et.al., [4] use CNN as the base model of architecture for their study.

Dataset and features

The dataset has 566 donor and acceptor pairs. They were obtained from literature. The features in the dataset are VOC, JSC, PCE, SMILES representation of donor, SMILES representation of acceptor. Only SMILES representation was used for this study. Figure 2 shows the first few rows of the dataset. For this study PCE and SMILES-D columns were used.



VOC	JSC	PCE	SMILES-D	SMILES-A
0.86	5.67	1.36	CCCCCCc1c(C#CC2=C3CCC4N3[Zn@@]35n6c2ccc6C(=C2N...	[C@@]123[C@@]4([C@@H]5[C@@H]6[C@@H]7[C@H]1[C@H...
0.84	5.56	1.24	CCCCCCc1c(C#CC2=C3CCC4N3[Zn@@]35n6c2ccc6C(=C2N...	c12[C@]34c5c6c1c1c7c8c6c6c9c5c5c%10[C@@]3(c3c%...
0.82	10.83	3.16	CCCCCCc1c(/C=C/c2sc(c(c2CCCCCC)CCCCCC)C=C(C#N)...	c12[C@]34c5c6c1c1c7c8c6c6c9c5c5c%10[C@@]3(c3c%...
0.80	6.40	2.10	CCCC[C@H](Cn1c2cc(ccc2c2c1c1c(c3c2n(C[C@H](CCC...	[C@@]123[C@@]4([C@@H]5[C@@H]6[C@@H]7[C@H]1[C@H...
0.82	8.00	2.70	CCCC[C@H](Cn1c2cc(ccc2c2c1c1c(c3c2n(C[C@H](CCC...	[C@@]123[C@@]4([C@@H]5[C@@H]6[C@@H]7[C@H]1[C@H...

Figure 2: First five rows of the dataset.

After the dataset was obtained, exploratory data analysis was performed. The minimum and maximum PCE values in the dataset were 0.003 and 12.08. The median PCE value was 3.1. Histogram (Fig.3a) and boxplot (Fig.3b) of the dataset showed that the most datapoints had PCE value in the range 1.75 - 5.3.

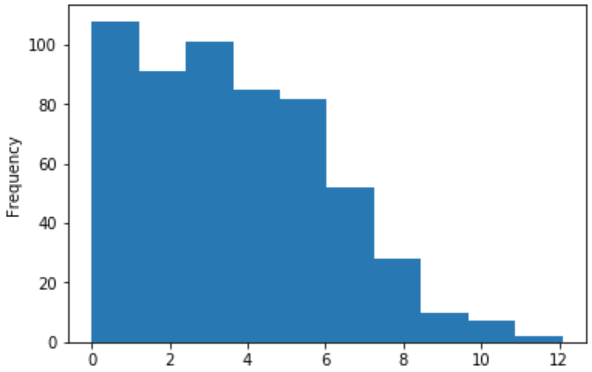


Figure 3a: Frequency of PCE values

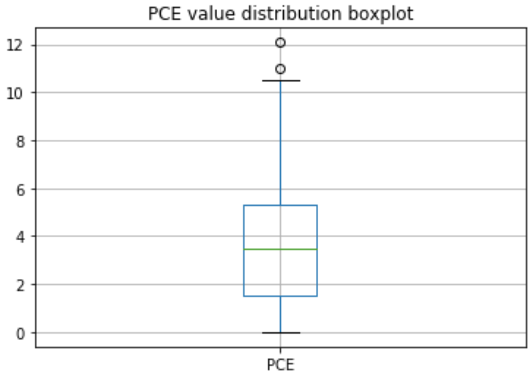


Figure 3b: Boxplot of PCE values in the dataset

Methods

SMILES is the 'Simplified Molecular Input Line Entry System', which is used to translate a chemical's three-dimensional structure into a string of symbols that can be used by computers. It is a linear representation that can be fed to a deep learning model for predicting the output which in our study was the PCE value for regression task and label for classification task.

RNN based deep learning models have proved to be very successful in identifying patterns in sequential and time-series data. It has had significant improvement in understanding Natural Language Processing related tasks in the last decade. Since the dataset has SMILES representation, which is a string of characters (letters and symbols) we decided to use LSTM which is an improved RNN based model. LSTM is used as the main model on which our architecture would be based.

The smiles data was first preprocessed prior to feeding it to the neural net. Tokenization was done at character level. Padded sequences were generated with post padding and zero masking before feeding to the LSTM. The first model used for prediction was a deep neural net of four dense layers with dropout layers and relu as activation function. The architecture of this model is shown in Figure 4. For training and evaluation Mean Squared Error (MSE) and Mean Absolute Error (MAE) were used as loss function. Same model was used with and without K-fold validation. Details are provided in experiments section.

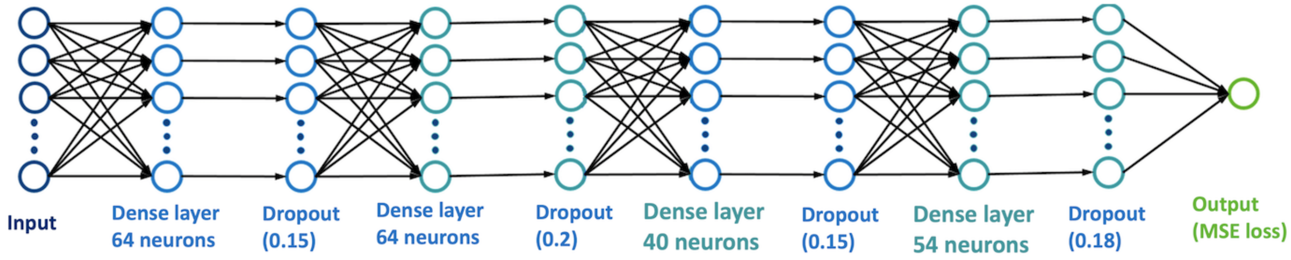


Figure 4: Model 1- Deep neural net for regression task for predicting PCE value.

For the same task a second model was used. Two LSTM layers with dropout layers and the final layer was a dense layer with one node for predicting the PCE value. Relu was used as activation function. And Mean Squared Error (MSE) and Mean Absolute Error (MAE) were used for training and evaluation. The architecture of this model is shown in Figure 5.

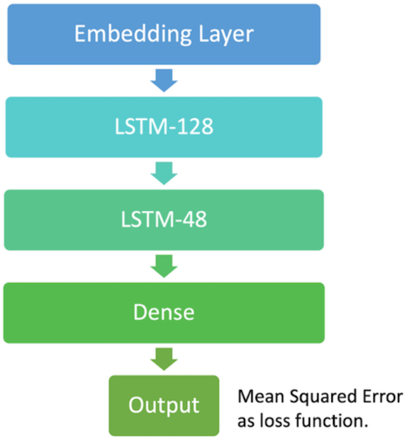


Figure 5: Model 2- LSTM model for regression task for predicting PCE value.

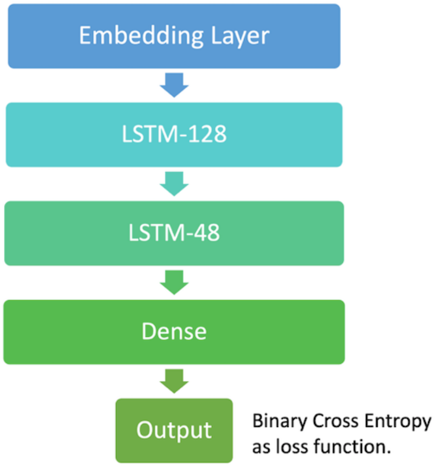


Figure 6: Model 3- LSTM model for classification task for predicting high or low efficiency OPVs.

The second task for in this study is classification. For classification the labels were generated based on the statistical analysis done during the exploratory data analysis. Based on the frequency of PCE values, the dataset was divided into two sections- 'High performance OPVs' and 'Low performance OPVs'. The threshold value was chosen as the 3.15, this divides the dataset into two classes with similar number of datapoints in each. After embedding, the SMILES data is fed to the LSTM model. This model has two LSTM layers with dropouts and one dense layer as the final layer. This dense layer has one node with sigmoid activation function for predicting the class label as 1 (high performance OPV) or 0 (low performance OPV). Binary cross Entropy loss was used for this model. Architecture of this model is shown in Figure 6.

Experiments and Results

For the regression task Model-1 was trained for 500 epochs. The MSE was 5.18 and MAE was 1.86. Training and Validation set MAE values during training is shown in figure 7. The same model was used with k-fold cross validation with k=5. This time the model trained for 300 epochs. Even with lesser number of epochs, the MSE and MAE values improved to 4.603 and 1.747 respectively. The training and validation set MAE values during training is shown in figure 8.

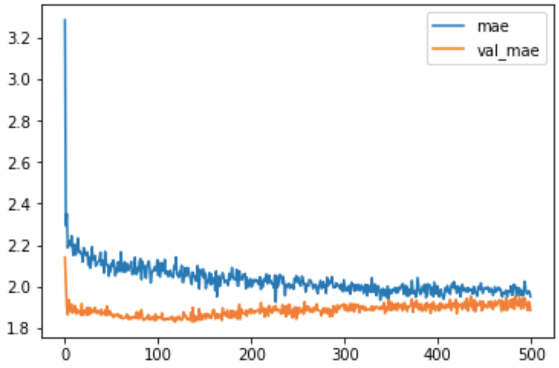


Figure 7: MAE without cross validation

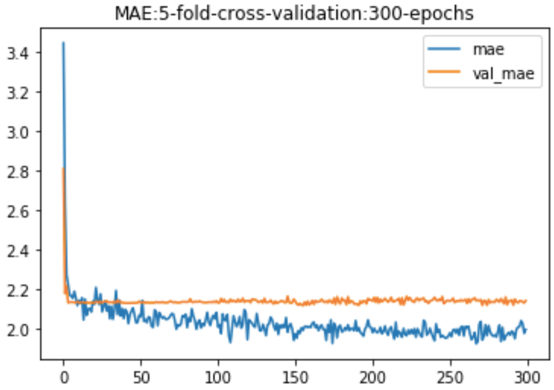


Figure 8: MAE with k=5 cross validation

Model-2 (LSTM based for prediction task), the model was trained for 250 epochs. The MSE and MAE values reduce to 3.78 and 1.50 respectively. R2 was calculated for this model, and its value was 0.2609. Table 1 represents the performances of Model-1 and Model-2 based experiments. Figure 9 shows the predicted vs ground truth PCE plot.

Model	**MAE**	**MSE**	**R2**
Model-1 (DNN)	1.86	5.18	-
Model-1 (with cross validation)	1.747	4.603	-
Model-2 (LSTM)	1.50	3.78	0.2609

Table 1: Results of experiments for prediction task.

For high performance OPV classification task Model-3 was used. The evaluation was done with accuracy. Two experiments with different number of epochs were tested. Results are shown in table 2. The training and validation set accuracy for the two experiments is shown in figures 10 and 11.

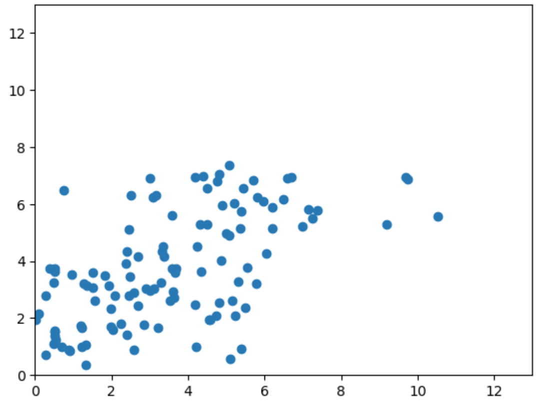


Figure 9: actual vs predicted PCE value by Model-2

Model	**Epochs**	**Accuracy**
Model-3 (LSTM)	100	67%
Model-3 (LSTM)	250	71%

Table 2: Model-3 results

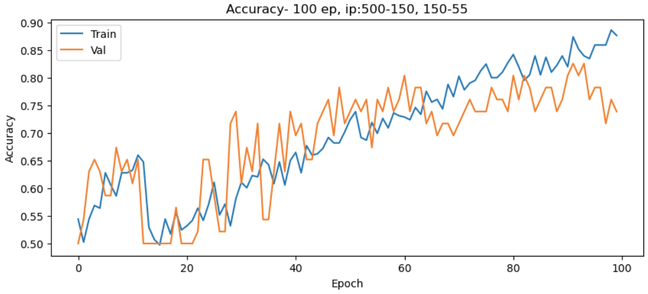


Figure 10: Accuracy of training and validation set plot for LSTM based Model-3 for 100 epochs.

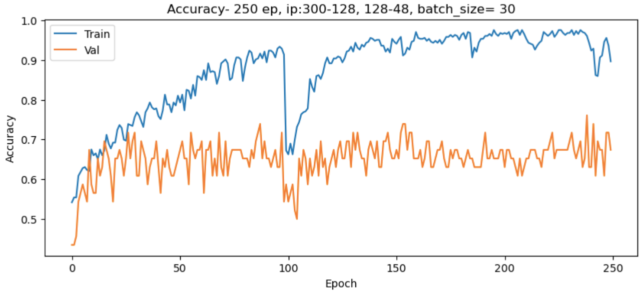


Figure 11: Accuracy of training and validation set plot for LSTM based Model-3 for 250 epochs.

Model	**Accuracy**
Model-3 (100 epochs)	67 %
Model-3 (250 epochs)	71 %
[3] BP	61.5 %
[3] DNN	52.5 %
[3] RF	67 %
[3] SVM	54 %
Best	71 %

Table -3: Results comparison for classification

Model	**MAE**	**MSE**	**R2**
Model-1	1.86	5.18	-
Model-1 (K-fold)	1.74	4.6	-
Model-2	1.50	3.78	0.26
[2] BiLSTM	1.48	3.48	0.29
[2] GNN	1.45	3.29	0.33
[2] AFP	1.67	4.41	0.12
[2] SVR	1.31	2.68	0.45
[2] RF	1.31	2.86	0.41

Table-4: Results comparison for regression.

Conclusion/Future work

The results of this study are comparable with results obtained in 2 and 3 for same tasks using SMILES representation. We compare our results for classification with [3], where the size of dataset is 1700 whereas our dataset size is 566. Even so, our model performs better than the best performing model for classification task of [3]. For prediction task, we compare our results with [2]. In [2] the authors use two different datasets, HOPV which has 343 usable datapoints and CEPDB, from 2.3 million datapoints only 25000 are used for their study. Whereas in case of CEPDB dataset, their results are far better. This suggests that we need to include more datapoints in order to train our model better. Our results are very close/comparable to results with some models used for HOPV dataset.

This study is an initial step towards designing high performance OPVs. Our future directions would be to gather more datapoints and actually design high performance OPVs using a generative model like Variational Auto Encoders.

References

1. J. Hachmann, R. Olivares-Amaya, S. Atahan-Evrenk, C. Amador-Bedolla, R. S. Sánchez-Carrera, A. Gold-Parker, L. Vogt, A. M. Brockway, and A. 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.
2. Eibeck, A., Nurkowski, D., Menon, A., Bai, J., Wu, J., Zhou, L., Mosbach, S., Akroyd, J., & Kraft, M. (2021). Predicting Power Conversion Efficiency of Organic Photovoltaics: Models and Data Analysis. ACS Omega, 6, 23764 - 23775.
3. Sun, W., Zheng, Y., Yang, K., Zhang, Q., Shah, A.A., Wu, Z., Sun, Y., Feng, L., Chen, D., Xiao, Z., Lu, S., Li, Y., & Sun, K. (2019). Machine learning–assisted molecular design and efficiency prediction for high-performance organic photovoltaic materials. Science Advances, 5.
4. Hirohara, M., Saito, Y., Koda, Y., Sato, K., & Sakakibara, Y. (2018). Convolutional neural network based on SMILES representation of compounds for detecting chemical motif. BMC Bioinformatics, 19.