A Data Science Approach for Prediction of Organic Photovoltaic Cell Material

Kailei Liu, Zonglun Lee, Sijin Luozhong, Chih-Wei Hsu
Department of Chemical Engineering, University of Washington, Seattle 98105, U.S.
Data Intensive Research Enabling Clean Technologies 2019

Molecular Engineering & Sciences Institute

PVC STRONG

Introduction & Background Information

Predicting Organic Photovoltaic Cell Material (OPVCM) is a python package that can predict Power Conversion Efficiency(pce) of an organic material in PV-Cell based on user's input data. The predicted model is built based on correlations between pce and molecular features (bond type, functional group, heteroatom and etc.).

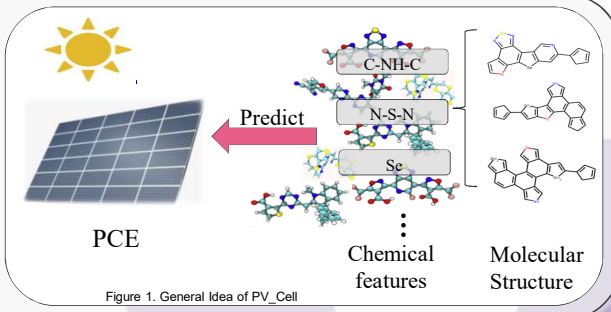
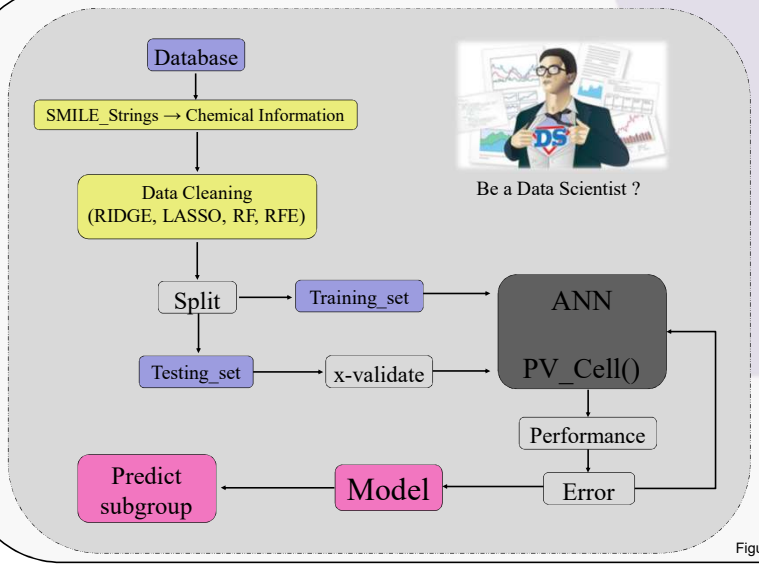


Figure 1. General Idea of PV_Cell

Dataset

All data is retrieved from The Harvard Clean Energy Project Database (HCEPDB). Sixty features (C-NH-C, N-S-N, Se……) are sorted to be used in this package.

Design



Be a Data Scientist ?

Our code apply several packages which are available online such as RDKit, Keras, and SciKit-Learn.

Moreover, we have also built custom functions to format data and wrap all components into a single call.

Figure 2. Flow Chart of PV_Cell code

Data Cleaning

We used the following packages to shrink the size of features and to analyze the weight of each feature:

- ✓ RIDGE (Ridge Regression)
- ✓ LASSO (Least Absolute Shrinkage and Selection Operator)
- ✓ Random Forest
- ✓ RFE (Recursive Feature Elimination)

We use MSE to compare and evaluate these methods for final decision making.

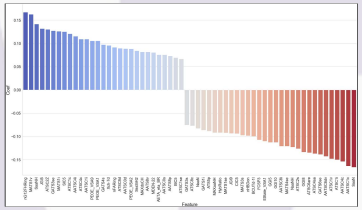


Figure 3. Selected Feature Coefficients in this Model

Training the ANN

Sklearn and Keras are used for training the ANN.

Parameter Grid	
Number of Hidden Layers	3
Number of Units in each Hidden Layer	30, 15, 8
Number of Epochs	200, 200, 200

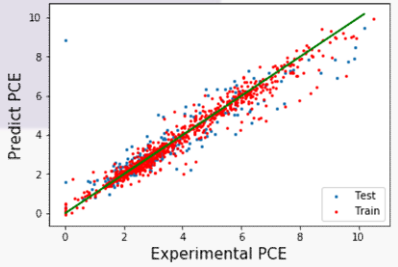


Figure 5. Parity Plot: predicted PCE vs actual PCE

Artificial Neural Network

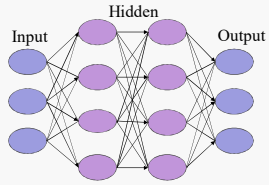


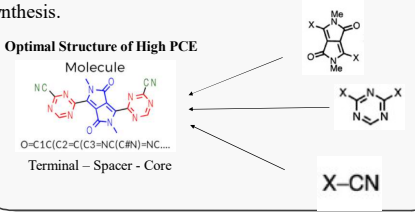
Figure 4. Graphical representation of an ANN with double hidden layers

The ANN system is a framework that several machine learning algorithms can work together and process complex data input. It takes a vector of input features and passes through layers and nodes which can transform input signals into outputs that predicted quantity. ANNs has been trained by adjusting the weights with every channels and nodes.

Our input is built with 60 features, which are sorted from Data Cleaning.

Future Work

Using this package, we can predict optimal structure of an organic compound simply from input of SMILES_str. A useful application is to screen the common patterns in organic compound (which can be categorized into Ter.-Spa.-Core) indicated by their PCE. This model provides researchers a practical approach to design their chemical synthesis.



References

1. Steven A. Lopez, Benjamin Sanchez-Lengeling, Julio de Goes Soares and Alán Aspuru-Guzik, "Design Principles and Top Non-Fullerene Acceptor Candidates for Organic Photovoltaics", 2017 Cell Press
2. Hiirotomo Moriwaki, Yu-Shi Tian, Norihito Kawashita and Tatsuya Takagi, "Mordred: a molecular descriptor calculator", 2018 Journal of Cheminformatics
3. Johannes Hachmann, Roberto Olivares-Amaya, Sule Atahan-Evrenk, Carlos Amador-Bedolla, Roel S. S´anchez-Carrera, Aryeh Gold-Parker, Leslie Vogt, Anna M. Brockway and Al´an Aspuru-Guzik, "The Harvard Clean Energy Project: Large-Scale Computational Screening and Designof Organic Photovoltaics on the World Community Grid", 2011 Harvard Library

DIRECT
Data Intensive Research Enabling Clean Technologies

Acknowledgements

- Special thanks to Professor David Beck and the excellent teaching assistant group in DIRECT program.
- We would also like to thank the Harvard Clean Energy Project for providing the dataset.

CEP – the Harvard Clean Energy Project

CHEMICAL ENGINEERING
UNIVERSITY of WASHINGTON