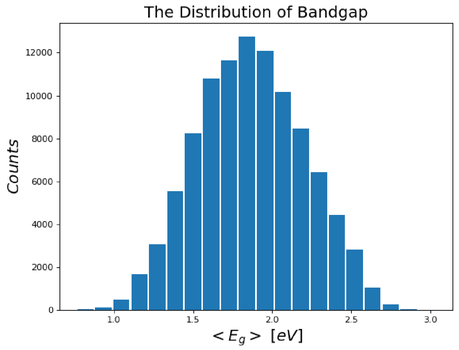
Prediction of Bandgap for Organic Photovoltaic Materials by Machine Learning Method

Yuhuan Meng*, Liang Xu**, Zhi Peng**, Hongbo Qiao**.

* Molecular Engineering and Sciences Institute ** Department of Materials Science and Engineering

Motivation

Bandgap is an important parameter for all the photovoltaic materials. Efficient and accurate prediction of bandgap helps us to screen semiconductors for different application and explore new photovoltaic materials.



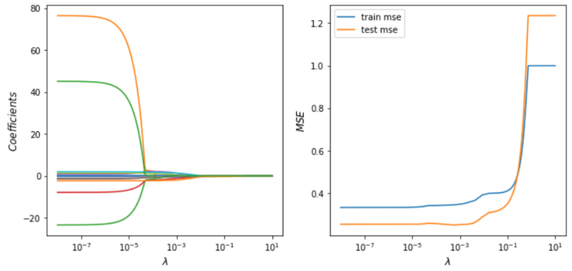
Bandgap is an intrinsic property, and related to the molecular structure of organic semiconductors.

Goals:
Build a reliable model to predict the bandgap of organic semiconductors

1. Data Mining and Cleaning
   - Calculation of molecular descriptors
   - Determining the predictors
2. Machine Learning Models
   - Screening models
   - Optimization models
3. A Wrap Function
   - Input SMILES string
   - Output predicted bandgap

Data Mining and Cleaning

- SMILES String in RDkit
- Descriptors Calculation by Mordred
- Initial Screen: Pearson Correlation
- LASSO Regression
- Final Five Predictors



Start: 1406 descriptors were calculated

Pearson correlation>0.65(13 descriptors)

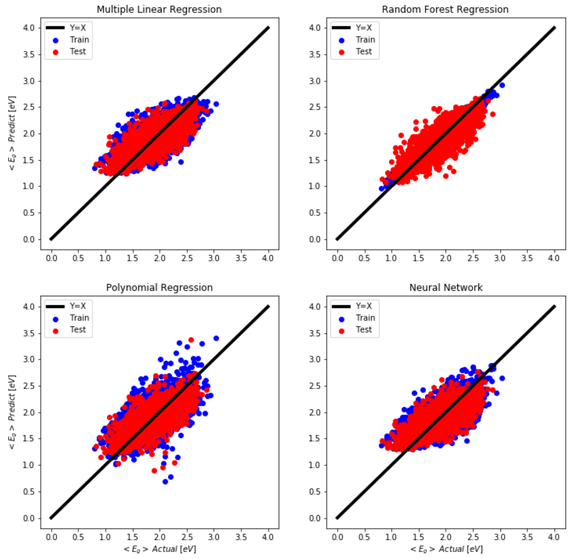
Lasso regression: alpha = 10^{-5}, absolute value of coefficients > 1(5 descriptors)

10k SMILES strings were used to determine important features. After two double screening, original features Descriptors Calculation by Mordred were reduced to 5 features, with mean square error about 0.2.

Machine Learning Models

Model Selection:

- Multiple Linear Regression
- Polynomial Regression
- Random Forest Regression
- Neural Network Model



Error	Multiple Linear	Random Forest	Polynomial	Neural Network
MSE	0.045093	0.035795	0.050361	0.173221
MAE	0.165998	0.142503	0.172879	0.334963
MAPE	0.092812	0.079263	0.095950	0.188373
R^2	0.593371	0.677222	0.545867	0.566512
Kfold	0.590672	0.681907	0.590672	-0.062011

According the error calculation, random forest algorithm was found to the best algorithm to build prediction model due the lowest error value and highest validation score.

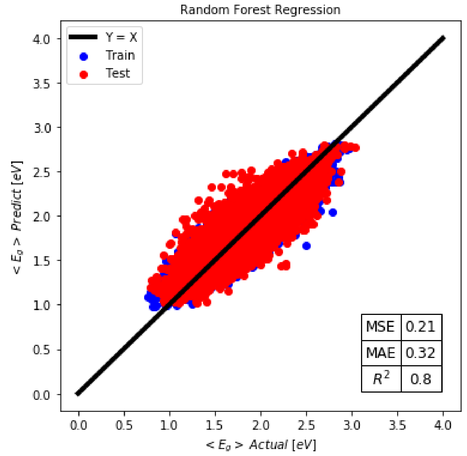
Model Optimization:

- Hyperparameter
  max_depth": [3, None]
  min_samples_split: [2,30]
  bootstrap : [True, False]
- Result: min_samples_split: 15
  other : default

DIRECT
Data Intensive Research
Enabling Clean Technologies

Results

As a Result, we use 100k dataset to build final prediction model based on random forest regression with good performances and was build with good performance and excellent statistic data.



Conclusions

1. The bandgap of organic semiconductors can be predicted by machine learning efficiently and accurately.
2. Random forest regressor gives quite low test set mean error 0.32eV
3. The models need to be further optimized.

CLEAN ENERGY INSTITUTE
UNIVERSITY of WASHINGTON