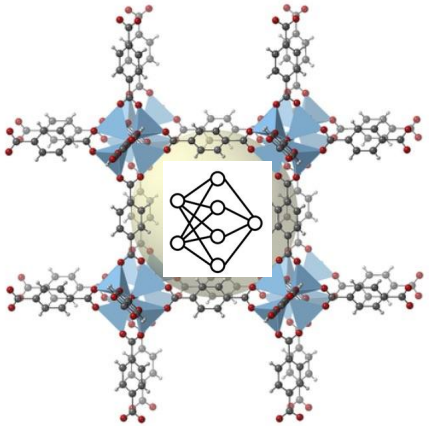
Predicting Band Gaps of Next-generation Battery Materials *via* Machine Learning

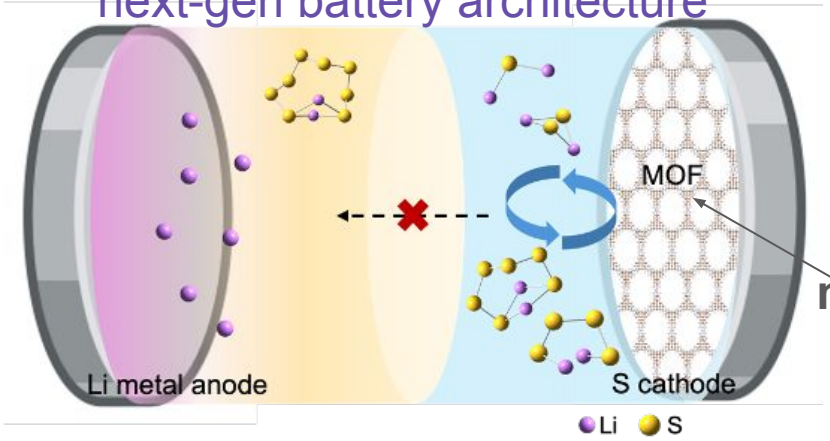
Avinash Karamchandani, Dorisa Tabaku, Qinying Chen,
Simran Kaur, Sadisha Nanayakkara

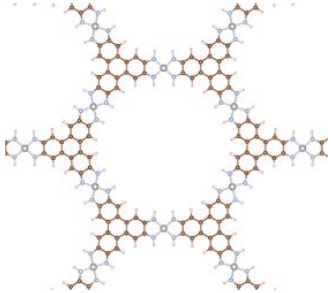


Spring 2025 Data Science Boot Camp
Erdös Institute

Our material of interest and predicting the properties

next-gen battery architecture





metal organic frameworks (MOFs) are important components

conductive

band gap (**property**)

state-of-the-art computations

DFT calculations

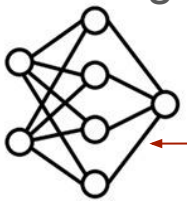
(computationally expensive: ~1 week / MOF)

cost
scalability
resources

$\hat{H}\Psi = E\Psi$

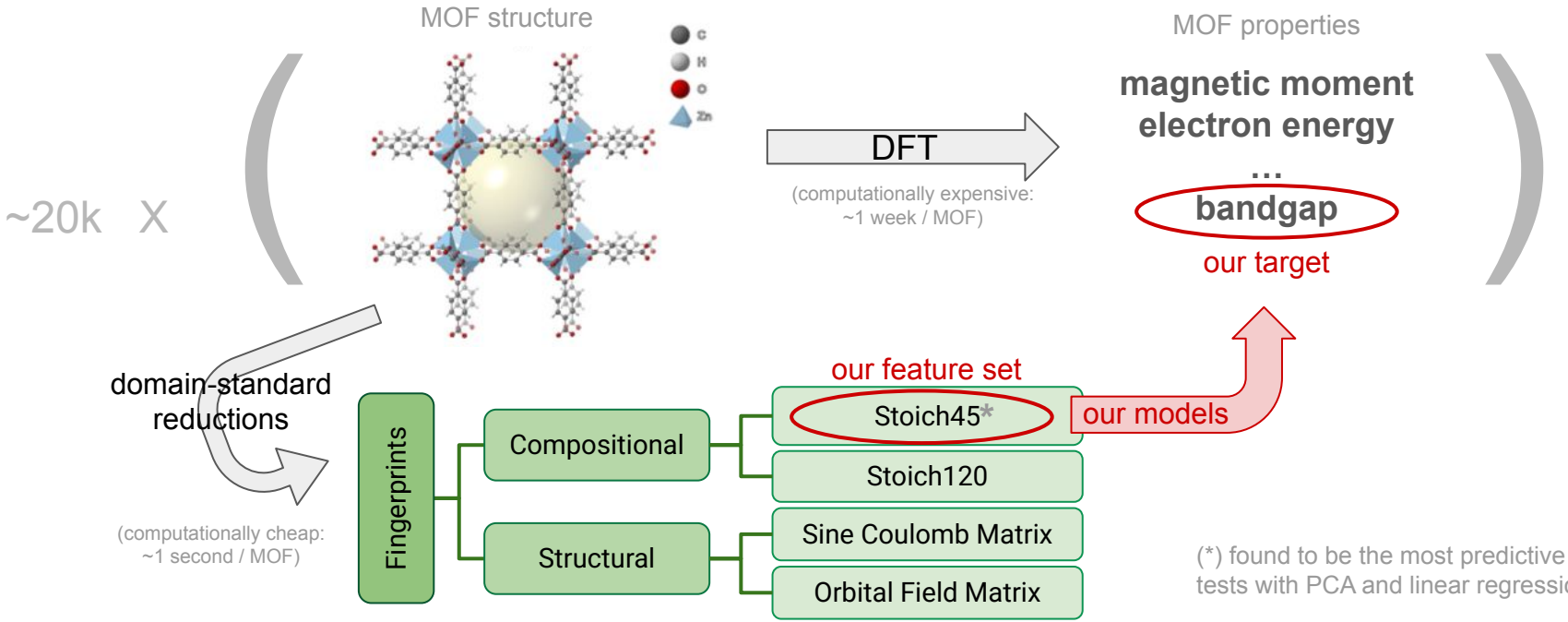
(computationally cheap: ~1 second / MOF)

Machine Learning



Our Dataset: the QMOF database

We use the QMOF database (Rosen et al. (2021)): ~20k theoretical MOF structures and their DFT-derived properties.



(*) found to be the most predictive from early tests with PCA and linear regression.

Features in the Stoich45 ‘fingerprint’

the 5 * 9 = 45 stoichiometric quantities of the form:

The { **mean**
geometric mean
standard deviation
minimum
maximum } of the { **atomic numbers**
group numbers
period numbers
electronegativities
electron affinities
melting points
boiling points
densities
ionization energies }

of the atomic elements that make up the MOF.

(The mean, geometric mean, and standard deviation are weighted by number of atoms of each type in the MOF unit cell.)

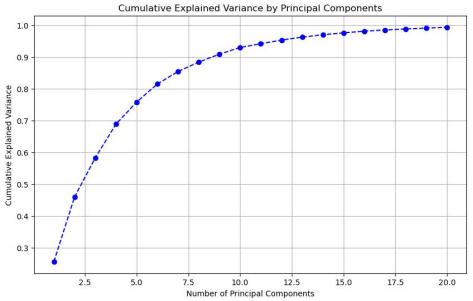
Note: these are derived only from the chemical formula of the MOF (e.g. $Zn_8C_{48}H_{26}O_{26}$) and known elemental properties that appear e.g. on the periodic table.

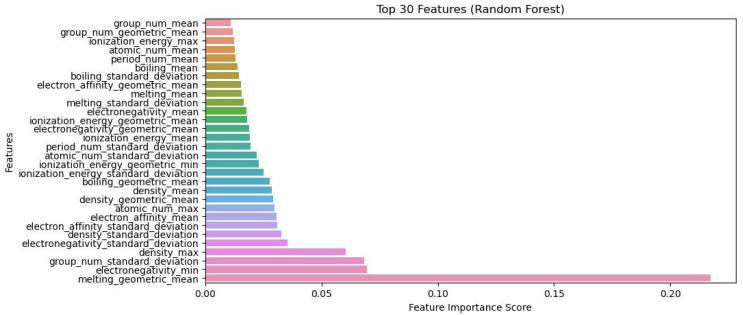
Feature Engineering

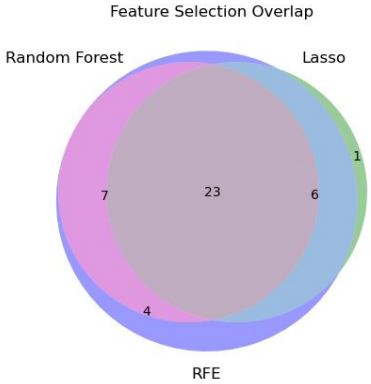
Total no. of features = 45

1. Principal Component Analysis: Top 8 PCs
2. Other approaches:
   a. Random Forest - Recursive Feature Elimination: Top 40
   b. Random Forest Feature Importance: Top 30
   c. Lasso (L1 regularization) for feature selection: Top 30

- Overlapping features = 23







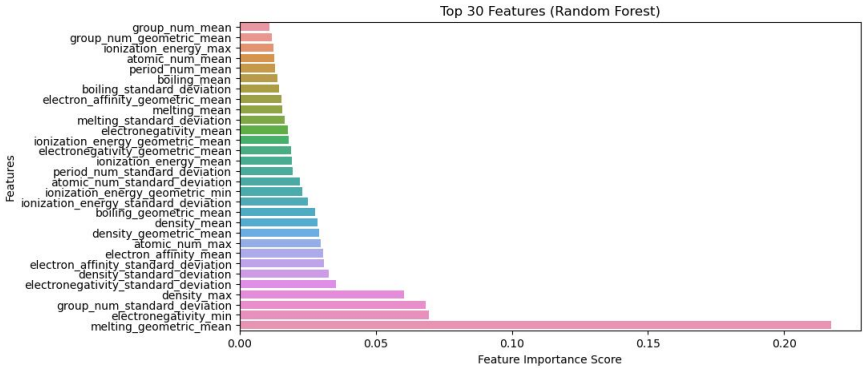
Choice of final feature set out of PCA and intersection dataset

Factors to consider:

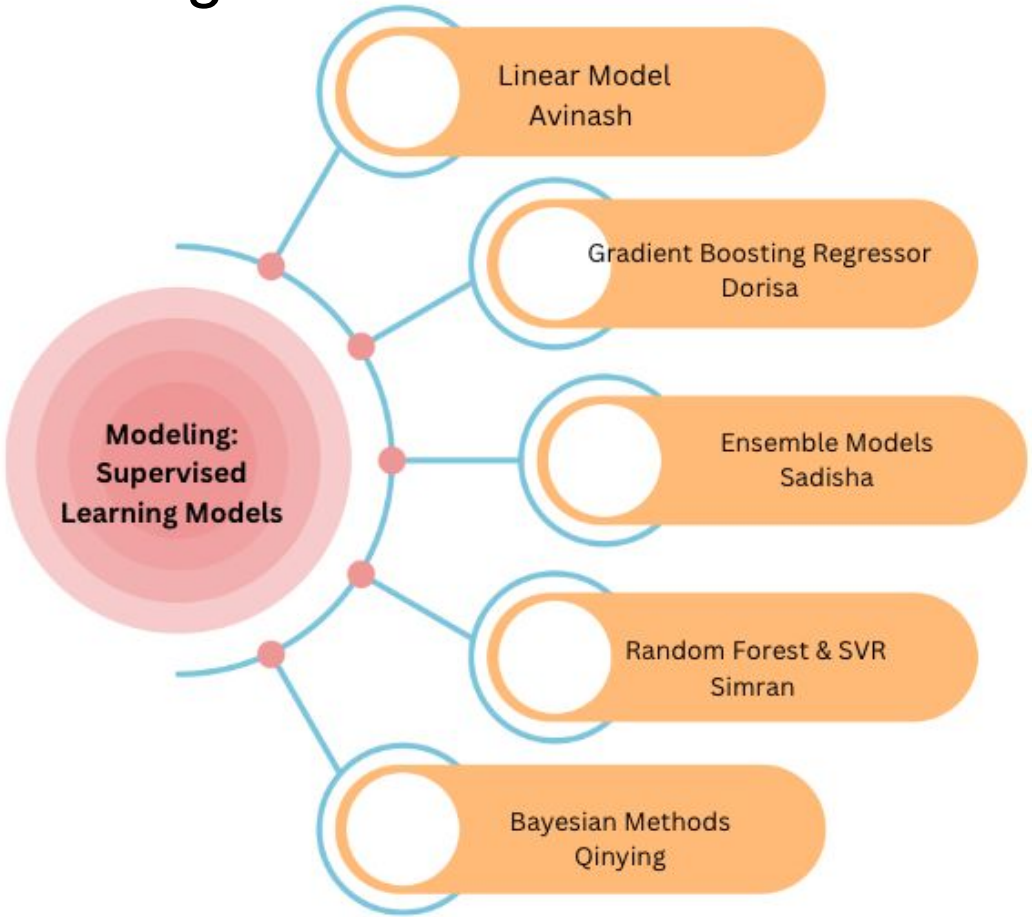
- **Predictive Performance:** Regression accuracy
- **Computational Efficiency:** Training time and complexity.
- **Interpretability**

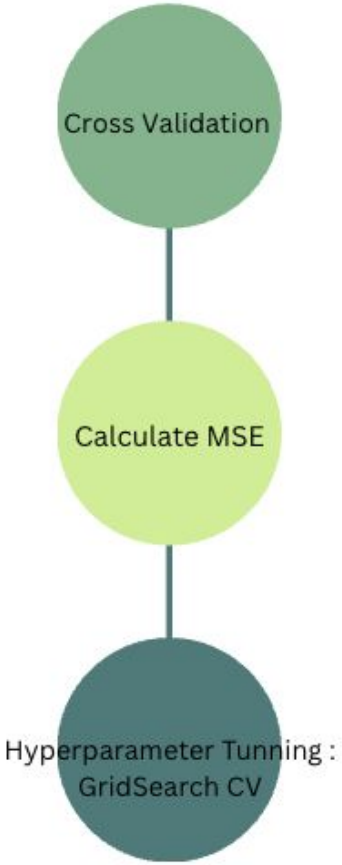
Final Dataset: 23 **Intersection Features**

	PCA Features	**Intersection Features**
Model Performance (Random Forest)	RMSE: 0.7684 R^2: 0.4930	RMSE: 0.7063 R^2: 0.5716
Training Time	5.50 seconds	12.41 seconds

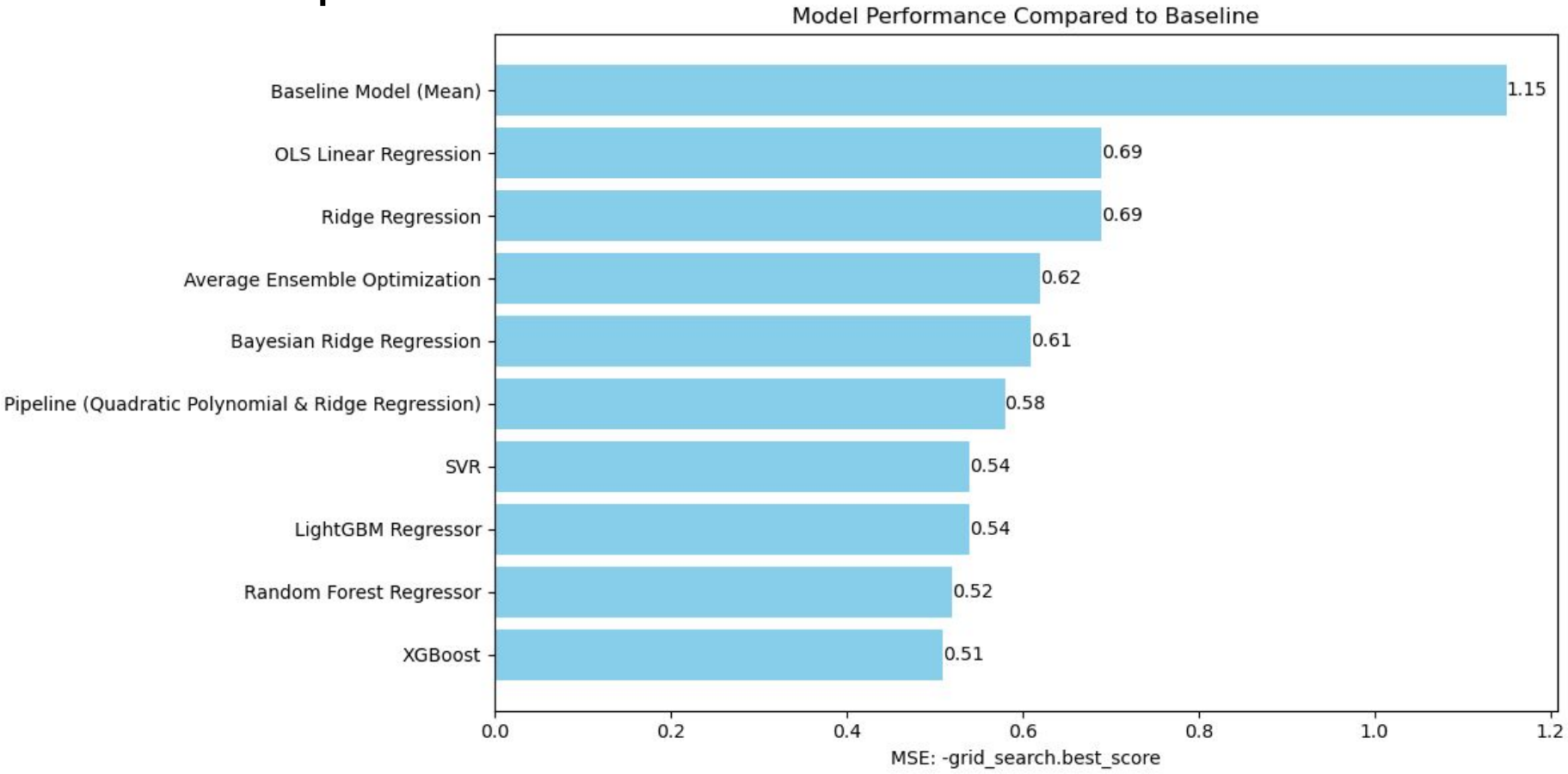


Modeling


Modeling: Supervised Learning Models
Linear Model
Avinash
Gradient Boosting Regressor
Dorisa
Ensemble Models
Sadisha
Random Forest & SVR
Simran
Bayesian Methods
Qinying


Cross Validation
Calculate MSE
Hyperparameter Tunning :
GridSearch CV

Model comparison



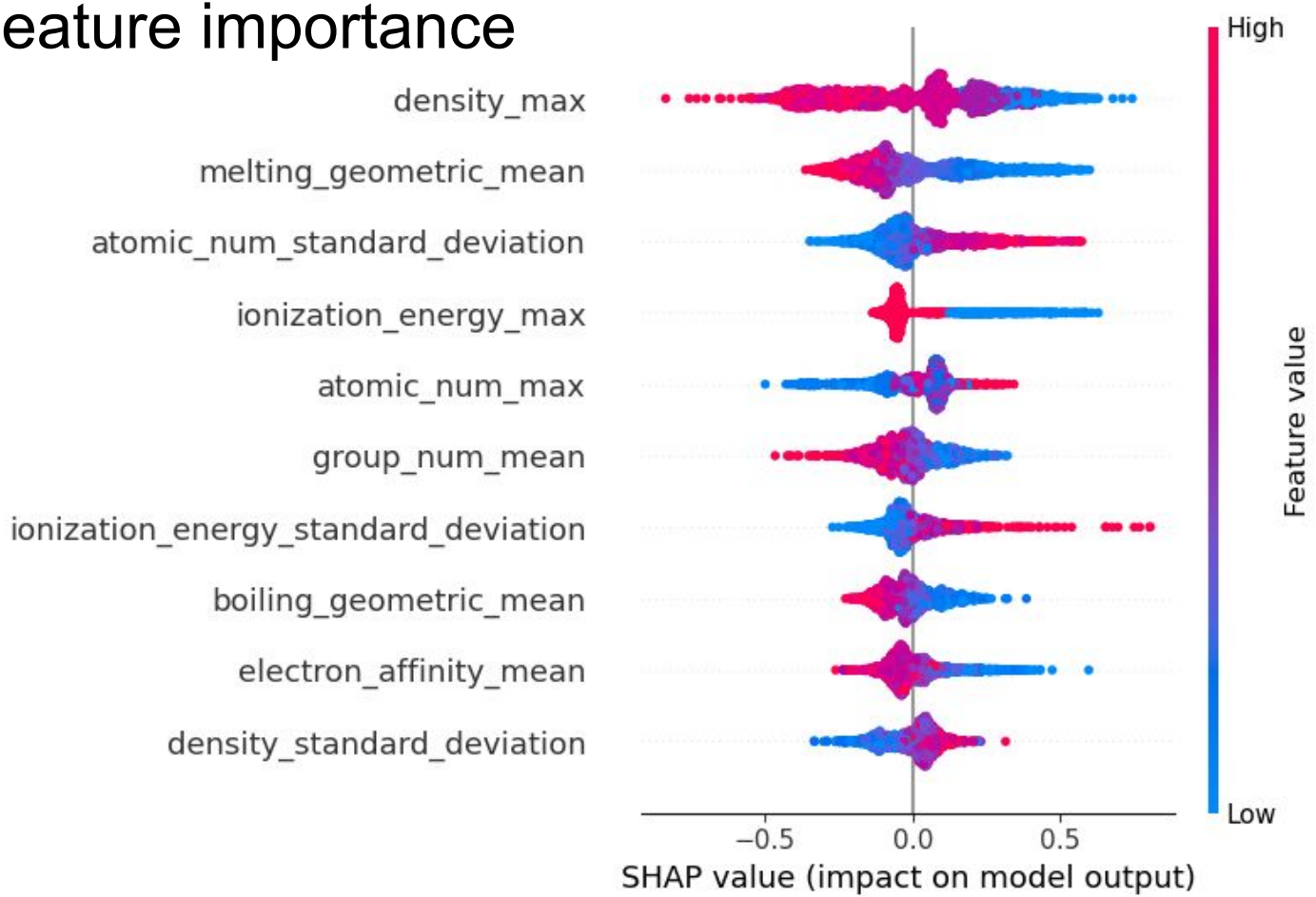
Best model

XGBoost Prediction Performance

Metric	Value
MSE	0.51
MAE	0.53

Baseline Mean Model MSE = 1.13, MAE = 0.8 (n = 2097)
~55% improvement from the baseline

Feature importance



Conclusions

Generated feature set for training via feature engineering

Successfully trained a machine learning model to predict band gap

Room for improvement

Future work

- Use more sophisticated feature sets (*i. e.*, structure-sensitive)
- Implement neural networks and deep learning

Acknowledgements

- Roman Holowinsky, Steven Gubkin, and Alec Clott – Erdös Institute
- Viraj Meruliya for his guidance and support as our project mentor
- QMOF Database (made available on the Materials Project) and its original authors