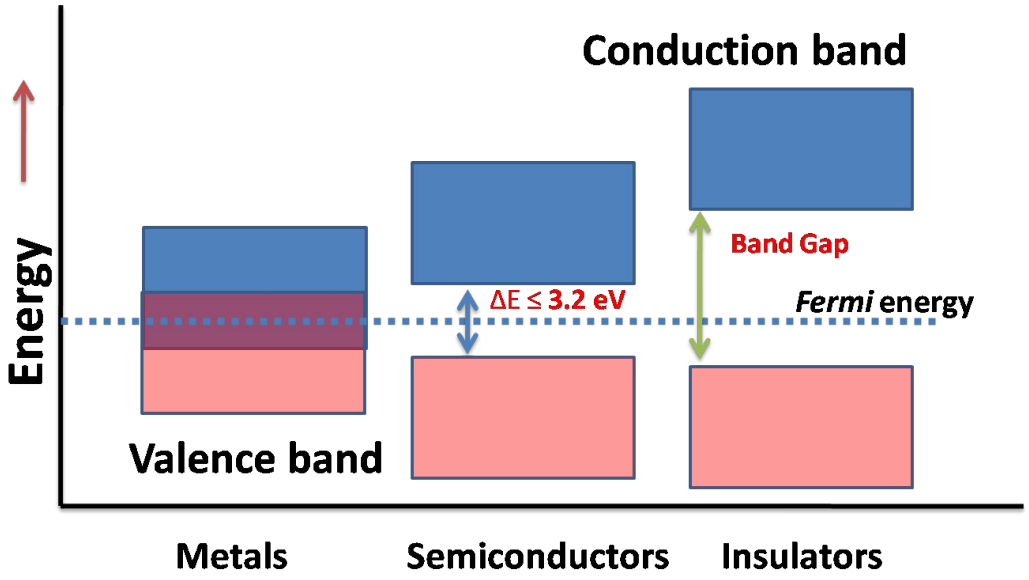
# Predicting the Band Gap of Metal Oxides - a ML Study

Coenradt Taylor
ML4MSD
12/01/25

# Model Overview

**Target Property: Band Gap**

- Fundamental electronic property that determines a material's conductivity and optical behavior
- Dependent on composition and structure
- Accurate prediction enables accelerated materials discovery – guides experimental efforts



# Model Overview

**Materials Class: Metal Oxides**

- Inorganic compounds formed by metal cations bonded with oxygen anions
- Widely used in magnetics, energy applications, catalysis
- Typically exhibit band gaps in the semiconductor to insulator regime
- Predicting their band gaps is essential for designing new functional materials

# Model Overview

**Data Source: Open Quantum Materials Database (OQMD)**

- Repository of 1,317,811 real and hypothetical materials
- Data from experimental sources and high-throughput DFT computational workflows
- Can be accessed via an API, enabling automated querying

# Background

**Goal**

- Replace expensive DFT/experiments with fast ML models.

**Common Inputs**

- Composition-based features (Magpie, ElemNet), structure-based descriptors (CIF-derived, SOAP, Ewald).

**Typical Models**

- Random Forest, SVR, Gradient Boosting, Kernel methods, Graph Neural Networks, etc.

**Databases**

- DFT-computed: Materials Project, AFLOW, Jarvis-DFT, C2DB, and OQMD

**Key Limitations**

- DFT underestimates gaps, skewed datasets, missing structures, inconsistent labels.



# Background

**Composition featurized models**

- Better when structures are noisy or inconsistent in databases

**Structure featurized models**

- Outperform only when high-quality structures are available.

**Typical MAEs**

- ~0.3–0.5 eV for oxides; higher for small datasets.

**Challenges**

- DFT label noise, structure sensitivity, imbalanced data

**Opportunities**

- Larger curated datasets, GNNs, better labels, uncertainty quantification

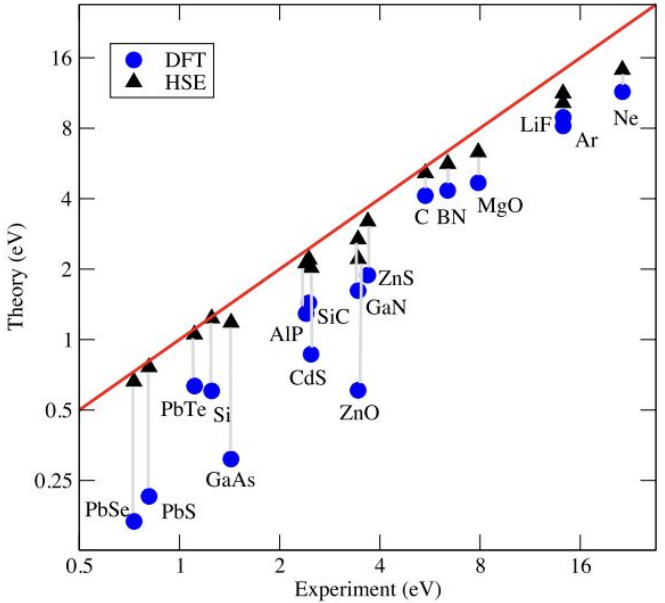



**Figure 1.** Comparison of experimental and theoretical bandgaps of semiconductors and insulators, calculated using DFT (PBE) and hybrid (HSE03) functionals. After Paier *et al* [19, 20].

Hafner, Jürgen. "Ab Initio Density-Functional Calculations in Materials Science: From Quasicrystals over Microporous Catalysts to Spintronics." Journal of Physics: Condensed Matter, vol. 22, no. 38, Sept. 2010, p. 384205. DOI.org (Crossref), https://doi.org/10.1088/0953-8984/22/38/384205.

# Model Parameters

**Querying/filtering**

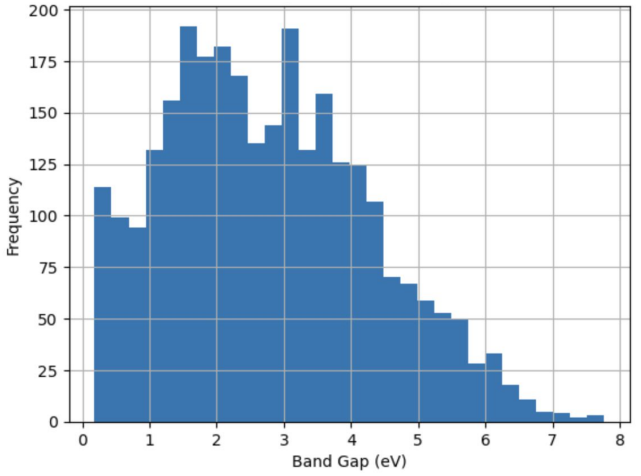
10,000 oxygen-containing compounds

⬇

3601 Me+O with band gap>0

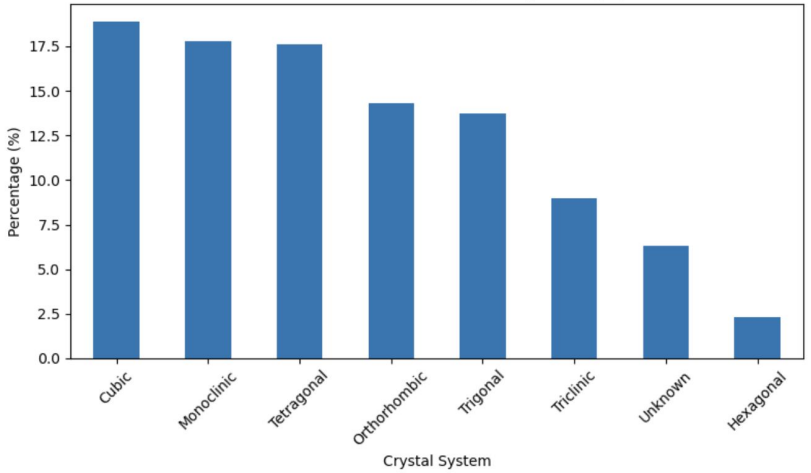
⬇

2835 after NaN+duplicate cleaning




**Final Dataset**

- All unique compositions
- 141 unique space groups
  - Cubic-structured *Pm-3m* most abundant with 378
- Average band gap of 2.78 eV



# Model Parameters

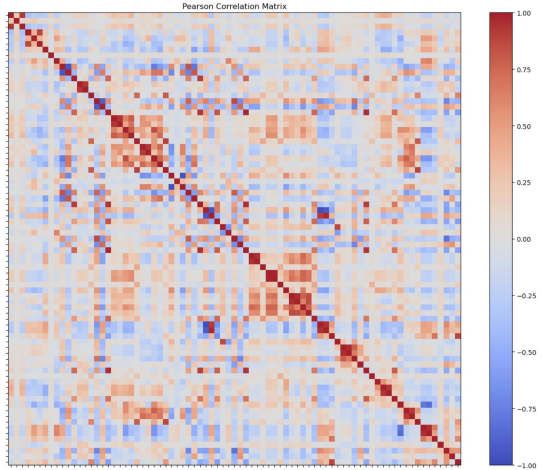
## Model Training

- Used Pymatgen to convert unit cell and atomic site location data into usable ‘structure’ column
- 70/20/10 randomized train/val/test split
- Random Forest regressor model

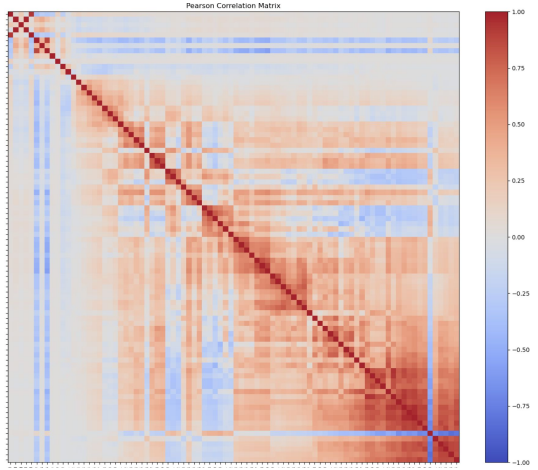
## Two featurizers

- Matminer – Element Property (*compositional*)
  - 79 features after cleaning
- DScribe – Ewald Sum Matrix (*structural*)
  - 86 features after cleaning
- After RFE, only ~20 features needed for each

Element Property




Ewald Sum Matrix



# Hyperparameter Optimization

**Parameter selection**

- Random Forest model: ~15 possible hyperparameters
- 5 selected hyperparameters
  - *n_estimators* – number of trees
  - *max_depth* – maximum depth of each tree
  - *max_features* – number of features considered at each split
  - *min_samples_split* – minimum samples to attempt a split
  - *min_samples_leaf* – minimum samples required in a leaf
- 243 permutations tested for each featurizer

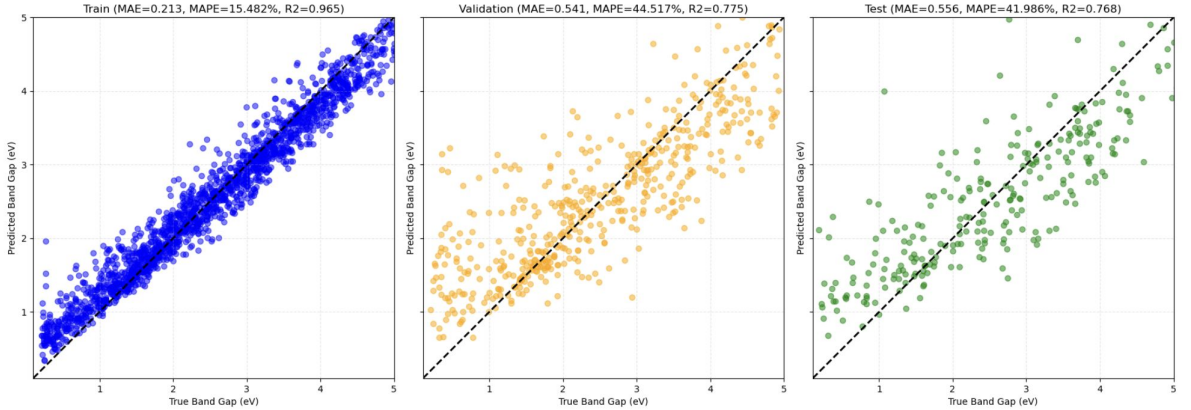
| | Tested | Element Property model | Ewald Sum Matrix model |
|---|---|---|---|
| n_estimators | [100, 300, 500] | 500* | 500* |
| max_depth | [None, 10, 20] | 20* | 20* |
| max_features | ["sqrt", 0.5, None] | 0.5 | None* |
| min_samples _split | [2, 5, 10] | 2* | 2* |
| min_samples _leaf | [1, 2, 4] | 1* | 1* |
| **train_MAE** | | **0.209** | **0.258** |
| **validation_MAE** | | **0.542** | **0.731** |

* = Either min/max value tested - risk of overfitting

# Results

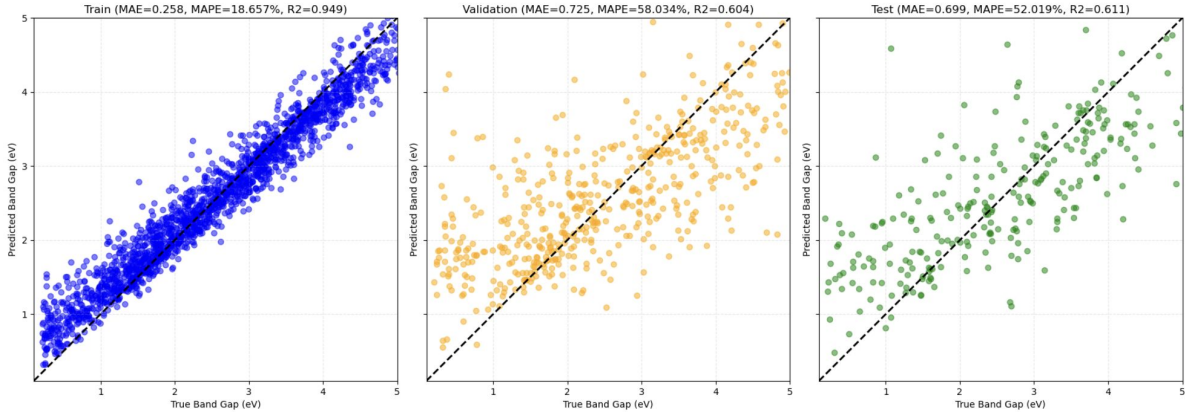
## Composition featurized

| | Element Property model | | |
|---|---|---|---|
| | MAE | MAPE | R2 |
| Train | 0.213 | 15.48% | 0.965 |
| Validation | 0.541 | 44.51% | 0.775 |
| Test | 0.556 | 41.98% | 0.768 |




## Structure featurized

| | Ewald Sum Matrix model | | |
|---|---|---|---|
| | MAE | MAPE | R2 |
| Train | 0.258 | 18.65% | 0.949 |
| Validation | 0.725 | 58.03% | 0.604 |
| Test | 0.699 | 52.01% | 0.611 |




- Dummy model MAE=1.304

# Discussion

**Performance**

- Relatively poor: >0.5 MAE, >40% MAPE on both models
- Compositional outperformed structural model
  - Compositional can generally perform better w/ limited datasets
  - Less sensitive to noise, overfitting

**Challenges**

- Dataset too small: long query times (> 1 hr for 10,000 samples)
- Suspect filtering: likely not all samples in final dataset were metal oxides
  - OQMD also contains hypothetical compounds
- Different data splitting could have improved performance
  - However, dataset statistics do not show better option
- More advanced models than RF needed
  - Gradient boost, neural networks, support vector regression, etc.

# Thank you!