Predicting Band Gap in 2D Materials using OQMD

Two-dimensional materials are an important class of materials since their tunable electronic properties, such as band gap, constitute a backbone for applications in nanoelectronics and optoelectronics. To study these properties efficiently, this project develops a dataset of 2D-like compounds, based on transition-metal chalcogenides, filtered from the OQMD database and featurized by simple composition and structural descriptors. A Random Forest regression model, trained on this dataset, has a moderate predictive accuracy (MAE ≈ 0.43 eV), testifying that simple, lightweight descriptors can capture meaningful trends in the band gaps; however, it also suggests that larger datasets or richer features are needed for improved performance.

1. Literature Overview

Research in applying machine learning to predict the band gaps of inorganic and two-dimensional (2D) materials has been surging in recent years. A number of studies illustrate both the promise of machine learning for high-throughput screening and the trade-offs involved in using different types of descriptors. The three papers below offer complementary perspectives that helped inform the design of this project.

A study by Vishwesh Venkatraman in 2021 evaluated the performance of composition-based machine learning models on a large dataset of inorganic compounds. It showed that using only elemental property descriptors such as mass or electronegativity allows for reasonably accurate bandgap predictions across diverse chemistries. However, the study also emphasized some key limitations. Composition only models cannot distinguish structural polymorphs, subtle coordination changes, or packing differences, all of which can strongly influence electronic properties such as band gap [1]. This shows a tradeoff commonly seen in machine learning. Composition models are easy to compute and scale well but may miss critical structure-driven variability. This motivated me to consider additional structural descriptors complementary to composition.

Focusing specifically on 2D and layered materials, Minh Tuan Dau et al. showed that in the case of electronic structure properties of 2D materials, descriptor engineering, or using a combination of both chemistry-based features and structure-derived descriptors, substantially outperforms the prediction capability. Their method showed that purely composition-based approaches considerably suffer with the geometries, anisotropy, and interlayer interactions characteristic of 2D systems [2]. By incorporating structural descriptors into the machine learning model alongside composition, a larger variance fraction in band gap and electronic properties was captured, showing that for 2D materials, geometry plays a role. Given this, I chose to combine composition and structural featurizers rather than adopt a purely composition-based approach.

Finally, in the context of purely 2D materials, Yu Zhang et al. conducted a study using machine learning to predict the band gaps of known 2D materials. Training models on data derived from 2D structures, they could demonstrate that machine learning achieves reasonable predictive performance for such systems, reinforcing the concept that band-gap prediction for a sufficiently homogeneous class of materials is doable [3]. The work demonstrates that machine learning can be an efficient screening tool in the discovery of 2D materials, especially when supported by domain-specific filtering and feature choices.

Taken together, these works show a logical evolution and provide guidance for the design choices in this project. Their results suggest that my approach of applying composition and structural featurizers to a chemically filtered 2D subset balances interpretability, computational efficiency, and relevance to 2D material physics, while considering the limitations established in previous works.

2. Final Dataset

The dataset used in this study was collected from the Open Quantum Materials Database API, or OQMD REST API, which provides DFT calculations on thousands of crystalline and quantum materials. I downloaded 8,000 entries from the formation-energy endpoint, as the example method provided by OQMD used the same endpoint, and further restricted the set to only those containing all information necessary for feature extraction: specifically, a calculated band gap, a full unit-cell description, and an atomic site list. Standard data-cleaning practices removed rows where any of these fields were missing. Each remaining valid entry was then reconstructed into a pymatgen structure object using the reported lattice vectors and atomic positions. In OQMD, there is no clear way to sort for 2D materials. This is why a chemically motivated filter was applied to isolate materials likely to exhibit 2D-like behavior. In this model, a material was considered '2D' if its composition contained at least one transition metal and one chalcogen atom. This filter captures a broad family of materials related to transition-metal chalcogenides, which frequently form layered or weakly bonded structures. No geometric c-axis threshold was used, since this proved too restrictive and eliminated many qualifying materials. To clean by c-axis ratios to b and c, it shrunk the dataset to 6 points. These structures in this dataset represent a chemically rich and structurally varied set of 2D-like materials appropriate for my model and machine learning.

3. Dataset Statistics

The final dataset consists of a chemically diverse set of chalcogenide-rich structures from DFT calculations within the OQMD. All entries possessed valid crystalline structures with complete atomic coordinates and lattice parameters. Bandgap values range from roughly 0 to 4 eV, but most lie between 0 and 2 eV. The compositions consist of several hundred unique formulae consisting of transition metals, chalcogens, and an occasional additional species. The dataset was split into an approximate 70, 20, and 10 percent split for training, validation, and testing. This yielded 266 training samples, 76 validation samples, and 39 test samples. These divisions were verified to possess consistent bandgap distributions, ensuring that the model was evaluated on data representative of the full set.

4. Model Training

Two featurizers were used in the training of the model: composition and structure based. The composition featurizer utilized the pymatgen Composition object to calculate descriptors such as the average atomic number, the average atomic mass, and the number of distinct elements in the formula. These features reflect basic chemical characteristics of the material, such as electronic density, atomic complexity, and periodic trends. The structural featurizer extracts geometric descriptors from the pymatgen Structure object. This gave lattice structure data such as a, b, c, the unit cell volume, density, c/a value, and b/a value. The c/a and b/a values are important for anisotropic or layered systems. Altogether, the model has two featurizers with ten independent features to interpret each structure.

A RandomForestRegressor was chosen to model the relationship between the features and band gaps. Random forests are well adapted to materials datasets since they handle nonlinear relationships, have no requirement for scaling of features, and are resistant to overfitting when appropriately tuned. Additionally, RandomForestRegressors were used several times throughout the course and especially in the ML Pipeline building lectures. Several values of the number of decision trees were tested, and the best performance found on the validation set was obtained with 200 trees. The model was trained on 266 training samples, and hyperparameters were selected based on validation MAE. Band-gap values were used directly as the regression target without normalization since tree-based models are invariant to scaling.

5. Results

Summary table of model performance metrics (MAE, R2, MAPE on training and test sets). Don't forget to add a random baseline model for comparison.

Table 1: Performance Results

Performance Results			
Set	**MAE (eV)**	**R^2**	**MAPE**
Training	0.1637	0.9012	19.41%
Validation	0.5511	-0.0707	104.59%
Test	0.4342	0.2862	77.50%
Baseline	0.599	-	-

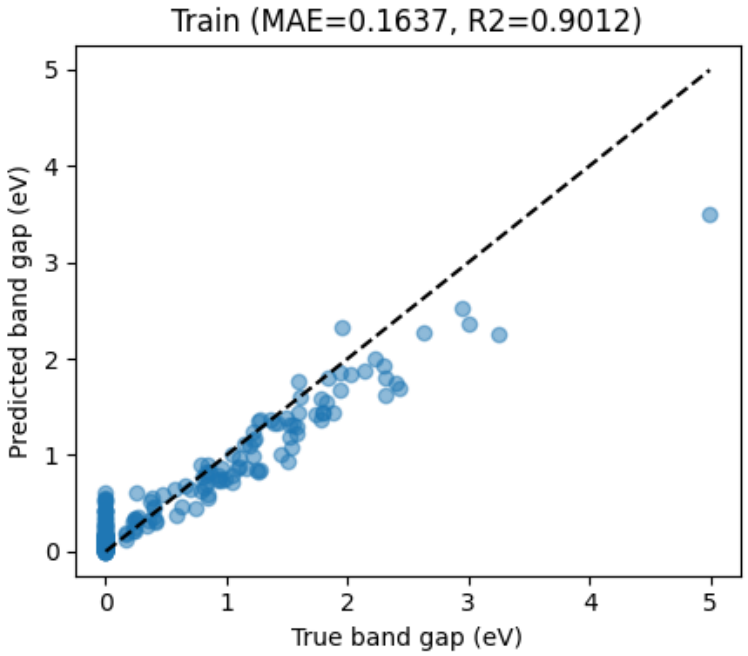

Figure 1: Training Set Results

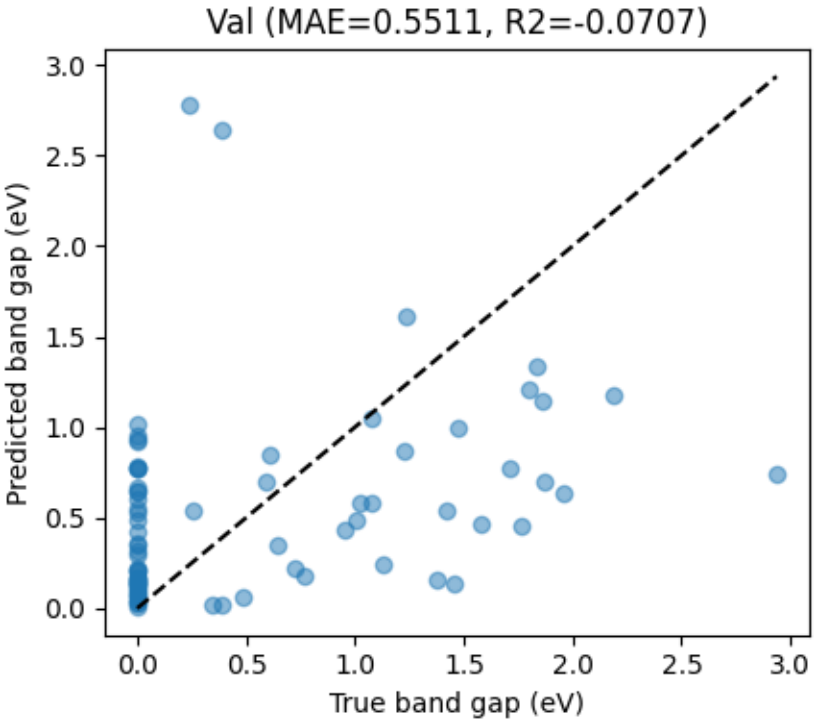

Figure 2: Validation Set Results

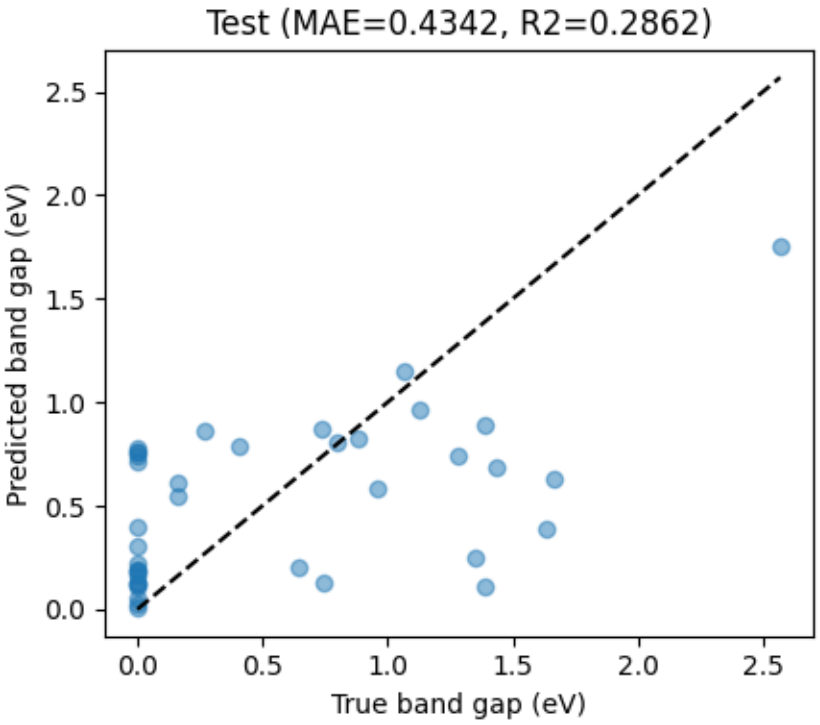

Figure 3: Test Set Results

6. Discussions

The performance results show the positives and negatives of using interpretable features for determining certain materials properties. The features used only grasp a small number of aspects for a material's unique electronic and bonding properties. By using only composition and geometry, the model does not have the ability to differentiate between small, subtle variations that have large effects on properties such as band gaps. For example, the model cannot differentiate between coordination (octahedral versus trigonal prismatic, etc.), spacing between layers, or coupling. With this, however, the model performed better than the baseline prediction as the information that was fetched from OQMD, cleaned, and featurized was sufficient to describe the data and train the model.

While validation and test values are modest, they are within the typical range for small material data sets and band-gap prediction tasks (0.4-0.6 eV) [1]. The performance variation during validation likely is a reflection of the small number of samples in that subset and intrinsic noise in properties calculated by DFT. More importantly, the training results and test MAE show that the model generalizes better than random guessing or mean prediction.

Some avenues for further improvement in performance would be incorporating more chemically relevant descriptors such as electronegativity differences, valence-electron counts, or orbital level features that provide a richer representation of bonding and electronic structure. Also, extending this data set by relaxing the chemical filter or by including chalcogenide-related materials such as oxides or halides would increase the statistical strength of the model. More advanced models, including graph neural networks such as CGCNN or MEGNet, would likely capture more subtle structural relationships. Nonetheless, the objective of this project was to construct a full, interpretable machine-learning pipeline using a curated materials subset, and these results successfully demonstrate the feasibility and value of this approach.

7. References

[1] Venkatraman, Vishwesh. "The utility of composition-based machine learning models for band Gap Prediction." Computational Materials Science, vol. 197, Sept. 2021, p. 110637, https://doi.org/10.1016/j.commatsci.2021.110637.

[2] Dau, Minh Tuan, et al. "Descriptor engineering in machine learning regression of electronic structure properties for 2D materials." Scientific Reports, vol. 13, no. 1, 3 Apr. 2023, https://doi.org/10.1038/s41598-023-31928-7.

[3] Zhang, Yu, et al. "Bandgap prediction of two-dimensional materials using machine learning." PLOS ONE, vol. 16, no. 8, 13 Aug. 2021, https://doi.org/10.1371/journal.pone.0255637.