Coenradt Taylor

Predicting the Band Gap of Metal Oxides - a Machine Learning Study

1. Abstract

Machine learning (ML) is utilized as a powerful tool in predicting the band gap energies of metal oxides, a property critical for applications in electronics, integrated circuits, and catalysis. A supervised learning model was trained on density-functional-theory (DFT) computed band gaps from the Open Quantum Materials Database (OQMD). From this database, 2835 metal oxide materials were extracted with non-zero band gaps for use in model training. Using both composition and structure featurizers, the tested models achieved a mean-absolute percentage error value of 41.98% and 52.01%, respectively.

2. Literature Review:

ML has proven to be a viable alternative to first-principles calculations for estimating electronic band gaps in inorganic materials, including many classes of metal oxides. Continued research has shown that models trained on large DFT databases can show useful band gap trends across many chemical families, increasing the evaluation of semiconductor materials. ML offers large computational savings and comparable accuracy for these efforts.

The effectiveness of the ML models is dependent upon the dataset. Research efforts often utilize DFT databases such as the Materials Project, AFLOW, Jarvis-DFT, C2DB, and OQMD, which, among other materials classes, are abundant in metal oxide compositions[1][2]. However, these databases can rely on semi-local DFT functionals, which are known to have fundamental limitations. They often underestimate band gaps or struggle to predict properties for strongly correlated or ionic materials [3]. Additionally, these databases can have an incomplete sampling of the materials space. High-throughput DFT databases can contain ground-state crystal structures with certain approximations (ideal periodic crystals, no defects, perfect ordering, neglecting temperature effects, etc.), limiting the transferability to real-world applications[4].

There have been efforts to diversify modeling approaches. Composition-based featurized models are useful when structure is unavailable, can give useful information on chemical trends or non-structure dependent properties, and usually require lower amounts of data/computational power [5]. For band gap predictions, however, material structure is highly important. Changes in the arrangement of atoms can alter the band gap by changing the energy levels of the electrons. The band gap is also affected by material size, crystal lattice constant, and dimensionality. Many studies have investigated the utility of structure-based featurizers in predicting the band gaps of oxide and 2D materials [1][6]. Both a composition and a structure-based featurizer are used and compared in this paper.

Due to the limitations of DFT, recent efforts have focused on robustness, transfer learning, and uncertainty quantification. DFT datasets each contain internal variations; more robust ML pipelines seek to generalize across databases and account for label noise [7][8]. Transfer learning also employs pretraining a model on high-throughput DFT databases (MP,

OQMD) and then fine-tuning to smaller targets [9]. Uncertainty quantification is another strategy for having models report their confidence level to identify where DFT data or other predictions are inconsistent [10]. Though the study described in this paper does not incorporate these advancements, metal-oxide band-gap prediction using DFT databases requires models that better compensate for underestimation and inaccuracies.

ML band-gap models are now routinely used for high-throughput screening in sectors such as oxide photocatalysts, transparent conductors, and semiconductors. Experimental synthesis is guided by limiting composition and structure spaces. Key open challenges still persist in label quality and bias from DFT, model transferability across chemical spaces, and coupling ML predictions to experimental metrics (stability, defect physics). Addressing these will require more robust experimental-based datasets, models with uncertainty factors, and better integration of physics and ML.

3. Final Dataset

The final dataset consisted of 2835 metal oxide samples. The OQMD database was queried via its RESTful API, which utilizes URL queries as a fundamental mechanism. The query was limited to 10000 materials containing at least one oxygen atom, with Pymatgen subsequently utilized to filter for materials containing metal and nonzero band gaps. This resulted in a dataset of metal-oxide compounds, although some exceptions that may sparsely appear in the dataset are molecular complexes (discrete molecules, not ionic/covalent networks), oxygen-rich salts (don't contain O^{2-} oxidation state), or oxides of metalloids. A data-cleaning step to remove any duplicate compositions or entries with missing band gap values. Of the original 10000 structures, 2835 remained after final filtering and data-cleaning.

4. Dataset statistics

Of the final metal oxide materials, every entry contained a unique chemical formula. The dataset contains compositional data as well as structural data through information on the unit cell and atomic site locations; all entries contain structural data. There are 141 unique space groups, with the high-symmetry cubic-structured *Pm-3m* being the most abundant with 378 entries. A relative distribution of crystal systems is given below, with the most abundant systems being cubic (18.91%), monoclinic (17.81%), and tetragonal (17.60%). It is also noted that an *unknown* category exists, which could be due to symmetry detection failing, unstable structures, or extremely low-symmetry systems (which should go to triclinic P1, but could go unlabeled).

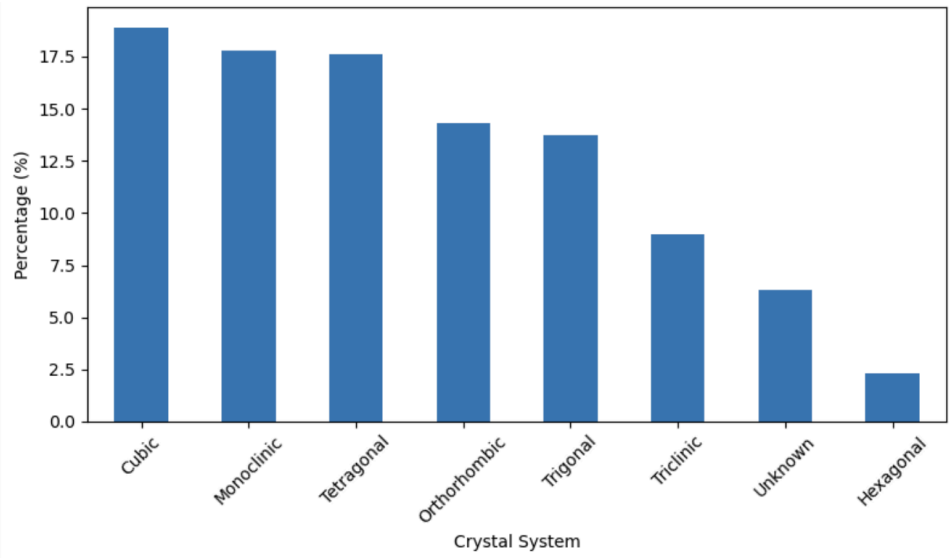

Figure 1: Relative Distribution of Crystal Systems

For the band gap target property, a distribution is given below. The average band gap is 2.78 eV, which is on the upper end of the semiconductor range and approaching the transition to an insulator. This makes sense, as the typical metal oxide band gap is 1.2-8.0 eV, depending on the specific oxide.

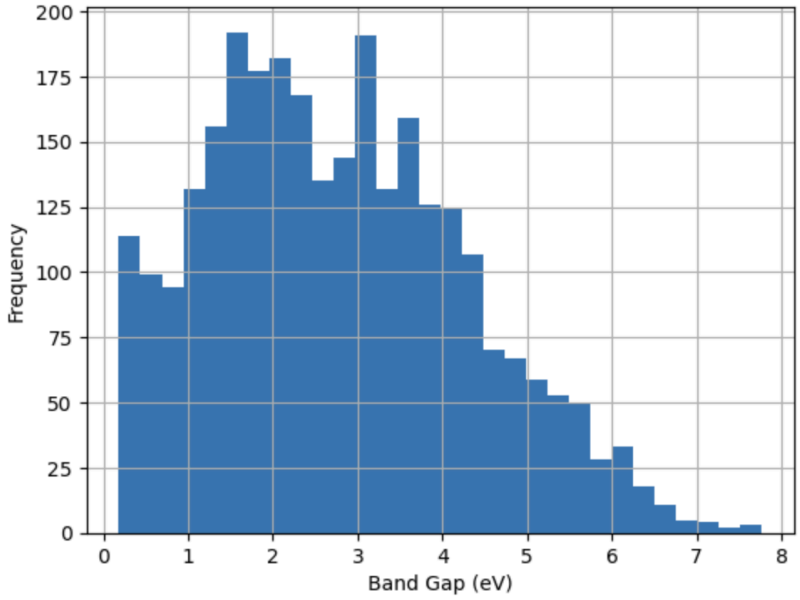

Figure 2: Histogram of Band Gaps

5. Model training

The chosen model was a Random Forest regressor model. This model works by training many decision trees on different random subsets of the data and features, then averaging their predictions to produce a better estimate than any single tree. The data were split randomly

between a 0.70, 0.20, 0.10 train/validation/test split. Both a composition and a structure-based featurizer were selected and compared. The composition featurizer was ElementProperty with Magpie preset features, which computes composition-weighted statistical properties of elemental attributes to translate a chemical formula into numeric features. The structure featurizer was the Ewald Sum Matrix (EWS), which encodes the long-range electrostatic interactions within a crystal structure via Coulomb energy contributions. For both featurizers, small variance and highly correlated columns were removed, leaving ~80 features for each. For each featurizer, three values for each of the following five hyperparameters (a total of 243 models) were tested and optimized: n_estimators, max_depth, max_features, min_samples_split, and min_samples_leaf.

	Tested	ElementProperty model	EwaldSumMatrix 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**

6. Results

	Dummy	ElementProperty model			EwaldSumMatrix model		
	MAE	MAE	MAPE	R2	MAE	MAPE	R2
Train	1.3041	0.213	15.48%	0.965	0.258	18.65%	0.949
Validation		0.541	44.51%	0.775	0.725	58.03%	0.604
Test		0.556	41.98%	0.768	0.699	52.01%	0.611

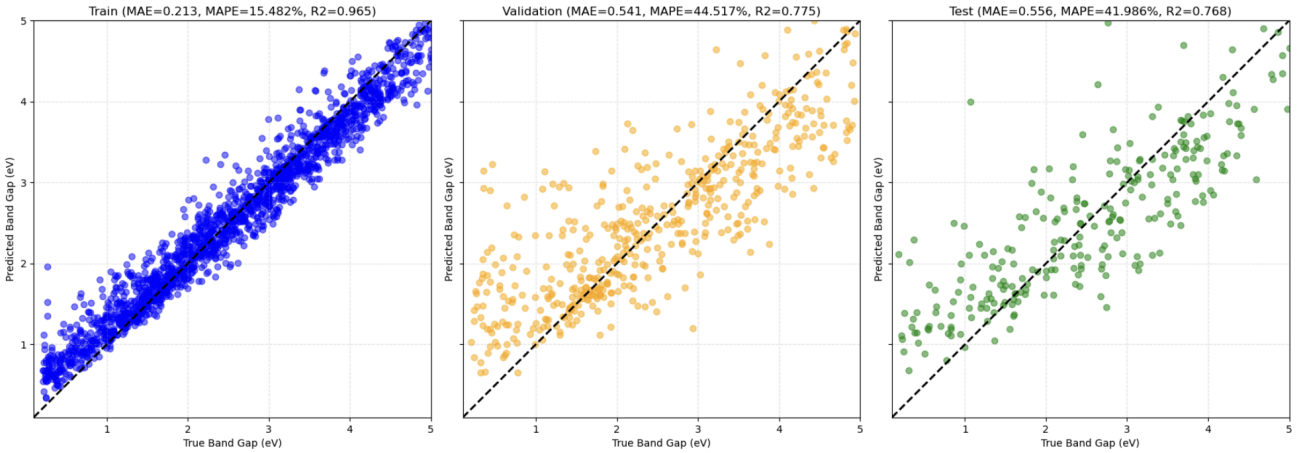

Figure 3: Parity plots for ElementProperty model

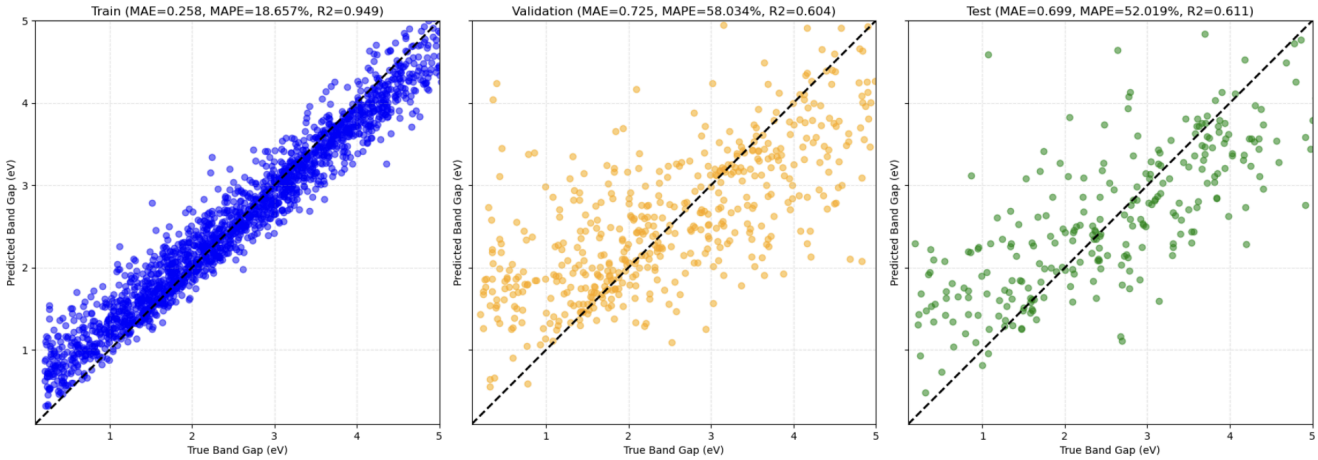

Figure 4: Parity plots for EwaldSumMatrix model

7. Discussion

The ML models performed relatively poorly. Even though neither model got close to a desired MAPE or 10-15%, the composition featurized model (MAPE=41.98%) outperformed the structure featurized model (MAPE=52.01%). This is possibly because composition featurizers rely on fewer and more generalizable features, and can succeed with a more limited dataset. Comparatively, structure featurizers attempt to encode detailed spatial interactions but can be much more sensitive to noise, inaccuracies, and are more susceptible to overfitting.

Properties of the dataset could also be a cause of poor performance. First, the final dataset size of 2,835 materials is modest for training robust models. The limited model size is mainly due to the querying time of the OQMD API, as it took nearly 90 minutes to query 10,000 materials. The limitations of the API also necessitated post-retrieval filtering, so much of the initially queried data was dropped. Second, OQMD, while comprehensive, contains DFT-computed data that can have inherent inaccuracies, especially in band gap predictions due to limitations of common exchange-correlation functionals. This uncertainty in the target values

may introduce noise, making it challenging for models to learn precise structure-property relationships. Additionally, the query parameters from OQMD could be a large source of error. The final dataset was built from originally querying for oxygen-containing samples, then using Pymatgen to filter for metal elements. This, however, does not mean all the remaining materials were metal oxides, and could have allowed for extraneous materials in the dataset.

Another important factor is the model of choice. The Random Forest regressor model may not be the best for capturing the complex, non-linear electronic interactions for band gap calculations. More advanced models, such as gradient-boost machines, along with a more effective structure-based featurizer, may better allow the model to generalize complex interactions.

Lastly, the random splitting of the dataset may have impacted model performance. Random splitting intends that it allows for representative samples of the overall data distribution, allowing for fair model evaluation without introducing bias from specific splitting groups. However, this may inadvertently cause chemically or structurally similar materials to appear across data splits, restricting the model's ability to generalize. Also, the queried dataset statistics do not show a better splitting option based on classes such as crystal system, point group, or space group; the data are too widely distributed among subcategories and do not allow for a 70-20-10 split.

While the models demonstrate moderate performance metrics for metal oxide band gaps, they are limited by innate challenges in accurately capturing complex structure-property relationships. Future improvements in dataset quality, featurization, and modeling approaches are needed to achieve more reliable and precise band gap predictions.

Sources cited:

[1] Zhang, Yu, et al. “Bandgap Prediction of Two-Dimensional Materials Using Machine Learning.” PLOS ONE, edited by Michael Loong Peng Tan, vol. 16, no. 8, Aug. 2021, p. e0255637. DOI.org (Crossref), https://doi.org/10.1371/journal.pone.0255637.

[2] Jung, Son Gyo, et al. “Automatic Prediction of Band Gaps of Inorganic Materials Using a Gradient Boosted and Statistical Feature Selection Workflow.” Journal of Chemical Information and Modeling, vol. 64, no. 4, Feb. 2024, pp. 1187–200. DOI.org (Crossref), https://doi.org/10.1021/acs.jcim.3c01897.

[3] Masood, Hassan, et al. “Enhancing Prediction Accuracy of Physical Band Gaps in Semiconductor Materials.” Cell Reports Physical Science, vol. 4, no. 9, Sept. 2023, p. 101555. DOI.org (Crossref), https://doi.org/10.1016/j.xcrp.2023.101555.

[4] Lu, Ziheng. “Computational Discovery of Energy Materials in the Era of Big Data and Machine Learning: A Critical Review.” Materials Reports: Energy, vol. 1, no. 3, Aug. 2021, p. 100047. DOI.org (Crossref), https://doi.org/10.1016/j.matre.2021.100047.

[5] Ward, Logan, et al. “A General-Purpose Machine Learning Framework for Predicting Properties of Inorganic Materials.” *Npj Computational Materials*, vol. 2, no. 1, Aug. 2016, p. 16028. *DOI.org (Crossref)*, https://doi.org/10.1038/npjcompumats.2016.28.

[6] Yao, Wen, et al. “Machine Learning Prediction of Bandgap and Formation Energy in Two-Dimensional Metal Oxides.” Physica B: Condensed Matter, vol. 717, Nov. 2025, p. 417821. DOI.org (Crossref), https://doi.org/10.1016/j.physb.2025.417821.

[7] Jain, Anubhav, et al. “Commentary: The Materials Project: A Materials Genome Approach to Accelerating Materials Innovation.” *APL Materials*, vol. 1, no. 1, July 2013, p. 011002. *DOI.org (Crossref)*, https://doi.org/10.1063/1.4812323.

[8] Kirklin, Scott, et al. “The Open Quantum Materials Database (OQMD): Assessing the Accuracy of DFT Formation Energies.” Npj Computational Materials, vol. 1, no. 1, Dec. 2015, p. 15010. DOI.org (Crossref), https://doi.org/10.1038/npjcompumats.2015.10.

[9] Buterez, David, et al. “Transfer Learning with Graph Neural Networks for Improved Molecular Property Prediction in the Multi-Fidelity Setting.” Nature Communications, vol. 15, no. 1, Feb. 2024, p. 1517. DOI.org (Crossref), https://doi.org/10.1038/s41467-024-45566-8.

[10] Wen, Mingjian, and Ellad B. Tadmor. “Uncertainty Quantification in Molecular Simulations with Dropout Neural Network Potentials.” Npj Computational Materials, vol. 6, no. 1, Aug. 2020, p. 124. DOI.org (Crossref), https://doi.org/10.1038/s41524-020-00390-8.