Accelerating Discovery of Biocompatible Transparent Semiconductors

Amala Vellappillil Biju Brent Thorne Dare Drouin Jehad Saif
Michael Amaraut Rachana Baskar *

December 13, 2025

Abstract

This project develops a multimodal machine learning pipeline to accelerate the discovery of biocompatible transparent semiconductors. By integrating structural, compositional, and electronic features from high-throughput Density Functional Theory (DFT) datasets, we predict and rank candidate materials with tunable bandgaps, high optical transparency, and low biological toxicity. Our approach combines reproducible preprocessing, interpretable model evaluation, and domain-specific filtering to support applications in optoelectronics, biosensors, and sustainable transparent semiconductors. Ultimately, the pipeline reduces research costs and accelerates experimental validation by focusing attention on the most promising candidates.

1 Introduction

Transparent semiconductors are critical for next-generation technologies in optoelectronics, biosensing, and sustainable energy.[1] However, discovering new candidates remains computationally expensive and often neglects biocompatibility. Our goal is to build a modular ML pipeline that accelerates discovery while ensuring reproducibility and safety.

2 Methodology

Our methodology is designed as a modular, reproducible pipeline that integrates domain-specific filtering, multimodal feature engineering, and advanced machine learning models. The central goal is to accelerate the discovery of biocompatible transparent semiconductors while maintaining interpretability and reproducibility at every stage.

2.1 Data Ingestion and Cleaning

We draw from multiple high-throughput Density Functional Theory (DFT) repositories, including NOMAD2018 [2] and JARVIS-DFT [3]. Each dataset undergoes standardized preprocessing to ensure consistency and reproducibility. Numeric features are normalized, categorical descriptors are encoded, chemical formulas are parsed into element counts, and atomic structures are converted into graph objects. All preprocessing steps are logged and version-controlled so that results can be traced back to their original configurations.

*Group 2

2.2 Exploratory Data Analysis

Exploratory data analysis (EDA) guided threshold selection, feature engineering, and model design. PCA on the NOMAD2018 [2] dataset revealed clusters aligned with crystal families and compositional patterns, with the first two components explaining 46% of the variance. Band gap classes distributed across these clusters confirmed the strong influence of structure on electronic properties. Nonlinear methods (UMAP, t-SNE) reinforced this, showing that materials with similar lattice parameters and bonding motifs cluster together, while distinct families formed separate islands [4] [5].

SHAP (SHapley Additive exPlanations) analysis identified volume density as the strongest driver of band gap predictions and lattice vector lengths, particularly the third vector, as key to formation energy [6]. These results validated the importance of structural descriptors, justified graph-based encodings, and supported removal of low-importance features to reduce noise.

EDA on JARVIS-DFT confirmed these trends. Band gap distributions showed that most materials are metallic, while formation energy emphasized the need for thermodynamic filtering. Element frequency analysis revealed oxygen dominance, followed by nitrogen and fluorine, consistent with oxide, nitride, and fluoride semiconductors. Space group distributions highlighted cubic symmetry with monoclinic and orthorhombic distortions associated with wider band gaps. PCA of filtered candidates showed tight clustering, suggesting shared underlying chemistry and physics.

Overall, EDA demonstrated that lattice geometry, symmetry, and volume density are central to predicting band gap and stability, directly informing the CandidateNet multimodal architecture.

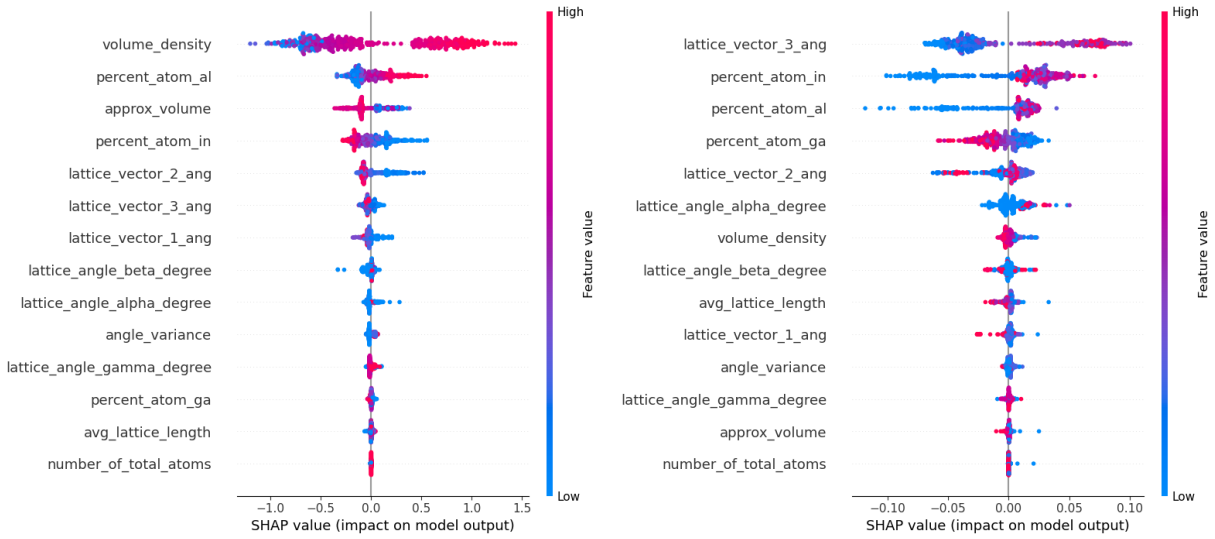

Figure 1: SHAP feature impacts on target features. Left: Band Gap. Right: Formation Energy.

2.3 Rule-Based Candidate Filtering

The NOMAD2018 [2] dataset, composed of ternary oxides of aluminum, gallium, and indium, provided a clean space for defining physically meaningful thresholds. Analysis of correlations between band gap, formation energy, and compositional ratios allowed us to distinguish semiconductors from metals and insulators. Specifically, band gaps below 0.5 eV correspond to metallic behavior, while gaps above 5 eV correspond to insulating materials. Transparent semiconductors consistently exhibited band gaps greater than 2.5 eV. Stability is best captured by an energy above hull threshold of 0.1 eV or less.

These NOMAD-derived thresholds were applied to the full JARVIS-DFT dataset, which contains more than 70,000 inorganic compounds. The filtering logic ensured that only physically plausible, stable, and potentially transparent candidates were retained for modeling. To prioritize practicality and safety, compounds containing Lead (Pb), Cadmium (Cd), Arsenic (As), and Mercury (Hg) were removed. This rule-based filtering step reduced noise and provided interpretable labels for supervised learning.

2.4 Multimodal Feature Embedding

After filtering, we constructed multimodal feature representations by integrating four distinct types of information: numeric descriptors (e.g., band gap, dielectric constants, and formation energy), categorical embeddings (e.g., space group, dimensionality, crystal system, and functional descriptors), formula embeddings encoding stoichiometric patterns across 89 elements into 32-dimensional vectors, and graph-based encodings capturing atomic connectivity and local bonding environments via message passing. These representations are concatenated into a unified vector, enabling the model to learn both global descriptors and fine-grained structural motifs relevant to transparent semiconductor behavior.

2.5 Model Training and Evaluation

Model training is carried out using CandidateNet, a multimodal neural network architecture. Depending on the dataset, we employ progressively more complex models: a simple multilayer perceptron (MLP) for tabular features, a CNN-style CandidateNet for categorical and formula embeddings, and a multimodal CandidateNetMultimodal that integrates graph neural networks for atomic structure. Training incorporates class imbalance weighting to account for the rarity of promising candidates, and evaluation metrics include ROC-AUC, precision, recall, and F1 score. Threshold calibration ensures balanced performance for deployment, with classification boundaries tuned to maximize discovery potential.

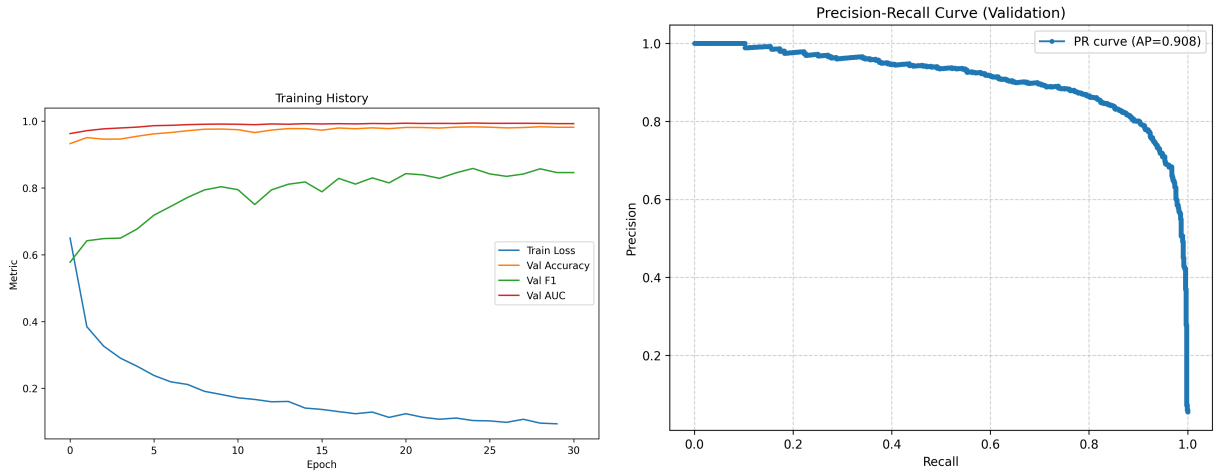

Figure 2: Model performance metrics. Left: Training history showing loss and validation metrics across epochs. Right: Precision–Recall curve with average precision (AP=0.908).

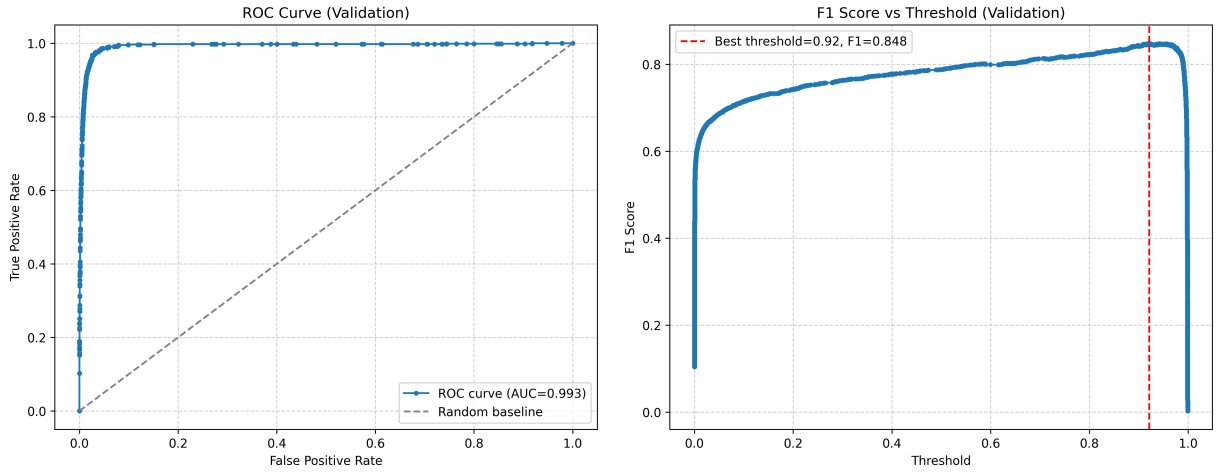

Figure 3: Classification performance. Left: ROC curve with area under the curve (AUC=0.993). Right: F1 score vs threshold (optimal threshold=0.92, F1=0.848).

2.6 Validation with OQMD

While NOMAD2018 [2] was used to establish interpretable thresholds and JARVIS-DFT served as the main dataset for multimodal modeling, the Open Quantum Materials Database (OQMD) [7] plays a complementary role in validation. OQMD contains over one million computed entries spanning diverse chemical and structural spaces. We validate candidate predictions by cross-checking stability and band gap values against high-quality DFT results. For candidates that have not yet been computed within NOMAD or JARVIS, OQMD provides a pathway to collect new DFT results. This ensures that our pipeline does not rely solely on model predictions but can be anchored in verified quantum-mechanical calculations. By integrating OQMD queries into the workflow, we extend the coverage of our candidate list, confirm reproducibility across datasets, and strengthen confidence in the discovery of novel biocompatible transparent semiconductors.

2.7 Disagreement Mining and Hypothesis Generation

A distinctive aspect of our methodology is the treatment of classification errors. Rather than discarding false positives and false negatives, we treat them as hypotheses for materials discovery. False positives are compounds that are predicted as promising but excluded by rule-based filters, and may represent novel transparent semiconductors outside conventional thresholds. False negatives materials flagged by rules but missed by the model highlight feature gaps or overlooked correlations. By mining these disagreements, we generate actionable leads for experimental validation.

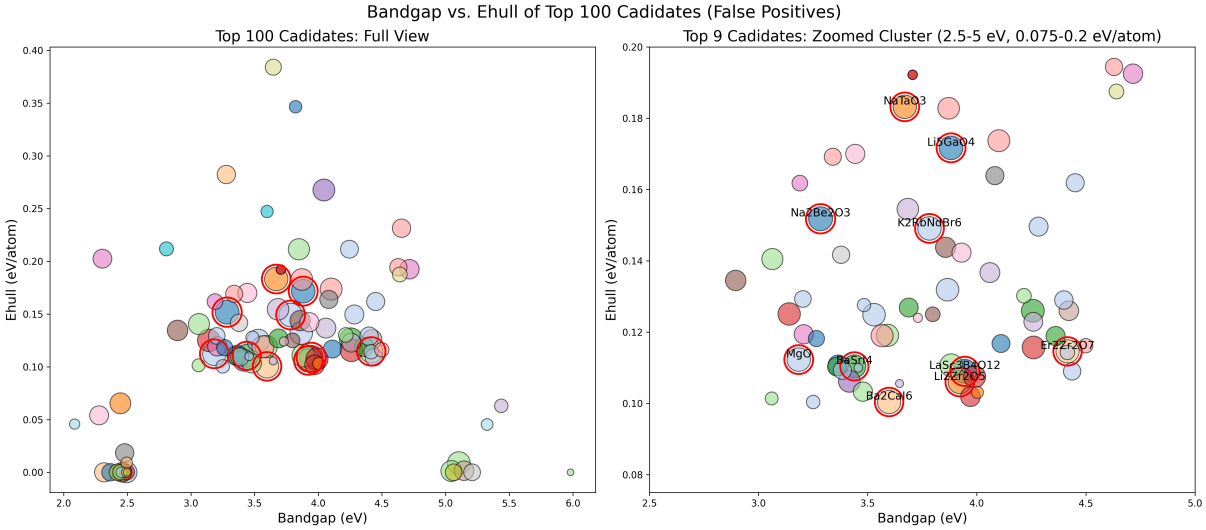

Figure 4: Side-by-side scatter plots of the top 100 false positive candidates showing bandgap vs. ehull. Left: full distribution. Right: zoomed view of the central cluster (bandgap 2.5–5 eV, ehull 0.075–0.2 eV/atom). Point size reflects normalized model probability, color encodes space group, and the top 10 formulas are labeled.

2.8 Lead Ranking and Reporting

Finally, the pipeline produces ranked candidate lists annotated with key metadata such as formula, band gap, stability, and toxicity flags. Transparent reports are generated with metrics, candidate tables, and provenance logs, enabling collaboration with chemists and materials scientists. This reproducible and extensible design ensures that the pipeline can scale to larger datasets, incorporate new descriptors, and evolve iteratively as experimental feedback is integrated.

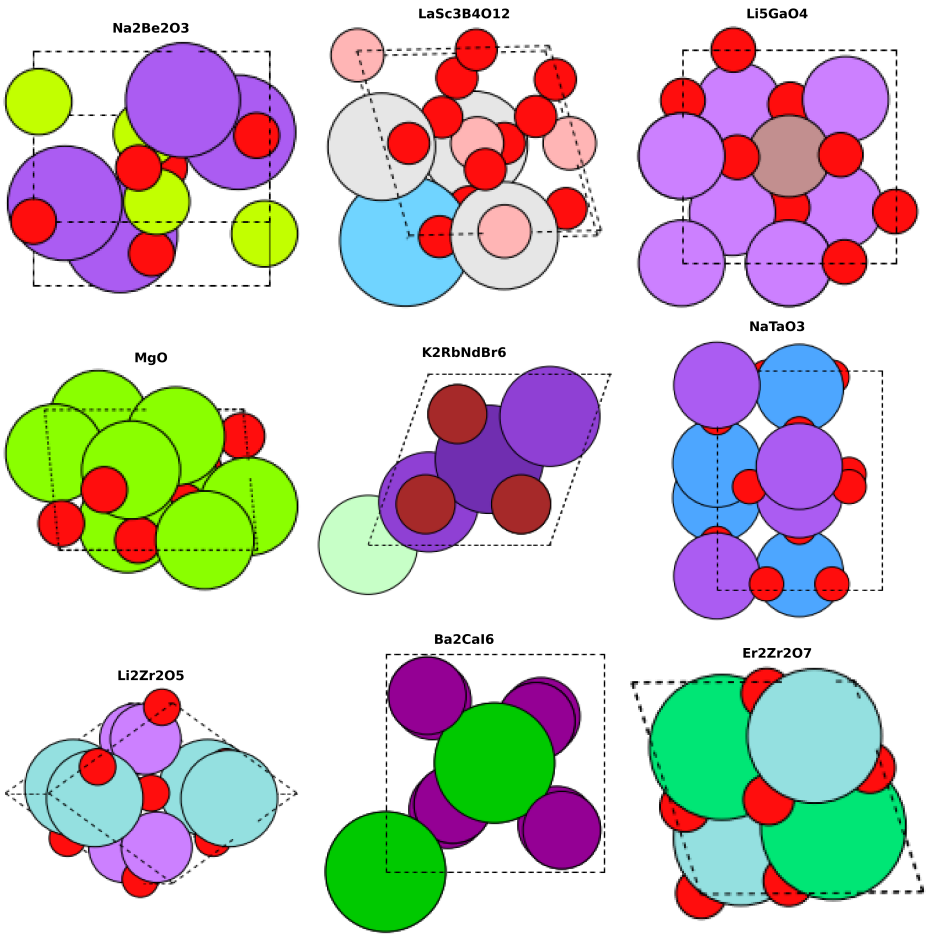

Figure 5: 3x3 grid of structural snapshots for the top 9 false positive candidates. Each panel shows an ASE-rendered atomic configuration with the chemical formula as the title. These candidates were selected by highest predicted probability, providing a visual overview of their structural diversity.

3 Results

3.1 Candidate Ranking

The pipeline produced a ranked list of candidate materials that satisfy the combined requirements of semiconductivity, transparency, and biocompatibility. Each of these compounds exhibits wide band gaps that place them in the range of transparent semiconductors, while also demonstrating thermodynamic stability and structural features like low intrinsic carrier concentration that makes doping control easier. Among the top candidates identified, $Bi_{10}Mo_3O_{24}$ [8] is particularly notable because the presence of d0 orbitals from Bi^{3+} cations makes it a potential candidate for p-type semiconductors, which are rare among semiconductor oxides. Its non-centrosymmetric structure and high thermal stability further enhance its suitability for thin-film growth and doping treatments. In contrast, Nb_9PO_{25} [9] and $Ba_3Nb_6Si_4O_{26}$, contains d0 orbitals from Nb^{5+} cations, making them strong candidate for n-type doping. Their wide band gaps and stability indicate promise for optoelectronic and photocatalytic applications. Finally, $K_4Hf_5O_{12}$ demonstrates structural viability for n-type doping and could be synthesized as thin film for transparent conductor applications.

Together, these candidates highlight the strength of the pipeline in identifying materials that not only meet the electronic and optical requirements but also offer practical pathways for doping

and experimental validation. The inclusion of both p-type and n-type candidates is particularly important, as balanced doping strategies are essential for building functional semiconductor devices.

Further, upon evaluating the candidates that our model predicted false positives, LaSc3B4O12 and NaTaO3 [10] could be potential transparent semi-conductors theoretically.

4 Novelty

The novelty of our methodology lies in the way it bridges traditional rule-based approaches with modern machine learning, creating a discovery pipeline that is both interpretable and exploratory. Rather than optimizing predictive accuracy, our framework is designed to generate hypotheses and identify materials that may fall outside conventional screening criteria.

First, our pipeline anchors discovery in physically interpretable thresholds derived from NOMAD dataset, while the multimodal neural network captures complex structural motifs beyond linear rules. Second, model disagreements are treated as discovery signals: false positives and false negatives are reframed as hypotheses that reveal candidates outside traditional thresholds or highlight gaps in the feature space. Third, our methodology accounts for polymorph diversity by analyzing crystal structures directly and incorporating graph-based encodings, allowing the pipeline to distinguish structural variants and identify which polymorphs yield desired transparency, conductivity, and biocompatibility. Finally, we embed provenance and reproducibility into every step through logged and version-controlled operations, ensuring that predictions can be validated by experimental collaborators.

Together, these elements distinguish our approach as a reproducible engine that unifies physical rules, structural learning, and hypothesis generation to accelerate the search for biocompatible transparent semiconductors.

5 Limitations

Although our pipeline demonstrates strong potential for accelerating the discovery of biocompatible transparent semiconductors, several limitations remain. First, the underlying datasets lack explicit biocompatibility descriptors. While the rule-based filters exclude known toxic elements such as Pb, Cd, As, and Hg, biological safety is inferred indirectly rather than modeled explicitly, limiting screening precision. Second, many candidates lack experimental validation although DFT predictions provide valuable guidance, synthesis and testing are required to confirm real-world performance. Moreover, scarcity of experimental data for less-studied compounds slows the transition from computational discovery to practical application. Finally, ambiguity in property values, such as band gaps that vary with the DFT functional or formation energies that differ across datasets, introduces uncertainty and complicates candidate ranking.

Addressing these limitations will require expanded datasets with biocompatibility-specific descriptors, and closer collaboration with experimental workflows to ensure that computationally identified candidates can be translated into real-world biocompatible transparent semiconductors.

6 Future Work

Several directions can further strengthen and extend the capabilities of our pipeline. First, systematic integration of the Open Quantum Materials Database (OQMD) will significantly expand compositional and structural coverage, enabling validation of current candidates against a broader DFT dataset and providing results for compounds absent from NOMAD and JARVIS. This will improve scalability to industrially relevant materials spaces and increase confidence in predictions.

Second, future work will focus on incorporating biocompatibility proxies. While our current approach excludes toxic elements through rule-based filters, it does not explicitly model biological safety, so developing proxy descriptors such as elemental toxicity scores, bioavailability indicators, or empirical heuristics—will allow us to rank candidates by both functional properties and likelihood of safe biomedical use.

Finally, closing the loop with experimentalists will allow synthesis and testing of high-ranking candidates, refinement of models using empirical feedback, and validation of reproducibility across computational and experimental domains. Together, these future efforts will expand the pipeline's scope, improve prediction reliability, and facilitate translation of computational discoveries into real-world biocompatible transparent semiconductors.

7 Conclusion

In this work, we developed a reproducible multimodal machine learning pipeline to accelerates the discovery of biocompatible transparent semiconductors. By combining rule-based filtering with advanced neural network architectures, we created a framework that is both interpretable and capable of learning complex structural and compositional relationships. Physically motivated thresholds derived from NOMAD dataset were applied to the JARVIS-DFT dataset, enabling the identification of stable, transparent, and non-toxic candidates. Validation through the Open Quantum Materials Database (OQMD) further strengthened confidence in the results by providing high-quality Density Functional Theory (DFT) calculations for candidates not previously computed.

A key innovation of this approach is treatment of model disagreements as opportunities for discovery rather than errors. By accounting for polymorph diversity and embedding reproducibility throughout the pipeline, our framework supports transparent and traceable predictions suitable for experimental follow-up. Overall, this work demonstrates how integrating physical intuition with multimodal machine learning can accelerate materials discovery and provides a scalable foundation for identifying safe and functional materials for future optoelectronic and biomedical applications.

References

[1] David S. Ginley and Christopher Bright. Transparent conducting oxides. *MRS Bulletin*, 25(8):15–18, 2000.

[2] Luca M. Ghiringhelli and et al. The nomad 2018 dataset. *Scientific Data*, 2018.

[3] Kamal Choudhary and et al. Jarvis-dft: A dataset for materials discovery. *npj Computational Materials*, 2020.

[4] Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection. *arXiv preprint arXiv:1802.03426*, 2018.

[5] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 2008.

[6] Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, 2017.

[7] James E. Saal and et al. The open quantum materials database (oqmd). *JOM*, 2013.

[8] Xuefan Wang, Yan Xiao, Wenjing Tan, Hongbo Huang, Daqing Yang, Ying Wang, and Bingbing Zhang. Computer-aided screening of bismuth molybdates nonlinear optical crystals γ-bi_2moo_6. *Inorganic Chemistry Frontiers*, 2024.

[9] A. J. Green, E. H. Driscoll, Y. Lakhdar, E. Kendrick, and P. R. Slater. Structural and electrochemical insights into novel wadsley roth $nb_7ti_{1.5}mo_{1.5}o_{25}$ and $ta_7ti_{1.5}mo_{1.5}o_{25}$ anodes for li-ion battery application. *Dalton Transactions*, 2023.

[10] Kootak Hong, Shaun Tan, Matthew J. McDermott, Tianyi Huang, Finn Babbe, Tim Kodalle, Max Gallant, Sehun Seo, Francesca M. Toma, Kristin A. Persson, Yang Yang, and Carolin M. Sutter-Fella. Shape-controlled $natao_3$ by flux-mediated synthesis. *Advanced Functional Materials*, 2022.

8 Appendix

A Definitions

Density Functional Theory (DFT). A quantum-mechanical method used to compute the electronic structure of materials. DFT provides estimates of band gaps, formation energies, total energies, and charge densities, serving as the foundational data source for this project.

Band Gap. The energy difference between the valence band maximum and the conduction band minimum. Materials with band gaps between 2.5–5 eV are typically considered transparent semiconductors, while gaps below 0.5 eV correspond to metallic behavior.

Formation Energy. The thermodynamic stability metric defined as the energy required to assemble a compound from its constituent elements in their reference states. Lower formation energies indicate more stable materials.

Energy Above Hull. A stability metric indicating how far a compound lies above the convex hull of competing phases. Materials with energy above hull less than 0.1 eV/atom are considered synthesizable and thermodynamically accessible.

Space Group. A symmetry classification describing all possible crystallographic symmetries in a material's atomic arrangement. Space group information encodes structural constraints relevant to optical and electronic properties.

Principal Component Analysis (PCA). A linear dimensionality reduction technique that projects high-dimensional data into a lower-dimensional space by maximizing variance.

UMAP (Uniform Manifold Approximation and Projection). A nonlinear dimensionality reduction method that preserves local and global structure.

t-SNE (t-distributed Stochastic Neighbor Embedding). A nonlinear visualization method that preserves local structure in high-dimensional data.

SHAP (SHapley Additive exPlanations). An interpretability framework that quantifies the contribution of each feature to a model's predictions.

Multimodal Learning. A machine learning paradigm in which models incorporate multiple complementary feature types—including numeric descriptors, categorical labels, chemical formulas, and atomic graphs—to improve prediction accuracy.

Graph Neural Network (GNN). A neural network architecture operating on graph-structured data. In materials informatics, atoms are represented as nodes and bonds or near-neighbor connections as edges, enabling the model to learn structural motifs and local bonding environments.

Compositional Embedding. A vector representation of a chemical formula where each element's contribution is weighted by stoichiometry. This encoding captures periodic trends and elemental interactions relevant to band gap and stability.

Energy–Volume Density. A derived feature that measures the amount of energy or mass per unit cell volume. It was identified through SHAP analysis as a strong predictor of band gap and formation energy.

CandidateNet. A multimodal neural network architecture developed for this project. It integrates numeric features, categorical embeddings, formula embeddings, and graph encodings to predict transparent semiconductor candidates.

Disagreement Mining. A hypothesis-generation technique in which false positives and false negatives are analyzed to identify overlooked candidate materials or limitations in rule-based filtering.