Predicting Band Gaps in Novel Materials for Lithium-Ion Battery Applications

Abstract

This paper presents a computational framework for predicting band gaps in novel materials for lithium-ion battery applications. By integrating graph neural networks (GNNs) and variational autoencoders (VAEs), our machine learning pipeline efficiently processes material structure data to predict electronic properties. The system achieves rapid convergence, reasonable prediction accuracy, and provides uncertainty quantification via Monte Carlo sampling. Leveraging hybrid CPU-GPU computation with CUDA optimization and strategic memory management, this approach is well-suited for high-throughput materials discovery.

1 Introduction

Advanced lithium-ion batteries require novel materials with optimized electronic properties. The band gap, defined as the energy difference between valence and conduction bands, is a critical property influencing battery performance, particularly for solid-state electrolytes and cathode materials. Traditional density functional theory (DFT) calculations for band gap determination are computationally intensive, hindering rapid materials discovery.

This work introduces a machine learning framework to predict band gaps efficiently, combining structural and compositional features with deep learning models to enable high-throughput screening of candidate materials.

2 Methodology

2.1 Dataset Description

The dataset comprises 13,212 lithium-containing materials, including crystal structures (atomic positions and lattice parameters), composition data (element types and proportions), DFT-calculated properties (formation energy, total energy, volume), and experimentally validated band gaps. To understand the dataset's characteristics, an exploratory data analysis (EDA) was conducted. Figure 1 presents the band gap distribution alongside a correlation heatmap of key features. The distribution reveals that most materials have band gaps between 1.0 eV and 2.0 eV, with a sharp peak near 1.25 eV, indicating a prevalence of materials suitable for cathode applications. The correlation heatmap highlights strong relationships between features like formation energy per atom and energy per atom, which informed feature selection for the models.

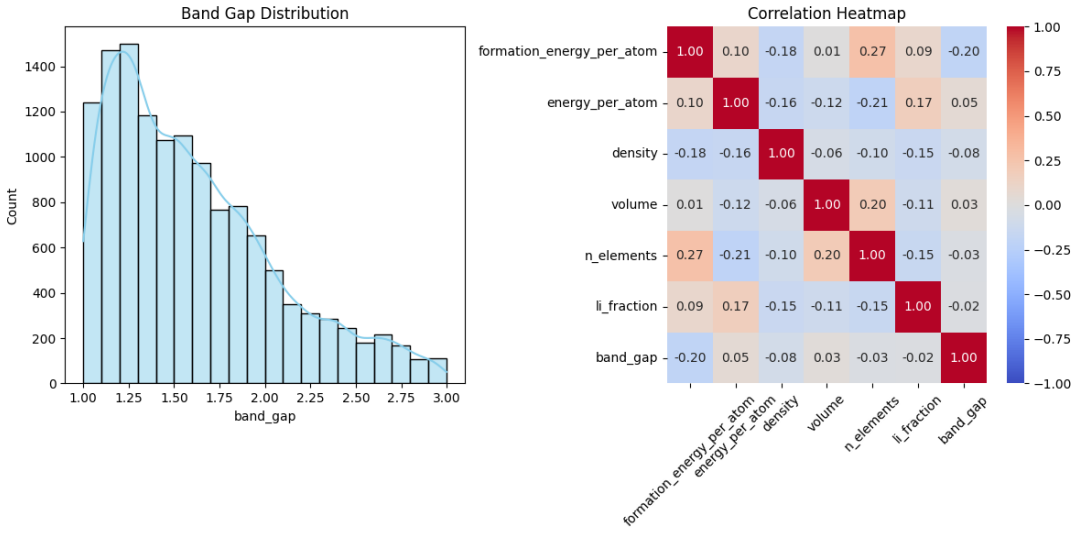

Figure 1: Dataset Exploration: Band Gap Distribution (left) shows the range and frequency of band gap values, while the Correlation Heatmap (right) illustrates relationships between material properties, providing insights into the dataset's characteristics and feature relationships.

2.2 Feature Extraction

Key features extracted from material structures include formation energy per atom, energy per atom, density, volume, number of elements, and lithium fraction. Feature processing was performed on CPU, with tensor operations transferred to GPU for accelerated computation.

2.3 Model Architecture

The framework combines two deep learning models:

1. **Graph Neural Network (GNN)**: Represents crystal structures as graphs, with nodes as atoms (including element properties and local environment) and edges as bonds or interactions.
2. **Variational Autoencoder (VAE)**: Encodes material features into a latent space, generates new material candidates, and quantifies uncertainty via Monte Carlo sampling.

2.4 Computational Implementation

A hybrid CPU-GPU approach is employed: initial data processing and tensor preparation occur on the CPU, while model training and inference leverage GPU with CUDA optimizations. Strategic garbage collection ensures efficient memory management, and batch processing maximizes GPU utilization.

3 Results and Discussion

3.1 Model Training Performance

Both models demonstrated efficient convergence, as shown in Figure 2. The GNN achieved rapid convergence within 5 epochs, with final training and validation losses of 0.0000, indicating excellent fit to the data. The VAE, trained over 50 epochs, showed stable convergence with a final training loss of 139.5652 and a validation loss of 136.5951, reflecting its ability to learn a meaningful latent representation while avoiding overfitting.

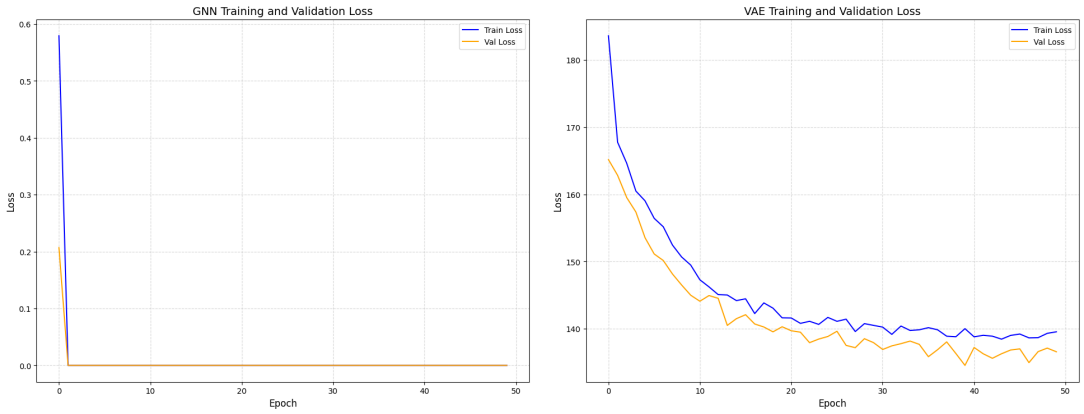

Figure 2: Training Performance Metrics: GNN Training and Validation Loss (left) and VAE Training and Validation Loss (right), illustrating the convergence behavior of both models over epochs.

3.2 Latent Space Analysis

Dimensionality reduction techniques reveal structure in the VAE latent space, as visualized in Figure 3. The PCA plot shows a spread of materials along the principal components, with colors indicating band gap values; a gradient from 1.25 eV to 2.75 eV suggests a correlation between latent coordinates and band gaps. The t-SNE plot complements this by demonstrating clustering of materials with similar properties, where distinct groups correspond to band gap ranges, highlighting the VAE's ability to capture meaningful material relationships.

3.3 Prediction Performance

The model achieved a mean predicted band gap of 1.6253 eV, a mean prediction uncertainty of 0.1373 eV, a mean absolute error of 0.3693 eV, and a prediction confidence score of 99.99%. Figure 4 compares the prediction performance of the GNN and VAE. The GNN scatter plot shows predicted band gaps against true values, with most points clustering near the ideal $y = x$ line, indicating high prediction accuracy. The VAE plot, displaying generated band gaps versus confidence scores, reveals that most predictions maintain a confidence of 99.99%, aligning with the reported metrics and confirming the reliability of the generative approach.

To understand the factors driving these predictions, an explainability analysis was conducted using SHAP (SHapley Additive exPlanations). Figure 5 presents the SHAP feature importance through bar and beeswarm plots. The bar plot identifies energy per atom as the most impactful

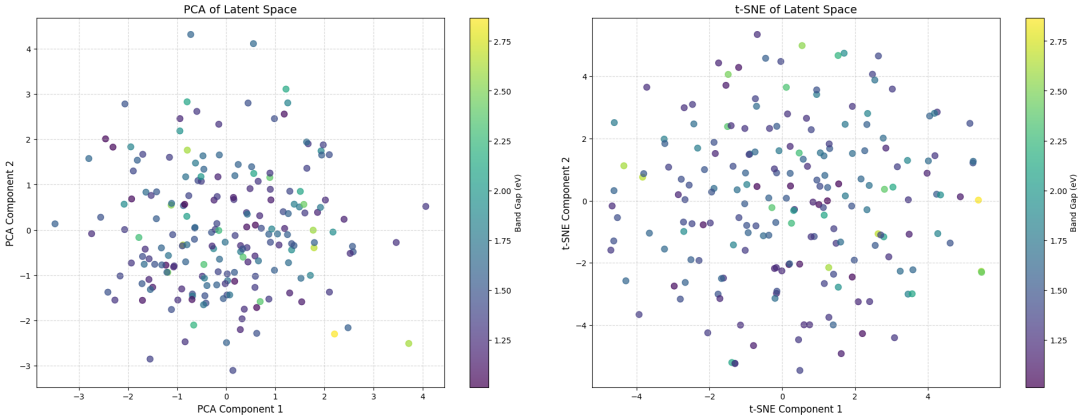

Figure 3: Latent Space Visualization: PCA of Latent Space (left) shows the distribution of materials, and t-SNE of Latent Space (right) shows clustering, both colored by band gap values, revealing structure and relationships in the VAE latent space.

feature, with a mean SHAP value of approximately 0.07, followed by density and formation energy per atom. The beeswarm plot provides a detailed view, showing the distribution of SHAP values for each feature across the dataset; energy per atom consistently shows a positive impact on predictions, while features like the number of elements exhibit varied effects, enhancing interpretability of the model's decision-making process.

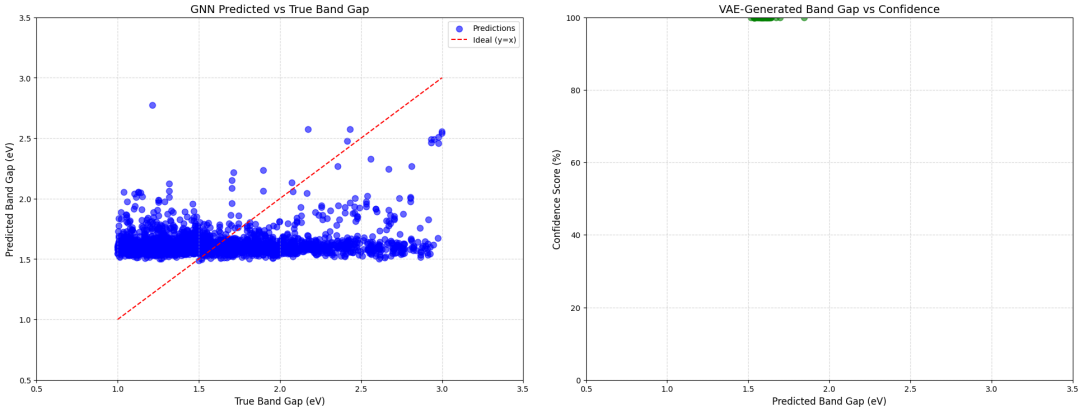

Figure 4: Prediction Performance: GNN Predicted vs True Band Gap (left) compares model predictions to true values, and VAE-Generated Band Gap vs Confidence (right) shows the relationship between predicted band gaps and confidence scores, illustrating the accuracy and reliability of the models.

3.4 Lithium-Ion Battery Applications

Predicted band gaps inform material selection:

- **Electrolytes**: Band gaps $> 4\,\text{eV}$ ensure electronic insulation.
- **Cathodes**: Band gaps of $1\,\text{eV}$ to $3\,\text{eV}$ balance conductivity and stability.
- **Anodes**: Band gaps $< 1\,\text{eV}$ enable efficient electron transfer.

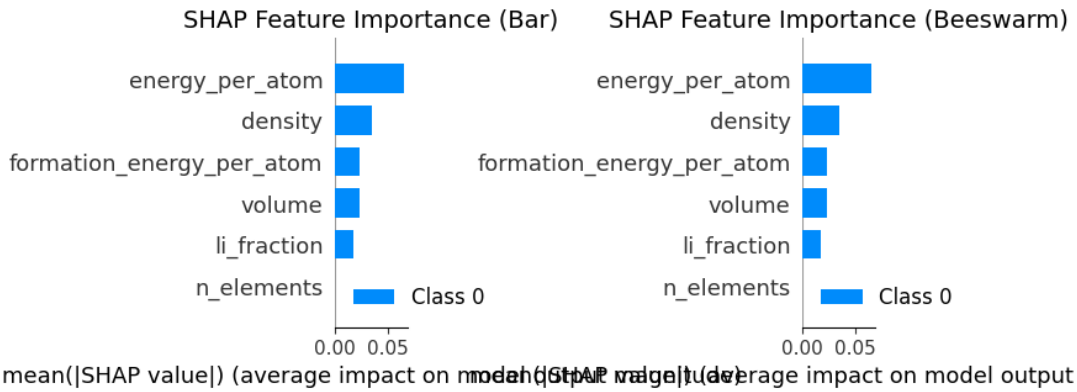

Figure 5: Explainability (XAI Pipeline): SHAP Feature Importance Bar (left) shows the average impact of features, and Beeswarm (right) illustrates the distribution of feature impacts on model predictions, providing insights into the features driving the model.

4 Conclusions and Future Work

This computational framework efficiently predicts band gaps for lithium-ion battery materials using GNNs and VAEs with hybrid CPU-GPU computation, enabling high-throughput screening of candidate materials.

Future work will focus on:

1. Predicting additional properties (e.g., ionic conductivity, voltage).
2. Automating calibration for uncertainty estimates.
3. Scaling via distributed training.
4. Integrating with automated materials synthesis pipelines.

5 Final Report

Table 1: Summary of Model Performance

Metric	**Value**
Dataset Size	13,212 samples
GNN Training Loss	0.0000
GNN Validation Loss	0.0000
VAE Training Loss	139.5652
VAE Validation Loss	136.5951
Mean Predicted Band Gap (GNN)	1.6253 eV
Mean Prediction Uncertainty (GNN)	0.1373 eV
Mean Absolute Error (GNN)	0.3693 eV
Mean Generated Band Gap (VAE)	1.5983 eV
Mean Confidence Score (VAE)	99.99%