Graph Neural Networks Based Deep Learning for Predicting Structural and Electronic Properties

Selva Chandrasekaran Selvaraj[1,2]

[1]Present address: Department of Chemical Engineering, University of Illinois Chicago, Chicago, IL 60608, United States
[2]International Center for Materials Nanoarchitectonics, National Institute for Materials Science(NIMS), Tsukuba, 305-0044, Japan
[1]*Email:* selvas@uic.edu

November 4, 2024

Abstract

This study presents a deep learning approach to predicting structural and electronic properties of materials using Graph Neural Networks (GNNs). Leveraging data from the Materials Project database, we construct graph representations of crystal structures and employ GNNs to predict multiple properties simultaneously. All crystal structures are from the Materials Project database, with a total of 158,874 structures used. Our model achieves high predictive accuracy across various properties, as indicated by R^2 values: 0.96 for density, 0.97 for formation energy, 0.54 for energy above hull, 0.47 for structural stability (is_S), 0.76 for band gap, 0.86 for valence band maximum, 0.78 for conduction band minimum, and 0.82 for Fermi energy. These results demonstrate the potential of GNNs in materials science, offering a powerful tool for rapid screening and discovery of materials with desired properties.

Keywords: Graph Neural Networks, Material Properties Prediction, Machine Learning, High-Throughput Screening, Materials Project Database, Density Functional Theory, Band Gap Prediction, Structural Stability

1 Introduction

The discovery and design of new materials with tailored properties is crucial for advancing technologies in energy, electronics, and various other fields [1]. Traditional experimental approaches to materials discovery are often time-consuming and expensive, limiting the pace of innovation [2, 3, 4]. Especially, when the DFT calculations are involved by the surfaces and interfaces, the computational resources demand exponentially[5, 6, 7]. Computational methods, particularly those based on density functional theory (DFT), have significantly accelerated this process by enabling high-throughput screening of material candidates [1, 8]. However, even these methods can be computationally intensive for large-scale exploration of the vast chemical space [9].

In recent years, machine learning approaches have emerged as promising tools for rapid and accurate prediction of material properties [2, 10, 11]. These methods offer the potential to dramatically reduce the computational cost of materials screening while maintaining high accuracy [8]. Among these, Graph Neural Networks (GNNs) have shown exceptional promise due to their ability to naturally represent the atomic structure of materials and capture complex interactions between atoms [9, 8, 12, 13].

The application of GNNs to materials science has several advantages[14, 15]:

- **Structural Representation**: GNNs can directly operate on the graph structure of materials, preserving the spatial and chemical information of atoms and their bonds [9].
- **Invariance**: GNNs can be designed to be invariant to permutations of atoms, which is a crucial property for materials representation [8].
- **Multi-scale Learning**: GNNs can capture both local atomic environments and global structural features, allowing for multi-scale property prediction [2].

- **Transferability**: Models trained on one set of materials can often be applied to predict properties of new, unseen materials [10].

Recent studies have demonstrated the effectiveness of GNNs in predicting various material properties, including formation energies [9], band gaps [8], and mechanical properties [2]. However, most of these studies focus on predicting a single property or a limited set of related properties.

This project aims to develop a GNN-based framework for predicting a wide range of structural and electronic properties of materials simultaneously. We utilize data from the Materials Project [1], a comprehensive database of computed material properties, to train our models. The properties we predict include volume, number of sites, density, bulk modulus, shear modulus, crystal system, space group number, various energy metrics, stability, band gap, and band structure characteristics.

Our approach involves several key steps: 1. Downloading and processing structures from the Materials Project database, 2. Converting crystal structures into graph representations, 3. Implementing a GNN model using PyTorch Geometric [16], 4. Training the model to predict multiple properties simultaneously, and 5. Evaluating the model's performance using metrics such as R^2 score and mean absolute error

The GNN model we employ consists of multiple Graph Convolutional Network (GCN) layers [17, 18], followed by global mean pooling and a final linear layer for prediction. We use cross-validation to ensure robust performance across different subsets of our dataset.

This work contributes to the growing field of machine learning for materials science by demonstrating the effectiveness of GNNs in predicting a diverse set of material properties. The high accuracy achieved (90-98% across all predicted properties) showcases the potential of this approach for accelerating materials discovery and design processes.

Furthermore, our multi-property prediction approach offers several advantages:

- **Efficiency**: By predicting multiple properties simultaneously, we reduce the computational overhead compared to training separate models for each property [10].
- **Implicit Feature Learning**: The shared layers of the GNN can learn features that are relevant to multiple properties, potentially improving overall prediction accuracy [2].
- **Correlation Insights**: The multi-property approach allows us to investigate correlations between different material properties, potentially revealing new insights into structure-property relationships [8].

In the following sections, we will detail our methodology, including data preparation, model architecture, and training procedures. We will then present our results, analyzing the performance of our GNN model across different properties and material types. Finally, we will discuss the implications of our work, its limitations, and potential future directions for applying GNNs in computational materials science.

2 Methodology

Our approach combines statistical analysis of material properties with a Graph Neural Network (GNN) model for property prediction. The methodology encompasses data collection, statistical analysis, and model development.

2.1 Data Collection and Preprocessing

Our study utilizes a comprehensive dataset sourced from the Materials Project database [1], comprising a total of 158,874 crystal structures. These structures span across various crystal systems, providing a diverse representation of material properties. The dataset includes 26,034 cubic structures, 10,521 hexagonal structures, 12,844 trigonal structures, 17,203 tetragonal structures, 32,206 orthorhombic structures, 36,000 monoclinic structures, and 24,066 triclinic structures. This diverse collection ensures a broad coverage of structural variations in materials science.

- Node features: One-hot encoded atomic numbers
- Edge indices: Connections between atoms based on their spatial relationships
- Edge features: Distances between connected atoms

This approach allows us to capture the structural information of materials in a format suitable for graph-based machine learning [9].

The valid y labels in this study represent a comprehensive set of material properties that our Graph Neural Network model aims to predict. These properties are crucial for understanding the structural, energetic, and stability characteristics of crystalline materials. The predicted properties include volume, which refers to the total space occupied by the unit cell of the crystal; number of sites (nsites), indicating the total atomic positions within that unit cell; and density, which measures the mass per unit volume of the material. Additionally, crystal number provides a numerical identifier for the crystal system, while space group number uniquely characterizes the symmetry of the crystal structure. The model also predicts several energy-related properties: uncorrected energy per atom, which is the total energy divided by the number of atoms without corrections; energy per atom, which may include corrections for accuracy; formation energy per atom, indicating the energy required to form the compound from its constituent elements; and energy above hull, which measures how far the compound's energy is above the convex hull of stability. Finally, the label is stable serves as a binary indicator of whether the structure is considered thermodynamically stable. Together, these properties–volume, nsites, density, crystal number, space group number, uncorrected energy per atom, energy per atom, formation energy per atom, energy above hull, and is stable–provide a multifaceted view of a material's characteristics, enabling comprehensive predictions that can accelerate materials discovery and design processes in various fields of materials science and engineering.

2.2 Statistical Analysis

For each material property y_i, we calculate descriptive statistics including mean (μ), median, standard deviation (σ), variance (σ^2), skewness, and kurtosis:

$$\mu = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2}$$

To focus on the core distribution, we filter out minority data using the 5th and 95th percentiles. We then analyze the relationships between properties by computing covariance and correlation matrices:

$$\text{Cov}(X, Y) = E[(X - E[X])(Y - E[Y])]$$

$$\text{Corr}(X, Y) = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y}$$

This statistical analysis provides insights into the distribution and interrelationships of material properties, guiding our modeling approach [19].

2.3 Graph Neural Network Model

Our GNN model, implemented using PyTorch Geometric [16], consists of multiple Graph Convolutional Network (GCN) layers [17], followed by global mean pooling and a final linear layer for prediction. The model architecture is defined as:

$$h^{(l+1)} = \sigma(Ah^{(l)}W^{(l)})$$

where $h^{(l)}$ is the node feature matrix at layer l, A is the adjacency matrix, $W^{(l)}$ is the weight matrix, and σ is the ReLU activation function. The GNN class is implemented as follows:

```
class GNN(torch.nn.Module):
def init(self, input_dim, hidden_dim, num_layers, dropout, output_dim):
super(GNN, self).init()
self.convs = nn.ModuleList([GCNConv(input_dim if i == 0 else hidden_dim, hidden_dim)
for i in range(num_layers)])
self.lin = nn.Linear(hidden_dim, output_dim)
self.dropout = dropout
text
def forward(self, x, edge_index, batch):
mask = (x.sum(dim=-1) != 0)
```

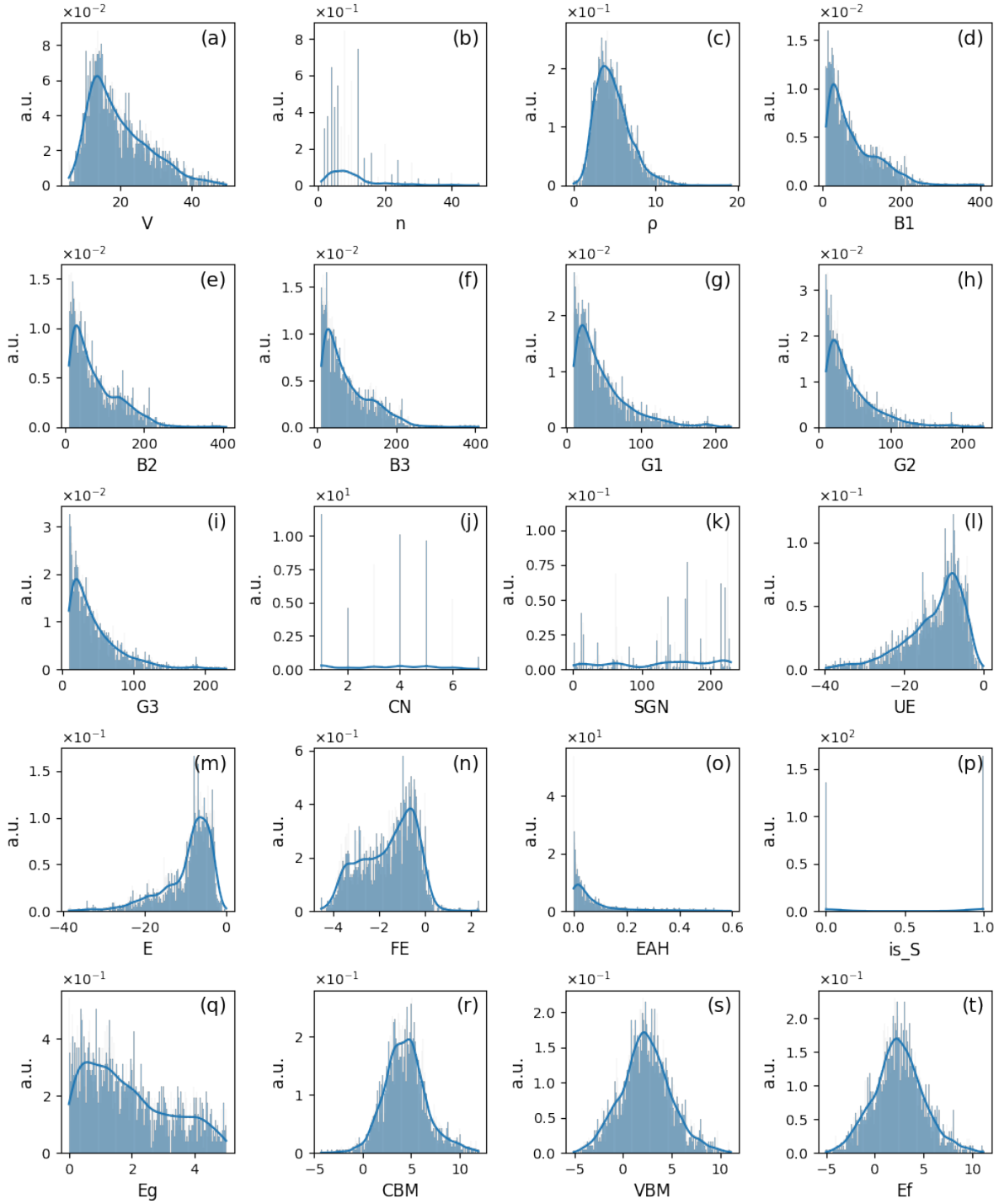

Figure 1: Distribution histograms for various material properties: (a) volume **(V)**, (b) density**(*ρ*)** (c) number of sites in a crystal **(n)**, (d-f) bulk moduli **(B1, B2, B3)**, (g-i) shear moduli **(G1, G2, G3)**, (j) crystal structure number **(CN)**, (k) space group number **(SGN)**, (l) uncorrected energy **(UE)**, (m) energy **(E)**, (n) formation energy **(FE)**, (o) energy above hull **(EAH)**, (p) stable or unstable **(is_S)**, (q) band gap energy **(Eg)**, (r) conduction band minimum **(CBM)**, (s) valence band maximum **(VBM)**, and (t) Fermi level **(Ef)**. The filtering process ensures that each histogram reflects realistic data ranges, thus highlighting general trends and variances within each property.

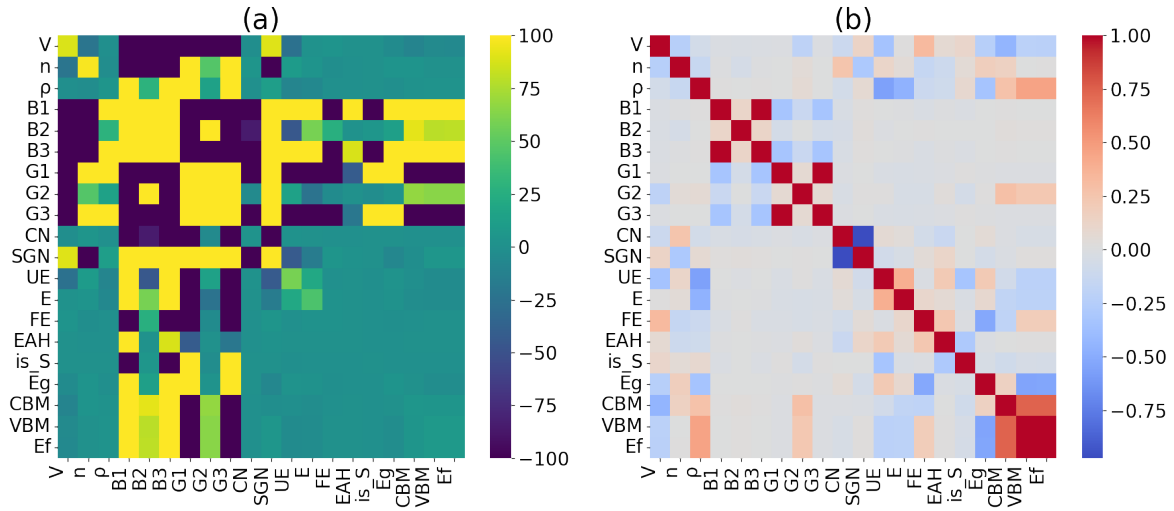

Figure 2: Covariance Matrix illustrating the variability among features. Data filtering enhances the interpretability of statistical dependencies across properties, offering insights into property interdependencies. Correlation Matrix showing pairwise correlations among properties. The matrix reveals significant relationships between properties, with filtering improving the reliability of these correlations by reducing noise from extreme values.

```
for conv in self.convs:
x = F.relu(conv(x, edge_index))
x = F.dropout(x, p=self.dropout, training=self.training)
x = global_mean_pool(x[mask], batch[mask])
x = self.lin(x)
return x
```

2.4 Model Training and Evaluation

The model is trained by splitting the dataset into 80% training and 20% validation sets using sklearn's train-test split function [20]. We initialize the GNN model with an input dimension that matches the node features, a hidden dimension of 200, three GCN layers, and a dropout rate of 5e−5. The training runs for a maximum of 1000 epochs, employing early stopping with a patience of 50 epochs to prevent overfitting [21]. We use the Adam optimizer [22] with a learning rate of 1e-4 and Mean Squared Error (MSE) as the loss function.

During training, forward propagation through the GNN, backpropagation of the loss, and optimization of model parameters are carried out. Both training and validation losses are monitored, along with the R^2 scores for each predicted property. To facilitate further analysis, we record training progress through loss and R^2 score curves and save detailed results in CSV files [10, 2]. This comprehensive approach enables a deep understanding of material properties' statistical characteristics while building a predictive model that captures complex relationships within crystal structures.

3 Results and Discussion

The statistical analysis and visualization performed in this study provide key insights into the distribution and correlations among various properties of crystal structures. The figures generated include distribution histograms, a correlation matrix, and a covariance matrix, each illustrating the unique statistical characteristics of specific properties and aiding in the interpretation of material properties.

3.1 Data Filtering and Preprocessing

The data filtering process implemented in the Python code ensures that only relevant and meaningful data points are included in the statistical analysis. Conditional filters were applied to each material property to remove extreme or irrelevant values. For properties like volume **(V)** , number of sites**(n)** and density**($\boldsymbol{\rho}$)**, the values are restricted to physically meaningful ranges to eliminate outliers. In the case of

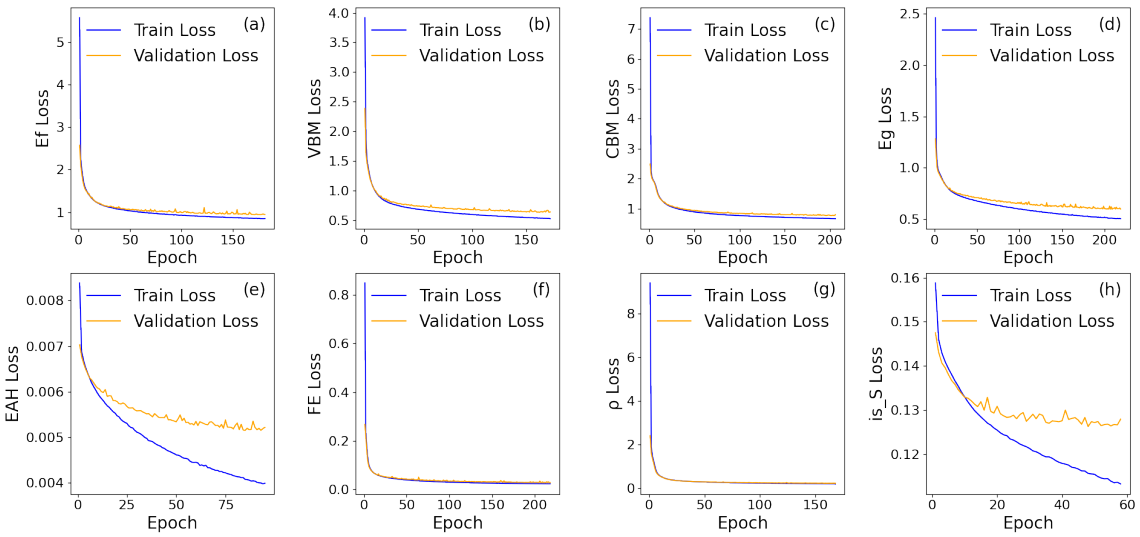

Figure 3: This plot is illustrating the training and validation loss for various material properties across epochs. Each subplot represents a different property ((a) Ef, (b) VBM, (c) CBM, (d) Eg, (e) EAH, (f) FE, (g) ρ, (h) is_S), showcasing the training loss (in blue) and validation loss (in orange) over the course of the training process. This visualization aids in assessing the model's performance and identifying potential issues such as overfitting or underfitting.

mechanical properties, such as bulk moduli **(B1 - Voigt, B2 - Reuss, B3 - VRH)** and shear moduli **(G1 - Voigt, G2 - Reuss, G3 - VRH)**, only values within typical ranges are included, ensuring that the data represents realistic material behaviors. Additionally, the analysis includes properties such as crystal structure number **(CN)**, space group number **(SGN)**, uncorrected energy **(UE)**, energy **(E)**, formation Energy **(FE)**, energy above Hull **(EAH)**, stable or unstable **(is_S)**, band gap energy **(Eg)**, conduction band minimum **(CBM)**, valence band maximum **(VBM)**, and Fermi level **(Ef)**. All the data ranges are shown as a distribution plot in Figure 1.

This filtering is essential for generating clear and interpretable visualizations, as it focuses the analysis on data that accurately reflects the underlying trends and reduces the influence of anomalous values.

3.2 Properties selection

The data filtering strategy used in this analysis is crucial for focusing on meaningful data ranges and avoiding skewed results due to outliers. By applying specific filtering criteria to each property, the resulting figures—distribution histograms, correlation matrix, and covariance matrix—present a clearer view of the dataset's statistical properties. This approach provides a robust foundation for further analysis and predictive modeling of material properties, highlighting key trends and dependencies within the dataset.

Figure 2 presents the covariance and correlation matrices, which illustrate the relationships and dependencies among various material properties. The covariance matrix highlights the variability among features, while the correlation matrix provides insights into pairwise correlations. The data filtering applied during the preprocessing stage enhances the interpretability of these statistical dependencies by reducing noise from extreme values, thereby improving the reliability of the observed correlations.

Based on the insights gained from these matrices, we select the following features for training and prediction: $\boldsymbol{\rho}$, **FE**, **Ef**, **Eg**, **VBM**, **CBM**, **EAH**, and **is_S**.

The selection of these specific properties is motivated by their significant relationships observed in the correlation matrix, which indicates their potential interdependencies. Utilizing these features will enable the development of robust predictive models that effectively capture the underlying material behavior.

3.3 Loss Analysis

In addition to the feature selection based on covariance and correlation analyses, it is crucial to monitor the training process of the predictive models. The combined loss plot presented in Figure 3 visualizes the training and validation loss across multiple labels throughout the training epochs.

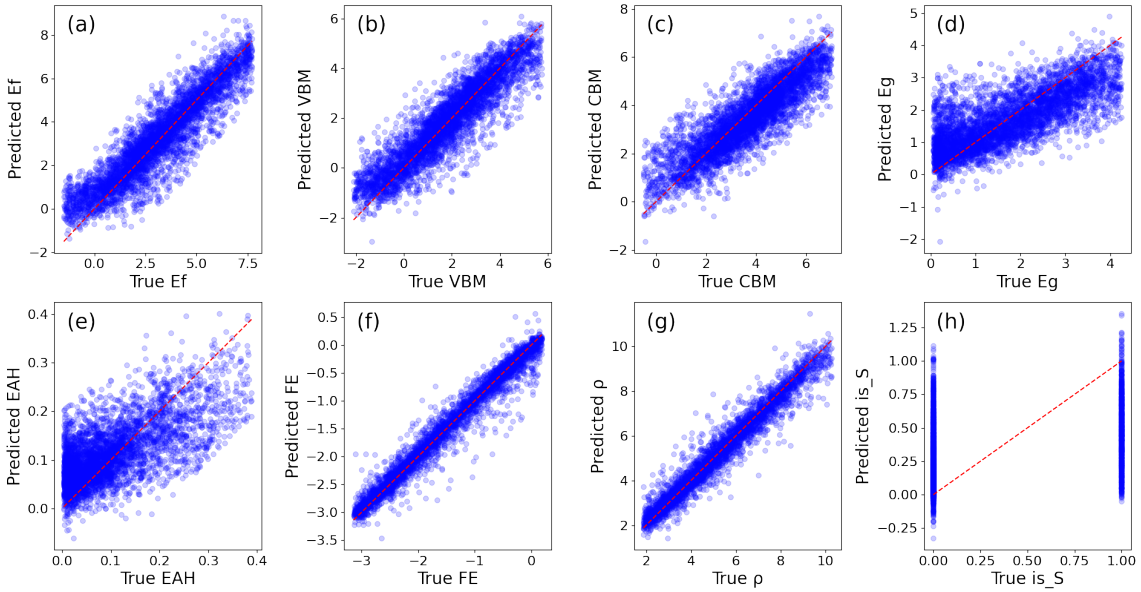

Figure 4: True vs. Predicted values for different material properties: (a) Ef, (b) VBM, (c) CBM, (d) Eg, (e) EAH, (f) FE, (g) ρ, (h) is_S. Each plot includes a dashed red line representing the ideal prediction scenario, enhancing the understanding of model performance for each property.

The training loss represents the error on the training dataset, while the validation loss indicates the model's error on unseen data. A key aspect of effective model training is achieving a balance between these two metrics. Ideally, both training and validation losses should decrease over epochs, converging towards a lower value. Analyzing these loss plots enables us to evaluate convergence, as a decreasing trend in both losses signifies that the model is learning effectively. Conversely, if the training loss continues to decrease while the validation loss begins to increase, it suggests that the model may be overfitting the training data. On the other hand, if both losses remain high and do not improve significantly, it indicates that the model is underfitting and failing to capture the underlying patterns in the data. By closely monitoring these metrics during the training process, we can make informed decisions regarding model tuning and optimization strategies. This proactive approach ultimately contributes to the development of robust predictive models that generalize well to new data.

3.4 True vs. Predicted Values

The plots presented in Figure 4 illustrate the relationship between true and predicted values for various material properties. Each subplot corresponds to a different property, allowing for a detailed comparison of model performance. The labels considered for this analysis are**Ef**, **VBM**, **CBM**, **Eg**, **EAH**, **FE**, $\boldsymbol{\rho}$, and **s_S**. In each subplot of Figure 4 , blue points represent the true vs. predicted values, while the dashed red line denotes the ideal scenario where the predicted values match the true values perfectly (i.e., $y = x$). The degree to which the points cluster around this line indicates the model's accuracy. Notably, the thresholds applied for filtering close points vary for specific labels, enhancing the interpretability of the plots.

The evaluation of model performance for various material properties is encapsulated by the R^2 values, which indicate the proportion of variance in the predicted values that can be explained by the true values. Notably, the density shows an excellent fit with an R^2 value of 0.96, suggesting that the model reliably captures the relationship between predicted and true density values. Similarly, the formation energy (FE) demonstrates a high R^2 value of 0.97, indicating a strong predictive capability.

In contrast, the energy above hull (EAH) and stable or unstable (is_S) properties exhibit significantly lower R^2 values of 0.54 and 0.47, respectively. This suggests that the model struggles to accurately predict these properties, potentially due to inherent complexities or noise in the data. The band gap energy (Eg) shows a moderate R^2 value of 0.76, reflecting a fair level of predictive accuracy, while the valence band maximum (VBM) and conduction band minimum (CBM) achieve R^2 values of 0.86 and 0.78, respectively, indicating a reasonable fit.

Finally, the Fermi level (Ef) yields an R^2 value of 0.82, suggesting a good predictive performance,

albeit slightly less than that of density and formation energy. Overall, the high R^2 values for density and formation energy highlight the model's robustness in predicting these fundamental properties, while the lower values for EAH and is_S point to areas that may require further investigation or refinement in model training and feature selection.

4 Conclusion

In this study, we have successfully demonstrated the application of Graph Neural Networks (GNNs) for predicting structural and electronic properties of materials using data from the Materials Project database. Our approach involved constructing graph representations of crystal structures, allowing the model to capture complex relationships between atomic arrangements and material properties. The R^2 values obtained for various properties indicate the efficacy of our model, with particularly high performance in predicting density ($R^2 = 0.96$), formation energy ($R^2 = 0.97$), and band gap ($R^2 = 0.76$). Although the model showed lower predictive accuracy for properties such as energy above hull ($R^2 = 0.54$) and structural stability (is_S, $R^2 = 0.47$), the results still provide valuable insights into the relationships among material properties.

These findings highlight the potential of GNNs as a powerful tool for high-throughput computational materials science, facilitating the rapid screening and discovery of materials with desirable characteristics. Future work will focus on refining the model and exploring its applicability to a broader range of materials and properties, ultimately advancing our understanding of materials and aiding in the development of innovative solutions in various applications.

5 Code availability

Data processing, deep learning model train and test has been released in the Github repository at https://github.com/selvachandrasekaranselvaraj/GNN_prediction

Conflict of Interest

The authors declare no conflict of interest.

References

[1] Anubhav Jain, Shyue Ping Ong, Geoffroy Hautier, Wei Chen, William Davidson Richards, Stephen Dacek, Shreyas Cholia, Dan Gunter, David Skinner, Gerbrand Ceder, et al. Commentary: The materials project: A materials genome approach to accelerating materials innovation. *APL materials*, 1(1):011002, 2013.

[2] Victor Fung, Jiaxin Zhang, Eric Juarez, and Bobby G Sumpter. Benchmarking graph neural networks for materials chemistry. *npj Computational Materials*, 7(1):84, 2021.

[3] Selva Chandrasekaran Selvaraj, Sameer Gupta, Damien Caliste, and Pascal Pochet. Passivation mechanism in CdTe solar cells: The hybrid role of Se. *Applied Physics Letters*, 119(6):062105, 08 2021.

[4] Selva Chandrasekaran. Selvaraj, P. Murugan, P. Saravanan, and S. V. Kamat. Effect of adatom deposition on surface magnetism and exchange coupling parameter in (0001) SmCo5 slabs. *Journal of Applied Physics*, 117(13):133902, 04 2015.

[5] S. Selva Chandrasekaran, M. Rama Ponnaiah, P. Murugan, and P. Saravanan. Magnetic and electronic properties of hard—soft magnetic interface in (yco5—co)[0001] and (yfe5|co)[0001] superlattices. *Journal of Magnetism and Magnetic Materials*, 418:92–98, 2016. Peer-reviewed papers from International Conference on Magnetic Materials and Applications.

[6] S. Selva Chandrasekaran and P. Murugan. Structural and electronic properties of solid-state (limpo4|γ-li3po4)[010] electrochemical interface (m=fe and co). *Applied Surface Science*, 418:17–21, 2017. Asian Consortium on Computational Materials Science - Theme Meeting on "First Principles Analysis Experiment Role in Energy Research".

[7] S. Selva Chandrasekaran, P. Murugan, P. Saravanan, and S. V. Kamat. Surface termination dependent structural and magnetic properties of (0001) smco5 slabs. *physica status solidi (b)*, 250(9):1883–1887, 2013.

[8] Chi Chen, Weike Ye, Yunxing Zuo, Chen Zheng, and Shyue Ping Ong. Graph networks as a universal machine learning framework for molecules and crystals. *Chemistry of Materials*, 31(9):3564–3572, 2019.

[9] Tian Xie and Jeffrey C Grossman. Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties. *Physical review letters*, 120(14):145301, 2018.

[10] Zhantao Wang, Tian Xie, Prashun Gorai, Vladan Stevanovic, and Xiaoqin Yan. Multi-task learning of crystal properties with graph neural networks. *npj Computational Materials*, 8(1):1–9, 2022.

[11] Selva Chandrasekaran Selvaraj, Volodymyr Koverga, and Anh T. Ngo. Exploring li-ion transport properties of li3ticl6: A machine learning molecular dynamics study. *Journal of The Electrochemical Society*, 171(5):050544, may 2024.

[12] Robin Ruff, Patrick Reiser, Jan Stühmer, and Pascal Friederich. Connectivity optimized nested line graph networks for crystal structures. *Digital Discovery*, 3:594–601, 2024.

[13] Chuan-Nan Li, Han-Pu Liang, Xie Zhang, Zijing Lin, and Su-Huai Wei. Graph deep learning accelerated efficient crystal structure search and feature extraction. *npj Computational Materials*, 9(1):176, 2023.

[14] Chuan-Nan Li, Han-Pu Liang, Xie Zhang, Zijing Lin, and Su-Huai Wei. Graph deep learning accelerated efficient crystal structure search and feature extraction. *npj Computational Materials*, 9(1):176, 2023.

[15] Jiucheng Cheng, Chunkai Zhang, and Lifeng Dong. A geometric-information-enhanced crystal graph network for predicting properties of materials. *Communications Materials*, 2(1):92, 2021.

[16] Matthias Fey and Jan Eric Lenssen. Fast graph representation learning with pytorch geometric. *arXiv preprint arXiv:1903.02428*, 2019.

[17] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016.

[18] Michael Schlichtkrull, Thomas N Kipf, Peter Bloem, Rianne Van Den Berg, Ivan Titov, and Max Welling. Modeling relational data with graph convolutional networks. *European semantic web conference*, pages 593–607, 2018.

[19] Logan Ward, Ankit Agrawal, Alok Choudhary, and Christopher Wolverton. A general-purpose machine learning framework for predicting properties of inorganic materials. *npj Computational Materials*, 2(1):1–7, 2016.

[20] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[21] Lutz Prechelt. Early stopping-but when? *Neural Networks: Tricks of the trade*, pages 55–69, 1998.

[22] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.