CRYSTAL STRUCTURE BASED AI/ML APPROACH FOR MATERIAL MAGNETIC ORDERING PREDICTION CONFERENCE SUBMISSIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

AI/ML models are useful for the prediction of materials' properties. Recently, there has been great progress in developing Neural Network models to predict the properties of solids, molecules, and crystals, avoiding expensive and time-consuming laboratory testing. This paper studies both Graph Neural Network (GNN) and Large Language Model (LLM)-based models to predict material magnetic ordering in four classes: Non-magnetic, Ferromagnetic, Antiferromagnetic, and Ferrimagnetic. Material crystal structure information is fed into a GNN and also converted to semantic text descriptions and tokenized as input to a LLM. The models are trained, validated, and tested on a dataset which is a combination of Materials Project dataset, Magndata dataset, and a set of uranium compounds with magnetic ordering determined in laboratory testing, with an 80/10/10 (training/validation/testing) data split. The study features two stage training process to train feature extraction and classification separately, which is necessary to handle class imbalance in the dataset due to limited availability of data for certain classes (AFM and FiM). In the first stage, the feature extraction is trained; in the second stage, the feature extraction is frozen and the classification head is trained for the final classification task. Based on the two stage training, the proposed LLM and ALIGNN-LLM multi-modal approaches can achieve an average 90% overall accuracy on the testing dataset across the four classes, and high (80% - 90%) per-class precision, recall and F1-macro scores.

1 INTRODUCTION

Material properties are critical to engineering since they indicate how materials perform under different thermal, mechanical and electrical conditions. However, experimental determination of material properties is time-consuming and costly. To mitigate this problem, Density Functional Theory (DFT) was proposed in Callaway & March (1984); Parr & Yang (1989); March (1999) to mathematically model quantum mechanical behavior of atoms and predict material properties. Though this mathematical approach is very popular, its prediction accuracy is heavily dependent on known approximation functions, and it cannot simulate complicated structures accurately.

Recent research has focused on the development of AI/ML Graph Neural Network models to accurately predict material properties, reducing traditional laboratory testing. Schütt et al. (2017) created SchNet, one of the first GNNs capable of modeling local interactions in the molecule without requiring atomic positions to lie on a fixed grid (a challenge faced by convolution neural networks (CNN)). These interactions were captured using interaction blocks that aggregated information from neighboring atoms and would be weighted atom-wise to produce a final prediction. Xie & Grossman (2018) introduced the Crystal Graph Convolutional Neural Network (CGCNN), which adapted the previous framework for crystalline compounds, encoding atoms as nodes and bonds as edges in a crystal graph and producing predictions that match the accuracy of DFT-computed properties across a broad range of crystal types. Louis et al. (2020) introduced the Global Attention Graph Neural Network (GATGNN) model, which further differentiates the contributions of individual atoms by introducing a global attention layer and augmenting graph-attention layers and incorporates temperature, pressure, and entropy as global state inputs. In 2021, Choudhary & DeCost (2021) introduced

the Atomistic Line Graph Neural Network (ALIGNN) model, one of the first models to effectively incorporate bond angle data as input, which achieved top prediction precision at time of proposal.

Since the introduction of transformer architecture, the huge success of transformer-based large language models (LLMs) proposed by Vaswani et al. (2017) to process language-based problems demonstrates the potential of utilizing a transformer-based approach to tackle problems in the domain of material properties prediction. Yan et al. (2024) proposed a transformer designed specifically for crystalline materials, employing invariant geometric descriptors of Euclidean distances and angles and equivariant vector representation. Walker et al. (2021) proposed MatBERT (Bidirectional Encoder Representations from Transformers (BERT)), which uses contextualized word representations and semantically rich sentence embeddings for material prediction. Rubungo et al. (2023) proposed LLM-Prop, which leverages the general purpose learning capabilities of large language models (LLMs) to predict properties of crystals from their text descriptions. Li et al. (2025) proposed a hybrid approach to use pretrained GNN and LLM models to encode material information including crystal structure and a tree-based ML regressor to perform property prediction.

In this paper, we present models incorporating GNN and LLM to tackle the problem of material magnetic ordering prediction using material compound and crystal information which was studied in Broyles et al. (2024). The dataset used in the paper is a combination of compounds from Material Project, Magndata and uranium compounds measured in lab testing. The class samples are highly imbalanced, which poses a challenging problem for existing GNN and LLM models. A standalone ALIGNN model, standalone LLM (GPT-2 was selected for this study) and ALIGNN-LLM multi-modal approach inspired by Li et al. (2025) were selected as baselines. ALIGNN is used to extract information from atoms and crystal structure, while LLM extracts information from material semantic text descriptions. The outputs of the GNN and LLM are then fed, individually or fused, into the classification block. To improve classification precision on minority classes AFM and FiM, a two stage process training feature extraction and classification separately was used and results show this approach significantly improves performance on minority classes relative to baselines. Moreover, LLM is shown to perform better than ALIGNN at feature extraction. The LLM and ALIGNN-LLM multi-modal approaches show the best overall accuracy and per-class precision, recall and F1 scores using the two stage training process.

2 METHODS

2.1 MODEL ARCHITECTURE

The model architecture includes LLM and ALIGNN as feature extractors and a classification head to perform classification. ChemNLP is used to convert material compound and crystal structure information into text descriptions and ALIGNN converts material compound and crystal structure information into a graph and line graph representation. The ALIGNN and LLM output are then fed, individually or fused, to the classification head. The model architecture is shown in Fig. 1.

2.2 ALIGNN FEATURE EXTRACTOR

Material structure information is first extracted in Pymatgen Structure JSON format and then encoded using Jarvis Atom functionality jar. Then, it is converted to a graph and line graph (which is the required input format for ALIGNN) via edge and angle embedding using RBF expansion and MLP (Multi-Layer Perceptron). ALIGNN includes the following two graphs: a Graph Convolutional Network (GCN) using atoms as nodes and bonds as edges, and a line GCN using bonds as nodes and bond angle triplets as edges. In this manner, more structural information can be captured and used to predict more accurately.

Taking MP-149 (silicon) as an example, for the GCN, atoms serve as the nodes and atomic bonds are the edges. CGCNN (Crystal Graph Convolutional Neural Network) embedding is used to represent each atom, and provides a 92-dimensional feature vector for each atom based on its properties from the periodic table. There is one node per Si atom; thus, the atom features have dimensions [2, 92]. Each atom has 28 neighbors within a cutoff of 5Å, and for each neighbor there is one directed ingoing and outgoing edge; hence, there are 56 directed bonds for each atom, and 56 edges per node for the GCN.

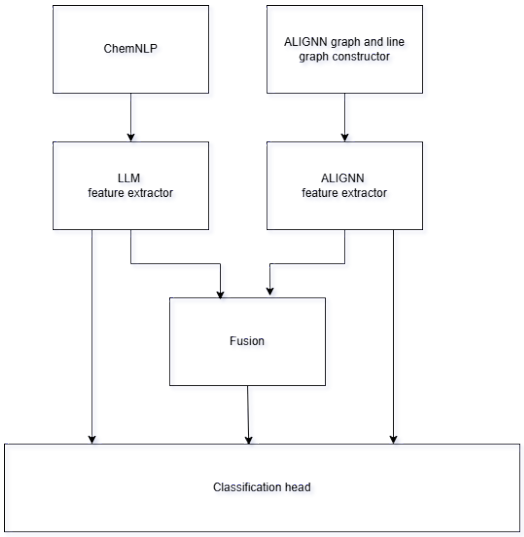

Figure 1: Architecture of ALIGNN, LLM and multi-modal classifiers

For the line GCN, the atom bonds act as the nodes and angle triplets act as the edges. To determine the number of angle triplets, each atom has 28 incoming and outgoing bonds, and thus the number of bonds including each atom is 784. Hence, for each bond (which includes two atoms), the number of angle triplets is $784 \cdot 2 = 1568$; the total is 1544 after removing duplicates for both atoms.

The constructed graphs are then fed to ALIGNN, which acts as a feature extractor. (The full ALIGNN algorithm is shown in Fig. 2; please refer to Choudhary & DeCost (2021) for details.)

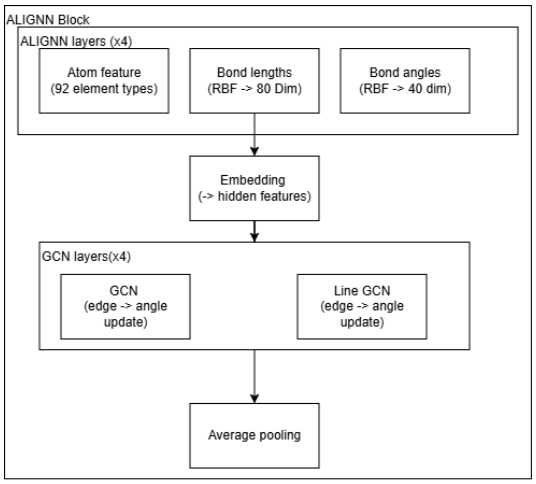

Figure 2: ALIGNN architecture

2.3 CHEMNLP

The Material compound is first converted to a text description based on the ChemNLP implementation in che. As an example, the information of MP-149 from Materials Project is converted to the following text description:

"The chemical information includes: The chemical has an atomic formula of Si with a prototype of A; Its molecular weight is 28.09 g/mol; The atomic fractions are Si: 1.0, and the atomic values X and Z are 1.9 and 14, respectively. The structure information includes: The lattice parameters

are 3.85, 3.85, 3.85 with angles 60.0, 60.0, 60.0 degrees; The top K XRD peaks are found at 28.4, 47.2, 56.0, 69.0, 76.2 degrees; The material has a density of 2.313 g/cm³; The bond distances are as follows: Si-Si: 2.36, 3.85."

2.4 LLM (GPT-2) FEATURE EXTRACTOR

In this study, the GPT-2 model gpt (Fig. 3) is used in the LLM block. During testing, the base GPT-2 model with 124.64M parameters was found to be capable of handling the size of the dataset and complexity of the classification problem. More complicated LLMs such as Meta-Llama met with 7B to 65B parameters can be used if a larger dataset is available and/or if the model is used to address more complex multi-property regression prediction problems.

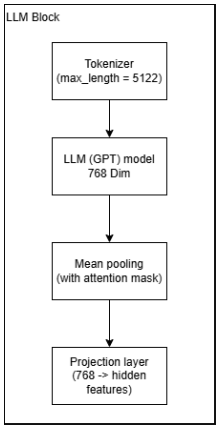

Figure 3: GPT-2 architecture

2.5 INFORMATION FUSION AND CLASSIFICATION BLOCK

The information fusion block is used to combine the extracted features of ALIGNN and LLM when using multi-model approach. The LLM and ALIGNN are concatenated together along the innermost dimension (size of 256). After concatenation, the innermost dimension has a size of 512, which is then reduced to 256 via a linear (fully connected) layer.

The classification block takes the fused information from the information fusion block and performs the following computations sequentially: Linear (fully connected) layer with layer normalization and SiLU activation, Linear (fully connected) layer to output result vector in size of 4, and Multinomial logistic regression (softmax) to obtain the final classification result.

3 DATA PREPARATION AND TRAINING

3.1 DATA PREPARATION

The data preparation processing is shown in Fig. 4.

A preliminary dataset consisting of 210,579 compounds was extracted from 3 sources: database.xlsx from Broyles et al. (2024), Materials Project from mpr, and Magndata from mag. database.xlsx includes 1064 compounds, Magndata includes 231 compounds, and MP includes 209,284 compounds. When there is duplication on certain material compounds in different sources, experimentally-verified data is prioritized over Materials Project data.

Compounds without an indicated magnetic ordering (i.e. non-magnetic, ferromagnetic, ferrimagnetic, or antiferromagnetic) were filtered out. The remaining dataset contained 200,431 compounds, with the following distribution of magnetic orderings as shown in Table 1:

As seen in Table 1, the distribution of data is unequal - there is significantly more data for the non-magnetic class than the others. For classification problems, optimal performance of AI/ML

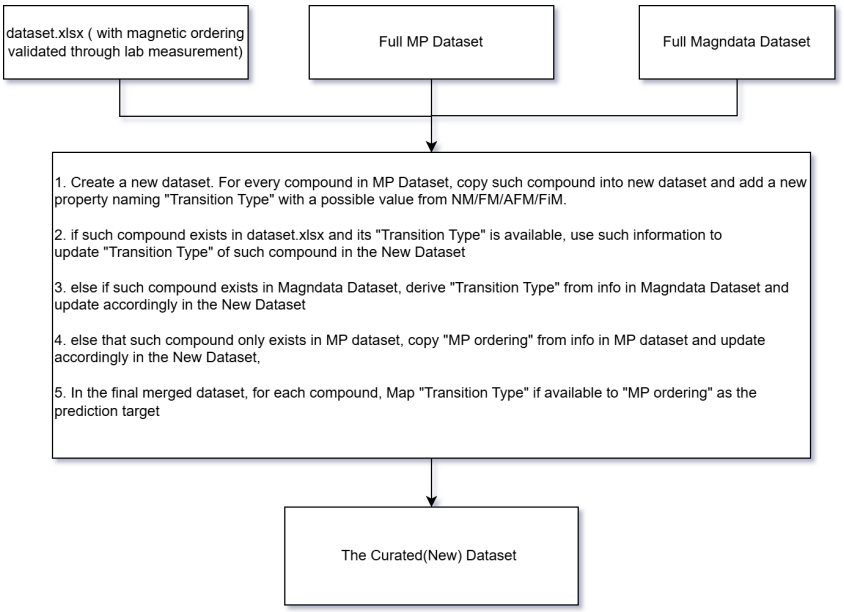

Figure 4: Data preparation

Table 1: Distribution of data in final dataset

Magnetic Ordering	**Number of data entries**	**Proportion of complete dataset**
All (across all 4 classes)	200,431	100%
Non-magnetic	119,225	59.48%
Ferromagnetic	64,253	32.06%
Antiferromagnetic	3,709	1.85%
Ferrimagnetic	13,244	6.61%

approaches requires a relatively balanced proportion of classes in the dataset. This poses a big challenge and the two stage training process is introduced to tackle this problem effectively. Furthermore, magnetic ordering information contained in database.xlsx and Magndata was experimentally determined, whereas it was determined via DFT+U calculations for Materials Project. As DFT+U calculations are approximations, it is less accurate than experimental validation. However the laboratory validated materials only have a total of 1295 samples which would significantly negatively impact the statistical meaning of this study if the experiment is limited to these samples alone.

3.2 Training Procedure

The full dataset is split into training, validation, and testing sets with 80%, 10%, and 10% of total dataset used for each respectively; the percentage of classes in each are identical as in the full dataset. During training, the training and validation accuracy are evaluated after each epoch and the model with the highest validation accuracy up to that epoch is saved. After training, the best model is tested on the testing dataset to determine its final classification performance. Training/validation/testing were performed on DUAL Nvidia 3090 GTX GPU.

Training lasts for 30 epochs, with AdamW as the optimizer. Class-weighted Cross Entropy Loss is the loss function, with class weights computed inversely based on the percentage of NM/FM/AFM/FiM respectively (the smaller the sample size, the larger the class weights in loss computation - this penalizes errors in minority classes more to address imbalance in the dataset). The empirical value of learning rate are 0.001 for ALIGNN and 0.0001 for LLM and ALIGNN-LLM as it is found ALIGNN needs a larger learning rate to converge.

Table 2: Per-class precision, recall and F1 for GNN models On Test Dataset

Class	ALIGNN (P/R/F1)	CGCNN v2019 (P/R/F1)	coGN (P/R/F1)
Non-magnetic	96%/88%/91.8%	93%/92%/93%	94%/93%/93%
Ferromagnetic	74%/83%/78.2%	79%/64%/70%	79%/61%/69%
Antiferromagnetic	32%/42%/36.3%	16%/64%/26%	16%/53%/24%
Ferrimagnetic	54%/55%/54.5%	47%/59%/52%	40%/64%/49%

4 Results and Analysis

4.1 GNN Model Performance Validation

To evaluate ALIGNN performance on the problem of the paper, two others GNN models, CGCNN v2019 Xie & Grossman (2018) and coGN Ruff et al. (2024), which are top performing models from the Materials Project matbench v0.1 matbench_mp_is_metal task leaderboard Dunn et al. (2020) were tested using a gcnn_keras implementation Reiser et al. (2021). The aforementioned training/validation/test data splits and training configuration were used. As shown in Table 2, ALIGNN shows better overall performance (F1) than the other two.

4.2 Baseline Models - Performance Validation

Training and validation results for ALIGNN-only, LLM-only and ALIGNN-LLM multi-modal approach are presented in Fig. 5, Fig. 6 and Fig. 7. All three approaches converge rapidly: the training and validation accuracy quickly exceed 90% and 80% respectively and plateau at around 30 epochs.

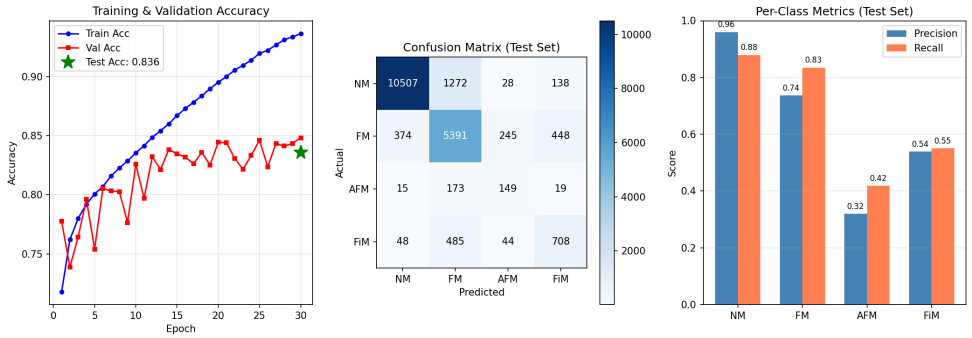

Figure 5: ALIGNN training, validation and testing results

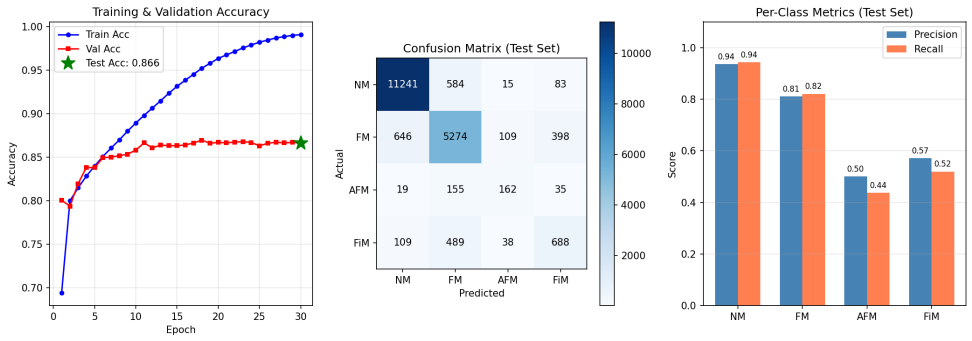

Figure 6: LLM training, validation and testing results

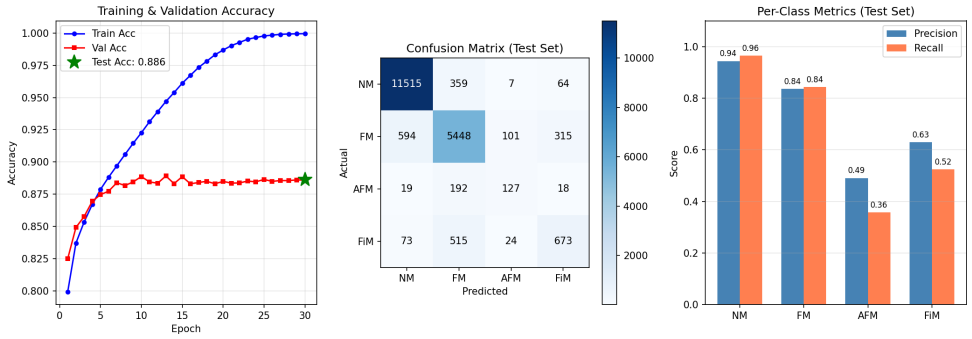

Figure 7: ALIGNN-LLM multi-modal training, validation and testing results

The testing results for each model are shown in Fig. 5, Fig. 6 and Fig. 7 with confusion matrix, precision and recall for each class.

The ALIGNN-LLM multi-modal approach achieves the best overall model accuracy across all four classes, obtaining an accuracy of 88.6% compared to 83.6% from ALIGNN alone and 86.6% from LLM alone.

The per-class testing precision, recall and F1 are shown in Table 3. As demonstrated, the ALIGNN-LLM multi-modal approach achieves comparable to better classification precision and recall compared to standalone ALIGNN and LLM across most classes (with the exception of the AFM class). Of the four classes, AFM has the lowest classification precision and recall for each model.

Table 3: Per-class precision, recall and F1 for all three models on test dataset

Class	ALIGNN-only	LLM-only	ALIGNN-LLM multi-modal
Non-magnetic	96%/88%/91.8%	94%/94%/94.0%	94.8%/96.1%/95.4%
Ferromagnetic	74%/83%/78.2%	81%/82%/81.5%	83.3%/85.2%/84.2%
Antiferromagnetic	32%/42%/36.3%	50%/44%/46.8%	46.9%/36.5%/41.1%
Ferrimagnetic	54%/55%/54.5%	57%/52%/54.4%	64.2%/53.4%/58.3%

Three forms of ”randomly guessing” labels are investigated: the accuracy of ”Uniform Guessing”, ”Proportional Guessing” and ”Zero Rule” can be calculated as 25%, 46.13% and 59.48% respectively (see Section A). This clearly demonstrates the strength of the models over random guessing.

4.3 Two Stage Training Performance Validation

To tackle the problem of lowest classification precision and recall on minority classes, a two stage training process is proposed in the paper: in the first stage, the LLM and ALIGNN are trained as feature extractors using identical training settings as in Section 4.2 for 30 epochs. In the second stage, the LLM and ALIGNN feature extractors are frozen and its output is used to train the classification head for five epochs. For multi-modal approach, both feature extractors are frozen and only the classification head is trained.

Table 4: Per-class precision, recall and F1 for all three models (2 stage trained) on test dataset

Class	ALIGNN-froze	LLM-frozen	ALIGNN-LLM frozen
Non-magnetic	92.7%/94.6%/93.7%	98.9%/98.8%/98.8%	98.0%/99.0%/98.5%
Ferromagnetic	80.9%/90.0%/80.5%	96.3%/96.4%/96.3%	95.4%/94.9%/95.2%
Antiferromagnetic	37.1%/33.4%/35.2%	91.3%/89.0%/90.2%	88.7%/76.9%/82.4%
Ferrimagnetic	60.2%/53.8%/56.8%	90.5%/91.6%/91.0%	88.9%/86.5%/87.7%

The accuracy of trained models is evaluated against the test dataset. The results are shown in Table 4: frozen LLM after two stage training performs the best, with significant improvement on minority classes AFM and FiM relative to the baseline. Contrastingly, frozen ALIGNN after two stage training provides little to no improvement. This indicates the parameter number difference between LLM and ALIGNN is the main cause: LLM (GPT-2) with many more parameters (124.64M in Table 7) allows extraction of more information especially for minority classes, while ALIGNN with much fewer parameters (4.29M in Table 7) will primarily extract features from majority classes and thus captures less information for minority classes. However, the large number of parameters in LLM makes it difficult to optimize while training with the classification head, so training both separately via two stage training allows for independent optimization of weights in each block and improves performance in both majority and minority classes. As for frozen ALIGNN, the model does not effectively extract information for minority classes, so second stage training of the classification head does not help much. For the multi-modal approach, ALIGNN's less accurate extracted features for minority classes introduces error, resulting in worse performance of the multi-modal approach relative to LLM-frozen on AFM and FiM. To account for any impact from information fusion, various fusion methods besides simple concatenation were tested including 1) projecting features from LLM and ALIGNN into the same dimension and adding a layer to align the projected features, and later adding alignment layer loss into the second-stage training loss; and 2) using multi-head attention layer instead of simple concatenation. Test results show different fusion mechanisms have very minor impact on classification results vs simple concatenation, thus confirming the above analysis.

4.4 Precision-Recall AUC Analysis

To analyze the model performance on test data with unbalanced classes, model prediction probability values for each class (i.e. output of the final Softmax layer) are logged during the test process. Afterwards, the precision-recall (PR) curve for each class is plotted by choosing a class (one vs rest) and computing precision and recall for this class while iterating the decision threshold from low to high. The PR curve shows the tradeoff between precision and recall. Subsequently, the Area Under Curve (AUC) is calculated to determine how well the model maintains high precision while increasing recall. A higher AUC is preferred, as having high precision and recall is more ideal than high precision or recall alone. The LLM-frozen PR curve (Fig. 8) and AUC show it provides very good results on all four class classification.

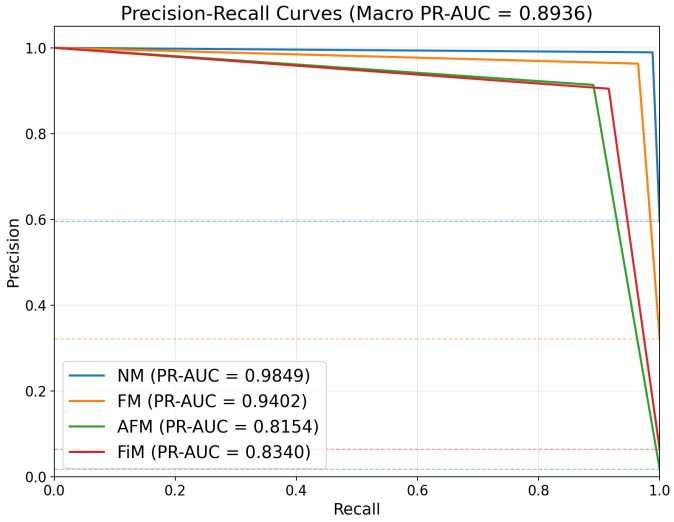

Figure 8: LLM-frozen PR curve with AUC

4.5 Numerical Stability Analysis

LLM frozen model with two stage training was trained and tested three times from scratch to evaluate numerical stability. Table 5 results and confidence interval in Table 6 show good numerical stability.

Table 5: Per-class precision, recall and F1 for LLM frozen model on three runs

Class	Run 1	Run 2	Run 3
NM	98.9%/98.8%/98.9%	99.0%/98.7%/98.8%	98.9%/98.8%/98.88%
FM	96.2%/96.4%/96.3%	96.1%/96.4%/96.3%	96.3%/96.2%/96.27%
AFM	91.8%/88.7%/90.3%	92.1%/88.5%/90.2%	91.3%/88.7%/90.3%
FiM	90.4%/91.4%/90.9%	89.8%/91.8%/90.8%	89.8%/91.5%/90.67%

Table 6: Per-class metrics with 95% confidence intervals on LLM-frozen model

Class	Mean (P/R/F1)	[95% CI] (P/R/F1)
NM	0.9896 / 0.9885 / 0.9891	[0.9887, 0.9904] / [0.9877, 0.9894] / [0.9889, 0.9892]
FM	0.9628 / 0.9635 / 0.9632	[0.9617, 0.9639] / [0.9619, 0.9651] / [0.9621, 0.9642]
AFM	0.9185 / 0.8867 / 0.9023	[0.9128, 0.9242] / [0.8760, 0.8973] / [0.8995, 0.9052]
FiM	0.9006 / 0.9144 / 0.9075	[0.8921, 0.9091] / [0.9124, 0.9164] / [0.9036, 0.9113]

4.6 Model Complexity Analysis

The complexity of the model including parameters and MACs (Multiply-Accumulate Operations) is shown in Table 7. As shown, GPT-2 requires much more computational power (MACs) and memory/storage (parameters) than ALIGNN. For edge devices with limited computational capacity, ALIGNN can be a good choice as it offers a good trade-off between resource requirement and performance. Using a smaller LLM and increasing ALIGNN parameters could make the multi-modal approach a better trade-off choice between model complexity and classification accuracy. This will be investigated in future studies.

Table 7: ALIGNN LLM Model Complexity Breakdown

	Parameter	MACs
Total	129.26M	47.42 GMACs
GPT-2 Block (seq_len=512)	124.64M	43.55 GMACs
ALIGNN Block	4.29M	3.87 GMACs
Concat Fusion and Classifier Block	132.87K	133 KMACs

5 Conclusion and Future Work

This paper presents a number of ALIGNN- and LLM-based approaches to predict material magnetic ordering. Due to the imbalance of class samples in the dataset, the baseline LLM, ALIGNN and multi-modal models all struggle to provide accurate classification on minority classes. To solve this problem, a two stage based training is proposed to use frozen feature extractor in combination with training on classification head. This approach significantly improves the classification precision of the LLM-only and multi-modal approaches on minority classes; the LLM-only model reach 90+% on all classes.

In future work, this approach is to be adapted to handle regression problems or multi-properties prediction with the use of more complicated LLM models. Moreover, this approach can be applied to problems of material generation by utilizing the decoding capabilities of the LLM block to generate new materials with desired properties.

REFERENCES

Alignn-bert. URL https://github.com/Jonathanlyj/ALIGNN-BERT-TL-crystal.

gpt2 model. URL https://huggingface.co/openai-community/gpt2.

jarvis.core.atoms. URL https://jarvis-tools.readthedocs.io/en/master/autoapi/jarvis/core/atoms/index.html.

magndata cif. URL https://www.cryst.ehu.es/magndata/.

llama model. URL https://huggingface.co/meta-llama.

Material project. URL https://next-gen.materialsproject.org/materials.

Christopher Broyles, William Charles, and Sheng Ran. Structure-driven prediction of magnetic order in uranium compounds. *Phys. Rev. Mater.*, 8:114405, Nov 2024. doi: 10.1103/PhysRevMaterials.8.114405. URL https://link.aps.org/doi/10.1103/PhysRevMaterials.8.114405.

J. Callaway and N. H. March. Density functional methods: Theory and applications. *Solid State Physics*, 38:135–221, 1984. doi: https://doi.org/10.1016/S0081-1947(08)60313-6.

K. Choudhary and B. DeCost. Atomistic line graph neural network for improved materials property predictions. *npj Computational Materials 7, 185*, 2021. doi: https://doi.org/10.1038/s41524-021-00650-1.

Alexander Dunn, Qi Wang, Alex Ganose, Daniel Dopp, and Anubhav Jain. Benchmarking materials property prediction methods: the matbench test set and automatminer reference algorithm. *npj Computational Materials*, 6(1):138, Sep 2020. ISSN 2057-3960. doi: 10.1038/s41524-020-00406-3. URL https://doi.org/10.1038/s41524-020-00406-3.

Youjia Li, Vishu Gupta, Muhammed Nur Talha Kilic, Kamal Choudhary, Daniel Wines, Wei-keng Liao, Alok Choudhary, and Ankit Agrawal. Hybrid-llm-gnn: integrating large language models and graph neural networks for enhanced materials property prediction. *Digital Discovery*, 2025.

Steph-Yves Louis, Yong Zhao, Alireza Nasiri, Xiran Wang, Yuqi Song, Fei Liu, and Jianjun Hu. Graph convolutional neural networks with global attention for improved materials property prediction. *Phys. Chem. Chem. Phys.*, 22:18141–18148, 2020. doi: 10.1039/D0CP01474E. URL http://dx.doi.org/10.1039/D0CP01474E.

N H March. *Electron Correlation in the Solid State*. PUBLISHED BY IMPERIAL COLLEGE PRESS AND DISTRIBUTED BY WORLD SCIENTIFIC PUBLISHING CO., London, 1999. doi: https://doi.org/10.1142/p174.

R. G. Parr and W. Yang. *Density-Functional Theory of Atoms and Molecules*. Oxford University Press, Oxford, 1989. doi: https://doi.org/10.1093/oso/9780195092769.001.0001.

Patrick Reiser, Andre Eberhard, and Pascal Friederich. Graph neural networks in tensorflow-keras with raggedtensor representation (kgcnn). *Software Impacts*, pp. 100095, 2021. ISSN 2665-9638. doi: https://doi.org/10.1016/j.simpa.2021.100095. URL https://www.sciencedirect.com/science/article/pii/S266596382100035X.

Andre Niyongabo Rubungo, Craig Arnold, Barry P. Rand, and Adji Bousso Dieng. Llm-prop: Predicting physical and electronic properties of crystalline solids from their text descriptions, 2023. URL https://arxiv.org/abs/2310.14029.

Robin Ruff, Patrick Reiser, Jan Stühmer, and Pascal Friederich. Connectivity optimized nested line graph networks for crystal structures. *Digital Discovery*, 3(3):594–601, March 2024. ISSN 2635-098X. doi: 10.1039/d4dd00018h. URL https://doi.org/10.1039/d4dd00018h.

Kristof Schütt, Pieter-Jan Kindermans, Huziel Enoc Sauceda Felix, Stefan Chmiela, Alexandre Tkatchenko, and Klaus-Robert Müller. Schnet: A continuous-filter convolutional neural network for modeling quantum interactions. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/303ed4c69846ab36c2904d3ba8573050-Paper.pdf.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

Nicholas Walker, Amalie Trewartha, Haoyan Huo, Sanghoon Lee, Kevin Cruse, John Dagdelen, Alexander Dunn, Kristin Persson, Gerbrand Ceder, and Anubhav Jain. The impact of domain-specific pre-training on named entity recognition tasks in materials science. *Available at SSRN 3950755*, 2021.

Tian Xie and Jeffrey C. Grossman. Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties. *Phys. Rev. Lett.*, 120:145301, Apr 2018. doi: 10.1103/PhysRevLett.120.145301. URL https://link.aps.org/doi/10.1103/PhysRevLett.120.145301.

Keqiang Yan, Cong Fu, Xiaofeng Qian, Xiaoning Qian, and Shuiwang Ji. Complete and efficient graph transformers for crystal material property prediction. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=BnQY9XiRAS.

A Appendix

Random Guessing Algorithms

Three "random guessing" algorithms are employed as a baseline in this paper against which the ALIGNN-LLM multi-modal approach is compared: uniform guessing, proportional guessing, and the Zero Rule.

For uniform guessing, the algorithm predicts each of the four classes (NM, FM, AFM, FiM) with equal probability (i.e. 25% chance to predict each class). The accuracy of this algorithm is 25%.

For proportional guessing, the algorithm predicts each class with a probability equal to its prevalence in the dataset. For instance, NM samples make up 59.48% of the dataset, so the algorithm will guess NM with a probability of 59.48%. This algorithm has an accuracy of 46.13%:

$$\begin{aligned} P_{\text{proportional}} &= \sum_{i \in \{\text{NM, FM, AFM, FiM}\}} P(\text{algorithm guesses class } i) \cdot P(\text{correct label is class } i) \\ &= 59.48\% \cdot 59.48\% + 32.06\% \cdot 32.06\% + 1.85\% \cdot 1.85\% + 6.61\% \cdot 6.61\% \\ &= 46.13\% \end{aligned}$$

The Zero Rule algorithm guesses the class with the largest proportion of the dataset each time. For the dataset used in this study, NM makes up the largest percentage of the data at 59.48%; hence, the model guesses NM each time. The accuracy of this model is 59.48%.

Reproducibility statement

Open access (Github) to code and database.xlsx (the dataset of uranium compounds) in this paper will be available after paper decision under the MIT license. The training procedure is provided in details in section 3.2.

AI USE STATEMENT

We only used generative AI tools to validate the algorithm implementation by writing unit tests and performing code review and have not used generative AI tools for all other aspects of research and writing work related to the paper.