MATMMFUSE: MULTI-MODAL FUSION MODEL FOR MATERIAL PROPERTY PREDICTION

Abhiroop Bhattacharya
Department of Electrical Engineering
École de technologie supérieure
Montréal, H3C 1K3, Canada
{abhiroop.bhattacharya.1@ens.etsmtl.ca

Sylvain G. Cloutier
Department of Electrical Engineering
École de technologie supérieure
Montréal, H3C 1K3, Canada
{SylvainG.Cloutier@etsmtl.ca

ABSTRACT

The recent progress of using graph based encoding of crystal structures for high throughput material property prediction has been quite successful. However, using a single modality model prevents us from exploiting the advantages of an enhanced features space by combining different representations. Specifically, pre-trained Large language models(LLMs) can encode a large amount of knowledge which is beneficial for training of models. Moreover, the graph encoder is able to learn the local features while the text encoder is able to learn global structure information. In this work, we propose Material Multi-Modal Fusion(**MatMMFuse**), a fusion based model which uses a multi-head attention mechanism for the combination of structure aware embedding from the Crystal Graph Convolution Network (CGCNN) and text embeddings from the SciBERT model. We train our model in an end-to-end framework using data from the Materials Project Dataset. We show that our proposed model shows an improvement compared to the vanilla CGCNN and SciBERT model for all four key properties- formation energy, band gap, energy above hull and fermi energy. Specifically, we observe an improvement of 40% compared to the vanilla CGCNN model and 68% compared to the SciBERT model for predicting the formation energy per atom. Importantly, we demonstrate the zero shot performance of the trained model on small curated datasets of Perovskites, Chalcogenides and the Jarvis Dataset. The results show that the proposed model exhibits better zero shot performance than the individual plain vanilla CGCNN and SciBERT model. This enables researchers to deploy the model for specialized industrial applications where collection of training data is prohibitively expensive.

1 INTRODUCTION

Machine learning (ML) has been popular as a potent and adaptable technique in the hunt for materials targeting a wide range of applications, especially when a thorough investigation of the materials space is required(Schmidt et al., 2019; Chen et al., 2020). With the continuous expansion of high-throughput density functional theory(DFT) datasets and the ongoing development of ML algorithms, it is anticipated that the use of ML for materials discovery will increase even more(Saal et al., 2013; Draxl & Scheffler, 2019; Jain et al., 2013). Historically, structural descriptors that meet rotational and translational invariance had been used for encoding the crystal structures, ranging from Coulomb matrixFaber et al. (2015) and atom-centered symmetry functions (ACSFs) to smooth overlap of atomic positions (SOAP)(Behler, 2011; De et al., 2016).

First proposed more than 15 years ago, Graph Neural Networks(GNNs)(Scarselli et al., 2008; Gori et al., 2005) have drawn more interest lately in material informatics as a way to overcome static descriptor limitations by learning the representations on adaptable graph-based inputs(Li et al., 2024). Such GNNs have been implemented to predict materials in complex systems including surfaces(Palizhati et al., 2019; Back et al., 2019) and periodic crystal arrangements(Chen et al., 2019; Xie & Grossman, 2018). The GNN models effectively encode and utilize the structure of the lattice. Particularly, the CGCNN model(Xie & Grossman, 2018) has shown exemplary performance in encoding the structure property relation while handling periodic boundary conditions. However, the graph convolution based models require a large training dataset to learn generalizable

structure property mapping. Moreover, the instances of model failure are difficult to understand and interpret(Fung et al., 2021). Most importantly, GNN models are unable to incorporate global structural information like crystal symmetry, space group number and rotational information.

Large Language Models (LLMs) provide a promising approach for knowledge discovery in materials science due to their generalization and transferability(Jablonka et al., 2023). Their success has motivated applications in structure-property relationship discovery, particularly through pre-trained domain-specific language models, which effectively capture latent knowledge from domain-specific literature. SciBERT, which has been trained on a scientific corpus of 3.17 billion tokens has shown remarkable performance across a diverse set of tasks(Beltagy et al., 2019). Compared to graph neural network (GNN) models, LLMs are able to incorporate global information such as space group and crystal symmetry. Combining the strength of the GNN based models with LLM models using multi modal data enhances the feature space, enabling the model to prioritize critical features from diverse latent embeddings. While several studies have explored the potential of LLMs to improve generalization, transferability, and few-shot learning, limited research has focused on integrating textual information from natural language with structural-aware learning from GNNs for crystal property prediction. Li et al. (2025) have used embedding concatenation for combining multiple modalities while, Ock et al. (2024) have combined the graph structure of crystals with X-ray diffraction patterns for augmenting the structure aware graph embedding with diffraction information. Lee et al. (2025) applied masked node prediction pretraining strategy to train a multi-modal model using a combination of text tokens and information from lattice neighbors. However, this architecture might result in locally valid but globally inconsistent structures. Das et al. (2023) have developed CrysMMNet which uses concatenation to combine multiple modalities. These models have shown that using multi-modal data with fusion models allows the model to leverage the enhanced feature space. Concatenation uses static connections between modalities and the model design does not focus on cross model connections. While, the proposed model uses cross attention which enables the model to focus on long range dependencies across modalities. Moreover, compared to concatenation, cross attention gives clear attention weights that can be interpreted. To the best of our knowledge, this is the first work, which explores a multi-head attention mechanism to combine structure aware and context aware embeddings to improve prediction and zero shot performance for the prediction of material properties for inorganic crystals.

In this work, we propose, **Mat**erial **M**ulti-**M**odal **Fus**ion(**MatMMFuse**), a fusion model which uses a multi-head attention based combination of structure aware embedding of the Crystal Graph Convolution Network (CGCNN)(Xie & Grossman, 2018) and text embeddings of SciBERT(Beltagy et al., 2019). Importantly, we train our model in an end-to-end framework using data from the Materials Project Dataset. We show that MatMMFuse performs in line with state of the art models for four key properties- formation energy, band gap and Fermi Energy. We observe an improvement of 35% and 68% respectively compared to the plain vanilla versions of the model for predicting the formation energy per atom. Furthermore, we demonstrate the zero shot performance of the trained model on small curated datasets of Perovskites, Chalcogenides and the Jarvis Dataset. The primary contributions of this paper are:

- Introduction of a multi-head cross attention based fusion approach for accurate material property prediction.
- Efficiently using multi-modal data to combine structure aware and context aware information to combine local and global information.
- Improved zero shot performance for specialized materials like Perovskites and Chalcogenides.

2 Proposed Model Architecture

The following section describes the architecture of our multi-modal framework. Given a dataset of inorganic crystals denoted by $D = [(S, T), P]$ where S, T and P denote the structure information in CIF format, the text description and the material property respectively. The model trains the parameters of the Graph encoder (G_θ) and the BERT encoder (B_θ) to learn the function $f_\theta \rightarrow P$. The figure 1 captures the model schematic. Each section of the model is explained below.

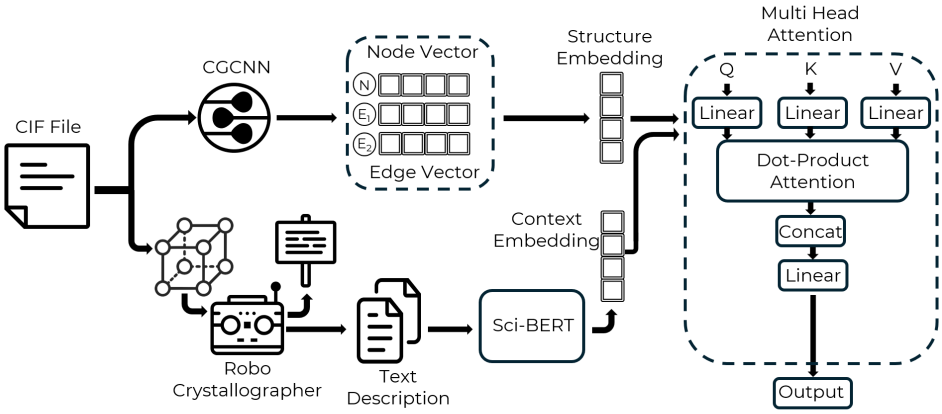


Figure 1: The figure provides an overview of MatMMFuse. The CGCNN model generates a structure aware embedding while the SciBERT model generates a context aware embedding which are combined using a multi-head attention mechanism.

2.1 Graph Encoder

For this model, the material structure from the crystallographic information file(CIF) is encoded as a graph $G(V, E)$ using the CGCNN model where, the atoms are the nodes V and the bonds between the atoms are encoded as the edges E. In addition to the graph topology, the node attributes capture the different properties of the atom such as group, position in the periodic table, electro-negativity, first ionization energy, covalent radius, valence electrons, electron affinity and atomic number. For each atom i and it's neighbor $j \in \mathcal{N}(I)$, the convolution updates the atom's feature vector h_i as follows:

$$h_i^{(l+1)} = h_i^{(l)} + \sum_{j \in \mathcal{N}(i)} \sigma\left(z_{i,j}^{(l)} W_f^{(l)} + b_f^{(l)}\right) \odot g\left(z_{i,j}^{(l)} W_s^{(l)} + b_s^{(l)}\right) \tag{1}$$

where, the feature vector of atom i at layer l is denoted by $h_i^{(l+1)}$. The concatenation of $h_i^{(l)}$ and $h_i^{(j)}$ and the edge features $e_{i,j}$ is $z_{i,j}^{(l)}$. $W_f^{(l)}$ and $W_s^{(l)}$ denote the learnable weight matrices. Similarly, $b_f^{(l)}$ and $b_s^{(l)}$ denote the bias terms. The element wise multiplication is represented by $\odot$ with σ and g denoting the activation functions. After L graph convolution layers, the graph level representation is obtained by global pooling where in h_G denotes the graph level embedding and N denotes the number of atoms.

$$h_G = \frac{1}{N} \sum_{i=1}^{N} h_i^{(L)} \tag{2}$$

2.2 Text Encoder

The textual description of the CIF files are generated using the Robocrystallography(Ganose & Jain, 2019) framework. We leverage the scientific knowledge encoded in the pretrained SciBERT modelBeltagy et al. (2019) followed by a projection layer. For an input sequence $X = (x_1, x_2, \cdots, x_n)$, the self attention mechanism uses the Query Matrix, Key Matrix and Value Matrix denoted by Q,K and V respectively. These are linear projections using the corresponding learnable weight matrices.

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right) V \tag{3}$$

It is important to note that BERT uses a multi-head attention mechanism.

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, \ldots, \text{head}_h)W_O \tag{4}$$

A fully connected feed forward network is used with a ReLU activation function and $W_1, W_2$and b_1, b_2 learnable weight matrices and biases respectively with the final output obtained by stacking different transformer layers.

$$\text{FFN}(x) = \text{ReLU}(xW_1 + b_1)W_2 + b_2 \tag{5}$$

The model has 12 transformer layers for encoding with 768 hidden dimensions and 12 attention heads. The model has been pre-trained on a 1.14 million papers from Semantic scholar resulting in a total of 3.17 billion tokens.

2.3 Multi-Head Cross Attention Fusion for Joint Embedding

The model uses a multi-head cross attention based framework for combining the embeddings generated by the LLM model(h_t) and the structure aware embedding generated by the GNN ((h_s). The entire framework is trained in a supervised end-to-end manner. This is a key advantage of the proposed approach because this enables the model to focus on the important sections from the structure aware embedding and the text based embedding.

$$Q = W_q h_t, \quad K = W_k h_s, \quad V = W_v h_s \tag{6}$$

$$\text{attention\_scores} = \frac{QK^T}{\sqrt{d}} \tag{7}$$

$$\text{attention\_weights} = \text{softmax}(\text{attention\_scores}) \tag{8}$$

$$\text{combined} = \text{attention\_weights} \cdot V \tag{9}$$

The combined embedding is passed through a fully connected layer for the final prediction.

$$y = W_o \cdot \text{combined} + b_o \tag{10}$$

3 Experimentation

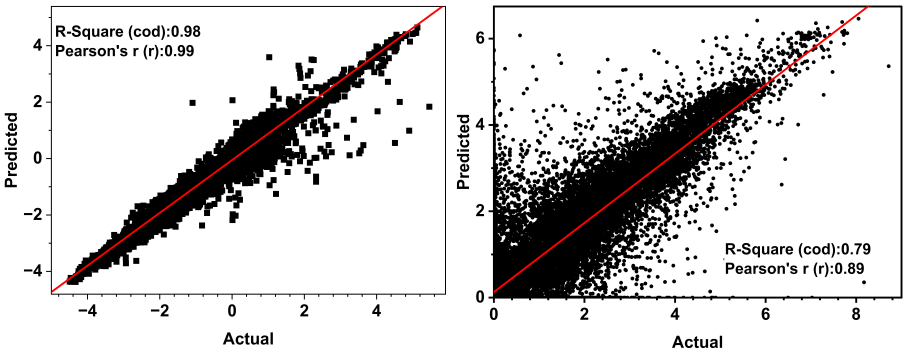
We use a Nvidia RTX 4090 graphics processing unit (GPU) to run our experiments. The framework is implemented using the Pytorch library version(Paszke et al., 2017).

3.1 Dataset

For model training and assessment, we leverage the widely used Materials Project dataset(Jain et al., 2013). We focus on four important material properties: the formation energy per atom, the energy above the hull, the fermi energy and the Band Gap. We use 95582 crystal structures with a 80%,10%,10% train, validation and test split. For the CGCNN model, we directly use the data in crystallographic file(CIF) format. We use RoboCrystallographer(Ganose & Jain, 2019) to convert the CIF file to text files. These text files are the input for the SciBert LLM model. The distribution of the target variables and text descriptions are available in the Appendix. For evaluating the zero shot performance of the model, we use the Cubic Oxide Perovskites, Chalcogenides and a subset of the JARVIS dataset. The distribution of the target variables and text descriptions are available in the Appendix.

3.2 Experimentation Overview

We perform experiments in two paradigms. Firstly, *In-domain* wherein we use the traditional approach to train MatMMFuse on examples from the Materials Project dataset. The model is trained in an end-to-end supervised manner. Secondly, we use the trained model to predict the material property of materials with specialized applications without explicitly training on the respective datasets. This paradigm is known as *Zero Shot*. This is intended to be used for specially curated small datasets for materials with specific industrial applications.



(a) Formation Energy per atom (E_f) (b) Band Gap(B_g)

Figure 2: The scatter plot presents the actual versus predicted values for (a) Formation Energy per atom, (b) Band Gap. The model tends to incorrectly predict the values for band gap when the values are close to zero.

4 Results and Discussions

4.1 In-Domain

We have used MatMMFuse to predict four key material properties. We evaluate the performance of the model for four important material properties - formation energy per atom(E_f), Fermi Energy(E_g) and the Band Gap(B_g). We use AdamW with a cosine learning rate scheduler and warmup. The trained model is then used to predict the formation energy for Perovskites, Chalcogenides and a subset of the Jarvis Dataset in a zero shot paradigm. We observe an improvement of 40% compared to the CGCNN model and 68% compared to the SciBERT model for the formation energy per atom. However, for the energy above hull, MatMMFuse performs marginally better than SciBERT with a 6.7% improvement and a 58.5% improvement over the CGCNN model. The total Fermi energy also shows a similar pattern with a 30% improvement over the vanilla versions of both the models. Machine learning models have struggled with predictions of the band gap for crystals(Zhuo et al., 2018) for which the proposed model has a marginal improvement of around 1% compared to the other models. We hypothesize that the improvement across all the properties is occurring due to the ability of MatMMFuse to selectively combine both local structural information and global information such as space group and symmetry using the attention mechanism.

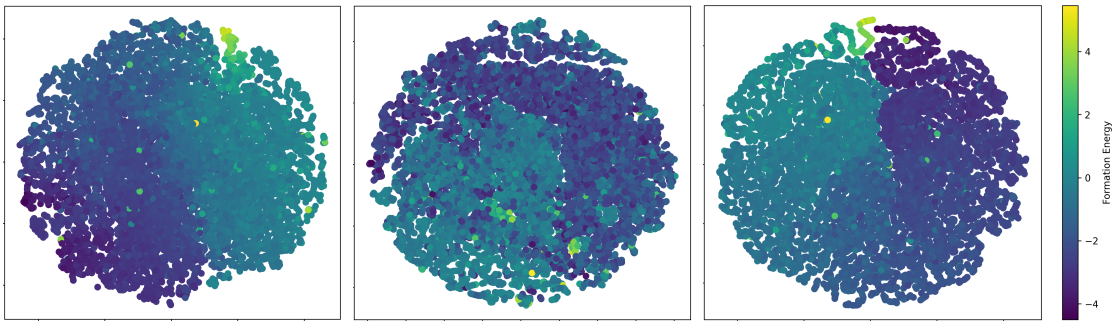
Table 1: Benchmarking model performance. The lower the error the better the model performance.

	Mean Absolute Error(MAE)			
	Formation Energy (eV/atom)	Fermi Energy (eV)	Energy Above Convex Hull (eV/atom)	Band Gap (eV)
CGCNN	0.042	0.60	0.071	0.37
SciBERT	0.081	0.59	0.031	0.38
MatMMFuse	**0.025**	**0.44**	**0.029**	**0.31**

To further investigate the results, we are comparing the plots of the actual versus predicted values for the formation energy per atom and the band gap for the test dataset 2. For formation energy, we observe that the predictions are aligned with the actual values with a R^2 of 0.97 while, for Band Gap we can clearly see that the model predicts a higher value when the actual value is close to zero.

The t-distributed stochastic neighbor method (t-SNE)(Van der Maaten & Hinton, 2008) allows us to understand the decision boundaries and segregation of data points in the high dimensional embedding using 2D plots. The Figure 3 depicts a combined structure-composition latent space for the trained materials, in which points within a grouping are anticipated to have similarities in both their atomic structures and elemental compositions. We see comparable clustering in the latent space. In the t-SNE plot of MatMMFuse we observe that the dark and light colored ones are segregated in different

clusters with lobe-structured decision boundaries which shows that the learned embedding is able to discern between crystals with high formation energies and ones with low formation energy. We observe decision boundaries in the embedding generated by the SciBERT model as well but the points are not clustered. The embedding generated by the graph encoder does not have clear clustering or decision boundaries.



(a) (b) (c)

Figure 3: The t-SNE plot of the embedding from the embedding for the test dataset from Materials Project, with each point representing an individual crystal. Colors for each point are associated with their formation energies. a CGCNN embedding, b SciBERT embedding , c MatMMFuse embedding

4.2 Zero Shot Performance

A key challenge in material science is the lack of large datasets for specialized applications. Most materials with specialized applications such as photovoltaic cells and battery, do not have large datasets with DFT calculated material properties to enable training of data hungry deep learning models. In this section, we demonstrate that the trained MatMMFuse model can be used for predicting the material properties for small curated datasets in a zero shot manner. The proposed attention-based method for combining embeddings leads to an improvement in the zero shot performance of the model. Attention allows the model to dynamically weight and combine embeddings based on the relevance to task enabling the model to focus on the most informative features from each embedding. In the table 2, we compare the zero shot performance of MatMMFuse for predicting the energy of Perovskites, Chalcogenides and a small subset of the Jarvis dataset with the vanilla CGCNN and SciBERT models.

4.2.1 Cubic Oxide Perovskites

ABO_3 perovskites are viewed as promising resistive-type gas sensors (Ishihara, 2009). It is important to remember that there are 2704 observations in the dataset which is insufficient for training large GNN or LLM models. MatMMFuse achieves a MAE of 1.28 on the test dataset which is 10% lower than the CGCNN model and 55% lower than the SciBERT model.

4.2.2 Chalcogenide Perovskites

For photovoltaic applications researchers have proposed Chalcogenide perovskites of the form $AB(S, Se)_3$ because of their stability, non-toxicity, and lead-free composition(Basera & Bhattacharya, 2022). To test our model, we repeated the same experiment for a dataset for $AB(S, Se)_3$ perovskites. The dataset has 1621 observations. Nonetheless, MatMMFuse achieves a low MAE of 1.05, lower by 21% and 27% as compared to the CGCNN and SciBERT models respectively.

4.2.3 JARVIS

The JARVIS (Joint Automated Repository for Various Integrated Simulations) dataset(Choudhary et al., 2020) is a high-throughput materials database developed by the National Institute of Standards and Technology(NIST). The dataset encompasses a wide array of materials properties, computed using density functional theory (DFT) simulations. MatMMFuse achieves a MAE of 0.078 which is 48% lower than the CGCNN and the 59% lower than the SciBERT model.

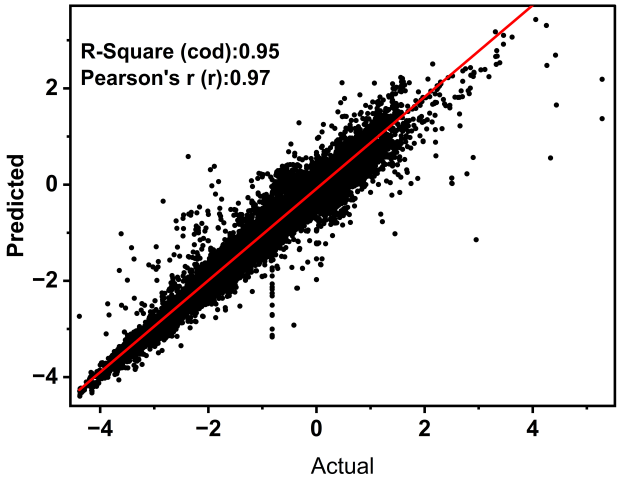


Figure 4: The scatter plot presents a comparison between the actual and the predicted values of formation energy per atom for the JARVIS dataset.

The actual versus predicted curve in Figure4 shows that the predicted and actual values are aligned with a R^2 of 0.94. However, there are a number of data points around -0.9 eV/atom which have a much lower prediction.

Table 2: Zero Shot performance for the Energy per atom. The lower the error the better the model performance.

	Energy $MAE(eV/atom)$		
	$CGCNN$	$SciBERT$	$Proposed$
Perovskites(ABO_3)	1.42	2.84	**1.28**
Chalcogenides(ABS_3)	1.33	1.44	**1.05**
JARVIS	0.15	0.19	**0.08**

4.3 Ablation Studies

This section presents the ablation studies performed by changing, adding or removing the key parts or inputs of the model architecture.

4.3.1 Encoded domain knowledge

To prove our hypothesis that MatMMFuse is able to leverage the encoded knowledge in the LLM Model, we have run experiments by using variations of the BERT model as the text encoder for predicting the formation energy per atom. The alternate models used are ALBERT(Lan et al., 2019), RoBERT(Masala et al., 2020), DeBERT(Sergio & Lee, 2021) and DistillBERT(Sanh et al., 2019). Due to the knowledge of material science encoded in the MatSciBERT model, we observe that it outperforms all models closely followed by SciBERT model. It is important to note that ALBERT shows a sharp deterioration in model performance. We posit that this might be due to two reasons. Firstly, ALBERT shares parameters across all transformer layers, reducing model size but limiting the model's capacity to learn distinct representations at different levels of abstraction. Secondly, it uses token order prediction as compared to next token prediction used in other BERT Models. The plot5 presents a comparison of model performance for different BERT models.

Further to this, we also observe a similar improvement in the zero shot performance of the model on the specialized cubic oxide Perovskite, Chalcogenides and the JARVIS dataset which is shown in table3

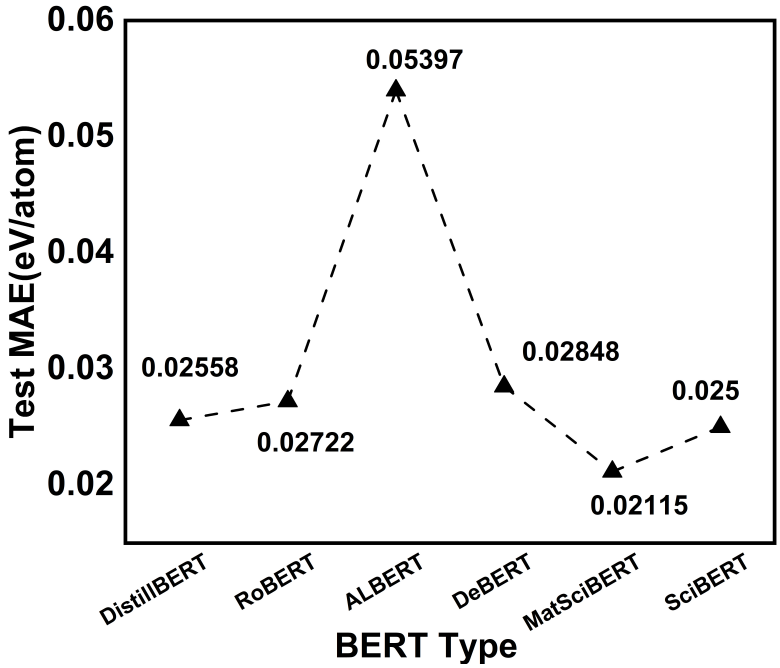


Figure 5: The plot compares the performance of different BERT models for encoding the text representation. MatSciBERT has the best performance and ALBERT has the worst performance.

Table 3: Zero Shot performance of the MatSciBERT model for the Formation Energy per atom. The lower the error the better the model performance.

	$MAE(eV/atom)$	
	$MATSciBERT$	$SciBERT$
Perovskites(ABO_3)	2.26	**1.28**
Chalcogenides(ABS_3)	**1.33**	0.98
JARVIS	**0.037**	0.08

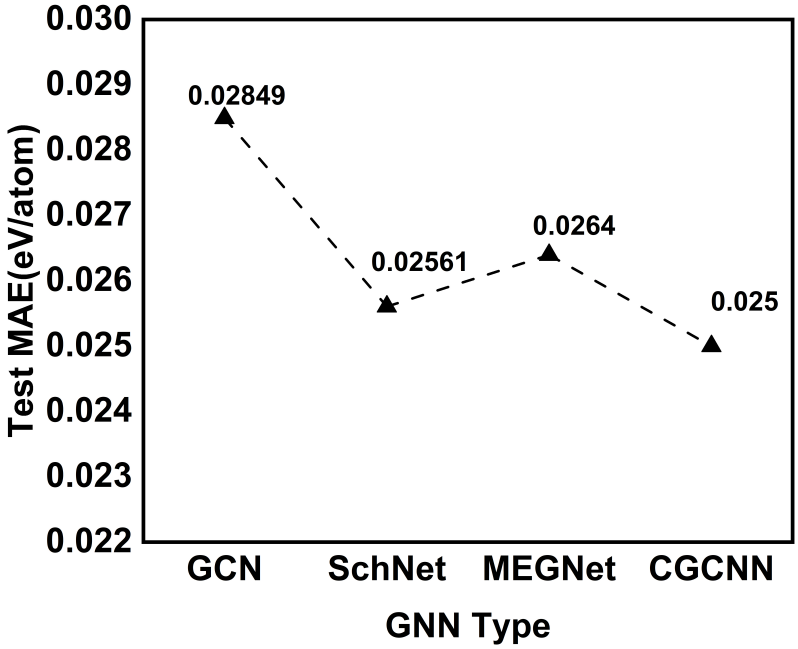


Figure 6: The plot compares the performance of different GNN models for encoding the lattice structure. CGCNN gives the optimum tradeoff between performance and efficiency.

We have decided to use SciBERT model because it has more interdisciplinary knowledge which leads to a more broader scientific context. Especially, for applications in biomedicine and energy. ABO_3 perovskites are used for solar cells and therefore SciBERT outperforms MatsciBERT in such specialized applications.

4.3.2 Encoded lattice structure

The encoding of the crystal lattice structure using different graph encoding models results in different ways of capturing the complex relationships within crystal structures. We used SchNet(Schütt et al., 2018), MEGNet(Chen et al., 2019), CGCNN and Graph convolution networks(GCN) for the analysis. Vanilla GCN architectures are not designed to incorporate periodic boundary conditions. On the other hand, SchNET and CGCNN explicitly incorporate crystal periodicity. CGCNN uses discretized bins for edge features while SchNet uses continuous radial basis functions for smooth distance representation. CGCNN includes more extensive information about the crystal structure and is computationally more efficient compared to SchNET which uses continuous-filter convolutions with filter-generating networks that create customized filters for each atomic interaction based on distance. Unlike other models, MEGNet uses global state variables such as unit cell parameters which makes it more expressive but also more computationally expensive. We found that CGCNN gives the optimum tradeoff between performance and efficiency. The plot6 presents a comparison of model performance for different GNN based encoder models.

4.3.3 Multi-Head Attention Module

A comprehensive ablation study was performed on the different sub-modules of the attention based fusion mechanism for allowing the model to focus on the specific parts of the structure aware and context aware embeddings. We observe that using a multi-head attention method considerably improves the performance. Using a layer-norm to normalize the layer outputs increases the stability of the training and helps the model converge. For small datasets, including dropout prevents overfitting and helps the model generalize better. Interestingly, including a residual layer does not lead to

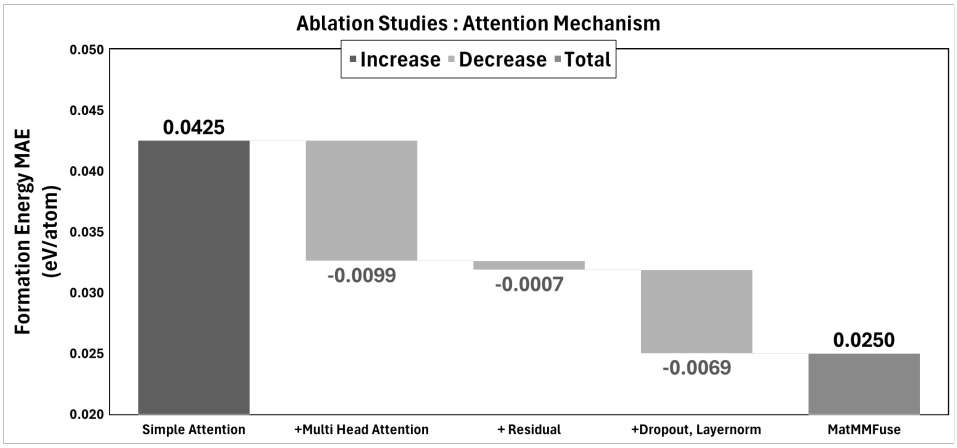


Figure 7: The waterfall chart shows the effect of adding individual components to improve the attention based fusion method.

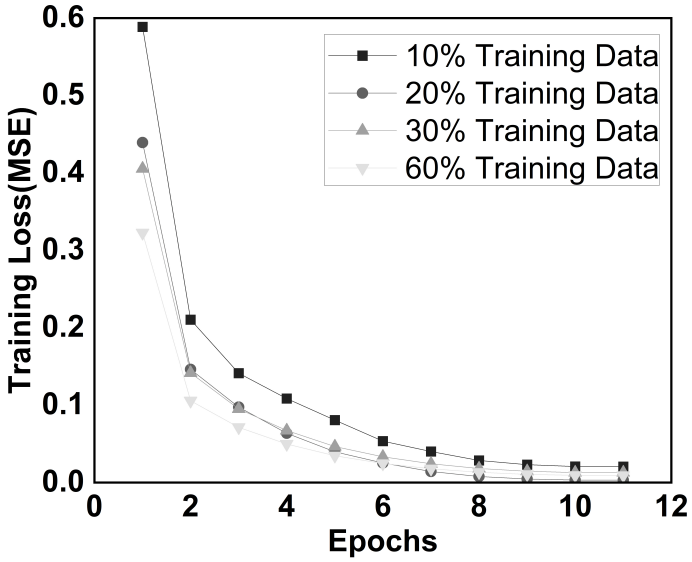


Figure 8: The chart shows that MatMMFuse is relatively robust to reduction in training data.

a significant improvement in model performance.The waterfall chart7 captures the effect of each change on the model performance.

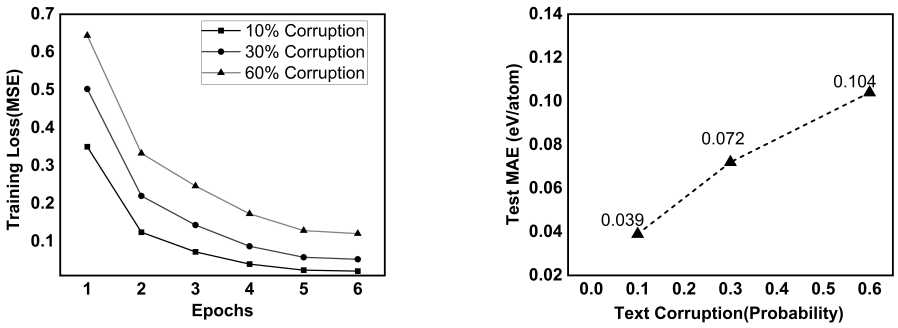
4.3.4 Robustness to Training data size

As reported in literature, reducing the size of the training data reduces the performance of the CGCNN model(Xie & Grossman, 2018). Interestingly, we observe that using an enhanced feature space which uses multiple modalities improves the robustness of the model to reduction in training data. The plot8 shows that the model is able to converge to a low training loss .

4.3.5 Corruption of text input

Corruption of text input remains a limitation of BERT models(Jin et al., 2020). Moreover, Robocrystallographer might lead to corrupted text output if there are aberrations in the CIF file (Ganose &

Jain, 2019). Thus, we have studied the effect of different levels of text corruption on the performance of the model for predicting the formation energy per atom. For corrupting the text, we have deleted random characters, added random punctuations and performed random word substitutions. The level of corruption has been controlled by using the probability of corruption. There is a significant decrease in model performance as captured by the plot. The plot9 captures the degradation in model performance in training and inference with the increase in corruption of the text input.



(a) (b)

Figure 9: The figure (9a) and figure (9b) shows the effect of the corruption of the input text on the training loss and the test loss respectively. The model performance deteriorates significantly with corruption in text.

5 CONCLUSION

This paper explores a multi-modal fusion model for predicting material properties. The Material Multi-Modal Fusion(**MatMMFuse**) model uses a multi-head cross attention based method for combining the embedding from graph neural network and a LLM model. The CGCNN model has been selected to encode the lattice structure as a graph encoding while, the SciBERT model has been used to encode the text descriptors. The SciBERT model already posses domain specific scientific knowledge which is helpful for generating meaningful embeddings. The enhanced feature space with the attention mechanism allows the model to selectively focus on key features from the structure aware graph embedding and the context aware embedding. The graph encoder focuses on local information while the text encoder is able to learn global information such as symmetry and space group. The results show that the proposed model is able to outperform both the plain vanilla versions of CGCNN and SciBERT models by 35% and 68% respectively for predicting the formation energy per atom. We observe an improvement for the Energy above Hull and the Fermi energy as well. Further, we observe a marginal improvement for the prediction of Band Gap which is aligned to the state of the art. Interestingly, we demonstrate that the zero shot performance of the model is better than the vanilla CGCNN and SciBERT models for cubic oxide perovskites, chalcogenide perovskites and a subset of the JARVIS datasets which is an important step for specialized uses cases. Analyses of the t-SNE plots show that our model is able to generate embeddings which have clear lobe-shaped decision boundaries and similar material properties are clustered together. Finally, we believe the ability of LLM models to use text based inputs for probing the underlying mechanism of the model to understand specific points of failure provides a tool to analyze the structure property relationships in crystalline solids.

Limitations and Future scope of work: The model is unable to accurately predict the band gap for near zero values. A possible explanation might be the lack of experimental data. MatMMFuse has been designed to work only with CIF Files and thus, performance might be improved with grounding in experimental data. Importantly, quantum effects become more dominant at very small energy gaps. Thus, a possible area of improvement would be the explicit incorporation of quantum effects. Furthermore, we believe that additional modalities might lead to an improvement in the model performance. The cross attention operation scales quadratically with sequence length which makes it computationally expensive for long sequences. Also, it is possible for one modality to

dominate the training process leading to imbalance. Thus, it would be interesting to explore alternate approaches for a balanced integration of modalities.

6 REPRODUCIBILITY

Data Availability: The Material project dataset analyzed during this study is available at https://next-gen.materialsproject.org/. The cubic oxide perovskite and chalcogenides datasets are available in the Computational Material Repository(https://cmr.fysik.dtu.dk). The JARVIS dataset is available at https://jarvis.nist.gov/.
Code availability: The code is available in the Github repository- https://github.com/AbhiroopBhattacharya/MatMMFuse . Moreover, the pseudocode is also provided in the Appendix section.

7 ACKNOWLEDGMENTS

SGC thanks the Canada Research Chair and the NSERC Discovery programs for their support.

AUTHOR CONTRIBUTIONS

The concept and methodology were planned and done by AB and SGC. The first version of manuscript was written by AB. The manuscript was reviewed and commented by SGC.

REFERENCES

Seoin Back, Junwoong Yoon, Nianhan Tian, Wen Zhong, Kevin Tran, and Zachary W Ulissi. Convolutional neural network of atomic surface structures to predict binding energies for high-throughput screening of catalysts. *The journal of physical chemistry letters*, 10(15):4401–4408, 2019.

Pooja Basera and Saswata Bhattacharya. Chalcogenide perovskites (abs3; a= ba, ca, sr; b= hf, sn): An emerging class of semiconductors for optoelectronics. *The Journal of Physical Chemistry Letters*, 13(28):6439–6446, 2022.

Jörg Behler. Atom-centered symmetry functions for constructing high-dimensional neural network potentials. *The Journal of chemical physics*, 134(7), 2011.

Iz Beltagy, Kyle Lo, and Arman Cohan. Scibert: A pretrained language model for scientific text. *arXiv preprint arXiv:1903.10676*, 2019.

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.

Chi Chen, Yunxing Zuo, Weike Ye, Xiangguo Li, Zhi Deng, and Shyue Ping Ong. A critical review of machine learning of energy materials. *Advanced Energy Materials*, 10(8):1903242, 2020.

Kamal Choudhary, Kevin F Garrity, Andrew CE Reid, Brian DeCost, Adam J Biacchi, Angela R Hight Walker, Zachary Trautt, Jason Hattrick-Simpers, A Gilad Kusne, Andrea Centrone, et al. The joint automated repository for various integrated simulations (jarvis) for data-driven materials design. *npj computational materials*, 6(1):173, 2020.

Kishalay Das, Pawan Goyal, Seung-Cheol Lee, Satadeep Bhattacharjee, and Niloy Ganguly. Crys-mmnet: multimodal representation for crystal property prediction. In *Uncertainty in Artificial Intelligence*, pp. 507–517. PMLR, 2023.

Sandip De, Albert P Bartók, Gábor Csányi, and Michele Ceriotti. Comparing molecules and solids across structural and alchemical space. *Physical Chemistry Chemical Physics*, 18(20):13754–13769, 2016.

Claudia Draxl and Matthias Scheffler. The nomad laboratory: from data sharing to artificial intelligence. *Journal of Physics: Materials*, 2(3):036001, 2019.

Felix Faber, Alexander Lindmaa, O Anatole Von Lilienfeld, and Rickard Armiento. Crystal structure representations for machine learning models of formation energies. *International Journal of Quantum Chemistry*, 115(16):1094–1101, 2015.

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

Alex M Ganose and Anubhav Jain. Robocrystallographer: automated crystal structure text descriptions and analysis. *MRS Communications*, 9(3):874–881, 2019.

Marco Gori, Gabriele Monfardini, and Franco Scarselli. A new model for learning in graph domains. In *Proceedings. 2005 IEEE International Joint Conference on Neural Networks, 2005.*, volume 2, pp. 729–734. IEEE, 2005.

Tatsumi Ishihara. Structure and properties of perovskite oxides. *Perovskite Oxide for Solid Oxide Fuel Cells*, pp. 1–16, 2009.

Kevin Maik Jablonka, Qianxiang Ai, Alexander Al-Feghali, Shruti Badhwar, Joshua D Bocarsly, Andres M Bran, Stefan Bringuier, L Catherine Brinson, Kamal Choudhary, Defne Circi, et al. 14 examples of how llms can transform materials science and chemistry: a reflection on a large language model hackathon. *Digital Discovery*, 2(5):1233–1250, 2023.

Anubhav Jain, Shyue Ping Ong, Geoffroy Hautier, Wei Chen, William Davidson Richards, Stephen Dacek, Shreyas Cholia, Dan Gunter, David Skinner, Gerbrand Ceder, et al. The materials project: A materials genome approach to accelerating materials innovation, apl mater. *Applied Physics Letter*, 2013.

Di Jin, Zhijing Jin, Joey Tianyi Zhou, and Peter Szolovits. Is bert really robust? a strong baseline for natural language attack on text classification and entailment. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pp. 8018–8025, 2020.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*, 2019.

Jaewan Lee, Changyoung Park, Hongjun Yang, Sungbin Lim, and Sehui Han. Cast: Cross attention based multimodal fusion of structure and text for materials property prediction. *arXiv preprint arXiv:2502.06836*, 2025.

Juanhui Li, Harry Shomer, Haitao Mao, Shenglai Zeng, Yao Ma, Neil Shah, Jiliang Tang, and Dawei Yin. Evaluating graph neural networks for link prediction: Current pitfalls and new benchmarking. *Advances in Neural Information Processing Systems*, 36, 2024.

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.

Mihai Masala, Stefan Ruseti, and Mihai Dascalu. Robert–a romanian bert model. In *Proceedings of the 28th International Conference on Computational Linguistics*, pp. 6626–6637, 2020.

Janghoon Ock, Joseph Montoya, Daniel Schweigert, Linda Hung, Santosh K Suram, and Weike Ye. Unimat: Unifying materials embeddings through multi-modal learning. *arXiv preprint arXiv:2411.08664*, 2024.

Aini Palizhati, Wen Zhong, Kevin Tran, Seoin Back, and Zachary W Ulissi. Toward predicting intermetallics surface properties with high-throughput dft and convolutional neural networks. *Journal of chemical information and modeling*, 59(11):4742–4749, 2019.

Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. *Neural Information Processing Systems*, 2017.

James E Saal, Scott Kirklin, Muratahan Aykol, Bryce Meredig, and Christopher Wolverton. Materials design and discovery with high-throughput density functional theory: the open quantum materials database (oqmd). *Jom*, 65:1501–1509, 2013.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019.

Franco Scarselli, Marco Gori, Ah Chung Tsoi, Markus Hagenbuchner, and Gabriele Monfardini. The graph neural network model. *IEEE transactions on neural networks*, 20(1):61–80, 2008.

Jonathan Schmidt, Mário RG Marques, Silvana Botti, and Miguel AL Marques. Recent advances and applications of machine learning in solid-state materials science. *npj Computational Materials*, 5 (1):83, 2019.

Kristof T Schütt, Huziel E Sauceda, P-J Kindermans, Alexandre Tkatchenko, and K-R Müller. Schnet–a deep learning architecture for molecules and materials. *The Journal of Chemical Physics*, 148(24), 2018.

Gwenaelle Cunha Sergio and Minho Lee. Stacked debert: All attention in incomplete data for text classification. *Neural Networks*, 136:87–96, 2021.

Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

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.

Ya Zhuo, Aria Mansouri Tehrani, and Jakoah Brgoch. Predicting the band gaps of inorganic solids by machine learning. *The journal of physical chemistry letters*, 9(7):1668–1673, 2018.

A APPENDIX

A.1 DATASET DESCRIPTION

A.1.1 MATERIALS PROJECT

We leverage the widely used Materials Project dataset (Jain et al., 2013). We focus on four important material properties: the formation energy per atom, the energy above the hull, the fermi energy and the Band Gap. The table 4 shows the distribution of the target variables.

Table 4: Summary Statistics for Target Variable for Materials Project Dataset

	Target Variable Statistics			
	Formation Energy (eV/atom)	Fermi Energy (eV)	Energy Above Convex Hull (eV/atom)	Band Gap (eV)
Mean	-1.66	3.069	0.022	0.874
Standard Deviation	1.009	2.776	0.244	1.514
Range	[-11.86, 5.45]	[-14.017, 19.41]	[0.00, 7.497]	[0.00, 17.891]
Median	-1.75	3.024	0.00	0.00

We use Robocrystallographer(Ganose & Jain, 2019) to convert the crystal data encoded in CIF file into text format. The distribution of the generated text descriptions are given below in the table5

Table 5: Summary Statistics for Text descriptions for Materials Project.

Text Description Statistics	
Average Length	741.4 words
Standard Deviation	1426.9 words
Range	[28, 49051] words

A.1.2 CUBIC OXIDE PEROVSKITES

We use the cubic oxide perovskite ABO_3 dataset from Computational material repository for evaluating the Zero shot performance of MatMMFuse. The summary statistics of the text descriptions are given in the table6.

Table 6: Summary Statistics for Text descriptions for Cubic Oxide perovskites($AB0_3$)

Text Description Statistics	
Average Length	136.6 words
Standard Deviation	31.6 words
Range	[79, 239] words

A.1.3 CHALCOGENIDES

Chalcogenide perovskites have the form $AB(S, Se)_3$. We have used them for evaluating the zero shot performance of MatMMFuse. The summary statistics of the text descriptions are tabulated in the table7.

A.2 PSEUDOCODE FOR IMPLEMENTING PROPOSED FRAMEWORK

MatMMFuse can be implemented using the following pseudocode.

Table 7: Summary Statistics for Text descriptions for Chalcogenide perovskites(ABS_3,$ABSe_3$)

Text Description Statistics	
Average Length	199.2 words
Standard Deviation	103.2 words
Range	[63, 1641] words

Algorithm 1 Fusion of Graph and Text Embeddings for Material Property Prediction

Require: CIF file $\mathcal{C}$, Text Description $\mathcal{T}$, Property Label y, Pretrained GNN $\mathcal{G}$, Pretrained Transformer $\mathcal{B}$, Attention Combiner $\mathcal{A}$, Learning Rate η, Cosine Warmup λ
Ensure: Trained Model for Property Prediction
Initialize model parameters θ
Split dataset into Train ($\mathcal{D}_{train}$), Validation ($\mathcal{D}_{val}$), and Test ($\mathcal{D}_{test}$)
for each epoch in $1, \ldots, N_{epochs}$ **do**
  for each batch $(\mathcal{C}_i, \mathcal{T}_i, y_i)$ in $\mathcal{D}_{train}$ **do**
    Extract Graph Features:
      Construct crystal graph G_i from CIF file $\mathcal{C}_i$
      Compute graph embedding: $h_G = \mathcal{G}(G_i)$
      Project embedding: $\tilde{h}_G = W_G h_G$
    Extract Text Features:
      Tokenize text: $X_T = \text{Tokenizer}(\mathcal{T}_i)$
      Compute transformer embedding: $h_T = \mathcal{B}(X_T)$
      Pool embedding: $h_T = \text{Mean}(h_T)$
      Project embedding: $\tilde{h}_T = W_T h_T$
    Fuse Representations using Attention:
      Compute query: $Q = W_Q \tilde{h}_T$
      Compute key: $K = W_K \tilde{h}_G$
      Compute value: $V = W_V \tilde{h}_G$
      Compute attention scores: $\alpha = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)$
      Compute attended representation: $h_{fused} = \alpha V$
      Apply residual connection: $h_{fused} = \text{LayerNorm}(h_{fused} + \tilde{h}_T)$
    Predict Property:
      $y_{\text{pred}} = \sigma(W_o h_{fused})$
    Compute Loss:
      $\mathcal{L} = \frac{1}{N} \sum_{i=1}^{N} (y_i - y_{\text{pred}})^2$
    Optimize Parameters:
      Compute gradients: $\nabla_\theta \mathcal{L}$
      Update parameters: $\theta \leftarrow \theta - \eta \cdot \lambda(t) \cdot \nabla_\theta \mathcal{L}$
  end for
  Evaluate on $\mathcal{D}_{val}$ and adjust learning rate η
end for
Test Model: Evaluate on $\mathcal{D}_{test}$