PAPER • **OPEN ACCESS**

MatMMFuse: Multimodal fusion model for material property prediction

To cite this article: Abhiroop Bhattacharya and Sylvain G Cloutier 2025 *Mach. Learn.: Sci. Technol.* **6** 045054

IOP Publishing Mach. Learn.: Sci. Technol. 6 (2025) 045054 https://doi.org/10.1088/2632-2153/ae1927

MACHINE LEARNING
Science and Technology

OPEN ACCESS

RECEIVED
3 June 2025

REVISED
30 September 2025

ACCEPTED FOR PUBLICATION
29 October 2025

PUBLISHED
25 November 2025

PAPER

MatMMFuse: Multimodal fusion model for material property prediction

Abhiroop Bhattacharya and Sylvain G Cloutier*

Department of Electrical Engineering, École de technologie supérieure, Montréal, Canada
* Author to whom any correspondence should be addressed.

E-mail: SylvainG.Cloutier@etsmtl.ca and abhiroop.bhattacharya.1@ens.etsmtl.ca

Keywords: deep learning, artificial intelligence, accelerated material design, multimodal

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 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 information such as space group and crystal symmetry. In this work, we propose Material MultiModal fusion, 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 joint automated repository for various integrated simulation 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 [1, 2]. 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 [3–5]. Historically, structural descriptors that meet rotational and translational invariance had been used for encoding the crystal structures, ranging from Coulomb matrix [6] and atom-centered symmetry functions to smooth overlap of atomic positions [7, 8].

First proposed more than 15 years ago, graph neural networks (GNNs) [9, 10] 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 [11]. Such GNNs have been implemented to predict materials in complex systems including surfaces [12, 13] and periodic crystal arrangements [14, 15]. The GNN models effectively encode and utilize the structure of the lattice. Particularly, the crystal graph convolution network (CGCNN) model [15] 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 [16]. 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 [17]. 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 [18]. Compared to 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 information from natural language with structural-aware learning from GNNs for crystal property prediction. Li *et al* [19] have used embedding concatenation for combining multiple modalities while, Ock *et al* [20] 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* [21] applied masked node prediction pretraining strategy to train a multimodal 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* [22] have developed CrysMMNet which uses concatenation to combine multiple modalities. These models have shown that using multimodal 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 CGCNN [15] and text embeddings of SciBERT [18]. 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 joint automated repository for various integrated simulation (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 multimodal 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 multimodal 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. Further, in the current setup, the text descriptions T are generated from the structure information S using the RoboCrystallographer framework [23]. The framework mimics the text description written by a human crystallographer given the structural information of an inorganic crystal. Robocrystallographer is explained in further detail in the appendix C. 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. A explanation of technical term have been provided in the appendix F.

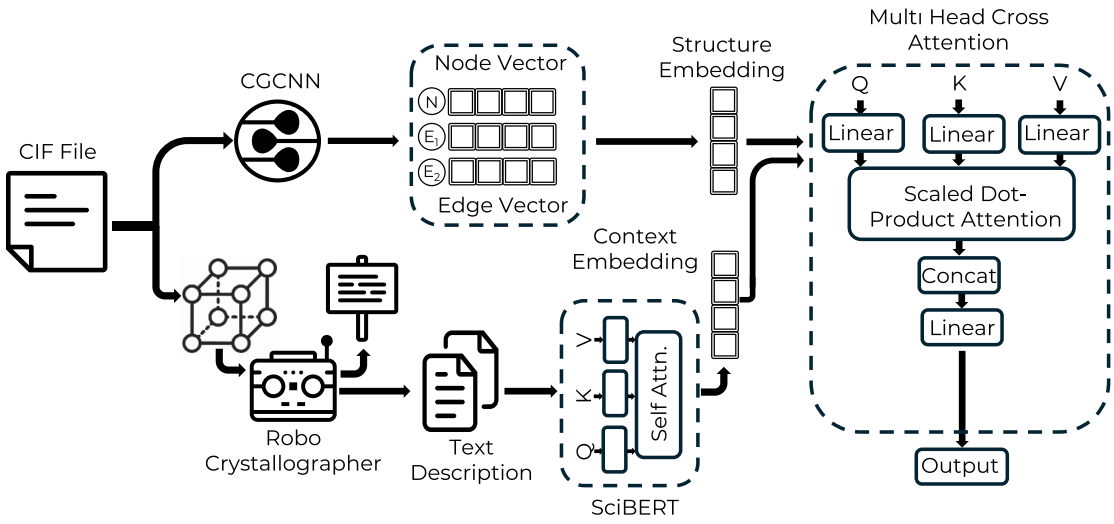


Figure 1. The figure provides an overview of MatMMFuse. The CGCNN model generates a structure aware embedding while the SciBERT model generates a context aware embedding which are combined using a multi-head cross attention mechanism.

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 is 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.

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

2.2. Text encoder

The description of the CIF files are generated using the Robocrystallography [23] framework. We leverage the scientific knowledge encoded in the pretrained SciBERT model [18] 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.

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

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

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

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.

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

The model has 12 transformer layers for encoding with 768 hidden dimensions and 12 attention heads. The architecture has been inherited from the pretrained SciBERT model which follows the BERT base configuration. 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

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. MatMMFuse applies the multi-head attention mechanism in two stages. First, as self-attention within SciBERT, and second, as cross-attention during the fusion process. Self-attention (in SciBERT) forms Q, K, and V from the same sequence, whereas our fusion applies cross-attention with text as queries and structure embeddings as keys/values. The attention mechanism is further explained in the appendix D.

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

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

$$\text{attention}_{\text{weights}} = \text{softmax}\left(\text{attention}_{\text{scores}}\right) \tag{8}$$

$$\text{combined} = \text{attention}_{\text{weights}} \cdot V. \tag{9}$$

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. Model training and evaluation

We use a Nvidia RTX 4090 graphics processing unit to run our experiments. The framework is implemented using the Pytorch library version [24]. We use AdamW [25] with a cosine warm-up schedule. The implementation uses SmoothL1Loss, which behaves differently near zero compared to a pure L1 penalty. Further, AdamW decouples weight decay from the gradient update, which improves regularization in large models. The AdamW optimizer and evaluation metrics have been explained in further detail in the appendix E.

3.1. Dataset

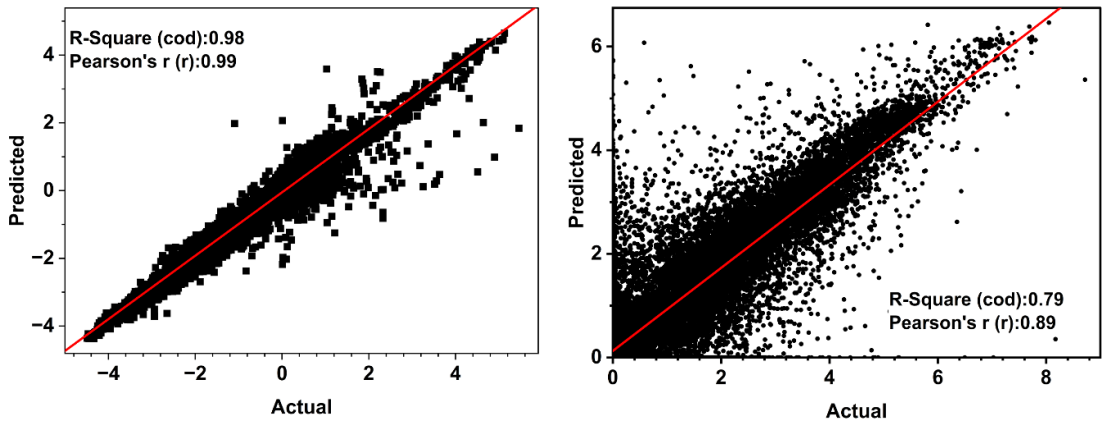
For model training and assessment, we leverage the widely used materials project dataset [5]. We focus on four important material properties: the formation energy per atom, the energy above the hull, the fermi energy and the band gap. We use 95582 crystal structures with a 80%,10%,10% train, validation and test split. For the CGCNN model, we directly use the data in CIF format. We use RoboCrystallographer [23] 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 the JARVIS dataset. The distribution of the target variables and text descriptions are available in the appendix A.1.

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. The model is trained on the materials project dataset. This trained model is then used to predict the material properties of Perovskites and Chalcogenides dataset without seeing any training examples. The model uses the structure-property relationships learned during training on the large material project data to predict the properties on the test Perovskite datasets. We benchmark the model against the unimodal graph (CGCNN) and text (SciBERT) models.

Table 1. Benchmarking model performance. The lower the error the better the model performance. The best results are highlighted in bold.

	Mean absolute error (MAE)			
	Formation energy (eV atom^{-1})	Fermi energy (eV)	Energy above convex hull (eV atom^{-1})	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**



(a) Formation Energy(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

The table 1 demonstrates the performance of MatMMFuse for the prediction of four important material properties—formation energy per atom (E_f), Fermi energy (E_g) and the band gap (B_g).We report the MAE (average absolute error) and R^2 (coefficient of determination) to study the model performance. 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 26% improvement over the vanilla versions of both the models. ML models have struggled with predictions of the band gap for crystals [26] for which the proposed model has an improvement of around 16% compared to the other models. We present the training and validation error curves for different training regimes in the appendix B.5 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.

To further investigate the results, the figure 2 compares the scatter plots of the actual versus predicted values for the formation energy per atom and the band gap for the test dataset. A detailed explanation of the statistical terms has been provided in the appendix F. 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.

To take a closer look at the incorrect predictions for band gap, we mapped the predicted values to the physical properties of the crystal. As shown in figure 3(a), MatMMFuse tends to mispredict the band gap for crystals which have both a low average atomic radius and high actual band gap. Although crystals with smaller average ionic radii generally have stronger orbital overlap and therefore smaller band gaps, this trend does not hold universally, which leads to mispredictions in certain cases. The mapping with other crystal properties have been presented in the appendix B.3. The plot 3(b) shows that although the model occasionally underestimates or overestimates the values in the test data as indicated by the orange outlier points. However, at a broader level, the model demonstrates strong generalization

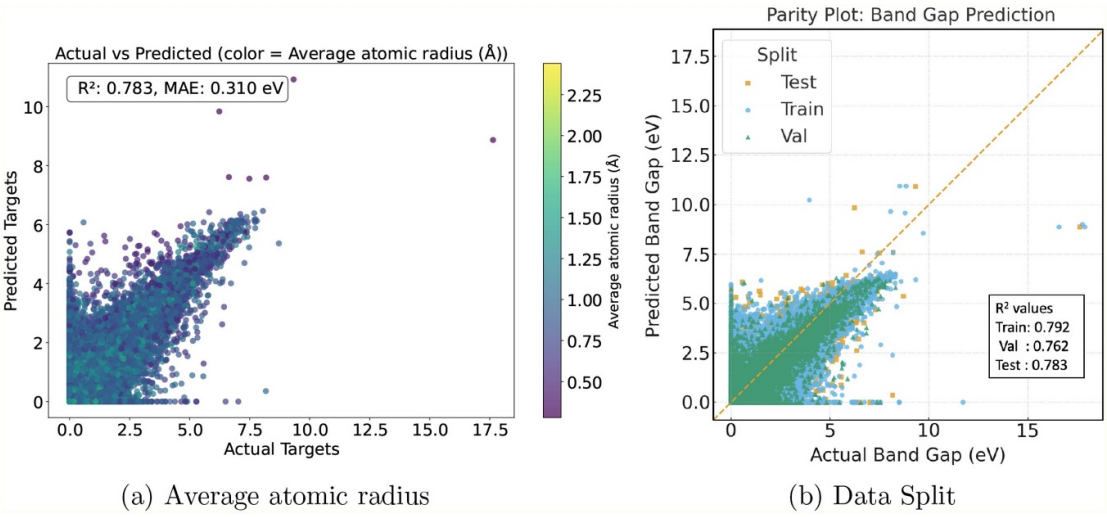


Figure 3. The parity plots study the the actual versus predicted values from different perspectives (a) property mapping to average atomic radius. (b) Training, validation and test data splits.

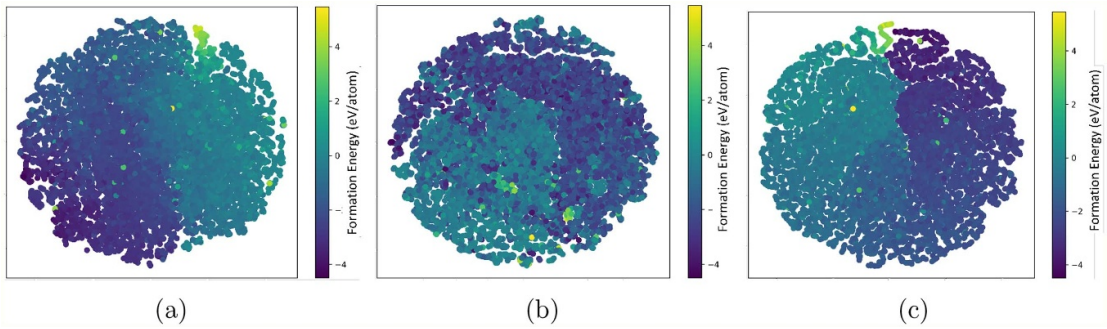


Figure 4. The t-SNE plot of the embedding from the embedding for the test dataset from materials project, with each point representing an individual crystal. Colors for each point are associated with their formation energies. (a) CGCNN embedding, (b) SciBERT embedding , (c) MatMMFuse embedding.

to unseen data. The colored parity plots for different test, validation and train dat splits for Formation energy per atom and Energy above convex hull are presented in the appendix B.2

The t-distributed stochastic neighbor method (t-SNE) [27] allows us to understand the decision boundaries and segregation of data points in the high dimensional embedding using 2D plots. The figure 4 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 similar pattern is observed for prediction of band gap. The t-SNE plots for band gap are included in the appendix B.4.

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

Table 2. Zero shot performance for predicting the energy. The lower the error the better the model performance.

	Mean absolute error (MAE)(eV)		
	CGCNN	SciBERT	Proposed
Perovskites(ABO_3)	1.42	2.84	**1.28**
Chalcogenides(ABS_3)	1.33	1.44	**1.05**

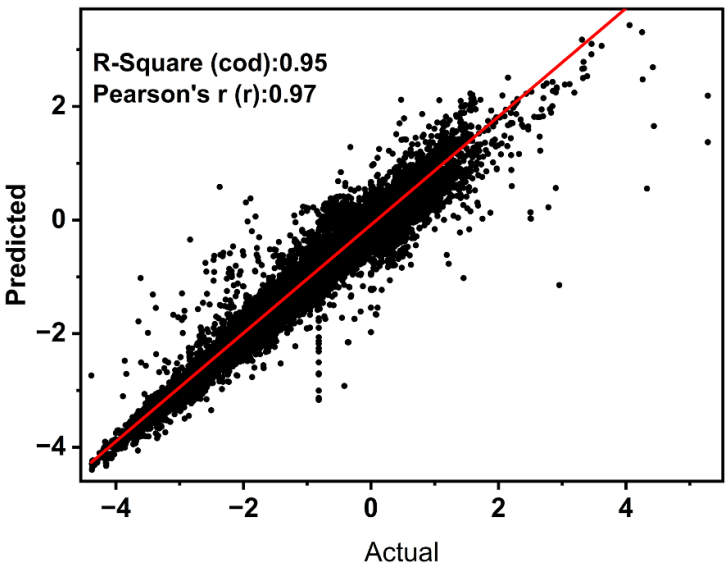


Figure 5. The scatter plot presents a comparison between the actual and the predicted values of formation energy per atom for the JARVIS dataset.

compare the zero shot performance of MatMMFuse for predicting the material property of Perovskites, Chalcogenides and the Jarvis dataset with the vanilla unimodal CGCNN and SciBERT models.

4.2.1. Cubic oxide perovskites

ABO_3 perovskites are viewed as promising resistive-type gas sensors [28]. The target variable is the formation energy per formula unit. 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 eV 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 [29]. To test our model, we repeated the same experiment for a dataset for $AB(S, Se)_3$ perovskites. We have used the relative energy as the target variable. This indicates the stability of a crystal by comparing with the lowest energy polymorph of the same chemical composition. The dataset has 1621 observations. Nonetheless, MatMMFuse achieves a low MAE of 1.05 eV, lower by 21% and 27% as compared to the CGCNN and SciBERT models respectively.

4.2.3. JARVIS

The JARVIS dataset [30] is a high-throughput materials database developed by the National Institute of Standards and Technology. The dataset encompasses a wide array of materials properties, computed using DFT simulations. MatMMFuse achieves a MAE of 0.078 eV/atom which is 48% lower than the CGCNN and the 59% lower than the SciBERT model. The actual versus predicted curve in figure 5 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^{-1} which have a much lower prediction.

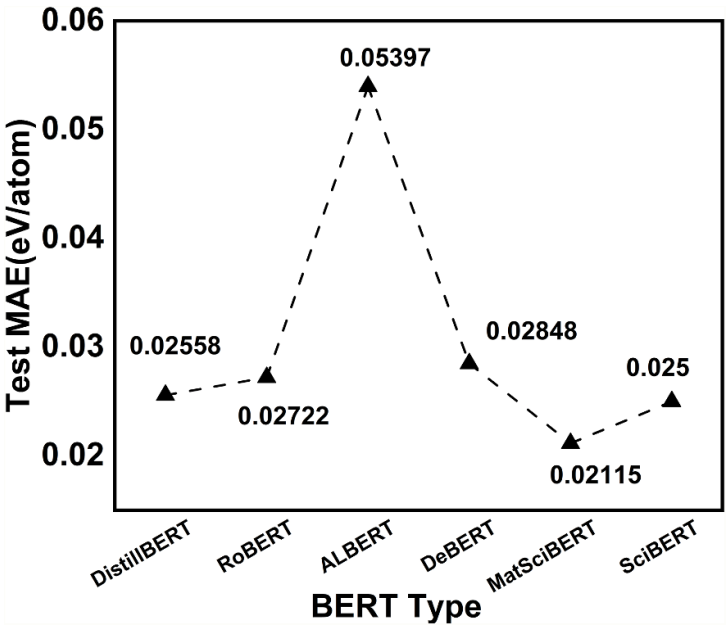


Figure 6. The plot compares the performance of different BERT models for encoding the text representation. MatSciBERT has the best performance and ALBERT has the worst performance.

Table 3. Zero shot performance of the MatSciBERT model for predicting the Energy. The lower the error the better the model performance.

	MAE (eV atom^{-1})	
	MATSciBERT	SciBERT
Perovskites (ABO_3)	2.26	**1.28**
Chalcogenides (ABS_3)	**1.33**	0.98

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. Additional ablation studies have been provided in the appendix B. They include the study of the training epochs and change in learning rate on the model performance.

4.3.1. Encoded domain knowledge

To prove our hypothesis that MatMMFuse is able to leverage the encoded knowledge in the BERT 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 [31], RoBERT [32], DeBERT [33] and DistillBERT [34]. 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 figure 6 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 table 3.

We decided to use SciBERT model because it has more interdisciplinary knowledge which leads to a 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 [35], MEGNet [14], 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

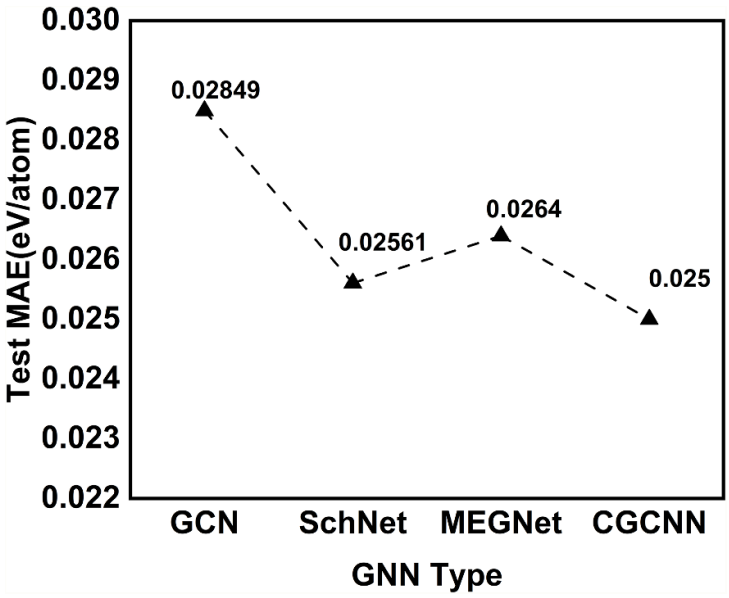


Figure 7. The plot compares the performance of different GNN models for encoding the lattice structure. CGCNN gives the optimum tradeoff between performance and efficiency.

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 figure 7 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 a significant improvement in model performance.The waterfall chart (figure 8) 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 [15]. 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 figure 9 shows that the model is able to converge to a low training loss.

4.3.5. K fold cross validation

We report the results for K fold validation ($K = 3$) of MatMMFuse for the prediction of band gap and Formation Energy per atom. We have reported the normalized MAE (nMAE) which is calculated by normalizing MAE by mean of the true values of the property. This enables us to compare error values across different properties. We also show the coefficient of determination(R^2). We observe as shown in figure 10 that there is no significant change in error values with the change in the split of the data. However, the error value is higher than the train and test errors in the single data split. We believe this might improve with training the model for each split to higher epochs and increasing the number of folds at a higher compute cost.

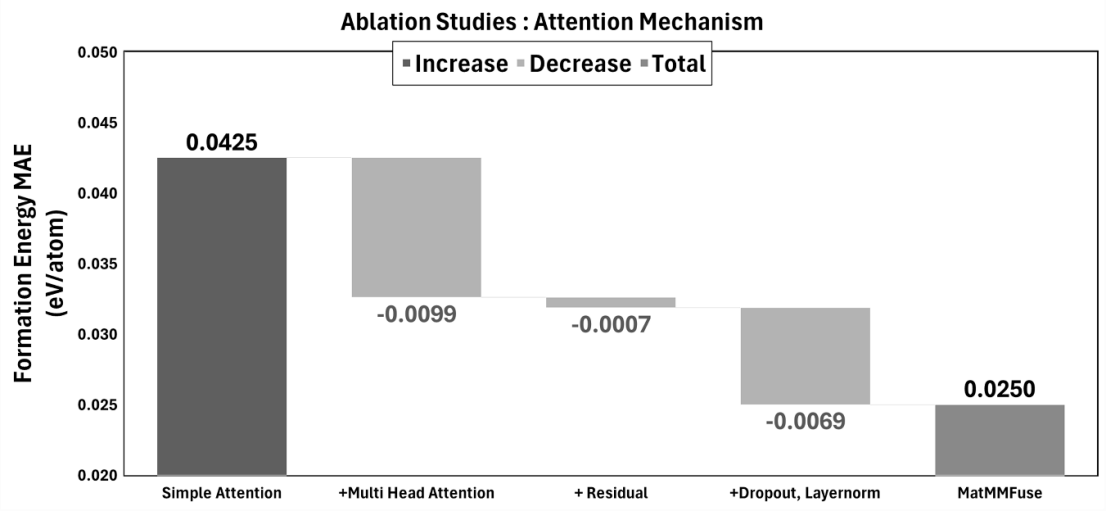


Figure 8. The waterfall chart shows the effect of adding individual components to improve the attention based fusion method.

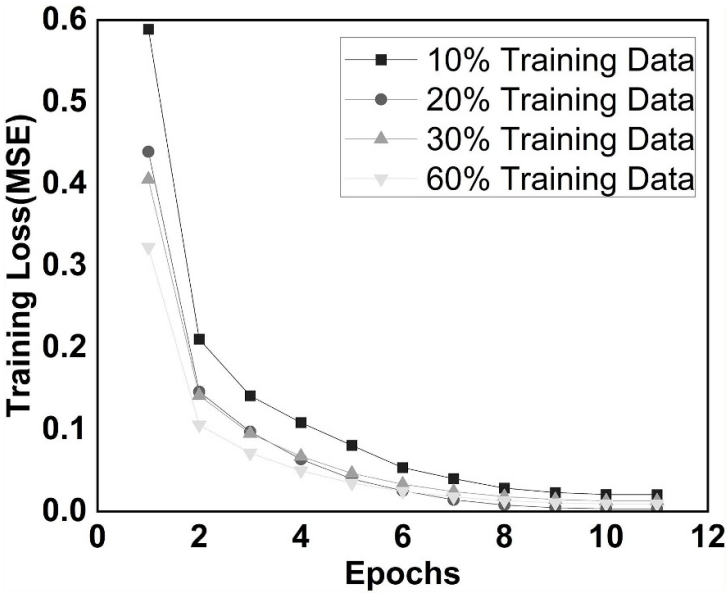


Figure 9. The chart shows that MatMMFuse is relatively robust to reduction in training data.

4.3.6. Corruption of text input

Corruption of text input remains a limitation of BERT models [36]. Moreover, Robocrystallographer might lead to corrupted text output if there are aberrations in the CIF file [23]. 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 figure 11 captures the degradation in model performance in training and inference with the increase in corruption of the text input.

5. Conclusion

This paper explores a multimodal fusion model for predicting material properties. The **MatMMFuse** model uses a multi-head cross attention based method for combining the embedding from GNN and a BERT 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

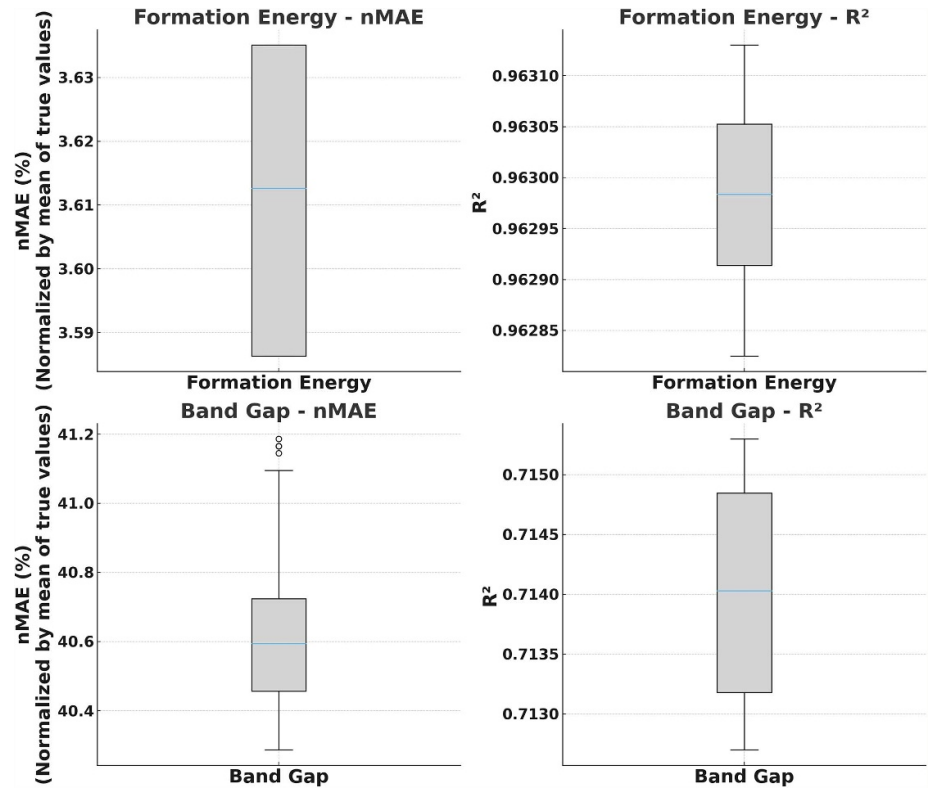


Figure 10. The box and whisker plot shows the nMAE and R^2 metrics for the prediction of band gap and formation energy per atom using K fold cross validation ($K = 3$).

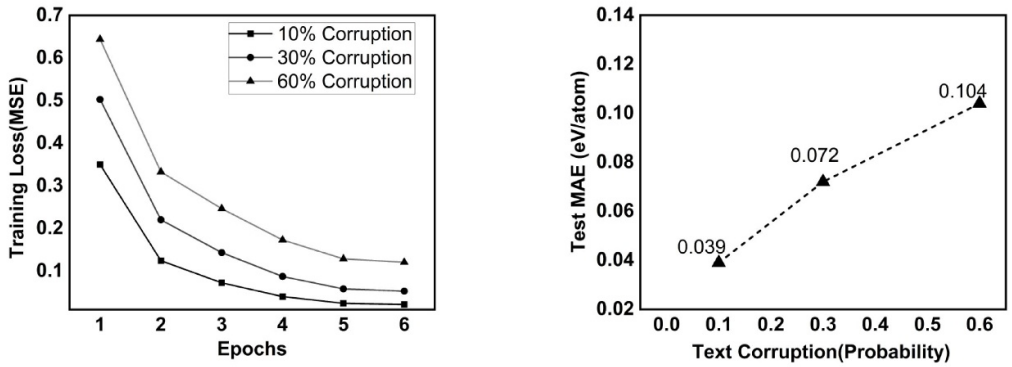


Figure 11. The figures 11(a) and (b) shows the effect of the corruption of the input text on the training loss and the test loss respectively. The model performance deteriorates significantly with corruption in text.

possess 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 the JARVIS datasets which is an important step for specialized uses cases. Analyzes 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.

Data availability statement

The data that support the findings of this study are openly available at the following URL/DOI: https://next-gen.materialsproject.org/, https://cmr.fysik.dtu.dk.

Acknowledgments

S G C acknowledges the NSERC-Discovery (Award : RGPIN-2022-03083) and Canada Research Chair (Award : CRC-2021-00490) programs for their financial support.

Code availability

The code is publicly available at the Github repository- https://github.com/AbhiroopBhattacharya/MatMMFuse. Moreover, the pseudocode is also provided in the appendix G.

Author contributions

Abhiroop Bhattacharya 0000-0003-2883-0754
Conceptualization (equal), Data curation (lead), Formal analysis (lead), Investigation (lead), Methodology (lead), Resources (equal), Software (lead), Validation (lead), Visualization (lead), Writing – original draft (lead), Writing – review & editing (equal)

Sylvain G Cloutier 0000-0003-0092-5241
Conceptualization (lead), Funding acquisition (lead), Project administration (lead), Resources (lead), Supervision (lead), Writing – review & editing (lead)

Appendix A

A.1. Dataset description

A.1.1. Materials project

We leverage the widely used materials project dataset [5]. 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 A1 shows the distribution of the target variables. Most of the crystals are thermodynamically stable compounds, as indicated by a mean formation energy of –1.66 eV atom^{-1} and a median of –1.75 eV atom^{-1}. The average energy above hull is only 0.022 eV atom^{-1}, the maximum value of 7.497 eV atom^{-1} reflects the presence of highly metastable phases. The Fermi energy distribution is centered near 3.07 eV with a substantial spread ($\sigma = 2.78$ eV) and extreme outliers (–14.02–19.41 eV), suggesting a diverse range of electronic structures. Band gaps are heavily skewed toward metallic behavior, with a median of 0.00 eV and more than half the compounds being gapless, although the range extends to 17.891 eV, encompassing wide-gap insulators. Overall, the materials project dataset covers a broad range of crystals.

We use Robocrystallographer [23] to convert the crystal data encoded in CIF file into text format. The distribution of the generated text descriptions are given below in the table A2.

Table A1. Summary statistics for target variable for materials project dataset.

	Target variable statistics			
	Formation energy (eV atom^{-1})	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

Table A2. Summary statistics for text descriptions for materials project.

Text description statistics	
Average length	741.4 words
Standard deviation	1426.9 words
Range	(28, 49 051) words

Table A3. 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

Table A4. 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

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 table A3. The average length of words is much smaller for the Perovskite and Chalcogenide datasets compared to the Materials project dataset with a smaller standard deviation. This indicates that MatMMFuse does not require detailed descriptions for accurate predictions.

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 table A4.

Appendix B. Analysis of results

B.1. Effect of epochs on model performance

The variation of the test loss with the number of epochs used for training the model is a good indicator of overfitting and model convergence. We observe that our model converges within a few epochs and changing the number of epochs does not affect the test performance. The figure B1 shows that the test loss is high when the model is trained only for 2 epochs. However, as we increase the number of epochs to 6 we observe that the model is able to generalize and achieve superior performance on the test dataset.

B.2. Generalization

To study the ability of the model to generalize, we have explored the shift in the test predictions compared to the true values for the prediction of formation energy per atom and energy above convex hull. The data points have been colored according to the train, test and validation splits. The figure B2 shows that MatMMFuse generalizes well to the unseen test dataset.

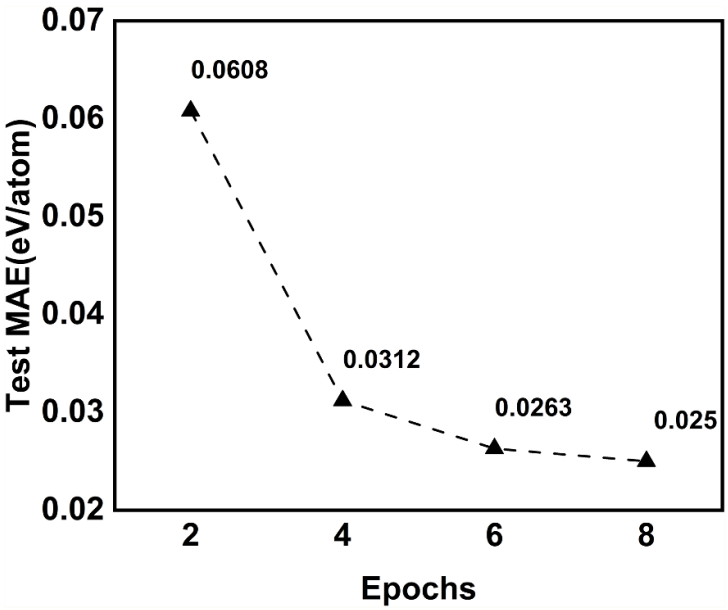


Figure B1. The line plot shows the generalization capability of MatMMFuse when trained for a low number of epochs. We observe that the model converges within a few epochs and generalizes well after 6 epochs.

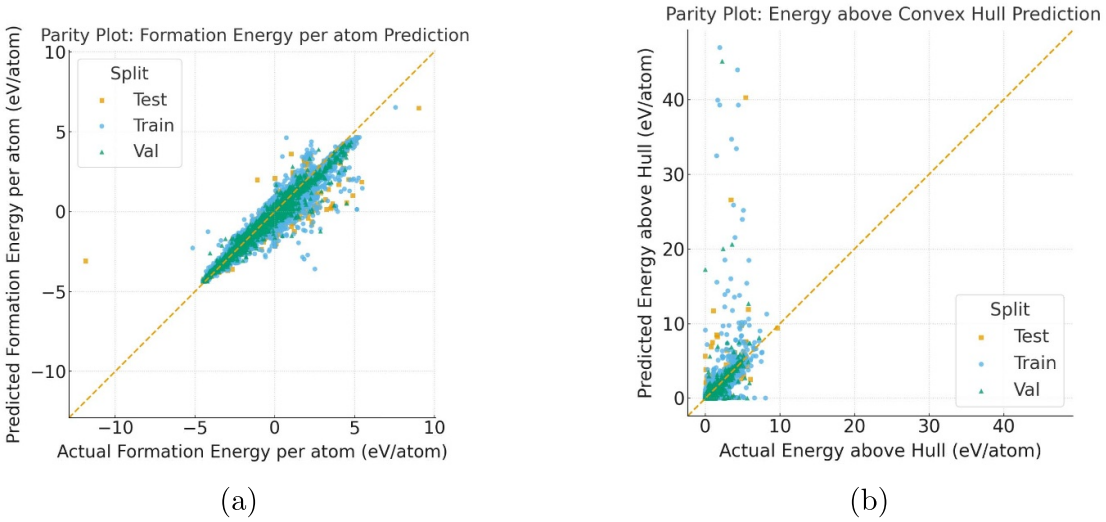


Figure B2. The colored parity plots for visualizing the ability of MatMMFuse to generalize to unseen Test and validation data. The plots have been colored according to the data split. (a) Formation energy per atom, (b) energy above convex hull.

B.3. Color mapped parity plots

Unit-cell volume and mass density quantify packing; higher density (smaller volume at fixed stoichiometry) generally strengthens interactions and alters phonon/electron transport coefficients. The number of atoms per primitive cell reflects structural complexity. Larger cells add vibrational and electronic modes and scattering channels, often reducing lattice thermal conductivity and broadening property dispersion. The number of distinct elements (chemical complexity) governs electronegativity contrast and bonding heterogeneity, enabling phase stabilization via configurational entropy and tuning electronic and ionic transport. The plots show the actual versus predicted values for prediction of band gap (figure B3) and the formation energy per atom (figure B4) for the crystals. Unfortunately, We do not observe a clear pattern in the behavior of outliers with the physical properties.

B.4. t-SNE

The figure B5 shows the t-SNE plot for the prediction of band gap on the materials project test data. We can observe that the supervised embedding has many cluster spread throughout the plot, the BERT

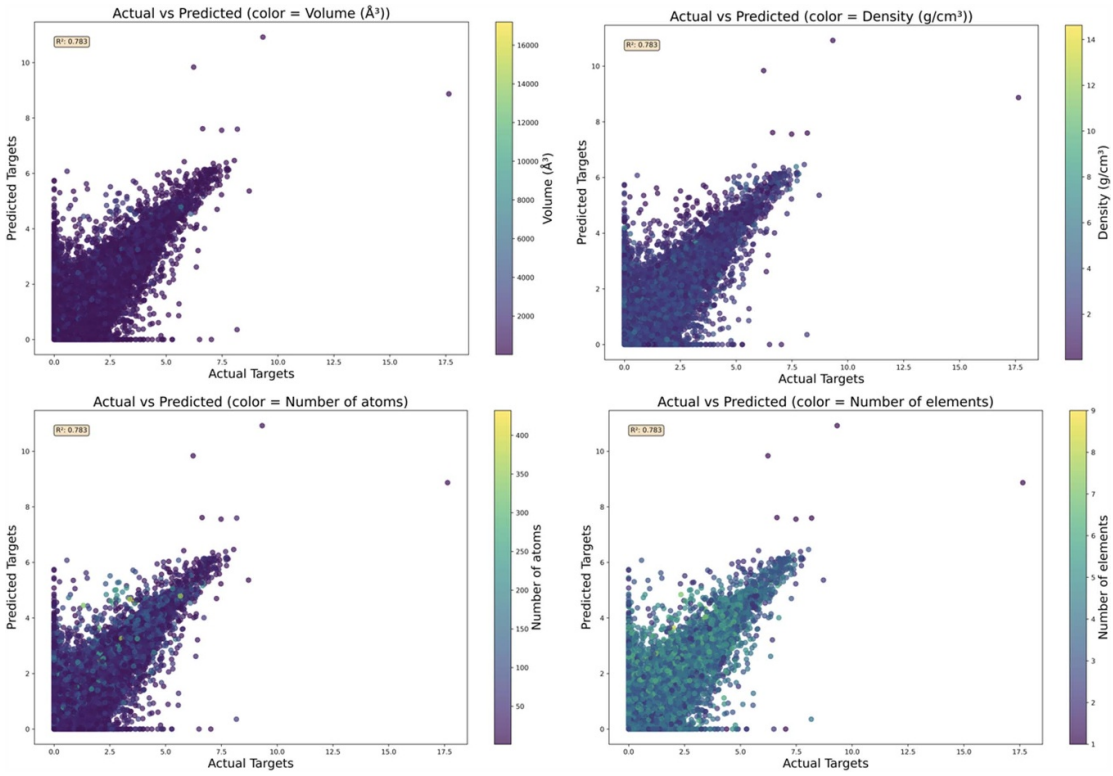


Figure B3. The line plot shows the predicted and true values for the band gap. The color mapping shows different physical properties of the crystal system.

model has denser clusters while MatMMFuse has clearer lobe shaped dense clusters for materials with lower band gap.

B.5. Grid search for learning rate

We have carried out a grid search of the learning rate using four values 1×10^{-03}, 1×10^{-04}, 1×10^{-05} and 1×10^{-06}. Interestingly, we observe that apart from the high learning rate of 1×10^{-03}, the model starts with a low loss after the first epoch and quickly converges. The figure B6(a) shows that the training is more unstable for higher learning rates of 1×10^{-03} and 1×10^{-04}. The learning rate of 1×10^{-06} is too low and leads to slower convergence for some material properties like band gap. Hence, we have selected 1×10^{-05} as the optimum learning rate. The figure B6(b) shows that model generalizes better when we train using lower learning rates.

Further, we have also investigated the effect of freezing the transformer parameters on model overfitting during model training by plotting the training and validation loss along with a rough indicator of overfitting using the difference Validation and train loss values. For this experiment, we initially freeze the transformer parameters and then unfreeze them after some epochs. The figure B7 shows that for the first scenario where we unfreeze the transformer after 3 epochs and train for a total of 10 epochs. In the second scenario, we unfreeze the transformer parameter after 5 epochs and train for 25 epochs. As we keep on increasing the epochs, the model starts overfitting. When we freeze the transformer parameters, the slope of the training curve is steeper compared to when we unfreeze the transformer.

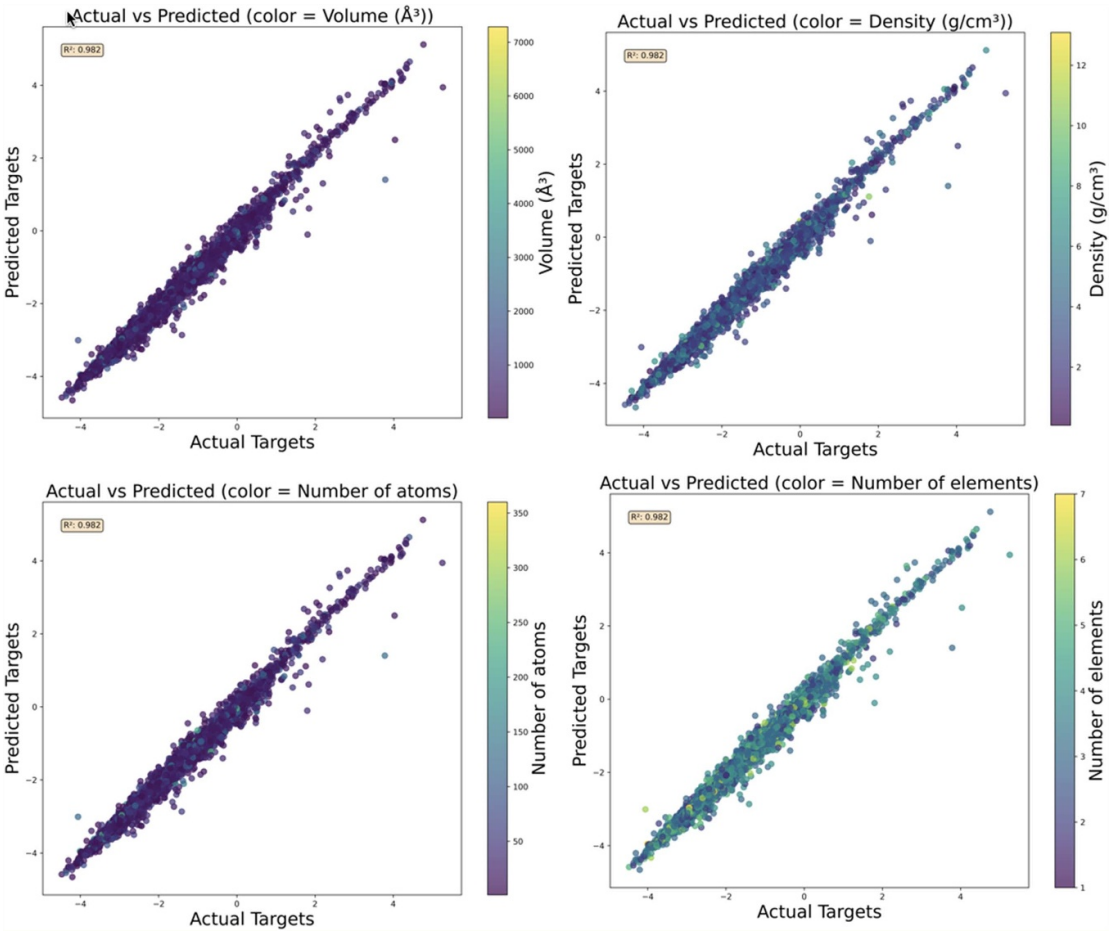
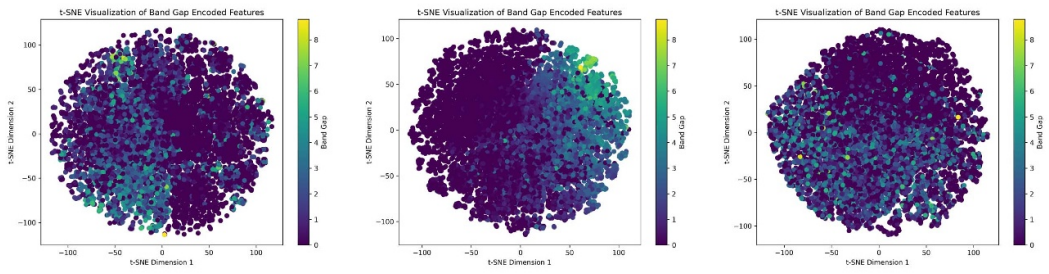
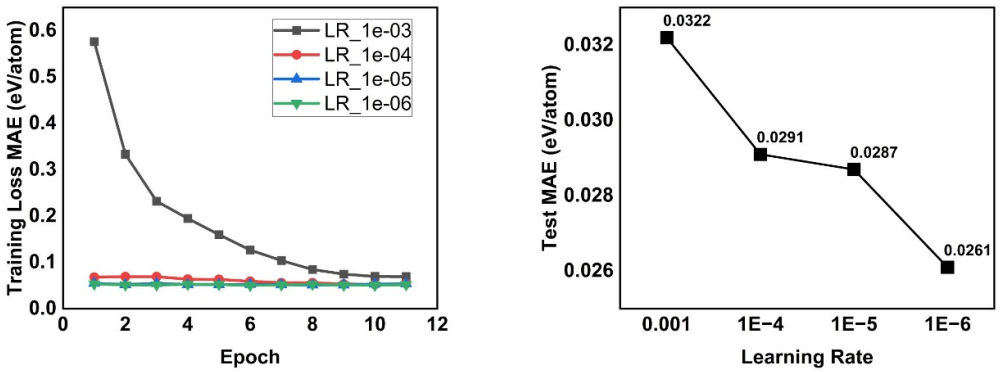


Figure B4. The line plot shows the predicted and true values for the formation energy per atom. The color mapping shows different physical properties of the crystal system.



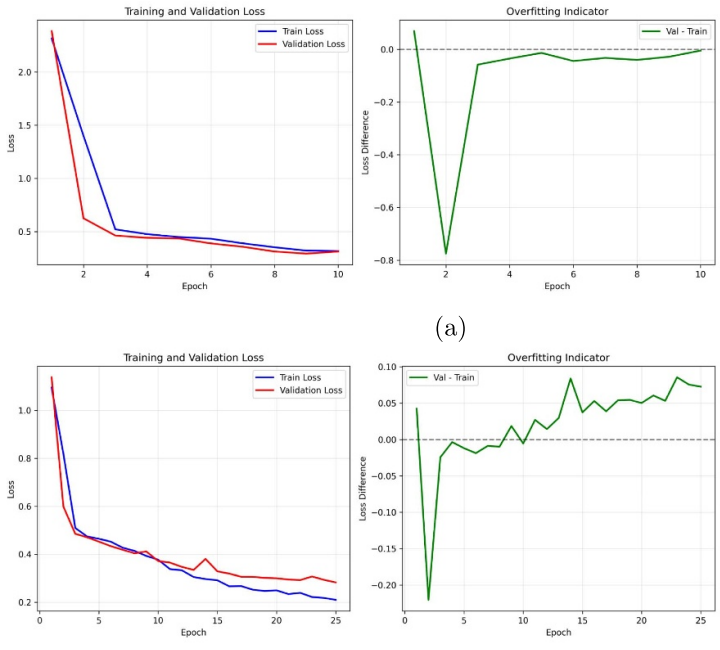
(a) (b) (c)

Figure B5. 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.



(a) (b)

Figure B6. The line plot B6(a) shows the convergence of training loss with change in learning rate. The plot B6(b) captures the change in test loss as we use models with different learning rates.



(a)

(b)

Figure B7. The line plot B7(a) shows that the learning curve behavior when we unfreeze the parameters after 3 epochs and train for 10 epochs while the plot B7(b) captures the model performance when we unfreeze the parameters after 5 epochs and train for 25 epochs leading to overfitting.

Appendix C. Robocrystallographer framework

Robocrystallographer is a tool that automatically generates descriptions of crystal structures from crystallographic information files (CIFs) [23]. These descriptions are designed to mimic how a human crystallographer would write about a structure, using standardized natural language templates. The CIF describing the structure of a crystal is used as input. Robocrystallographer parses the CIF to extract the space group, lattice parameters, symmetry operations, bond lengths and coordination environments. This information is written into descriptions using a rule based system. The text description of the crystal structure enables a LLM model to generate context aware embeddings. The text descriptions contains complementary information such as space group and symmetry operations which are not present in the graph descriptions. The primary limitation is that, for complex crystal systems, Robocrystallographer might generate text descriptions with errors and corrupted text.

Appendix D. Attention mechanism

The attention mechanism used in transformers relies on the Query (Q), Key (K) and Value (V) matrices. The query matrix represents what we are looking for in the data. The key matrix contains potential matches and the value matrix contains the actual data to be aggregated. First the transformer computes the similarity scores by calculating a dot product of the query and key matrices. This is then scaled by using the dimension of the key matrix. This is then converted into a probability distribution by using the softmax function. **Softmax** A function that converts vector of raw values into probabilities. This is multiplied with the value matrix to generate the context aware embedding. There are two types of attention mechanisms which are used with transformer architectures. The first one is *self attention* where the Q, K and V matrices are generated by using the projections of the same input vector into the three Q, K and V components. The other mechanism is *cross attention* wherein a different input vector is projected into the Q,K and V matrices. The manuscript uses self attention in the SciBERT Models to generate the context aware text embeddings and cross attention to dynamically fuse the graph and text embeddings. Often used in attention mechanisms to assign weights to inputs. Cross attention fusion allows the model to dynamically focus on different features unlike **Concat** wherein the embeddings from different modalities are combined end-to-end along a dimension.

Appendix E. AdamW optimizer

AdamW is an optimizer which improves on the standard Adam optimizer by decoupling weight decay from the gradient updates [25]. This helps prevent overfitting in large models by applying weight decay directly to the weights. The equations for the AdamW optimizer are explained below. We have an optimizer with a learning rate of η and β_1 and β_2 be the exponential decay rates for the moment estimates and λ be the coefficient for weight decay,

$$
\begin{aligned}
&\textbf{Given: } \theta_t \text{ (parameters at time } t),\\
&\quad g_t = \nabla_\theta \mathcal{L}_t(\theta_t) \text{ (gradient at time } t)\\
&\quad m_t = \beta_1 m_{t-1} + (1-\beta_1) g_t\\
&\quad v_t = \beta_2 v_{t-1} + (1-\beta_2) g_t^2\\
&\quad \hat{m}_t = \frac{m_t}{1-\beta_1^t}\\
&\quad \hat{v}_t = \frac{v_t}{1-\beta_2^t}\\
&\textbf{Weight decay step:} \quad \theta_t \leftarrow \theta_t - \eta \cdot \lambda \cdot \theta_t\\
&\textbf{Adam update step:} \quad \theta_{t+1} = \theta_t - \eta \cdot \frac{\hat{m}_t}{\sqrt{\hat{v}_t} + \epsilon}.
\end{aligned}
$$

Appendix F. Statistical terms

This section describes the statistical terms used in the manuscript. **Pearson correlation:** This measures the linear correlation between the actual and predicted values. The range of the values are from −1 to +1 with −1 indicating a strong inverse correlation and +1 indicating a strong positive correlation. **R^2 (coefficient of determination):** It is a statistical metric that measures how well a regression model explains the variability of the target variable,

$$R^2 = 1 - \frac{SS_{\text{res}}}{SS_{\text{tot}}} \tag{F1}$$

where SS_{res} is the residual sum of squares and SS_{tot} is the total sum of squares. An R^2 value of 1.0 indicates that the prediction explains all variability and a value close to 0 indicates that the model explains no variability beyond the mean.

Appendix G. Pseudocode for implementing proposed framework

MatMMFuse can be implemented using the following pseudocode.

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 $\mathcal{D}_{test}$

References

[1] Schmidt J, Marques M R, Botti S and Marques M A 2019 *npj Comput. Mater.* **5** 83
[2] Chen C, Zuo Y, Ye W, Li X, Deng Z and Ong S P 2020 *Adv. Energy Mater.* **10** 1903242
[3] Saal J E, Kirklin S, Aykol M, Meredig B and Wolverton C 2013 *JOM* **65** 1501–9
[4] Draxl C and Scheffler M 2019 *J. Phys. Mater.* **2** 036001
[5] Jain A *et al* 2013 *Appl. Phys. Lett.*
[6] Faber F, Lindmaa A, Von Lilienfeld O A and Armiento R 2015 *Int. J. Quantum Chem.* **115** 1094–101
[7] Behler J 2011 *J. Chem. Phys.* **134** 074106
[8] De S, Bartók A P, Csányi G and Ceriotti M 2016 *Phys. Chem. Chem. Phys.* **18** 13754–69
[9] Scarselli F, Gori M, Tsoi A, Hagenbuchner M and Monfardini G 2008 *IEEE Trans. Neural Netw.* **20** 61–80
[10] Gori M, Monfardini G and Scarselli F 2005 A new model for learning in graph domains *Proc. 2005 IEEE Int. Joint Conf. on Neural Networks* vol 2 (IEEE) pp 729–34
[11] Li J, Shomer H, Mao H, Zeng S, Ma Y, Shah N, Tang J and Yin D 2024 *Advances in Neural Information Processing Systems* vol 36
[12] Palizhati A, Zhong W, Tran K, Back S and Ulissi Z W 2019 *J. Chem. Inf. Model.* **59** 4742–9
[13] Back S, Yoon J, Tian N, Zhong W, Tran K and Ulissi Z W 2019 *J. Phys. Chem. Lett.* **10** 4401–8
[14] Chen C, Ye W, Zuo Y, Zheng C and Ong S P 2019 *Chem. Mater.* **31** 3564–72
[15] Xie T and Grossman J C 2018 *Phys. Rev. Lett.* **120** 145301
[16] Fung V, Zhang J, Juarez E and Sumpter B G 2021 *npj Comput. Mater.* **7** 84
[17] Jablonka K M *et al* 2023 *Dig. Discov.* **2** 1233–50
[18] Beltagy I, Lo K and Cohan A 2019 Scibert: a pretrained language model for scientific text *Proc. 2019 Conf. on Empirical Methods in Natural Language Processing and the 9th Int. Joint Conf. on Natural Language Processing (EMNLP-IJCNLP)* pp 3615–20
[19] Li Y, Gupta V, Kilic M N T, Choudhary K, Wines D, Liao W k, Choudhary A and Agrawal A 2025 *Dig. Discov.*
[20] Ock J, Montoya J H, Schweigert D, Hung L, Suram S K and Ye W 2024 *CoRR* **3** 842–68
[21] Lee J, Park C, Yang H, Lim S and Han S 2025 *CoRR* (arXiv:2502.06836)
[22] Das K, Goyal P, Lee S C, Bhattacharjee S and Ganguly N 2023 *Crysmmnet: Multimodal Representation for Crystal Property Prediction Uncertainty in Artificial Intelligence* (PMLR) pp 507–17
[23] Ganose A M and Jain A 2019 *MRS Commun.* **9** 874–81
[24] Paszke A, Gross S, Chintala S, Chanan G, Yang E, DeVito Z, Lin Z, Desmaison A, Antiga L and Lerer A 2017 *Neural Information Processing Systems*
[25] Loshchilov I and Hutter F 2017 *CoRR* (arXiv:1711.05101)
[26] Zhuo Y, Mansouri Tehrani A and Brgoch J 2018 *J. Phys. Chem. Lett.* **9** 1668–73
[27] Van der Maaten L and Hinton G 2008 *J. Mach. Learn. Res.* **9** 2579–605
[28] Ishihara T 2009 *Perovskite Oxide for Solid Oxide Fuel Cells* 1–16
[29] Basera P and Bhattacharya S 2022 *J. Phys. Chem. Lett.* **13** 6439–46
[30] Choudhary K *et al* 2020 *npj Comput. Mater.* **6** 173
[31] Lan Z, Chen M, Goodman S, Gimpel K, Sharma P and Soricut R 2020 Albert: a lite bert for self-supervised learning of language representations *Int. Conf. on Learning Representations*
[32] Masala M, Ruseti S and Dascalu M 2020 Robert–a romanian bert model *Proc. 28th Int. Conf. on Computational Linguistics* pp 6626–37
[33] Sergio G C and Lee M 2021 *Neural Netw.* **136** 87–96
[34] Sanh V 2019 Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter *Proc. 33rd Conf. on Neural Information Processing Systems (NIPS2019)*
[35] Schütt K T, Sauceda H E, Kindermans P J, Tkatchenko A and Müller K R 2018 *J. Chem. Phys.* **148**
[36] Jin D, Jin Z, Zhou J T and Szolovits P 2020 Is bert really robust? a strong baseline for natural language attack on text classification and entailment *Proc. AAAI Conf. on Artificial Intelligence* vol 34 pp 8018–25