CrysMMNet: Multimodal Representation for Crystal Property Prediction.

Abstract

Machine Learning models have emerged as a powerful tool for fast and accurate prediction of different crystalline properties. Exiting state-of-the-art models rely on a single modality of crystal data i.e crystal graph structure, where they construct multi-graph by establishing edges between nearby atoms in 3D space and apply GNN to learn materials representation. Thereby, they encode local chemical semantics around the atoms successfully but fail to capture important global periodic structural information like space group number, crystal symmetry, rotational information etc, which influence different crystal properties. In this work, we leverage textual descriptions of materials to model global structural information into graph structure to learn a more robust and enriched representation of crystalline materials. To this effect, we first curate a textual dataset for crystalline material databases containing descriptions of each material. Further, we propose CrysMMNet, a simple multimodal framework, which fuses both structural and textual representation together to generate a joint multimodal representation of crystalline materials. We conduct extensive experiments on benchmark datasets across ten different properties to show that CrysMMNet outperforms existing state-of-the-art baseline methods with a good margin. We also observe fusing textual representation with crystal graph structure provides consistent improvement for all the SOTA GNN models compared to their own vanilla version. We are going to share the textual dataset, that we have curated for both the benchmark material databases with the community for future use.

1 INTRODUCTION

In the recent past, we have witnessed a surge of interest in developing machine learning models Seko et al. [2015], Pilania et al. [2015], Lee et al. [2016], De Jong et al. [2016], Seko et al. [2017], Isayev et al. [2017], Ward et al. [2017], Lu et al. [2018], Im et al. [2019] for fast and accurate property prediction of crystalline materials. Crystalline materials are typically modeled by a minimal unit cell containing all the constituent atoms in different coordinates, repeated infinite times in 3D space on a regular lattice, which makes material structures periodic in nature. A key challenge in learning crystal representation is how to capture accurately global periodic structural information along with local chemical semantics. Recent state-of-the-art models Xie and Grossman [2018], Chen et al. [2019], Louis et al. [2020], Park and Wolverton [2020], Schmidt et al. [2021], Choudhary and DeCost [2021], Hsu et al. [2021], Das et al. [2022], Yan et al. [2022] construct multi-edge graphs for a 3D material structure where they create edges between nearby atoms within a pre-specified distance threshold in 3D space and apply GNN model to learn representations of crystal structures that are optimized for downstream property prediction tasks. Although existing variants of GNN models predict different crystal properties with high precision, they rely on a single modality of crystal data i.e crystal graph structure which limits the expressive power of these models. The architectural innovations of these approaches are primarily based on incorporating specific domain knowledge of the local bonding environment, such as explicitly encoding bond angle Choudhary and DeCost [2021], dihedral angle Hsu et al. [2021], etc. but they fail to incorporate crucial global periodic structural information like lattice constraint, space group number, crystal symmetry, rotational information, component 3D orientation, heterostructure information, etc, which will enrich its representation and subsequently aid the property prediction accuracy.

In this work, we propose to learn a more robust and enriched representation by using multi-modal data i.e graph structure and textual description of materials. One of the major advan-

Submitted to the 39th Conference on Uncertainty in Artificial Intelligence (UAI 2023). **To be used for reviewing only**.

tages of using the textual description of materials is it provides a diverse set of periodic structural information which is useful to estimate different crystal properties but difficult to incorporate explicitly into a graph structure. Leveraging textual modalities beyond graph structures of materials remains unexplored by the research community and to the best of our knowledge, there is no existing dataset containing textual descriptions of the materials. Hence, we first curate the textual dataset of two popular materials databases (Graph-based), Material Project (MP) and JARVIS, containing textual descriptions of each material of those databases. We used a popular tool robocrystallographer Ganose and Jain [2019] to generate descriptions for global crystal structures, which looks at the structural symmetry, local environment, and extended connectivity to generate a description that includes space group number, crystal symmetry, rotational information, component orientations, heterostructure information, etc.

Further, we propose, CrysMMNet (**Crys**tal **M**ulti-**M**odal **Net**work), a simple multi-modal framework for crystalline materials, which has two components: Graph Encoder and Text Encoder. Given a material, Graph Encoder uses its graph structure and applies GNN based approach to encode the local neighborhood structural information around a node (atom), and subsequently learn graph (crystal) representation. On the contrary, Text Encoder is a transformer-based model, which encodes the global structural knowledge from the textual description of the material and generates a textual representation. Finally, both graph structural and textual representation are fused together to generate a more enriched multimodal representation of materials, which captures both global and local structural knowledge and subsequently improves property prediction accuracy.

To show the merit of our proposed algorithm, we performed comprehensive experiments on two popular benchmark datasets, Materials Project and JARVIS-DFT, across ten diverse sets of properties and compare the results with popular state-of-the-art models. We observe that for all the properties CrysMMNet can achieve the lowest error in comparison with other baseline models. In addition, our results demonstrate that multi-modal representation learning helps to achieve even better improvements when the dataset is sparse. We also perform some ablation studies to investigate the expressiveness of textual representation and robustness of multimodal representation on different GNN architectural choices. Result shows, textual representations alone are not expressive enough to learn the structure-property relationship of the materials. Moreover, fusing both graph structural and textual representation together leads to substantial performance improvements for all the state-of-the-art GNN models compared to their vanilla versions. We also investigate the influence of local compositional information and global material structural knowledge encoded through textual representation and found for all crystal properties both local and global knowledge improves the downstream property prediction accuracy. We are going to share the textual dataset, that we have curated for both the benchmark material databases with the community for future use.

2 BACKGROUND AND RELATED WORK

Crystal Representation. The structure of a crystalline material can be modeled by a minimum unit cell, repeated infinite times in three-dimensional (3D) euclidean space on a regular lattice, which makes the crystalline structure periodic in nature. As mentioned in Xie et al. [2021], Yan et al. [2022], for a given crystal we can describe its unit cell by two matrices: Feature Matrix ($\mathbf{X}$) and Coordinate Matrix ($\mathbf{C}$). Feature Matrix $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n]^T \in R^{n \times d}$ denotes atomic feature set of the material, where $\mathbf{x}_i \in R^d$ corresponds to the d-dimensional feature vector of i-th atom. On the other hand, Coordinate Matrix $\mathbf{C} = [\mathbf{c}_1, \mathbf{c}_2, ..., \mathbf{c}_n]^T \in R^{n \times 3}$ denotes atomic coordinate positions, where $\mathbf{c}_i \in R^3$ corresponds to cartesian coordinates of i-th atom in the unit cell. Further, there is an additional lattice matrix $\mathbf{L} = [\mathbf{l}_1, \mathbf{l}_2, \mathbf{l}_3]^T \in R^{3 \times 3}$, which describes how a unit cell repeats itself in the 3D space towards $\mathbf{l}_1, \mathbf{l}_2$ and $\mathbf{l}_3$ direction to form the periodic 3D structure of the material. Formally, a given crystal can be defined as $\mathbf{M} = (\mathbf{X}, \mathbf{C}, \mathbf{L})$ and we can represent its infinite periodic structure as

$$\hat{\mathbf{C}} = \{\hat{\mathbf{c}}_i | \hat{\mathbf{c}}_i = \mathbf{c}_i + \sum_{j=1}^{3} k_j \mathbf{l}_j\}; \; \hat{\mathbf{X}} = \{\hat{\mathbf{x}}_i | \hat{\mathbf{x}}_i = \mathbf{x}_i\} \quad (1)$$

where $k_1, k_2, k_3, i \in Z, 1 \leq i \leq n$.

Crystal Property Prediction using GNNs. Graph neural network (GNN) based approaches have been very effective to encode structural information of the crystal materials into enriched embedding space so that it can predict different crystal properties with high accuracy. CGCNN Xie and Grossman [2018] is the first proposed model, which represents 3D crystal structure as an undirected weighted multi-edge graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{X}, \mathcal{F})$ where $\mathcal{V}$ denotes the set of nodes(atoms) in the unit cell of material and $\mathcal{E} = \{(u, v, k_{uv})\}$ denotes a multi-set of node pairs and k_{uv} denotes number of edges between a node pair (u, v). $\mathcal{X} = \{(x_u | u \in \mathcal{V})\}$ denotes the node feature set, which includes different chemical properties like electronegativity, valance electron, covalent radius, etc. Finally, $\mathcal{F}_i = \{\{s^k\}_{(u,v)} | (u, v) \in \mathcal{E}, k \in \{1..k_{uv}\}\}$ denotes the multi-set of edge weights where s^k corresponds to the kth bond length between a node pair (u, v), which signifies the inter-atomic bond distance between two atoms. Further, CGCNN develops a graph convolution neural network to update node features based on their local chemical and structural environment.

Following CGCNN, there are a lot of subsequent studies Chen et al. [2019], Louis et al. [2020], Park and Wolverton [2020], Schmidt et al. [2021], where authors proposed differ-

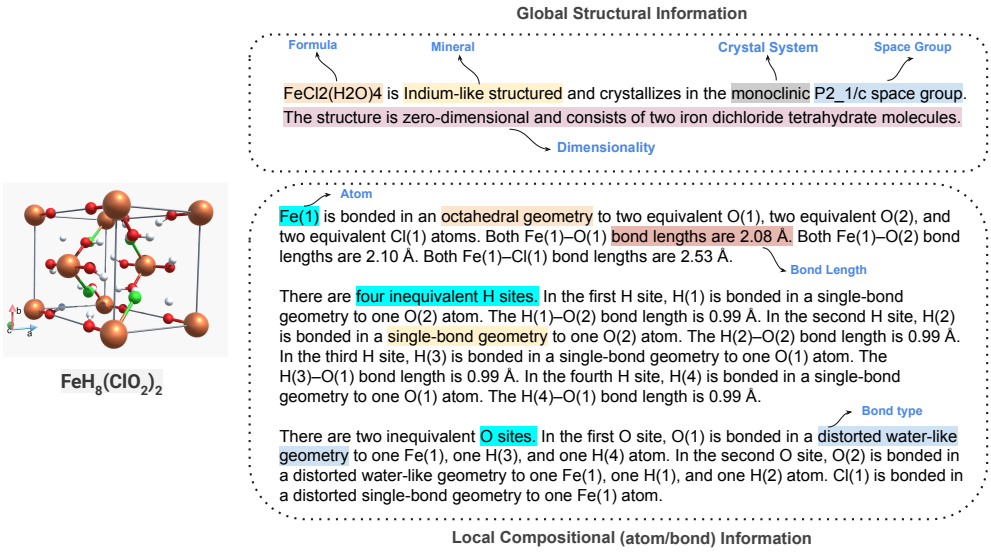

Figure 1: Textual description of $\mathbf{FeH_8(ClO_2)_2}$ material from JARVIS dataset generated by Robocrystallographer. The generated text contains both local chemical compositional information related to atom/bonds (like site coordination, geometry, polyhedral connectivity, tilt angles) and global structural knowledge (like mineral type, space group information, symmetry, and dimensionality).

ent variants of GNN architectures for effective crystal representation learning. Through multiple layers of graph convolutions, these models can implicitly encode many-body interactions. Further, ALIGNN Choudhary and DeCost [2021] explicitly captures many-body interactions by incorporating bond angles and local geometric distortions into the GNN encoding module to enhance property prediction accuracy and became SOTA for a large range of properties. Recently, transformer-based architecture Matformer Yan et al. [2022] is proposed to learn the periodic graph representation of the material, which is invariant to periodicity and can capture repeating patterns explicitly. Matformer marginally improves the performance compared to ALIGNN, however, is much faster than it. Moreover, scarcity of labeled data makes these models difficult to train for all the properties, and recently, some key studies Jha et al. [2019], Das et al. [2022, 2023] have shown promising results to mitigate this issue using transfer learning, pre-training, and knowledge distillation respectively.

3 METHODOLOGY

In this section, we first discuss the insights on the textual dataset that we have curated for two popular crystalline databases and explain the local compositional and global periodic information we are able to encode using textual representation, which is difficult to incorporate explicitly into a graph structure. Then we give a detailed overview of our proposed muti-modal framework CrysMMNet that generates joint embedding for materials, which facilitates accurate property prediction.

3.1 TEXTUAL DATASET

Leveraging textual modalities beyond the conventional graph structure of the materials to capture both local atomic and global periodic knowledge remains largely unexplored by the research community. To the best of our knowledge, there is no existing dataset containing textual descriptions of the materials. Hence, we first curate the textual dataset for two popular material databases JARVIS and Material Project(MP), containing textual descriptions for each material of those databases. Conventionally, in these databases, the periodic structure of the materials is represented in Crystallographic Information File (CIF File). We use Robocrystallographer Ganose and Jain [2019], which is a free utility, to generate a textual description of the material from the CIF file. Robocrystallographer decomposes crystal structures into local compositional (site coordination, geometry, polyhedral connectivity, and tilt angles) and global structural (mineral type, space group information, symmetry, dimensionality) components (Figure 1) and output this information in three formats: JSON for machine use, human-readable text, and machine learning format. In this work, we use human-readable text for collecting textual datasets, which are easily interpretable and resembles a human description of the crystal structure.

Local compositional information describes local chemical environments around different atoms and inter-atomic bonds in a unit cell. It provides a detailed description of different sites of the materials, like atomic compositions of different sites, site coordination, inter-atomic connectivity through chemical bonds, bond type, and length. Further, the geometry of each site is mentioned and the presence

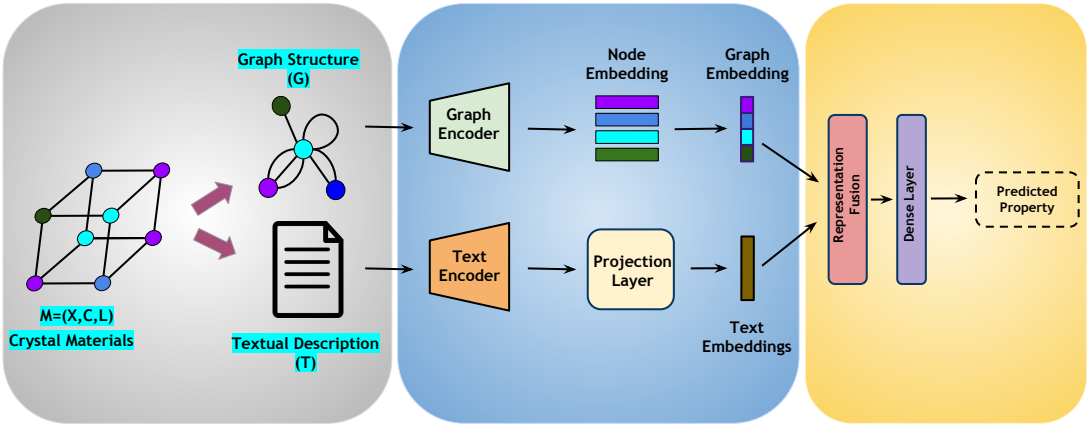

Figure 2: Overview of our adopted methodology CrysMMNet. Given Crystal Material (M), we use two modalities - Graph Structure ($\mathcal{G}$) and Textual Description ($\mathcal{T}$). Graph structure ($\mathcal{G}$) is passed through a graph encoder to generate graph embedding ($\mathcal{Z}_\mathcal{G}$). Textual Description ($\mathcal{T}$) is fed through a text encoder followed by a projection layer to generate text embedding ($\mathcal{Z}_\mathcal{T}$). Finally, both the representations are fused together and predict the crystal properties based on joint modeling of the input modalities.

of corner-sharing tetrahedra connectivity is specified. On the contrary, **global structural** information illustrates the global environment i.e. periodic structure and orientation of the material in 3D space. The most useful information it provides is regarding crystal symmetry, which includes the specific space group and crystal system the material belongs to. Space group is used to describe the symmetry of a unit cell of the crystal material in 3D space. In materials science literature there are 230 unique space groups and each crystal (graph) has a unique space group number. Further based on the space group level information can classify a crystal graph into 7 broad groups of crystal systems like Triclinic, Monoclinic, Orthorhombic, Tetragonal, Trigonal, Hexagonal, and Cubic. Moreover, it contains the mineral type of the material and the dimensionality of the crystal structure. Minerals are naturally occurring, inorganic substances with a specific chemical composition and a crystalline structure. The most common types include silicates (which contain silicon and oxygen), carbonates (which contain carbon and oxygen), sulfates (which contain sulfur and oxygen), halides (which contain a halogen element), oxides (which contain oxygen and one or more other elements), and sulfides (which contain sulfur and one or more other elements). Examples of minerals in each category include quartz, calcite, gypsum, halite, hematite, and pyrite. The chemical composition and crystal structure of a mineral determine its properties, such as its hardness, color, and cleavage. Further, dimensionality of a material is a significant global feature that refers to the number of dimensions that a particular component of the material spans. The dimensionality of a bonded cluster of atoms can be determined by calculating the rank of the subspace spanned by the central atom and its periodically connected neighbors.

A comprehensive understanding of both local and global environments is necessary for robust prediction of material properties. For example, in the case of formation energy, the local chemical environment, such as atom composition, bond length, and bond angles, plays a crucial role in determining the electronic and geometric structure of the material, which directly affects its formation energy. A slight variation in the local environment can result in significant changes in the electron density and, subsequently, the energy required to form the materials. Similarly, the global chemical environment, like the space group, has a profound impact on the formation energy by controlling the arrangement of atoms within the material. Different space groups are associated with different crystal structures and packing arrangements, which can lead to different formation energies. Moreover, the study by Larsen *et. al* Larsen et al. [2019] showed that the formation energies of layered materials can be related to their dimensionality, highlighting the importance of considering this feature in the investigation of materials.

3.2 MULTI-MODAL FRAMEWORK

Next, we propose a simple, yet effective multi-modal framework, CrysMMNet, for graph and textual embedding of materials, which realizes material dataset as $D = \{(\mathcal{G}, \mathcal{T}), \mathcal{Y}\}$, where $\mathcal{G}$,$\mathcal{T}$ and $\mathcal{Y}$ denote multi-graph structure, textual description and property value the material respectively. In our multi-modal architecture, the goal is to learn a function $f_\theta(\mathcal{G}, \mathcal{T})$

$$f_\theta : (\mathcal{G}, \mathcal{T}) \rightarrow \mathcal{Y} \tag{2}$$

By design, CrysMMNet (as shown in Figure 2) is composed of three modules: graph encoder $M_V(\mathcal{G}) \rightarrow \mathcal{Z}_\mathcal{G}$, text encoder $M_L(\mathcal{T}) \rightarrow \mathcal{Z}_\mathcal{T}$, and joint embedding model $E(\mathcal{Z}_\mathcal{G}, \mathcal{Z}_\mathcal{T}) \rightarrow \mathcal{Z}$, where $\mathcal{Z}_\mathcal{G}$,$\mathcal{Z}_\mathcal{G}$ and $\mathcal{Z}$ are graph-level, textual and multimodal embedding respectively. Next, we explain each part of the CrysMMNet framework in detail.

3.2.1 Graph Encoder:

CrysMMNet adopts a GNN architecture inspired by ALIGNN Choudhary and DeCost [2021] as Graph Encoder, to encode the chemical, structural, and bond angular information of a crystal graph $\mathcal{G}$. We derive additional line graph $\mathcal{L}(\mathcal{G})$ from the crystal graph $\mathcal{G}$ to describe the angles between the edges in $\mathcal{G}$, where nodes and edges in line graph $\mathcal{L}(\mathcal{G})$ correspond to inter-atomic bonds and bond angles. We denote h_i^l, $e_{i,j}^l$ and $t_{i,j,k}^l$ as l-th layer representation for i-th atom, $\{i,j\}$-th bond, and $\{i,j,k\}$-th angle (triplet) respectively. Graph encoder alternates edge-gated graph convolution layers between $\mathcal{L}(\mathcal{G})$ and $\mathcal{G}$ to propagate bond angular information through inter-atomic bond representation to atom embedding and vice versa. Specifically, at the (l)-th layer, given the line graph $\mathcal{L}(\mathcal{G})$, we apply Gated Graph ConvNet (GatedGCN) Dwivedi et al. [2020] to update triplet representation and generate bond messages m as follows :

$$
\begin{aligned}
t_{i,j,k}^{l+1} &= t_{i,j,k}^l + \gamma\Big(BN\Big(A_{lg}^l e_{i,j}^l + B_{lg}^l e_{j,k}^l + C_{lg}^l t_{i,j,k}^l\Big)\Big) \\
\hat{t}_{i,j,k}^{l+1} &= \frac{\sigma(t_{i,j,k}^{l+1})}{\sum_{(j,m)\in N_{i,j}} \sigma(t_{i,j,m}^{l+1}) + \epsilon} \\
m_{i,j}^l &= e_{i,j}^l + \gamma\Big(BN\Big((W_{lg}^l e_{i,j}^l + \sum_{\substack{(j,k)\\ \in N_{i,j}}} \hat{t}_{i,j,k}^{l+1} \odot V_{lg}^l e_{j,k}^l\Big)\Big)
\end{aligned}
\tag{3}
$$

Further, we apply another GatedGCN on the crystal graph $\mathcal{G}$ and update bond and atom features as follows :

$$
\begin{aligned}
e_{i,j}^{l+1} &= e_{i,j}^l + \gamma\Big(BN\Big(A_g^l h_i^l + B_g^l h_j^l + C_g^l m_{i,j}^l\Big)\Big) \\
\hat{e}_{i,j}^{l+1} &= \frac{\sigma(e_{i,j}^{l+1})}{\sum_{k\in N_i} \sigma(e_{i,k}^{l+1}) + \epsilon} \\
h_i^{l+1} &= h_i^l + \gamma\Big(BN\Big((W_g^l h_i^l + \sum_{j\in N_i} \hat{e}_{i,j}^{l+1} \odot V_g^l h_j^l\Big)\Big)
\end{aligned}
\tag{4}
$$

where σ is the sigmoid function, ϵ is a small fixed constant for numerical stability, $\odot$ is the Hadamard product, BN is batch normalization and γ is the activation function where we use Sigmoid Linear Unit (SiLU). $A_{lg}^l, B_{lg}^l, C_{lg}^l, V_{lg}^l, W_{lg}^l$ are learnable parameters of GatedGCN applied on $\mathcal{L}(\mathcal{G})$ and $A_g^l, B_g^l, C_g^l, V_g^l, W_g^l$ are learnable parameters of GatedGCN applied on $\mathcal{G}$. We apply L such layers of aggregation and update in Graph Encoder and return the final set of node embeddings $\mathcal{H} = \{h_1, ..., h_{|\mathcal{V}|}\}$, where $h_i := h_i^L \in R^d$ represents the final embedding of node i. We subsequently use a symmetric aggregation function (AvgPool) to generate graph-level representation $\mathcal{Z}_\mathcal{G} \in R^{d'}$ (we set d' as 256) of the crystal material M.

$$
\mathcal{Z}_\mathcal{G} = \sum_{i=1}^{|\mathcal{V}|} h_i^L \tag{5}
$$

3.2.2 Text Encoder:

As a text encoder, we adopt a pre-trained MatSciBERT model, which is a domain-specific language model for materials science, followed by a projection layer. MatSciBERT is effectively a pre-trained SciBERT model on a scientific text corpus of 3.17B words, which is further trained on a huge text corpus of materials science containing around 285 M words, using domain adaptive pretraining objective proposed by Gururangan et al. [2020]. We feed textual description of material $\mathcal{T}$ and extract embedding of [CLS] token $\mathcal{Z}_{CLS} \in R^{768}$ as a representation of the whole text. Further. we pass $\mathcal{Z}_{CLS}$ through a projection layer (two-layer neural network) to generate the textual embedding for the material $\mathcal{Z}_\mathcal{T} \in R^d$

$$
\mathcal{Z}_\mathcal{T} = W_2(g(W_1 \mathcal{Z}_{CLS})) \tag{6}
$$

We use standard non-linear function ReLU(·) as g(·), $W_1 \in R^{768\times128}$ and $W_2 \in R^{128\times d}$ are parameter matrix that project $\mathcal{Z}_{CLS}$ to embedding space R^d. In our experiment, we set d as 64.

3.2.3 Joint Embedding Model:

The graph encoder encodes local structural and chemical semantics around atoms in a unit cell of the material, whereas the text encoder captures global periodic knowledge from the textual description. Further, in the joint embedding model, we fuse both the representations $(\mathcal{Z}_\mathcal{G}, \mathcal{Z}_\mathcal{T})$ together into a single multi-modal representation $\mathcal{Z} := (\mathcal{Z}_\mathcal{G} \oplus \mathcal{Z}_\mathcal{T}) \in R^{(d'+d)}$, which can now capture both local and global structural semantics of the material. We tried different ways to fuse both the embeddings like sum, average, concatenation $(\oplus)$ and found concatenation performs best.

Further, we pass this multi-modal representation $\mathcal{Z}$ through a multi-layer perceptron which predicts the value of the properties. We train CrysMMNet end to end to optimize the following mean square error(MSE) loss :

$$
\mathcal{L}_{MSE} = \|\hat{\mathcal{Y}} - \mathcal{Y}\|^2 \tag{7}
$$

where $\hat{\mathcal{Y}}$ and $\mathcal{Y}$ are predicted and true property values respectively. Note, while training CrysMMNet we freeze the

Property	Unit	CIFID	CGCNN	SchNet	MEGNET	GATGNN	ALIGNN	Matformer	CrysMMNet
Formation Energy	eV/atom	0.140	0.063	0.045	0.047	0.047	0.033	0.0325*	**0.028**
Bandgap(OPT)	eV	0.30	0.20	0.19	0.145	0.17	0.14	0.137*	**0.114**
Bandgap(MBJ)	eV	0.53	0.41	0.43	0.34	0.51	0.31	0.30*	**0.209**
Total Energy	eV/atom	0.24	0.078	0.047	0.058	0.056	0.037	0.035*	**0.034**
Bulk Moduli(Kv)	GPa	14.12	14.47	14.33	15.11	14.32	10.40*	11.21	**6.86**
Shear Moduli(Gv)	GPa	11.98	11.75	10.67	13.09	12.48	9.48*	10.76	**6.73**

Table 1: Summary of the prediction performance (MAE) of CrysMMNet and different state-of-the-art models for different properties in JARVIS-DFT Dataset. The best performance is highlighted in bold and the second-best results are highlighted with *.

weights of MatSciBERT and don't tune it further. Fine-tuning MatSciBERT with CrysMMNet training for a specific property will add more computational overhead as it will increase the number of parameters significantly. This provides scope for further investigation and we keep it as future work.

4 EXPERIMENTAL RESULTS

In this section, we begin by describing the experimental setup which includes the benchmark datasets used for evaluation, alternate baseline approaches, and implementation details. Then we evaluate the performance of CrysMMNet in comparison with different SOTA property predictors on the downstream property prediction tasks using two popular material benchmark datasets. Next, we present the empirical evaluation results of our proposed framework in limited training data settings. Finally, we conduct some ablation studies to demonstrate the expressiveness and robustness of textual representation and the importance of global and local knowledge encoded in textual embedding.

4.1 EXPERIMENTAL SETUP

To evaluate the effectiveness of CrysMMNet, we conduct experiments on two benchmark material datasets, Materials Project Jain et al. [2013] (MP 2018.6.1) and JARVIS-DFT Choudhary et al. [2020] (2021.8.1), which comprises some important physical properties obtained with high-throughput DFT calculations. MP 2018.6.1 consists of 69,239 materials whereas JARVIS-DFT consists of 55,722 materials. We curated textual datasets for both datasets using robocrystallographer with a textual description of each material as described in subsection 3.1. We choose seven state of the art algorithm for crystal property prediction CIFID Choudhary et al. [2018], CGCNN Xie and Grossman [2018], SchNet Schütt et al. [2017], MEGNET Chen et al. [2019], GATGNN Louis et al. [2020], ALIGNN Choudhary and DeCost [2021] and Matformer Yan et al. [2022]. To avoid any deterioration of the performance of the baseline algorithms due to insufficient hyperparameter tuning, we report the property prediction results from the respective papers of the baseline models.

We use four convolution layers of the graph encoder module and pre-trained MatSciBERT followed by a two-layer neural network (projection layer) as the text encoder module in CrysMMNet. We train it for 300 epochs using AdamW Loshchilov and Hutter [2017] optimizer with normalized weight decay of 10^{-5} and keep the batch size as 64. We schedule the learning rate according to the one-cycle policy Smith [2018] with a maximum learning rate of 0.001. We keep embedding dimensions of the graph and text encoder as 64 and 256 respectively. We perform the experiments in shared servers having Intel E5-2620v4 processors which contain 16 cores/thread and four GTX 1080Ti 11GB GPUs each.

4.2 DOWNSTREAM TASK EVALUATION

JARVIS-DFT Dataset. To evaluate CrysMMNet, we first conduct experiments on the JARVIS-DFT dataset, which is a widely used large-scale material benchmark containing 55,722 crystals. Following previous state-of-the-art works, we choose six crystal properties including formation energy, bandgap(OPT), bandgap(MBJ), total energy, bulk moduli, and shear moduli for the downstream property prediction task. We use 80%,10%, and 10% train, validation, and test split for all the properties as used by ALIGNN. We report mean absolute error (MAE) of the predicted and actual value of a particular property for test data in table 1 to compare the performance of CrysMMNet and different participating methods. We observe that CrysMMNet outperforms every baseline model across all the properties with a significant margin. In specific, we observe 13.84%, 2.8%, 16.78%, 30.33%, 34%, and 29% improvements compared to the competing second-best baseline model for formation energy, total energy, bandgap(OPT), bandgap(MBJ), bulk moduli, and shear moduli respectively, which shows the effectiveness of multimodal representation capturing both local chemical semantics and global periodic structural knowledge towards crystal property prediction.

Materials Project (MP) Dataset. We further use another benchmark material dataset, Materials Project-2018.6.1,

Property	Unit	CGCNN	SchNet	MEGNET	GATGNN	ALIGNN	Matformer	CrysMMNet
Formation Energy	eV/atom	0.031	0.033	0.030	0.033	0.022	0.021*	**0.020**
Bandgap	eV	0.292	0.345	0.307	0.280	0.218	0.211*	**0.197**
Bulk Moduli(Kv)	log(GPa)	0.047	0.066	0.060	0.045	0.051	0.043*	**0.040**
Shear Moduli(Gv)	log(GPa)	0.077	0.099	0.099	0.075	0.078	0.073*	**0.062**

Table 2: Summary of the prediction performance (MAE) of CrysMMNet and different state-of-the-art models for different properties in The Materials Project dataset. The best performance is highlighted in bold and the second-best results are highlighted with *.

Property	Train-set Size	CGCNN	ALIGNN	CrysMMNet
Bandgap(MBJ)	3634	0.522	0.483	**0.456**
Bulk Moduli	3936	14.98	14.13	**13.04**
Shear Moduli	3936	13.07	12.61	**11.57**

Table 3: : MAE values of CGCNN, ALIGNN and CrysMMNet for three different properties in the JARVIS-DFT dataset with 20% training instances. The best performance is highlighted in bold.

Property	CrysTextNet	CGCNN	ALIGNN
Formation Energy	0.447	0.063	0.0331
Total Energy	0.352	0.078	0.037
Bandgap(OPT)	0.595	0.20	0.142
Bandgap(MBJ)	0.849	0.41	0.31
Bulk Moduli(Kv)	21.98	14.47	10.4
Shear Moduli(Gv)	14.76	11.75	9.48

Table 4: Summary of experiments for the ablation study on the effectiveness of Textual Representation.

comprises 69,239 materials. Here we evaluate CrysMMNet with all state-of-the-art models using four crystal properties namely formation energy, bandgap, bulk moduli, and shear moduli. For formation energy and bandgap, we use 60000, 5000, and 4239 crystals as train, validation as test split as used by ALIGNN, whereas use 4664, 393, and 393 crystals as train, validation as test split for bulk and shear moduli as used by GATGNN. We report the mean absolute error (MAE) of the predicted and actual property value for test data in table 2 to compare the performance of CrysMMNet with different participating methods. Note, to maintain consistency with the results reported by baseline works, we report (GPa) values of bulk and shear moduli in table 1, whereas log(GPa) values in table 2. We observe that CrysMMNet outperforms every baseline model across all the properties with a significant margin. In specific, we observe 4.76%, 6.63%, 6.9%, and 15.06% improvements compared to the competing second-best baseline model for formation energy, bandgap(OPT), bulk moduli, and shear moduli respectively. Overall, the superior performances show the effectiveness of multi-modal representation in CrysMMNet.

Resuts on Limited Training Data. CrysMMNet performs well in limited data settings as well. With 4664 training samples only CrysMMNet achieves 9.8% and 20.5% improvements for bulk moduli and shear moduli respectively in the Materials Project dataset. Further, in JARVIS-DFT Dataset, we conduct an additional set of experiments for three different properties including bandgap(MBJ), bulk moduli, and shear moduli, where we have limited labeled data. More specifically, we take 20-10-10% training-validation-test data split and evaluate the performance of CGCNN, ALIGNN, and CrysMMNet in the table 3. We observe, CrysMMNet achieves improvement for all three properties compared to CGCNN and ALIGNN. Overall, these superior performances indicate the robustness of our model and its adaptive ability to tasks of various data scales.

4.3 ABLATION STUDY

In this subsection, We demonstrate the robustness of multimodal representation on different GNN architecture choices and the influence of textual modality on CrysMMNet performance, by designing the following set of ablation studies:

1. Is only textual information sufficient to infer better property prediction accuracy?
2. How robust is the multimodal representation on different GNN architecture choices for graph encoder?
3. What are the influences of global structural and local compositional knowledge from the textual datasets on property prediction performance?

In the following subsections, we will thoroughly discuss these.

Expressiveness of Textual Representation. First we are interested to understand whether textual representations are alone expressive enough, to encode atomic chemical and periodic structural semantics from the curated textual data and predict different properties precisely. We conduct an ablation experiment, where we consider only the text embeddings $\mathcal{Z}_{\mathcal{T}}$ of CrysMMNet (output of projection layer) and pass it alone through a multi-layer perceptron to predict the property value. We denote this model as CrysTextNet and compare it with state-of-the-art graph-based models on different properties

of the JARVIS-DFT dataset. We report the MAE for test data in table 4. We observe, for all the properties, test MAE is higher for the CrysTextNet model compared to state-of-the-art graph-based models like CGCNN and ALIGNN. In specific, for mechanical properties like bulk modulus and shear modulus, CrysTextNet works better (closer test MAE with competing GNN baselines) than properties like formation energy, band-gap, and total energy. This is because properties, such as formation energy, band gap, and total energy rely on microscopic chemical information which textual representation fails to encode. Instead, graph encodings include node features $\mathbf{x}_u$ ($u \in \mathcal{V}$) that are high-dimensional vectors with meaningful chemical quantities like electronegativity, group number, covalent radius, number of valence electrons, first ionization energy, etc. Furthermore, through message passing and aggregation in the graph convolution layer, GNN models capture many-body interactions among atoms in the material. On the other side, mechanical properties like bulk modulus and shear modulus are more dependent on structural information like lattice structure and symmetry of the material, which textual representations are able to capture. Overall, though textual representations can capture many useful local and global information about the materials, unlike graph structural models, they are not alone expressive enough to capture atomic chemical features and the structural connectivity between different atoms in the materials. Message passing and neighborhood aggregation between atoms through the GNN model are still very fundamental in learning the structure-property relationship of the materials.

Robustness of Textual Representation. Further, we investigate the robustness of textual representations on different crystal GNN encoders. We conduct an ablation study where we replace graph encode of CrysMMNet with popular crystal GNN variants, e.g, CGCNN, MEGNET, GATGNN and evaluate the performance. We set up the experiments with six properties of the JARVIS dataset and report the MAE values in table 5 for the baseline GNN models and different variants of CrysMMNet with different GNN architectural choices as graph encoder. We observe all these variants outperform corresponding vanilla GNN models with a good margin for all the properties, which shows textual representations are rich enough to encode global structural knowledge which aids the property prediction accuracy of any state-of-the-art GNN models.

Importance of Local and Global Material Knowledge . Finally, we are curious to understand the importance of local(atom/bond) compositional information and global material structural knowledge encoded through textual representation in CrysMMNet. Specifically, we conduct an ablation study, where we train CrysMMNet in two additional setups along with the conventional (Global+Local) setup for CrysMMNet. (a) Only Global: In this scenario, we take only global knowledge about the periodic structure of the materials as textual data to train CrysMMNet. (b) Only Local: In this scenario, we take only local compositional information about atoms and inter-atomic bonds as textual data to train CrysMMNet. We use six properties of JARVIS-DFT dataset for the experiment and report MAE in Table 6. We observe performance gain across all the properties using both global and local information combined as textual knowledge, compared to only global or local knowledge separately.

Property	CGCNN	CrysMMNet (CGCNN)	MEGNET	CrysMMNet (MEGNET)	GATGNN	CrysMMNet (GATGNN)
Formation Energy	0.063	0.046	0.076	0.060	0.077	0.064
Bandgap(OPT)	0.200	0.163	0.184	0.165	0.169	0.157
Bandgap(MBJ)	0.41	0.339	0.369	0.339	0.343	0.331
Total Energy	0.078	0.059	0.058	0.057	0.056	0.053
Bulk Moduli(Kv)	14.47	12.98	15.11	13.29	14.32	13.73
Shear Moduli(Gv)	11.75	10.71	13.09	11.86	12.48	12.04

Table 5: Summary of the prediction performance (MAE) of different state-of-the-art GNN models with textual representation for six different properties in The JARVIS-DFT Dataset. Model M is the SOTA baseline model and CrysMMNet(M) is a variant where we replace graph encoder with M.

Property	Global+Local	Only Global	Only Local
Formation Energy	**0.028**	0.039	0.039
Total Energy	**0.034**	0.042	0.046
Bandgap(OPT)	**0.1144**	0.1906	0.1468
Bandgap(MBJ)	**0.2092**	0.2162	0.2186
Bulk Moduli(Kv)	**6.86**	6.91	6.87
Shear Moduli(Gv)	**6.44**	6.73	6.88

Table 6: Summary of experiments for the ablation study on the importance of Local and Global Material Knowledge.

5 CONCLUSIONS

In this work, we address the limitation of state-of-the-art GNN models for crystal property prediction to capture global periodic structural information and leverage textual modalities beside graph structures to resolve the issue. To this end, we curate textual datasets of two popular benchmark databases containing textual descriptions of each material containing both local compositional and global structural information of a material. Further, we propose a simple yet effective multi-modal framework, CrysMMNet, for crystalline materials, which fuse both graph structural and textual representation together to generate a more enriched and robust multimodal representation for materials, which subsequently improves property prediction accuracy. Extensive experiments show CrysMMNet outperforms all the popular state-of-the-art baselines across ten diverse sets of properties on two popular datasets.

References

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

Kamal Choudhary and Brian DeCost. Atomistic line graph neural network for improved materials property predictions. *npj Computational Materials*, 7(1):1–8, 2021.

Kamal Choudhary, Brian DeCost, and Francesca Tavazza. Machine learning with force-field-inspired descriptors for materials: Fast screening and mapping energy landscape. *Physical review materials*, 2(8):083801, 2018.

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, Bidisha Samanta, Pawan Goyal, Seung-Cheol Lee, Satadeep Bhattacharjee, and Niloy Ganguly. Crysxpp: An explainable property predictor for crystalline materials. *npj Computational Materials*, 8(1):1–11, 2022.

Kishalay Das, Bidisha Samanta, Pawan Goyal, Seung-Cheol Lee, Satadeep Bhattacharjee, and Niloy Ganguly. Crysgnn: Distilling pre-trained knowledge to enhance property prediction for crystalline materials. *arXiv preprint arXiv:2301.05852*, 2023.

Maarten De Jong, Wei Chen, Randy Notestine, Kristin Persson, Gerbrand Ceder, Anubhav Jain, Mark Asta, and Anthony Gamst. A statistical learning framework for materials science: application to elastic moduli of k-nary inorganic polycrystalline compounds. *Scientific reports*, 6(1):1–11, 2016.

Vijay Prakash Dwivedi, Chaitanya K Joshi, Thomas Laurent, Yoshua Bengio, and Xavier Bresson. Benchmarking graph neural networks. 2020.

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

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A Smith. Don't stop pretraining: Adapt language models to domains and tasks. *arXiv preprint arXiv:2004.10964*, 2020.

Tim Hsu, Tuan Anh Pham, Nathan Keilbart, Stephen Weitzner, James Chapman, Penghao Xiao, S Roger Qiu, Xiao Chen, and Brandon C Wood. Efficient, interpretable graph neural network representation for angle-dependent properties and its application to optical spectroscopy. *arXiv preprint arXiv:2109.11576*, 2021.

Jino Im, Seongwon Lee, Tae-Wook Ko, Hyun Woo Kim, YunKyong Hyon, and Hyunju Chang. Identifying pb-free perovskites for solar cells by machine learning. *npj Computational Materials*, 5(1):1–8, 2019.

Olexandr Isayev, Corey Oses, Cormac Toher, Eric Gossett, Stefano Curtarolo, and Alexander Tropsha. Universal fragment descriptors for predicting properties of inorganic crystals. *Nature communications*, 8(1):1–12, 2017.

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

Dipendra Jha, Kamal Choudhary, Francesca Tavazza, Weikeng Liao, Alok Choudhary, Carelyn Campbell, and Ankit Agrawal. Enhancing materials property prediction by leveraging computational and experimental data using deep transfer learning. *Nature communications*, 10 (1):1–12, 2019.

Peter Mahler Larsen, Mohnish Pandey, Mikkel Strange, and Karsten Wedel Jacobsen. Definition of a scoring parameter to identify low-dimensional materials components. *Physical Review Materials*, 3(3):034003, 2019.

Joohwi Lee, Atsuto Seko, Kazuki Shitara, Keita Nakayama, and Isao Tanaka. Prediction model of band gap for inorganic compounds by combination of density functional theory calculations and machine learning techniques. *Physical Review B*, 93(11):115104, 2016.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Steph-Yves Louis, Yong Zhao, Alireza Nasiri, Xiran Wang, Yuqi Song, Fei Liu, and Jianjun Hu. Graph convolutional neural networks with global attention for improved materials property prediction. *Physical Chemistry Chemical Physics*, 22(32):18141–18148, 2020.

Shuaihua Lu, Qionghua Zhou, Yixin Ouyang, Yilv Guo, Qiang Li, and Jinlan Wang. Accelerated discovery of stable lead-free hybrid organic-inorganic perovskites via machine learning. *Nature communications*, 9(1):1–8, 2018.

Cheol Woo Park and Chris Wolverton. Developing an improved crystal graph convolutional neural network framework for accelerated materials discovery. *Physical Review Materials*, 4(6), Jun 2020. ISSN 2475-9953. doi: 10.1103/physrevmaterials.4.063801. URL `http://dx.doi.org/10.1103/PhysRevMaterials.4.063801`.

Ghanshyam Pilania, James E Gubernatis, and TJPRB Lookman. Structure classification and melting temperature prediction in octet ab solids via machine learning. *Physical Review B*, 91(21):214302, 2015.

Jonathan Schmidt, Love Pettersson, Claudio Verdozzi, Silvana Botti, and Miguel AL Marques. Crystal graph attention networks for the prediction of stable materials. *Science Advances*, 7(49):eabi7948, 2021.

Kristof Schütt, Pieter-Jan Kindermans, Huziel Enoc Sauceda Felix, Stefan Chmiela, Alexandre Tkatchenko, and Klaus-Robert Müller. Schnet: A continuous-filter convolutional neural network for modeling quantum interactions. *Advances in neural information processing systems*, 30, 2017.

Atsuto Seko, Atsushi Togo, Hiroyuki Hayashi, Koji Tsuda, Laurent Chaput, and Isao Tanaka. Prediction of low-thermal-conductivity compounds with first-principles anharmonic lattice-dynamics calculations and bayesian optimization. *Physical review letters*, 115(20):205901, 2015.

Atsuto Seko, Hiroyuki Hayashi, Keita Nakayama, Akira Takahashi, and Isao Tanaka. Representation of compounds for machine-learning prediction of physical properties. *Physical Review B*, 95(14):144110, 2017.

Leslie N Smith. A disciplined approach to neural network hyper-parameters: Part 1–learning rate, batch size, momentum, and weight decay. *arXiv preprint arXiv:1803.09820*, 2018.

Logan Ward, Ruoqian Liu, Amar Krishna, Vinay I Hegde, Ankit Agrawal, Alok Choudhary, and Chris Wolverton. Including crystal structure attributes in machine learning models of formation energies via voronoi tessellations. *Physical Review B*, 96(2):024104, 2017.

Tian Xie and Jeffrey C Grossman. Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties. *Phys. Rev. Lett.*, 120 (14):145301, 2018.

Tian Xie, Xiang Fu, Octavian-Eugen Ganea, Regina Barzilay, and Tommi Jaakkola. Crystal diffusion variational autoencoder for periodic material generation. *arXiv preprint arXiv:2110.06197*, 2021.

Keqiang Yan, Yi Liu, Yuchao Lin, and Shuiwang Ji. Periodic graph transformers for crystal material property prediction. In *The 36th Annual Conference on Neural Information Processing Systems*, 2022.