CrysGNN : Distilling pre-trained knowledge to enhance property prediction for crystalline materials.

Anonymous submission

Abstract

In recent years, graph neural network (GNN) based approaches have emerged as a powerful technique to encode complex topological structure of crystal materials in an enriched representation space. These models are often supervised in nature and using the property-specific training data, learn relationship between crystal structure and different properties like formation energy, bandgap, bulk modulus, etc. Most of these methods require a huge amount of property-tagged data to train the system which may not be available for different properties. However, there is an availability of a huge amount of crystal data with its chemical composition and structural bonds. To leverage these untapped data, this paper presents CrysGNN, a new pre-trained GNN framework for crystalline materials, which captures both node and graph level structural information of crystal graphs using a huge amount of unlabelled material data. Further, we extract distilled knowledge from CrysGNN and inject into different state of the art property predictors to enhance their property prediction accuracy. We conduct extensive experiments to show that with distilled knowledge from the pre-trained model, all the SOTA algorithms are able to outperform their own vanilla version with good margins. We also observe that the distillation process provides significant improvement over the conventional approach of finetuning the pre-trained model. We will release the pre-trained model along with the large dataset of 800K crystal graph which we carefully curated; so that the pre-trained model can be plugged into any existing and upcoming models to enhance their prediction accuracy.

Introduction

Fast and accurate prediction of different material properties is a challenging and important task in material science. In recent times there has been an ample amount of data driven works (Im et al. 2019; Isayev et al. 2017; Lu et al. 2018; Ward et al. 2017; De Jong et al. 2016; Seko et al. 2015; Pilania, Gubernatis, and Lookman 2015; Seko et al. 2017; Lee et al. 2016) for predicting crystal properties which are as accurate as theoretical DFT [Density functional Theory] based approaches (Orio, Pantazis, and Neese 2009), however, much faster than it. The architectural innovations of these approaches towards accurate property predictions come from incorporating specific domain knowledge into a deep encoding module. For example, in order to encode the neighbourhood structural information around a node (atom), GNN based approaches (Chen et al. 2019; Das et al. 2022; Louis et al. 2020; Park and Wolverton 2020; Xie and Grossman 2018) gained some popularity in this domain. Understanding the importance of many-body interactions, ALIGNN (Choudhary and DeCost 2021) incorporates bond angular information into their encoder module and became SOTA for a large range of property predictions. However, as different properties expressed by a crystalline material are a complex function of different inherent structural and chemical properties of the constituent atoms, it is extremely difficult to explicitly incorporate them into the encoder architecture. Moreover, data sparsity across properties is a known issue (Das et al. 2022; Jha et al. 2019), which makes these models difficult to train for all the properties. To circumvent this problem we adopt the concept of self-supervised pre-training (Chen et al. 2020b,a; Devlin et al. 2018; He et al. 2020; Hu et al. 2020a,b; Qiu et al. 2020; Trinh, Luong, and Le 2019; You et al. 2020) for crystalline materials which enables us to leverage a large amount of untagged material structures to learn the complex hidden features which otherwise are difficult to identify.

In this paper, we introduce a **graph pre-training method** which captures (a) connectivity of different atoms, (b) different atomic properties and (c) graph similarity from a large set of unlabeled data. To this effect, we **curate a new large untagged crystal dataset with 800K crystal graphs** and undertake a pre-training framework (named CrysGNN) with the dataset. CrysGNN learns the representation of a crystal graph by initiating **self-supervised loss at both node (atom) and graph (crystal) level**. At the **node** level, we pre-train the GNN model to **reconstruct** the **node features and connectivity** between nodes in a self-supervised way, whereas at the **graph** level, we adopt **supervised and contrastive** learning to learn structural similarities between graph structures using the **space group** and **crystal system** information of the materials respectively.

We subsequently **distill important structural and chemical information of a crystal from the pre-trained CrysGNN model** and pass it to the property predictor. The distillation process provides wider usage than the conventional pretrain-finetuning framework as transferring pre-trained knowledge to a property predictor and finetuning it requires a similar graph encoder architecture between the pre-trained model and the property predictor, which limits the knowledge transfer capability of the pre-trained model. On the other hand,

using knowledge distillation (Romero et al. 2014; Hinton et al. 2015), we can **retrofit** the pre-trained CrysGNN model into any existing state-of-the-art property predictor, irrespective of their architectural design, to improve their property prediction performance. Also experimental results (presented later) show that even in case of similar graph encoder, **distillation performs better than finetuning**.

With rigorous experimentation across two popular benchmark materials datasets, we show that distilling necessary information from CrysGNN to various property predictors results in substantial performance gains for GNN based architectures and complex ALIGNN model. The **improvements** range **from 4.19% to 16.20%** over several highly optimized SOTA models. We also perform **ablation studies** to investigate the influence of different pre-training losses in enhancing the SOTA model performance and observe significant performance gain employing the both node and graph-level pre-training, compared to node-level or graph-level pre-training separately. Also using both supervised and contrastive graph-level pre-training, we are able to learn more robust and expressive graph representation which enhances the property predictor performance. This also helps to achieve even **better improvements** when the **dataset is sparse**. Moreover, the property-tagged dataset suffers from certain biases as it is theoretically (DFT) derived, hence the property predictor also suffers from such bias. We found that on being trained with small amount of experimental data, the **DFT bias decreases** substantially.

Related Work

In recent times, data driven approaches (Lu et al. 2018; Isayev et al. 2017; Im et al. 2019; Ward et al. 2017; De Jong et al. 2016; Seko et al. 2015; Pilania, Gubernatis, and Lookman 2015; Seko et al. 2017; Lee et al. 2016) have become quite popular to establish relationship between the atomic structure of crystalline materials and their properties with very high precision. Especially, graph neural network (GNN) based approaches (Xie and Grossman 2018; Louis et al. 2020; Chen et al. 2019; Choudhary and DeCost 2021) have emerged as a powerful machine learning model tool to encode material's complex topological structure along with node features in an enriched representation space.

There are attempts to pre-train GNNs to extract graph and node-level representations. (Hu et al. 2020a) develops an effective pre-training strategy for GNNs, where they perform both node-level and graph-level pre-training on GNNs to capture domain specific knowledge about nodes and edges, in addition to global graph-level knowledge. Followed by this work, there has been several other works on self-supervised graph pre-training (Hu et al. 2020b; Qiu et al. 2020; You et al. 2020), which proposes different graph augmentation methods and maximizes the agreement between two augmented views of the same graph via a contrastive loss. In the field of crystal graphs, CrysXPP (Das et al. 2022) is the only model which comes close to a pre-trained model. In their work, an autoencoder is trained on a volume of un-tagged crystal graphs and the learned knowledge is (transferred to) used to initialize the encoder of CrysXPP, which is fine-tuned with property specific tagged data.

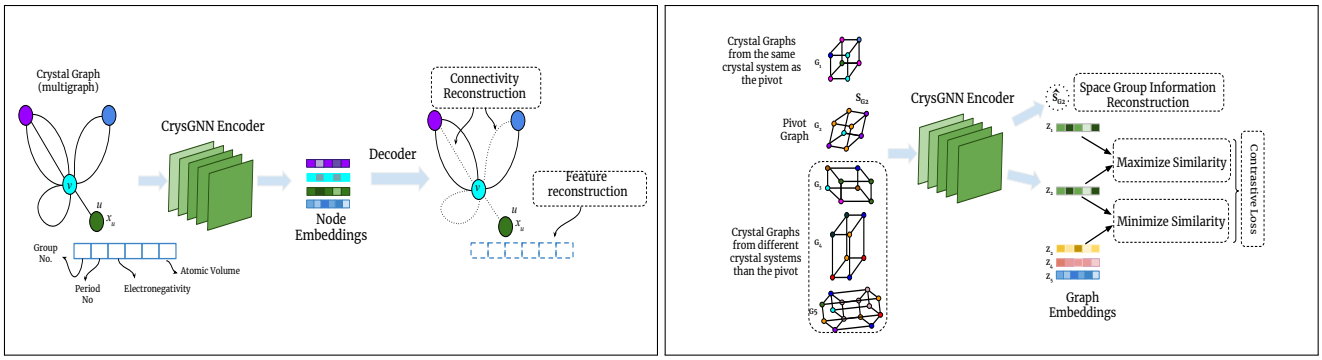
Although conceptually similar to the work done by Hu et al. (2020a), our work differs in the following three key aspects: (1) pre-training strategy proposed by Hu et al. is very effective for molecular dataset, but it is difficult to extend directly to crystalline material because structural semantics are different between molecules and materials (Xie et al. 2021). Molecules have non-periodic and finite structures, solid materials' structures are infinite and periodic in nature. (2) For graph-level pre-training, they adapted supervised graph-level property prediction using a huge amount of labelled dataset from chemistry and biology domain, which makes it less effective in several other domains like material science where property labeled data is extremely scarce. Also, a crucial step in graph-level prediction is to find graph structural similarity between two sets of graphs, which they do not explore but mention as a future work. We do not make use of supervised pre-training which requires a large amount of property tagged material data. Instead, we leverage the idea of structural similarity of materials belonging to the similar space group, and via contrastive loss and space group classification loss, we try to capture this similarity. (3) Finally they follow conventional pre-train finetuning framework, whereas in CrysGNN we incorporate the idea of knowledge distillation (Romero et al. 2014; Hinton et al. 2015) to distill important information from the pre-trained model and inject it into the property prediction process. By design, this knowledge distillation based approach is more robust and independent of the underlying architecture of the property predictor, thus it can enhance the performance of a diverse set of SOTA models.

Methodology

Formally, we first curate a huge amount of property un-tagged crystal graphs $\mathcal{D}_u = \{\mathcal{G}_i\}$ from various materials datasets to pre-train a deep GNN model f_θ, that learns intrinsic structural and chemical patterns of the crystal graphs. Further, we use a training set of property tagged crystal graphs $\mathcal{D}_t = \{\mathcal{G}_i, y_i\}$ for property prediction, which is smaller in volume and may or may not be disjoint from the original untagged set $\mathcal{D}_u$. We train any supervised property predictor $\mathcal{P}_\psi$ using $\mathcal{D}_t$ to predict the property value given the crystal graph structure. While training the property predictor, we incorporate the idea of knowledge distillation to distill important structural and chemical information from the pre-trained model. This knowledge may prove to be useful to a property predictor which now need not learn from scratch, but be armed with distilled knowledge from the pre-trained model. Hence in this section, we first describe the CrysGNN pre-training strategy, followed by the knowledge distillation and property prediction process.

CrysGNN Pre-training

We build a deep auto-encoder architecture CrysGNN, which comprises a graph convolution based encoder followed by an effective decoder. The autoencoder is (pre)trained end to end, using a large amount of property un-tagged crystal graphs $\mathcal{D}_u = \{\mathcal{G}_i\}$, where via node and graph-level self-supervised losses, the model can capture the structural and chemical information of the crystal graph data. First, we formalize the



(a) Node-level decoding

(b) Graph-level decoding

Figure 1: Overview of both node and graph-level decoding methods for CrysGNN. (a) In node-level decoding, node feature attributes and connectivity between nodes are reconstructed in a self-supervised way. (b) In graph-level decoding, G_2 is the pivot graph and G_1 is from the same crystal system (Cubic), whereas G_3, G_4, G_5 are from different crystal systems. First we reconstruct space group information of G_2, then through contrastive loss, CrysGNN will maximize similarities between positive pair (G_2, G_1) and minimize similarities between negative pairs (G_2, G_3), (G_2, G_4) and (G_2, G_5) in embedding space.

representation of a crystal 3D structure into a multi-graph structure, which will be an input to the encoder module.

Crystal Graph Representation. We realize a crystal material as a multi-graph structure $\mathcal{G}_i = (\mathcal{V}_i, \mathcal{E}_i, \mathcal{X}_i, \mathcal{F}_i)$ as proposed in (Das et al. 2022; Xie and Grossman 2018). $\mathcal{G}_i$ is an undirected weighted multi-graph where $\mathcal{V}_i$ denotes the set of nodes or atoms present in a unit cell of the crystal structure. $\mathcal{E}_i = \{(u, v, k_{uv})\}$ denotes a multi-set of node pairs and k_{uv} denotes number of edges between a node pair (u, v). $\mathcal{X}_i = \{(x_u | u \in \mathcal{V}_i)\}$ denotes the node feature set proposed by CGCNN (Xie and Grossman 2018). It includes different chemical properties like electronegativity, valance electron, covalent radius, etc. Finally, $\mathcal{F}_i = \{\{s^k\}_{(u,v)} | (u, v) \in \mathcal{E}_i, k \in \{1..k_{uv}\}\}$ denotes the multi-set of edge weights where s^k corresponds to the k^{th} bond length between a node pair (u, v), which signifies the inter-atomic bond distance between two atoms. Next, we formally define CrysGNN pretraining and knowledge distillation based property prediction strategy.

Self Supervision. We first develop a graph convolution (Xie and Grossman 2018) based encoding module, which takes crystal multi-graph structure $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{X}, \mathcal{F})$ as input and encodes structural semantics of the crystal graph into lower dimensional space. Each layer of convolution, follows an iterative neighbourhood aggregation (or message passing) scheme to capture the structural information within node's (atom's) neighbourhood. After L-layers of such aggregation, the encoder returns the final set of node embeddings $\mathcal{Z} = \{z_1, ..., z_{|\mathcal{V}|}\}$, where $z_u := z_u^L$ represents the final embedding of node u. Next, we design an effective decoding module, which takes node embeddings $\mathcal{Z}$ as input and learns local chemical features and global structural information through node and graph-level decoding, respectively. Decoding node-level information will enable CrysGNN to learn local domain specific chemical features and connectivity information around an atom, while decoding graph-level features will help CrysGNN capture global structural knowledge.

Node-Level Decoding. For node-level decoding (Figure-1(a)), we propose two self-supervised learning methods, where we reconstruct two important features that induce the local chemical environment of the crystal around a node (atom). For a given node u, we first reconstruct its node features x_u, which represent different chemical properties of atom u. Given a node embedding z_u, which is encoded based on neighbouring structure around atom u, we apply a linear transformation on top of z_u to reconstruct the node attributes. In crystalline graphs, node features correspond to different chemical properties associated with the constituent atoms through reconstructing these features CrysGNN captures local chemical semantics around that atom.
Further, we reconstruct local connectivity around an atom, where given node embeddings of two nodes u and v, we apply a bi-linear transformation module to generate combined transformed embedding of two nodes z_{uv}, which we pass through a feed forward network to predict the strength of association between two atoms. Through reconstructing this local connectivity around an atom, CrysGNN encodes the periodicity of the node i.e. the number of neighbours around it along with the relative position of its neighbours and their bond length.

Graph-level Decoding. We aim to capture periodic structure of a crystal material through graph-level decoding (Figure-1(b)). We specifically leverage two concepts in doing so. (a). **Space group** which is used to describe the symmetry of a unit cell of the crystal material. In materials science literature there are 230 unique space groups and each crystal (graph) has a unique space group number. (b). **Crystal system.** The space group level information can classify a crystal graph into 7 broad groups of crystal systems like Triclinic, Monoclinic, Orthorhombic, Tetragonal, Trigonal,

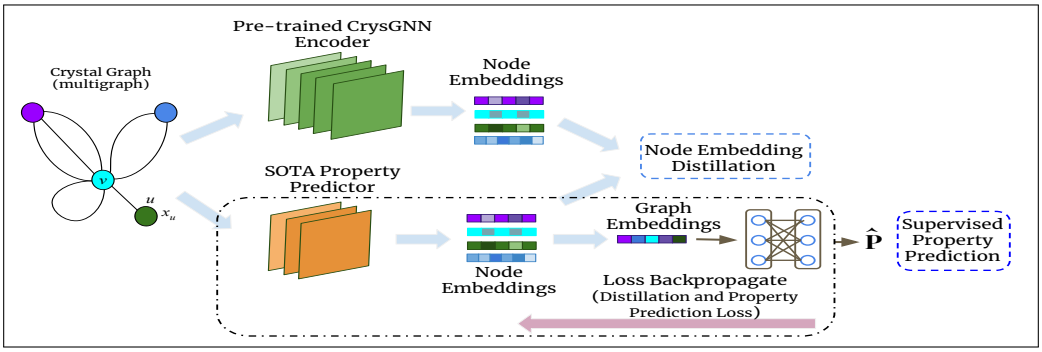


Figure 2: Overview of Property Prediction using Knowledge Distillation from CrysGNN.

Hexagonal, and Cubic. Several electronic and optical properties such as band gap, dielectric constant depend on the spacegroup and the crystal structure justifying its usage.

Given the set of node embeddings $\mathcal{Z} = \{z_1, ..., z_{|\mathcal{V}|}\}$, we use a symmetric aggregation function to generate graph-level representation $\mathcal{Z}_\mathcal{G}$. First, we pass $\mathcal{Z}_\mathcal{G}$ through a feed-forward neural network to predict the **space group number** of graph $\mathcal{G}$. Further, we develop a contrastive learning framework for pre-training of CrysGNN, where pre-training is performed by maximizing (minimizing) similarity between two crystal graphs belonging to the same (different) **crystal system** via contrastive loss in graph embedding space. A mini-batch of N crystal graphs is randomly sampled and processed through contrastive learning to align the positive pairs $\mathcal{Z}_{\mathcal{G}_i}, \mathcal{Z}_{\mathcal{G}_j}$ of graph embeddings, which belong to the same crystal system and contrast the negative pairs which are from different crystal systems. Here we adopt the normalized temperature-scaled cross-entropy loss (NT-Xent)(Sohn 2016; Van den Oord, Li, and Vinyals 2018; Wu et al. 2018) and NT-Xent for the i^{th} graph is defined:

$$\mathcal{L}_i = -log \frac{exp(sim(\mathcal{Z}_{\mathcal{G}_i}, \mathcal{Z}_{\mathcal{G}_j})/\tau)}{\sum_{k=1}^{K} exp(sim(\mathcal{Z}_{\mathcal{G}_i}, \mathcal{Z}_{\mathcal{G}_k})/\tau)} \quad (1)$$

where τ denotes the temperature parameter and $sim(\mathcal{Z}_{\mathcal{G}_i}, \mathcal{Z}_{\mathcal{G}_j})$ denotes cosine similarity function. The final loss $\mathcal{L}_{NTXent}$ is computed across all positive pairs in the minibatch. Overall we pre-train this deep auto-encoder architecture CrysGNN end to end to optimize the following loss :

$$\mathcal{L}_{pretrain} = \alpha\mathcal{L}_{FR} + \beta\mathcal{L}_{CR} + \gamma\mathcal{L}_{SG} + \lambda\mathcal{L}_{NTXent} \quad (2)$$

where $\mathcal{L}_{FR}, \mathcal{L}_{CR}$ are the reconstruction losses for node feature, and local connectivity, , $\mathcal{L}_{SG}$ is the space group supervision loss, $\mathcal{L}_{NTXent}$ is the contrastive loss and α, β, γ, λ are the weighting coefficients of each loss. We denote the set of parameters in CrysGNN model as θ and the pre-trained CrysGNN as f_θ.

Distillation and Property Prediction

We aim to retrofit the pre-trained CrysGNN model into any SOTA property predictor to enhance its learning process and improve performance. Hence we incorporate the idea of knowledge distillation to distill important structural and chemical information from the pre-trained model, which is useful for the downstream property prediction task, and feed it into the property prediction process. Formally, given the pre-trained CrysGNN model f_θ, any SOTA property predictor $\mathcal{P}_\psi$ and set of property tagged training data $\mathcal{D}_t = \{\mathcal{G}_i, y_i\}$, we aim to find optimal parameter values ψ^* for $\mathcal{P}$. We train $\mathcal{P}_\psi$ using dataset $\mathcal{D}_t$ to optimize the following multitask loss:

Task	Datasets	Graph Num.	Structural Info.	Properties Count	Data Type
Pre-training	OQMD	670K	✓	x	DFT Calculated
	MP	130K	✓	x	DFT Calculated
Property Prediction	MP 2018.6.1	69K	✓	2	DFT Calculated
	JARVIS(2018.6.1)	55K	✓	19	DFT Calculated
	OQMD-EXP	1.5K	✓	1	Experimental

Table 1: Datasets Details

$$\mathcal{L}_{prop} = \delta\mathcal{L}_{MSE} + (1-\delta)\mathcal{L}_{KD} \quad (3)$$

where $\mathcal{L}_{MSE} = (\hat{y}_i - y_i)^2$ denotes the discrepancy between predicted and true property values by $\mathcal{P}_\psi$ (**property prediction loss**). We define **knowledge distillation loss** $\mathcal{L}_{KD}$ to match intermediate node feature representation between the pre-trained CrysGNN model and the SOTA property predictor $\mathcal{P}_\psi$ as follows:

$$\mathcal{L}_{KD} = \|\mathcal{Z}_i^T - \mathcal{Z}_i^S\|^2 \quad (4)$$

where $\mathcal{Z}_i^T$ and $\mathcal{Z}_i^S$ denote intermediate node embeddings of the pre-trained CrysGNN and the property predictor $\mathcal{P}_\psi$ for crystal graph $\mathcal{G}_i$, respectively. Note, both $\mathcal{Z}_i^T$ and $\mathcal{Z}_i^S$ are projected on the same latent space. Finally, δ signifies relative weightage between two losses, which is a hyper-parameter to be tuned on validation data. During property prediction the pre-trained network is frozen and we backpropagate $\mathcal{L}_{prop}$ through the predictor $\mathcal{P}_\psi$ end to end.

Experimental Results

In this section, we evaluate how the distilled knowledge from CrysGNN enhances the performance of different state of the art property predictors on a diverse set of crystal properties from two popular benchmark materials datasets. We first briefly discuss the datasets used both in pre-training and downstream property prediction tasks. Then we report the

Property	CGCNN	CGCNN (Distilled)	CrysXPP	CrysXPP (Distilled)	GATGNN	GATGNN (Distilled)	ALIGNN	ALIGNN (Distilled)
Formation_Energy	0.039	**0.032**	0.041	**0.035**	0.096	**0.091**	0.026	**0.024**
Bandgap (OPT)	0.388	**0.293**	0.347	**0.287**	0.427	**0.403**	0.271	**0.253**
Formation_Energy	0.063	**0.047**	0.062	**0.048**	0.132	**0.117**	0.036	**0.035**
Bandgap (OPT)	0.200	**0.160**	0.190	**0.176**	0.275	**0.235**	0.148	**0.131**
Total_Energy	0.078	**0.053**	0.072	**0.055**	0.194	**0.137**	0.039	**0.038**
Ehull	0.170	**0.121**	0.139	**0.114**	0.241	**0.203**	0.091	**0.083**
Bandgap (MBJ)	0.410	**0.340**	0.378	**0.350**	0.395	**0.386**	0.331	**0.325**
Spillage	0.386	**0.374**	0.363	**0.357**	0.350	**0.348**	0.358	**0.356**
SLME (%)	5.04	**4.79**	5.11	**4.63**	5.05	**4.95**	4.65	**4.59**
Bulk Modulus (Kv)	12.45	**12.31**	13.61	**12.70**	11.64	**11.53**	11.20	**10.99**
Shear Modulus (Gv)	11.24	**10.87**	11.20	**10.56**	10.41	**10.35**	9.86	**9.80**

Table 2: Summary of the prediction performance (MAE) of different properties in Materials project (Top) and JARVIS-DFT (Bottom). Model M is the vanilla variant of a SOTA model and M (Distilled) is the distilled variant using the pretrained CrysGNN. The best performance is highlighted in bold.

results of different SOTA property predictors on the downstream property prediction tasks. Next, we illustrate the effectiveness of our knowledge distillation method compared to the conventional fine-tuning approach. We further conduct some ablation studies to show the influence of different pretraining losses in predicting different crystal properties and the performance of the system to sparse dataset. Finally, we demonstrate how distilled knowledge from the pre-trained model aids the SOTA models to remove DFT error bias, using very little experimental data.

Datasets

We curated around **800K untagged crystal graph data from two popular materials databases, Materials Project (MP) and OQMD**, to pre-train CrysGNN model. Further to evaluate the performance of different SOTA models with distilled knowledge from CrysGNN, we select **MP 2018.6.1 version of Materials Project** and **2021.8.18 version of JARVIS-DFT**, another popular materials database, for property prediction as suggested by (Choudhary and DeCost 2021). Please note, MP 2018.6.1 dataset is a subset of the dataset used for pre-training, whereas JARVIS-DFT is a separate dataset which is not seen during the pre-training. MP 2018.6.1 consists of 69,239 materials with two properties namely bandgap and formation energy, whereas JARVIS-DFT(2021.8.18) consists of 55,722 materials with 19 properties which can be broadly classified into two categories : 1) properties like formation energy, bandgap, total energy, bulk modulus, etc. which depend greatly on crystal structures and atom features, and 2) properties like ϵ_x, ϵ_y, ϵ_z, n-Seebeck, n-PF, etc. which depend on the precise description of the materials' electronic structure. In the following section, we will evaluate effectiveness of CrysGNN on the first class of properties. The impact of structural information is marginal on the second class of property hence all the SOTA perform poorly, there is however some modest improvement using CrysGNN; we have put the results and the discussion about it in the Appendix.

Moreover, all these properties in both Materials Project and JARVIS-DFT datasets are based on DFT calculations of chemicals. Therefore, to investigate how pre-trained knowledge helps to mitigate the DFT error, we also take a small dataset (Kirklin et al. 2015a, OQMD-EXP) containing 1,500 available experimental data of formation energy. Details of each of these datasets are given in Table 1.

Downstream Task Evaluation

To evaluate the effectiveness of CrysGNN, we choose four diverse state of the art algorithms for crystal property prediction, CGCNN (Xie and Grossman 2018), GATGNN (Louis et al. 2020), CrysXPP (Das et al. 2022) and ALIGNN (Choudhary and DeCost 2021). To train these models for any specific property, we adopt the multi-task setting discussed in equation 3 ,where we distill relevant knowledge from the pre-trained CrysGNN to each of these algorithms to predict different properties. We report mean absolute error (MAE) of the predicted and actual value of a particular property to compare the performance of different participating methods. For each property, we trained on 80% data, validated on 10% and evaluated on 10% of the data. We compare the results of distilled version of each SOTA model with its vanilla version (version reported in the respective papers), to show the effectiveness of the proposed framework.

Results. In Table 2, we report MAE of different crystal properties of Materials project and JARVIS-DFT datasets. In the distilled version of the SOTA models, while training the model, we distill information from the pre-trained CrysGNN model. We observe that the distilled version of any state-of-the-art model outperforms the vanilla model across all the properties. In specific, average improvement in CGCNN, CrysXPP, GATGNN and ALIGNN are 16.20%, 12.21%, 8.02% and 4.19%, respectively. These improvements are particularly significant as in most of the cases, the MAE is already low for SOTA models, still pretraining enables improvement over that. In fact, lower the MAE, higher the improvement. We calculate Spearman's Rank Correlation between MAE for each property across different SOTA models and their improvement due to distilled knowledge and found it to be very high (0.72), which supports the aforementioned observations. The average relative improvement across all properties for ALIGNN (4.19%) and GATGNN (8.02%) is lesser compared to CGCNN (16.20%) and CrysXPP (12.21%). A possible reason could be that ALIGNN and GATGNN are more complex models (more number of

parameters) than the pre-trained CrysGNN framework. Hence designing a deeper pre-training model or additionally incorporating angle-based information (ALIGNN) or attention mechanism (GATGNN) as a part of pre-training framework may help to improve further. This requires further investigation and we keep it as a scope of future work.

Comparison with Existing Pre-trained Models. We demonstrate the effectiveness of the knowledge distillation method vis-a-vis the conventional fine-tuning approaches. Note that the encoding architecture is same for CrysGNN, CGCNN, and CrysXPP. CrysXPP is very similar to a *pretrained-finetuned* version of CGCNN. Thus we compare distilled version of CGCNN with finetuned version of CrysGNN and CrysXPP (Das et al. 2022). Additionally, we consider Pretrain-GNN (Hu et al. 2019) which is a popular pretraining algorithm for molecules. We pre-train all the baseline models on our curated 800K untagged crystal data and fine-tune on seven properties in JARVIS dataset and report the MAE in Table 3. We feed multi-graph structure of the crystal material (as discussed in "Crystal Graph Representation") in Pretrain-GNN and try different combinations of node-level pre-training strategy along with the graph-level supervised pre-training (as suggested in (Hu et al. 2019)) and report the minimum MAE for any specific property. For finetuned CrysGNN, we take the pre-trained encoder of CrysGNN and feed a multilayer perceptron to predict a specific property. We observe that distilled CGCNN outperforms finetuned version of CrysGNN and both the baselines with a significant margin for all the properties. Pretrain-GNN performs the worst and the potential reason is - it is designed considering simple two-dimensional structure of molecules with a minimal set of node and bond features, which is hard to generalize for crystal materials which have very complex structure with a rich set of node and edge features.

Effectiveness on sparse training dataset. Finally, to demonstrate the effectiveness of the pre-training in limited data settings, we conduct additional set of experiments under different training data split. In specific, we vary available training data from 20 to 60 %, train different SOTA models and check their performance on test dataset. We observe that the distilled version of any SOTA model consistently outperforms its vanilla version even more in the limited training data setting, which illustrates the robustness of our pre-training framework. We report the MAE values of different baselines and their distilled version in Table 4.

Analysis of Different Pre-training Losses

We perform an ablation study to investigate the influence of different pre-training losses in enhancing the SOTA model performance. While pre-training CrysGNN (Eq. 2), we capture both local chemical and global structural information via node and graph-level decoding, respectively. Further, we are curious to know the influence of each of these decoding policies independently in the downstream property prediction task. In specific, we conduct the ablation experiments, where we pre-train CrysGNN with (a) only node-level decoding ($\mathcal{L}_{FR}$, $\mathcal{L}_{CR}$), (b) only graph-level decoding ($\mathcal{L}_{SG}$, $\mathcal{L}_{NTXent}$). Further, we perform ablations with individual graph-level losses, and pretrain with (c) removing $\mathcal{L}_{NTXent}$ (node-level with $\mathcal{L}_{SG}$ (space group)) and (d) removing $\mathcal{L}_{SG}$ (node-level with $\mathcal{L}_{NTXent}$(crystal system)). We train two baseline models, CGCNN and ALIGNN, with distilled knowledge from all the aforementioned variants of the pre-trained model and evaluate the performance on four crystal properties.

Experimental results are presented in Fig. 3. We can observe clearly that all the variants offer significant performance gain in all four properties using the combined node and graph-level pre-training, compared to node-level or graph-level pre-training separately. Only exception is formation energy, where only node-level pre-training produces less error compared to other variants, in both the baseline. Formation energy of a crystal is defined as the difference between the energy of a unit cell comprised of N chemical species and the sum of the chemical potentials of all the N chemical species. Hence pre-training at the node-level (node features and connection) is adequate for enhancing performance of formation energy prediction and incorporating graph-level information works as a noisy information, which degrades the performance. We also observe improvement in performance using both supervised and contrastive graph-level losses ($\mathcal{L}_{SG}$ and $\mathcal{L}_{NTXent}$), compared to using only one of

Property	CGCNN (Distilled)	CrysGNN (Finetuned)	CrysXPP	Pretrain -GNN
Formation_Energy	**0.047**	0.056	0.062	0.764
Bandgap (OPT)	**0.160**	0.183	0.190	0.688
Total_Energy	**0.053**	0.069	0.072	1.451
Ehull	**0.121**	0.130	0.139	1.112
Bandgap (MBJ)	**0.340**	0.371	0.378	1.493
Bulk Modulus (Kv)	**12.31**	13.42	13.61	20.34
Shear Modulus (Gv)	**10.87**	11.07	11.20	16.51
SLME (%)	**4.79**	5.45	5.11	9.85
Spillage	0.374	0.374	**0.363**	0.481

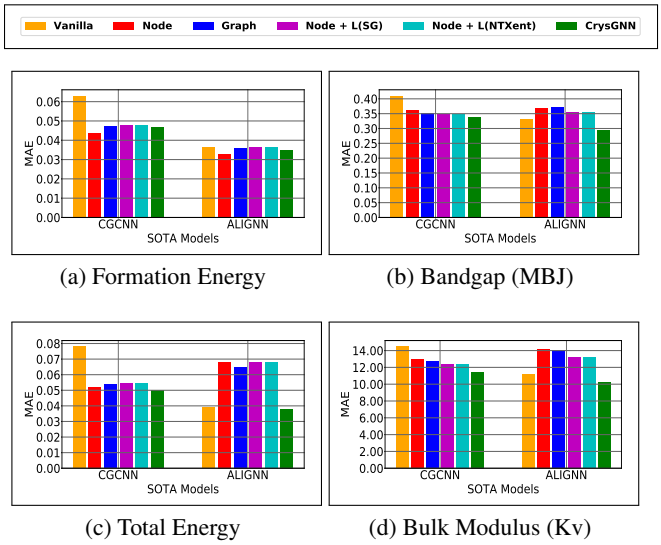
Table 3: Comparison of the prediction performance (MAE) of seven properties in JARVIS-DFT between CrysGNN and existing pretrain-finetune models, the best performance is highlighted in bold.

Property	Train-Val -Test(%)	CGCNN (Distilled)	CrysXPP (Distilled)	GATGNN (Distilled)	ALIGNN (Distilled)
Bandgap (MBJ)	20-10-70	0.453 (23.04)	0.450 (24.82)	0.521 (3.70)	0.485 (2.53)
	40-10-50	0.419 (21.41)	0.405 (18.40)	0.448 (2.81)	0.395 (2.20)
	60-10-30	0.364 (19.08)	0.360 (17.36)	0.439 (2.29)	0.380 (1.98)
Bulk Modulus (Kv)	20-10-70	16.26 (3.80)	14.25 (7.59)	14.19 (4.12)	14.06 (4.35)
	40-10-50	14.46 (2.36)	14.02 (7.34)	12.59 (3.00)	12.11 (2.89)
	60-10-30	14.05 (1.26)	13.73 (6.98)	11.75 (2.16)	11.01 (1.96)
Shear Modulus (Gv)	20-10-70	12.50 (10.01)	12.07 (9.86)	12.42 (3.20)	12.31 (3.15)
	40-10-50	11.54 (4.15)	11.01 (9.46)	11.23 (1.75)	10.67 (2.82)
	60-10-30	11.31 (3.74)	10.67 (9.35)	10.47 (1.69)	10.04 (1.95)
SLME (%)	20-10-70	6.62 (7.17)	5.90 (16.40)	6.02 (5.20)	6.27 (1.43)
	40-10-50	5.78 (5.90)	5.81 (15.67)	5.63 (2.60)	5.57 (1.42)
	60-10-30	5.24 (5.68)	4.84 (10.54)	5.34 (2.55)	4.82 (1.33)

Table 4: MAE values of distilled version of all the SOTA models for four different properties in JARVIS-DFT dataset with the increase in training instances from 20 to 60%. Relative improvement in the distilled model is mentioned in bracket.

Experiment Settings	CGCNN	CGCNN (Distilled)	CrysXPP	CrysXPP (Distilled)	GATGNN	GATGNN (Distilled)	ALIGNN	ALIGNN (Distilled)
Train on DFT Test on Experimental	0.265	0.244 (7.60)	0.243	0.225 (7.40)	0.274	0.232 (15.3)	0.220	0.209 (5.05)
Train on DFT and 20 % Experimental Test on 80 % Experimental	0.144	0.113 (21.7)	0.138	0.118 (14.2)	0.173	0.168 (2.70)	0.099	0.094 (5.60)
Train on DFT and 80 % Experimental Test on 20 % Experimental	0.094	0.073 (22.7)	0.087	0.071 (18.4)	0.113	0.109 (3.40)	0.073	0.069 (5.90)

Table 5: MAE of predicting experimental values by different SOTA models and their distilled versions with full DFT data and different percentages of experimental data for formation energy in OQMD-EXP dataset. Relative improvement in the distilled model is mentioned in bracket.



(a) Formation Energy (b) Bandgap (MBJ)

(c) Total Energy (d) Bulk Modulus (Kv)

Figure 3: Summary of experiments of ablation study on importance of different pre-training loss components on CrysGNN training and eventually its effect on CGCNN and ALIGNN models on four different properties (MAE for property prediction). (i) Vanilla: SOTA based model (without distillation) and all the other cases are SOTA models (distilled) from different pre-trained version of CrysGNN. (ii) Node : only node-level pre-training, (iii) Graph : only graph-level pre-training, (iv) Node + L(SG) : node-level and $\mathcal{L}_{SG}$, (v) Node + L(NTXent) : node-level and $\mathcal{L}_{NTXent}$ and (vi) CrysGNN : both node and graph-level pre-training.

them, which proves the learned representation via supervised and contrastive learning is more expressive that using any one of them. Moreover, in ALIGNN, with either node or graph-level pre-training separately, performance degrades across different properties. ALIGNN explicitly captures the three body interactions which drive its performance, to replicate that inclusion of both node and graph information is necessary.

Removal of DFT error bias using experimental data

One of the fundamental issues in material science is that experimental data for crystal properties are very rare (Kubaschewski, Alcock, and Spencer 1993; Bracht, Stolwijk, and Mehrer 1995; Turns 1995). Hence existing SOTA models rely on DFT calculated data to train their parameters. However, mathematical approximations in DFT calculation lead to erroneous predictions (error bias) compared to the actual experimental values of a particular property. Hence DFT error bias is prevalent in all SOTA models. (Das et al. 2022) has shown that pre-training helps to remove DFT error bias when fine-tuned with experimental data. Hence, we investigate whether SOTA models can remove the DFT error with distilled knowledge from pre-trained model, using a small amount of available experimental data. In specific, we consider OQMD-EXP dataset 1 to conduct an experiment, where we train SOTA models and their distilled variants with available DFT data and different percentages of experimental data for formation energy. We report the MAE of different SOTA models and their distilled variant in Table 5. We observe, with more amount of experimental training data, all the SOTA models are minimizing the error consistently. Moreover, with distilled knowledge from pre-trained CrysGNN, all SOTA models are reducing MAE further and we observe consistently larger degree of improvement with more amount of experimental training data in almost all the models.

Conclusion

In this work, we present a novel but simple pre-trained GNN framework, CrysGNN, for crystalline materials, which captures both local chemical and global structural semantics of crystal graphs. To pre-train the model, we curate a huge dataset of 800k unlabelled crystal graphs. Further, while predicting different crystal properties, we distill important knowledge from CrysGNN and inject it into different state of the art property predictors and enhance their performance. Extensive experiments on multiple popular datasets and diverse set of SOTA models show that with distilled knowledge from the pre-trained model, all the SOTA models outperform their vanilla versions. Extensive experiments show its superiority over conventional fine-tune models and its inherent ability to remove DFT-induced bias. The pretraining framework can be extended beyond structural graph information in a multimodal setting to include other important (text and image) information about a crystal which would be our immediate future work

References

Bracht, H.; Stolwijk, N.; and Mehrer, H. 1995. Properties of intrinsic point defects in silicon determined by zinc diffusion experiments under nonequilibrium conditions. *Physical Review B*, 52(23): 16542.

Chen, C.; Ye, W.; Zuo, Y.; Zheng, C.; and Ong, S. P. 2019. Graph networks as a universal machine learning framework for molecules and crystals. *Chem. Mater.*, 31(9): 3564–3572.

Chen, T.; Kornblith, S.; Norouzi, M.; and Hinton, G. 2020a. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, 1597–1607. PMLR.

Chen, T.; Liu, S.; Chang, S.; Cheng, Y.; Amini, L.; and Wang, Z. 2020b. Adversarial robustness: From self-supervised pretraining to fine-tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 699–708.

Choudhary, K.; and DeCost, B. 2021. Atomistic Line Graph Neural Network for improved materials property predictions. *npj Computational Materials*, 7(1): 1–8.

Choudhary, K.; Garrity, K. F.; Reid, A. C.; DeCost, B.; Biacchi, A. J.; Hight Walker, A. R.; Trautt, Z.; Hattrick-Simpers, J.; Kusne, A. G.; Centrone, A.; et al. 2020. The joint automated repository for various integrated simulations (JARVIS) for data-driven materials design. *npj Computational Materials*, 6(1): 1–13.

Das, K.; Samanta, B.; Goyal, P.; Lee, S.-C.; Bhattacharjee, S.; and Ganguly, N. 2022. CrysXPP: An explainable property predictor for crystalline materials. *npj Computational Materials*, 8(1): 1–11.

De Jong, M.; Chen, W.; Notestine, R.; Persson, K.; Ceder, G.; Jain, A.; Asta, M.; and Gamst, A. 2016. A statistical learning framework for materials science: application to elastic moduli of k-nary inorganic polycrystalline compounds. *Scientific reports*, 6(1): 1–11.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

He, K.; Fan, H.; Wu, Y.; Xie, S.; and Girshick, R. 2020. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9729–9738.

Hinton, G.; Vinyals, O.; Dean, J.; et al. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2(7).

Hu, W.; Liu, B.; Gomes, J.; Zitnik, M.; Liang, P.; Pande, V.; and Leskovec, J. 2019. Strategies for pre-training graph neural networks. *arXiv preprint arXiv:1905.12265*.

Hu, W.; Liu, B.; Gomes, J.; Zitnik, M.; Liang, P.; Pande, V.; and Leskovec, J. 2020a. Strategies for Pre-training Graph Neural Networks. In *International Conference on Learning Representations*.

Hu, Z.; Dong, Y.; Wang, K.; Chang, K.-W.; and Sun, Y. 2020b. Gpt-gnn: Generative pre-training of graph neural networks. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 1857–1867.

Im, J.; Lee, S.; Ko, T.-W.; Kim, H. W.; Hyon, Y.; and Chang, H. 2019. Identifying Pb-free perovskites for solar cells by machine learning. *npj Computational Materials*, 5(1): 1–8.

Isayev, O.; Oses, C.; Toher, C.; Gossett, E.; Curtarolo, S.; and Tropsha, A. 2017. Universal fragment descriptors for predicting properties of inorganic crystals. *Nature communications*, 8(1): 1–12.

Jain, A.; Ong, S. P.; Hautier, G.; Chen, W.; Richards, W. D.; Dacek, S.; Cholia, S.; Gunter, D.; Skinner, D.; Ceder, G.; et al. 2013. Commentary: The Materials Project: A materials genome approach to accelerating materials innovation. *APL materials*, 1(1): 011002.

Jha, D.; Choudhary, K.; Tavazza, F.; Liao, W.-k.; Choudhary, A.; Campbell, C.; and Agrawal, A. 2019. Enhancing materials property prediction by leveraging computational and experimental data using deep transfer learning. *Nature communications*, 10(1): 1–12.

Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kirklin, S.; Saal, J. E.; Meredig, B.; Thompson, A.; Doak, J. W.; Aykol, M.; Rühl, S.; and Wolverton, C. 2015a. The Open Quantum Materials Database (OQMD): assessing the accuracy of DFT formation energies. *npj Computational Materials*, 1(1): 1–15.

Kirklin, S.; Saal, J. E.; Meredig, B.; Thompson, A.; Doak, J. W.; Aykol, M.; Rühl, S.; and Wolverton, C. 2015b. The Open Quantum Materials Database (OQMD): assessing the accuracy of DFT formation energies. *npj Computational Materials*, 1(1): 1–15.

Kittel, C.; and McEuen, P. 2018. *Kittel's Introduction to Solid State Physics*. John Wiley & Sons.

Kubaschewski, O.; Alcock, C. B.; and Spencer, P. 1993. Materials Thermochemistry. Revised. *Pergamon Press Ltd, Headington Hill Hall, Oxford OX 3 0 BW, UK, 1993. 363*.

Lee, J.; Seko, A.; Shitara, K.; Nakayama, K.; and Tanaka, I. 2016. 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.

Louis, S.-Y.; Zhao, Y.; Nasiri, A.; Wang, X.; Song, Y.; Liu, F.; and Hu, J. 2020. Graph convolutional neural networks with global attention for improved materials property prediction. *Physical Chemistry Chemical Physics*, 22(32): 18141–18148.

Lu, S.; Zhou, Q.; Ouyang, Y.; Guo, Y.; Li, Q.; and Wang, J. 2018. Accelerated discovery of stable lead-free hybrid organic-inorganic perovskites via machine learning. *Nature communications*, 9(1): 1–8.

Orio, M.; Pantazis, D. A.; and Neese, F. 2009. Density functional theory. *Photosynthesis research*, 102(2-3): 443–453.

Park, C. W.; and Wolverton, C. 2020. Developing an improved crystal graph convolutional neural network framework for accelerated materials discovery. *Physical Review Materials*, 4(6).

Pilania, G.; Gubernatis, J. E.; and Lookman, T. 2015. Structure classification and melting temperature prediction in octet AB solids via machine learning. *Physical Review B*, 91(21): 214302.

Qiu, J.; Chen, Q.; Dong, Y.; Zhang, J.; Yang, H.; Ding, M.; Wang, K.; and Tang, J. 2020. Gcc: Graph contrastive coding for graph neural network pre-training. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 1150–1160.

Romero, A.; Ballas, N.; Kahou, S. E.; Chassang, A.; Gatta, C.; and Bengio, Y. 2014. Fitnets: Hints for thin deep nets. *arXiv preprint arXiv:1412.6550*.

Seko, A.; Hayashi, H.; Nakayama, K.; Takahashi, A.; and Tanaka, I. 2017. Representation of compounds for machine-learning prediction of physical properties. *Physical Review B*, 95(14): 144110.

Seko, A.; Togo, A.; Hayashi, H.; Tsuda, K.; Chaput, L.; and Tanaka, I. 2015. Prediction of low-thermal-conductivity compounds with first-principles anharmonic lattice-dynamics calculations and Bayesian optimization. *Physical review letters*, 115(20): 205901.

Sohn, K. 2016. Improved deep metric learning with multi-class n-pair loss objective. *Advances in neural information processing systems*, 29.

Trinh, T. H.; Luong, M.-T.; and Le, Q. V. 2019. Selfie: Self-supervised pretraining for image embedding. *arXiv preprint arXiv:1906.02940*.

Turns, S. R. 1995. Understanding NOx formation in non-premixed flames: experiments and modeling. *Progress in Energy and Combustion Science*, 21(5): 361–385.

Van den Oord, A.; Li, Y.; and Vinyals, O. 2018. Representation learning with contrastive predictive coding. *arXiv e-prints*, arXiv–1807.

Ward, L.; Liu, R.; Krishna, A.; Hegde, V. I.; Agrawal, A.; Choudhary, A.; and Wolverton, C. 2017. Including crystal structure attributes in machine learning models of formation energies via Voronoi tessellations. *Physical Review B*, 96(2): 024104.

Wu, Z.; Xiong, Y.; Yu, S. X.; and Lin, D. 2018. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3733–3742.

Xie, T.; Fu, X.; Ganea, O.-E.; Barzilay, R.; and Jaakkola, T. 2021. Crystal diffusion variational autoencoder for periodic material generation. *arXiv preprint arXiv:2110.06197*.

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

You, Y.; Chen, T.; Sui, Y.; Chen, T.; Wang, Z.; and Shen, Y. 2020. Graph contrastive learning with augmentations. *Advances in Neural Information Processing Systems*, 33: 5812–5823.

Related Work

Materials Property Prediction.

In recent times, data driven approaches (Lu et al. 2018; Isayev et al. 2017; Im et al. 2019; Ward et al. 2017; De Jong et al. 2016; Seko et al. 2015; Pilania, Gubernatis, and Lookman 2015; Seko et al. 2017; Lee et al. 2016) have become quite popular to establish relationship between the atomic structure of crystalline materials and their properties with very high precision. Specially, graph neural network (GNN) based approaches have emerged as a powerful machine learning model tool to encode material's complex topological structure along with node features in an enriched representation space. Models such as CGCNN (Xie and Grossman 2018) represent 3D crystal structure as a multi-graph and build a graph convolution neural network directly on the graph to update node features based on their local chemical and structural environment. GATGNN (Louis et al. 2020) incorporates the idea of graph attention network to learn the importance of different inter atomic bonds whereas MEGNet (Chen et al. 2019) introduces global state attributes for quantitative structure-state property relationship prediction in materials. While all these models implicitly represent many body interactions through multiple graph convolution layers, ALIGNN (Choudhary and DeCost 2021) explicitly captures many body interactions by incorporating bond angles and local geometric distortions.

Graph Pre-training Strategies

There are attempts to pre-train GNNs to extract graph and node level representations. (Hu et al. 2019) develops an effective pre-training strategy for GNNs, where they perform both node level and graph level pre-training on GNNs to capture domain specific knowledge about nodes and edges, in addition to global graph-level knowledge. GPT-GNN (Hu et al. 2020b) proposes a self-supervised attributed graph generation task to pre-train a GNN model, which captures structural and semantic properties of the graph. (Qiu et al. 2020) presents GCC, which leverages the idea of contrastive learning to design the graph pre-training task as subgraph instance discrimination, to capture universal network topological properties across multiple networks. Another recent work is GraphCL (You et al. 2020), which proposes different graph augmentation methods and maximises the agreement between two augmented views of the same graph via a contrastive loss. In the field of crystal graphs, CrysXPP (Das et al. 2022) is the only model which comes close to a pre-trained model. In their work, an autoencoder is trained on a volume of un-tagged crystal graphs and the learned knowledge is (transferred to) used to initialize the encoder of CrysXPP, which is fine-tuned with property specific tagged data.
To the best of our knowledge, CrysGNN is the first attempt to develop a deep pre-trained model for the crystalline materials and for that, we curated a new large untagged crystal dataset with 800K crystal graphs, We are going to share the pre-trained model along with the large dataset with the community, so that the pre-trained model can be plugged into any existing and upcoming models to enhance their prediction accuracy.

Features	Unit	Range	Feature Dimension
Group Number	-	1,2, ..., 18	18
Period Number	-	1,2, ..., 9	9
Electronegativity	-	0.5-4.0	10
Covalent Radius	pm	25-250	10
Valence Electrons	-	1,2, ..., 12	12
First Ionization Energy	eV	1.3–3.3	10
Electron Affinity	eV	-3–3.7	10
Block	-	s, p, d, f	4
Atomic Volume	cm^3/mol	1.5–4.3	10

Table 6: Description of different atomic features $\mathcal{X}_i$ and their unit, range and dimensions. All of them together forms 92 dimensional node feature vector $\mathcal{X}_i$.

Crystal Representation

We realize a crystal material as a multi-graph structure $\mathcal{G}_i = (\mathcal{V}_i, \mathcal{E}_i, \mathcal{X}_i, \mathcal{F}_i)$ as proposed in (Das et al. 2022; Xie and Grossman 2018). $\mathcal{G}_i$ is an undirected weighted multi-graph where $\mathcal{V}_i$ denotes the set of nodes or atoms present in a unit cell of the crystal structure. $\mathcal{E}_i = \{(u, v, k_{uv})\}$ denotes a multi-set of node pairs and k_{uv} denotes number of edges between a node pair (u, v). $\mathcal{X}_i = \{(x_u | u \in \mathcal{V}_i)\}$ denotes the node feature set proposed by CGCNN (Xie and Grossman 2018). It includes different chemical properties like electronegativity, valance electron, covalent radius, etc. Details of these node features are given in Table 6. Finally, $\mathcal{F}_i = \{\{s^k\}_{(u,v)} | (u, v) \in \mathcal{E}_i, k \in \{1..k_{uv}\}\}$ denotes the multi-set of edge weights where s^k corresponds to the k^{th} bond length between a node pair (u, v), which signifies the inter-atomic bond distance between two atoms.

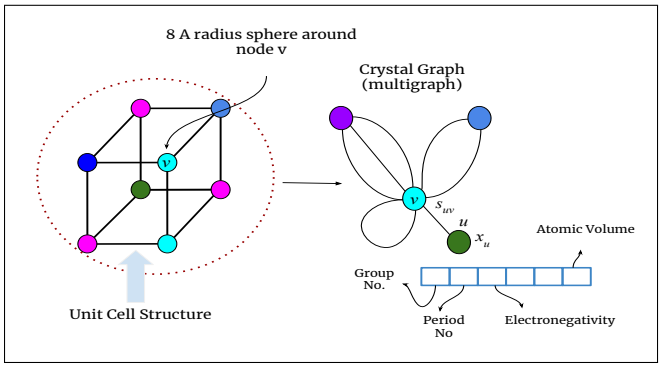


Figure 4: Multi-graph Representation from Crystal 3D Structure.

Details of GNN Architecture in CrysGNN

Here we describe GNN architectures used in CrysGNN for pre-training and property prediction model.

Encoder. We develop an graph convolution (Xie and Grossman 2018) based encoding module, which takes crystal multi-graph structure $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{X}, \mathcal{F})$ as input and encode structural semantics of the crystal graph into lower dimensional space. Considering L graph convolution layers, neighbour-

hood aggregation of the l-th layer is represented as :

$$z_u^{(l+1)} = z_u^{(l)} + \sum_{v,k} \sigma(h^{(l)}_{(u,v)_k} W_c^{(l)} + b_c^{(l)}) \odot g(h^{(l)}_{(u,v)_k} W_s^{(l)} + b_s^{(l)}) \tag{5}$$

where $h^{(l)}_{(u,v)_k} = z_u^{(l)} \oplus z_v^{(l)} \oplus s^{(k)}_{(u,v)}$. Here $s^{(k)}_{(u,v)}$ represents inter-atomic bond length of the k-th bond between (u, v) and $z_u^{(l)}$ denotes embedding of node u after l-th layer, which is initialized to a transformed node feature vector $z_u^0 := x_u W_x$ where W_x is the trainable parameter and x_u is the input node feature vector (Eq. 6). In Equation 5, $W_c^{(l)}, W_s^{(l)}, b_c^{(l)}, b_s^{(l)}$ are the convolution weight matrix, self weight matrix, convolution bias, and self bias of the l-th layer, respectively, $\oplus$ operator denotes concatenation, $\odot$ denotes element-wise multiplication, σ is the sigmoid function indicating the edge importance and g is a feed forward network. After L-layers of such propagation, $\mathcal{Z} = \{z_1, ..., z_{|\mathcal{V}|}\}$ represents the set of final node embeddings, where $z_u := z_u^L$ represents final embedding of node u.

Decoder. In the decoding module, we apply both node-level and graph-level self-supervised pre-training as discussed in **Node Level Decoding** and **Graph Level Decoding** sections.

Property Predictor. For the property prediction model, we use four diverse sets of SOTA models, which have different encoder choices and on top of that they have a regressor module to predict the property value. We retrofit CrysGNN model with each of these SOTA models using the modified multi-task loss (incorporating knowledge distillation loss with their property loss) as discussed in Equation 3 (main manuscript).

Crystal System and Space group Information

By definition crystal is a periodic arrangement of repeating "motifs"(e.g. atoms, ions). The symmetry of a periodic pattern of repeated motifs is the total set of symmetry operations allowed by that pattern. The total set of such symmetry operations, applicable to the pattern is the pattern's symmetry, and is mathematically described by a so-called Space Group. The Space Group of a Crystal describes the symmetry of that crystal, and as such it describes an important aspect of that crystal's internal structure.
In turn crystals across certain space groups show similarities among each other, hence they are divided among seven Crystal Systems : Triclinic, Monoclinic, Orthorhombic, Tetragonal, Trigonal, Hexagonal, and Cubic or Isometric system (Kittel and McEuen 2018).
Crystallographic Axes: The crystallographic axes are imaginary lines that we can draw within the crystal lattice. These will define a coordinate system within the crystal. We refer to the axes as **a**, **b**, **c** and angle between them as α, β and γ. By using these crystallographic axes we can define six large groups or crystal systems.

- **Cubic or Isometric system :** The three crystallographic axes are all equal in length and intersect at right angles to each other.

$$\boxed{a = b = c\ ;\ \ \alpha = \beta = \gamma = 90^\circ}$$

- **Tetragonal system :** Three axes, all at right angles, two of which are equal in length (a and b) and one (c) which is different in length (shorter or longer).

$$\boxed{a = b \neq c\ ;\ \ \alpha = \beta = \gamma = 90^\circ}$$

- **Orthorhombic system :** Three axes, all at right angles, all three have different length.

$$\boxed{a \neq b \neq c\ ;\ \ \alpha = \beta = \gamma = 90^\circ}$$

- **Monoclinic system :** Three axes, all unequal in length, two of which (a and c) intersect at an oblique angle (not 90 degrees), the third axis (b) is perpendicular to the other two axes.

$$\boxed{a \neq b \neq c\ ;\ \ \alpha = \gamma = 90^\circ \neq \beta}$$

- **Triclinic system :** The three axes are all unequal in length and intersect at three different angles (any angle but 90 degrees).

$$\boxed{a \neq b \neq c\ ;\ \ \alpha \neq \beta \neq \gamma \neq 90^\circ}$$

- **Hexagonal system :** Here we have four axes. Three of the axes fall in the same plane and intersect at the axial cross at 120°. These 3 axes, labeled a_1, a_2, and a_3, are the same length. The fourth axis, c, may be of *a* different length than the axes set. The c axis also passes through the intersection of the *a* axes set at right angle to the plane formed by the *a* set.

Details about datasets and different crystal properties.

OQMD: The Open Quantum Materials Database (OQMD) is a high-throughput database based on DFT calculations of chemicals from the Inorganic Crystal Structure Database (ICSD) and embellishments of regularly occurring crystal structures (Kirklin et al. 2015b). Comprehensive computations for significant structure types in materials research, such as Heusler and perovskite compounds, are among the distinguishing aspects of this database. The database is public and a free resource.

Materials project: This is another free and public resource database that contains crystal structures and calculated materials properties (Jain et al. 2013). The dataset is made from the results obtained with density functional theory based calculations. The dataset is composed of electronic structure, thermodynamic, mechanical, and dielectric properties. It also provides a visual web-based search interface.

JARVIS: JARVIS (Joint Automated Repository for Various Integrated Simulations) is a data repository that incorporates not only DFT-based calculations, but also data from classical force-fields and machine learning approaches (Choudhary et al. 2020). The databases are free and public as well.

Property	Unit	Data-size
Formation_Energy	$eV/(atom)$	69239
Bandgap (OPT)	eV	69239
Formation_Energy	$eV/(atom)$	55723
Bandgap (OPT)	eV	55723
Total_Energy	$eV/(atom)$	55723
Ehull	eV	55371
Bandgap (MBJ)	eV	18172
Bulk Modulus (Kv)	GPa	19680
Shear Modulus (Gv)	GPa	19680
SLME (%)	No unit	9068
Spillage	No unit	11377
ϵ_x (OPT)	No unit	44491
ϵ_y (OPT)	No unit	44491
ϵ_z (OPT)	No unit	44491
ϵ_x (MBJ)	No unit	16814
ϵ_y (MBJ)	No unit	16814
ϵ_z (MBJ)	No unit	16814
n-Seebeck	μVK^{-1}	23210
n-PF	$\mu W(mK^2)^{-1}$	23210
p-Seebeck	μVK^{-1}	23210
p-PF	$\mu W(mK^2)^{-1}$	23210

Table 7: Summary of different crystal properties in Materials Project (Top) and JARVIS-DFT (Bottom) datasets.

We curated around 800K untagged crystal graph data from two popular materials databases, Materials Project (MP) and OQMD, to pre-train CrysGNN model. Further to evaluate the performance of different SOTA models with distilled knowledge from CrysGNN, we select MP 2018.6.1 version of Materials Project and 2021.8.18 version of JARVIS-DFT, another popular materials database, for property prediction as suggested by (Choudhary and DeCost 2021). Please note, MP 2018.6.1 dataset is a subset of the dataset used for pre-training, whereas JARVIS-DFT is a separate dataset which is not seen during the pre-training. MP 2018.6.1 consists of 69,239 materials with two properties namely bandgap and formation energy, whereas JARVIS-DFT(2021.8.18) consists of 55,722 materials with 19 properties which can be broadly classified into two categories : 1) properties like formation energy, bandgap, total energy, bulk modulus, etc. which depend greatly on crystal structures and atom features, and 2) properties like ϵ_x, ϵ_y, ϵ_z, n-Seebeck, n-PF, etc. which depend on the precise description of the materials' electronic structure. Details of these properties is provided in Table 7

Moreover, all these properties in both Materials Project and JARVIS-DFT datasets are based on DFT calculations of chemicals. Therefore, to investigate how pre-trained knowledge helps to mitigate the DFT error, we also take a small dataset (Kirklin et al. 2015a, OQMD-EXP) containing 1,500 available experimental data of formation energy. Details of each of these datasets are given in Table 1.

Training Setup / Hyper-parameter Details / Computational Resources used.

We use five convolution layers of the encoder module to train CrysGNN and train it for 200 epochs using Adam (Kingma and Ba 2014) for optimization with a learning rate of 0.03. We keep the embedding dimension for each node as 64, batch size of data as 128, and equal weightage (0.25) for α, β, γ, and λ of Equation 2. During property prediction, we used the default configuration for hyper-parameters for all four vanilla SOTA models. However, while predicting properties by the distilled version of the SOTA models we keep node embedding dimension for each SOTA as 64, and train each model for 500 epochs using Adam (Kingma and Ba 2014) optimization and keep δ as 0.5. We perform the experiments in shared servers having Intel E5-2620v4 processors which contain 16 cores/thread and four GTX 1080Ti 11GB GPUs each.

Baseline Models.

To evaluate the effectiveness of CrysGNN, we choose following four diverse state of the art algorithms for crystal property prediction.

1. **CGCNN** (Xie and Grossman 2018) : This work generates crystal graphs from inorganic crystal materials and builds a graph convolution based supervised model for predicting various properties of the crystals.
2. **GATGNN** (Louis et al. 2020) : In this work, authors have incorporated a graph neural network with multiple graph-attention layers (GAT) and a global attention layer, which can learn efficiently the importance of different complex bonds shared among the atoms within each atom's local neighborhood.
3. **ALIGNN** (Choudhary and DeCost 2021) : In this work, authors adopted line graph neural networks to develop an alternative way to include angular information into convolution layer which alternates between message passing on the bond graph and its bond-angle line graph.
4. **CrysXPP** (Das et al. 2022) : In this work, the authors train an autoencoder (CrysAE) on a volume of un-tagged crystal graphs and then the learned knowledge is (transferred to) used to initialize the encoder of CrysXPP, which is fine-tuned with property specific tagged data.Also they design a feature selector that helps to interpret the model's prediction.

Baseline Implementations.

We used the available PyTorch implementations of all the baselines, viz, CGCNN[1], GATGNN[2], CrysXPP[3] and ALIGNN[4]. In order to retrofit CrysGNN to each SOTA model, we modified their supervised loss into the multi-task loss proposed in Equation 3 (main manuscript) and train each model.

[1]https://github.com/txie-93/cgcnn.git

[2]https://github.com/superlouis/GATGNN.git

[3]https://github.com/kdmsit/crysxpp.git

[4]https://github.com/usnistgov/alignn.git

Property	CGCNN	CGCNN (Distilled)	CrysXPP	CrysXPP (Distilled)	GATGNN	GATGNN (Distilled)	ALIGNN	ALIGNN (Distilled)
ϵ_x (OPT)	24.93	**23.86**	24.47	**23.33**	26.79	**25.23**	21.42	**21.37**
ϵ_y (OPT)	25.06	**24.08**	24.84	**23.98**	26.77	**25.02**	21.66	**21.25**
ϵ_z (OPT)	24.99	**23.85**	24.13	**23.76**	26.09	**24.81**	**20.51**	21.03
ϵ_x (MBJ)	27.18	**26.64**	28.40	**26.03**	**27.56**	27.61	24.76	**24.32**
ϵ_y (MBJ)	27.14	**25.80**	26.92	**25.34**	27.21	**26.89**	23.99	**23.83**
ϵ_z (MBJ)	28.38	**26.08**	26.76	**25.97**	25.50	**24.84**	**23.93**	24.59
n-Seebeck	50.32	**46.66**	48.79	**46.53**	52.03	**49.46**	42.85	**42.13**
n-PF	528.25	**506.41**	502.79	**498.43**	511.56	**505.22**	477.09	**474.88**
p-Seebeck	53.01	**49.15**	50.13	**48.67**	54.99	**50.72**	46.04	**45.19**
p-PF	513.57	**501.11**	493.35	**489.54**	516.84	**494.92**	460.11	**447.81**

Table 8: Summary of the prediction performance (MAE) of different properties (belong to second class) in JARVIS-DFT dataset where performance improvement is modest. Model M is the vanilla variant of a SOTA model and M (Distilled) is the distilled variant using the pretrained CrysGNN. All the models are trained on 80% data, validated on 10% and evaluated on 10% of the data. The best performance is highlighted in bold.

To ensure a fair comparison between all the baselines and perform the knowledge distillation between node embeddings, we have ensured node embedding dimension for each SOTA and CrysGNN as 64.

Additional Experimental Results

Evaluation on Electronic Properties.

Properties of crystalline materials can be broadly classified into two categories : 1) properties like formation energy, bandgap, total energy, bulk modulus, etc. which depend greatly on crystal structures and atom features, and 2) properties like ϵ_x, ϵ_y, ϵ_z, n-Seebeck, n-PF, etc. which depend on the precise description of the materials' electronic structure. We discuss performance of SOTA models with distilled knowledge from CrysGNN for the first class of properties in **"Downstream Task Evaluation"** section and report the MAE in Table 2. We further evaluate the effectiveness of CrysGNN for the second class of properties for both the vanilla and distilled model and report the MAE in Table 8. We observe, for this class of properties all the SOTA models had a higher MAE and though pre-training enhances the performance, the improvement is modest. In specific, average improvement in CGCNN, CrysXPP, GATGNN and ALIGNN are 6%, 4%, 3.6% and 0.8%, respectively. A potential reason is, electronic dielectric constant, Seebeck coefficients, and power factors all depend greatly on the precise description of the materials' electronic structure, which is neither captured by the SOTA models nor by the pre-trained CrysGNN framework explicitly. Hence error is high for these properties by SOTA models and injecting distilled structural information from the pre-trained model is able to achieve only modest improvements.

Effectiveness on sparse training dataset.

A concise version of this result is presented in the main paper, we here elaborate some details. To demonstrate the effectiveness of the pre-training in limited data settings, we conduct additional set of experiments under different training data split. In specific, we vary available training data from 20 to 60 %, train different SOTA models and check their performance on test dataset. We report the MAE values of different baselines and their distilled version in Table 9 for five different properties, for which available data is very limited. We observe that the distilled version of any SOTA model consistently outperforms its vanilla version even in the limited training data setting, which illustrates the robustness of our pre-training framework. Specifically, for CGCNN and GATGNN, the improvements are more for 20% training data, because these deep models suffer from data scarcity issue and with distilled knowledge from the pre-trained model they are able to mitigate the issue. ALIGNN achieves moderate improvements as before but surprisingly, we found the least improvement (additional improvement for 20% training data compared to other settings) for CrysXPP. CrysXPP itself pre-trains an autoencoder and transfers the learned information to the property predictor. Hence, the vanilla version performs well in the limited data settings.

Moreover, in some cases, we observe that the MAE values of the distilled version of a model using lesser training data is better than vanilla version of the model using more training data. For example, while predicting Bandgap (MBJ), CGCNN (Distilled) with 20% and 40% training data outperforms CGCNN with 40% and 60% training data, respectively. These results verify that even with lesser training data, the distilled information from the large pre-trained model gives performance comparable to using larger training data by the original model.

Ablation Study : Analysis of Different Pre-training Loss

A concise version of this result is presented in the main paper, we here elaborate some details. We perform an ablation study to investigate the influence of different pre-training losses in enhancing the SOTA model performance. While pre-training CrysGNN (Eq. 2), we capture both local chemical and global structural information via node and graph-level decoding, respectively. Further, we are curious to know the influence of each of these decoding policies independently in the downstream property prediction task. In specific, we conduct the ablation experiments, where we pre-train CrysGNN with (a)

Property	Train-Val -Test(%)	CGCNN	CGCNN (Distilled)	CrysXPP	CrysXPP (Distilled)	GATGNN	GATGNN (Distilled)	ALIGNN	ALIGNN (Distilled)
Bandgap (MBJ)	20-10-70	0.588	0.453* (23.04)	0.598	0.450* (24.82)	0.541	0.521 (3.70)	0.497	0.485 (2.53)
	40-10-50	0.532	0.419* (21.41)	0.496	0.405* (18.40)	0.462	0.448* (2.81)	0.404	0.395 (2.20)
	60-10-30	0.449	0.364 (19.08)	0.435	0.360 (17.36)	0.449	0.439 (2.29)	0.387	0.380 (1.98)
Bulk Modulus (Kv)	20-10-70	16.91	16.26 (3.80)	15.42	14.25* (7.59)	14.80	14.19 (4.12)	14.70	14.06 (4.35)
	40-10-50	14.81	14.46 (2.36)	15.13	14.02* (7.34)	12.98	12.59 (3.00)	12.47	12.11 (2.89)
	60-10-30	14.23	14.05 (1.26)	14.76	13.73 (6.98)	12.01	11.75 (2.16)	11.23	11.01 (1.96)
Shear Modulus (Gv)	20-10-70	13.89	12.50 (10.01)	13.39	12.07* (9.86)	12.83	12.42 (3.20)	12.71	12.31 (3.15)
	40-10-50	12.04	11.54* (4.15)	12.16	11.01* (9.46)	11.43	11.23 (1.75)	10.98	10.67 (2.82)
	60-10-30	11.75	11.31 (3.74)	11.77	10.67 (9.35)	10.65	10.47 (1.69)	10.24	10.04 (1.95)
SLME (%)	20-10-70	7.13	6.62 (7.17)	7.05	5.90* (16.40)	6.35	6.02 (5.20)	6.36	6.27 (1.43)
	40-10-50	6.14	5.78 (5.90)	6.89	5.81 (15.67)	5.78	5.63 (2.60)	5.65	5.57 (1.42)
	60-10-30	5.55	5.24 (5.68)	5.41	4.84 (10.54)	5.48	5.34 (2.55)	4.88	4.82 (1.33)
Spillage	20-10-70	0.424	0.389* (8.43)	0.422	0.403 (4.45)	0.406	0.392* (3.5)	0.402	0.392 (2.49)
	40-10-50	0.408	0.391* (4.28)	0.398	0.384 (3.57)	0.396	0.388 (1.94)	0.378	0.372 (1.72)
	60-10-30	0.395	0.380 (3.80)	0.379	0.368 (3.00)	0.382	0.377 (1.36)	0.362	0.357 (1.13)

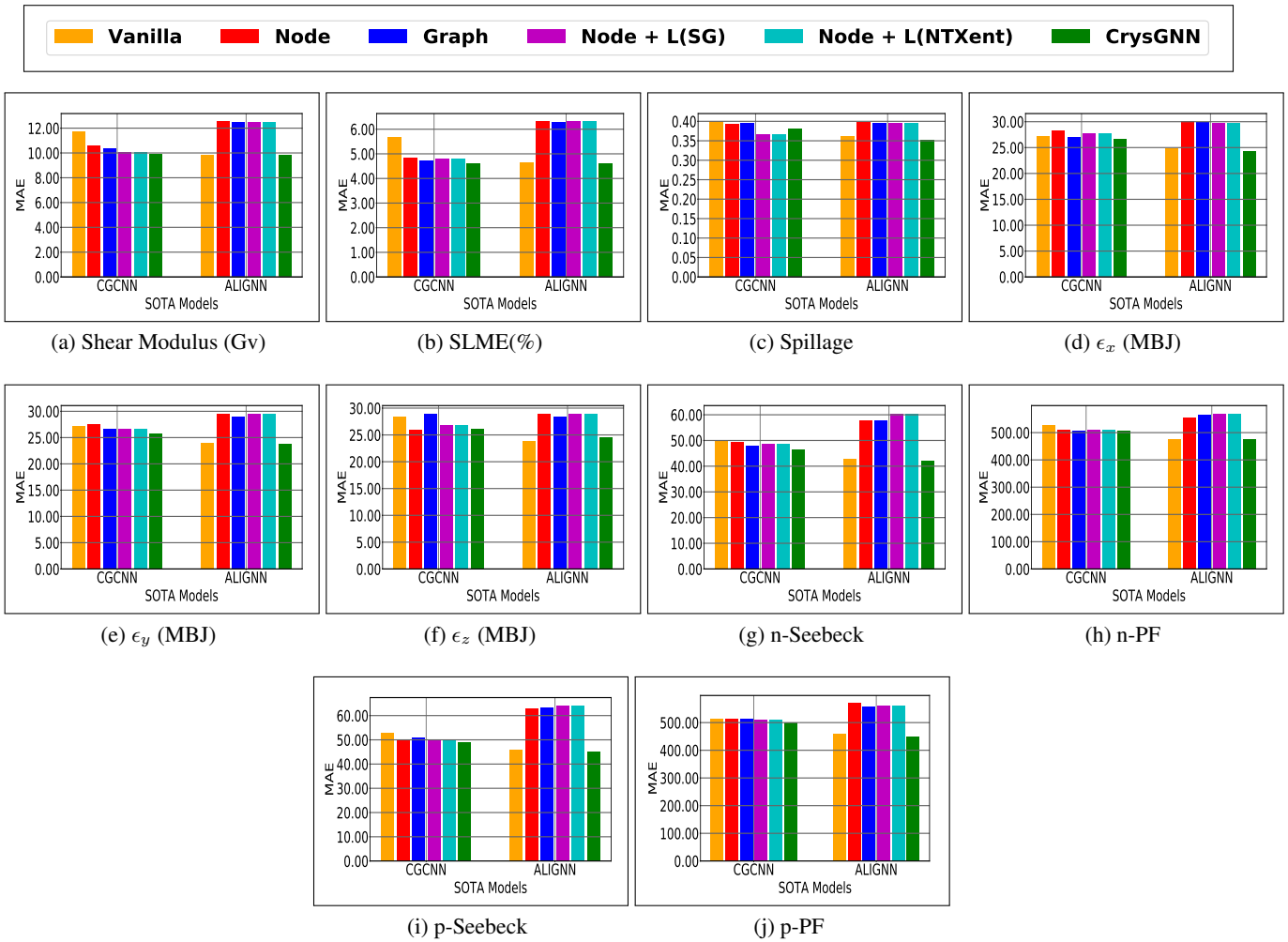
Table 9: MAE values of five different properties in JARVIS-DFT dataset with the increase in training instances from 20 to 60%. Model M is the vanilla variant of the SOTA model, M (Distilled) is distilled variant of it and relative improvement is mentioned in bracket. Distilled knowledge from CrysGNN improves the performance of all the baselines consistently. In few cases, we observe that MAE values of the distilled version of a model using even lesser training data is better than vanilla version of the same model using more training data and highlight it with *.

only node-level decoding ($\mathcal{L}_{FR}$, $\mathcal{L}_{CR}$), (b) only graph-level decoding ($\mathcal{L}_{SG}$, $\mathcal{L}_{NTXent}$). Further, we perform ablations with individual graph-level losses, and pretrain with (c) removing $\mathcal{L}_{NTXent}$ (node-level with $\mathcal{L}_{SG}$ (space group)) and (d) removing $\mathcal{L}_{SG}$ (node-level with $\mathcal{L}_{NTXent}$(crystal system)). We train two baseline models, CGCNN and ALIGNN, with distilled knowledge from all the aforementioned variants of the pre-trained model and evaluate the performance on four crystal properties. We report performance of both the experiments for four properties in Fig.3 and for ten different properties in Fig.5.

We can observe clearly that all the variants offer significant performance gain in all four properties using the combined node and graph-level pre-training, compared to node-level or graph-level pre-training separately. Only exception is formation energy, where only node-level pre-training produces less error compared to other variants, in both the baseline. Formation energy of a crystal is defined as the difference between the energy of a unit cell comprised of N chemical species and the sum of the chemical potentials of all the N chemical species. Hence pre-training at the node-level (node features and connection) is adequate for enhancing performance of formation energy prediction and incorporating graph-level information works as a noisy information, which degrades the performance. We also observe improvement in performance using both supervised and contrastive graph-level losses ($\mathcal{L}_{SG}$ and $\mathcal{L}_{NTXent}$), compared to using only one of them, which proves that the learned representation via supervised and contrastive learning is more expressive that using any one of them. Moreover, in ALIGNN, with either node or graph-level pre-training separately, performance degrades across different properties. ALIGNN explicitly captures the three body interactions which drive its performance, to replicate that inclusion of both node and graph information is necessary.

Limitations and Future Work.

Though our proposed pre-trained model is able to enhance the performance of all the state of the art baseline models, improvements for all types of SOTA models are not consistent and we found lesser improvement in case of complex models like GATGNN and ALIGNN. This provides scope for further investigation on designing a more complex and deeper pre-trained model. There are also scopes of exploring different graph representations, loss variations, distillation strategies, etc. Of course, the very fact that the simple model, CrysGNN, can provide substantial improvements, lays the foundation for such future exploration. Moreover, in this present work, we have focused on predicting crystal properties, which is a graph level regression task. However, the same framework can be used to explore the effect of pre-training on other graph level tasks (classification or clustering) or node level tasks which would be one of our future works.



(a) Shear Modulus (Gv) (b) SLME(%) (c) Spillage (d) ϵ_x (MBJ)

(e) ϵ_y (MBJ) (f) ϵ_z (MBJ) (g) n-Seebeck (h) n-PF

(i) p-Seebeck (j) p-PF

Figure 5: Summary of experiments of ablation study on importance of different pre-training loss components on CrysGNN training and eventually its effect on CGCNN and ALIGNN models on ten different properties from JARVIS DFT dataset(MAE for property prediction). (i) Vanilla: SOTA based model (without distillation) and all the other cases are SOTA models (distilled) from different pre-trained version of CrysGNN. (ii) Node : only node-level pre-training, (iii) Graph : only graph-level pre-training, (iv) Node + L(SG) : node-level and $\mathcal{L}_{SG}$, (v) Node + L(NTXent) : node-level and $\mathcal{L}_{NTXent}$ and (vi) CrysGNN : both node and graph-level pre-training.