CRYSGNN : DISTILLING PRE-TRAINED KNOWLEDGE TO ENHANCE PROPERTY PREDICTION FOR CRYSTALLINE MATERIALS.

Kishalay Das, Bidisha Samanta, Pawan Goyal, Niloy Ganguly
Department of Computer Science and Engineering
Indian Institute of Technology Kharagpur, India.
{kishalaydas@kgpian.,bidisha@,pawang@cse.,niloy@cse.}iitkgp.ac.in

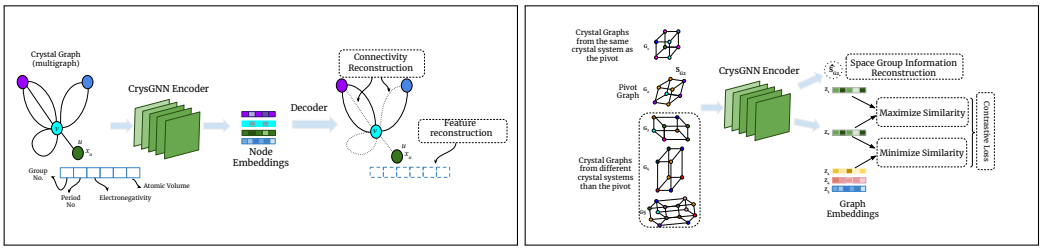
Seung-Cheol Lee & Satadeep Bhattacharjee
Indo Korea Science and Technology Center, Bangalore, India.
Joburg, South Africa
{seungcheol.lee,s.bhattacharjee}@ikst.res.in

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 a significant improvement over the conventional approach of finetuning the pre-trained model.

1 INTRODUCTION

Fast and accurate prediction of different material properties is a challenging and important task in material science. Though there has been an ample amount of data-driven works in recent times, 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 Xie & Grossman (2018); Chen et al. (2019); Louis et al. (2020); Park & Wolverton (2020); Schmidt et al. (2021) gained some popularity in this domain. Understanding the importance of many-body interactions, ALIGNN Choudhary & 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 Devlin et al. (2018); Trinh et al. (2019); Chen et al. (2020); Hu et al. (2020); Qiu et al. (2020); You et al. (2020) for crystalline materials which enables us to leverage a large amount of untagged material structures to



(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.

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[1]) 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 crystal 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.

2 METHODOLOGY

2.1 CRYSGNN PRE-TRAINING

We build a deep auto-encoder architecture CrysGNN, comprises of a graph convolution based encoder followed by an effective decoder which is (pre)trained end to end, using a large amount of property un-tagged crystal graphs $\mathcal{D}_u = \{\mathcal{G}_i\}$, which we have curated from various materials datasets.

[1]Source code, pre-trained model, and dataset of CrysGNN is made available at https://github.com/kdmsit/crysgnn

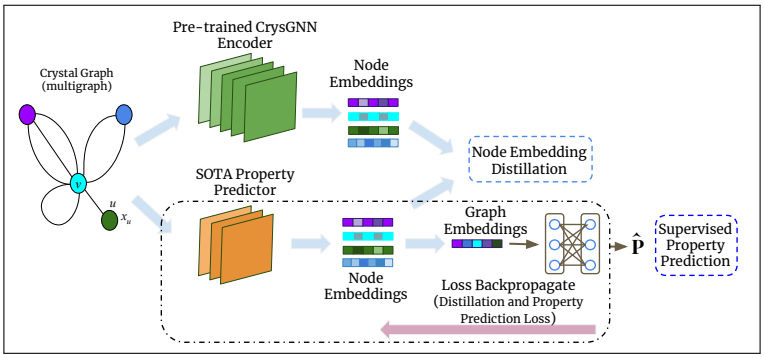


Figure 2: Overview of Property Prediction using Knowledge Distillation from CrysGNN.

2.1.1 SELF SUPERVISION.

We first develop a graph convolution Xie & 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.

Node-Level Decoding. For node-level decoding, we propose two self-supervised learning methods, where given an atom/node u we first reconstruct its node features x_u, which represent different chemical properties of atom u. 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.

Graph-level Decoding. We aim to capture periodic structure of a crystal material through graph-level decoding. We specifically leverage two concepts in doing so. (a). Space group and (b). Crystal system. 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 et al. (2018); Wu et al. (2018) and NT-Xent for the i^{th} graph is defined:

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

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 :

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

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.

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 (%)	5.040	**4.790**	5.110	**4.630**	5.050	**4.950**	4.650	**4.590**
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**	9.860	**9.800**

Table 1: 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.

2.2 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 (Fig-2). 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:

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

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:

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

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.

3 Experimental Results

3.1 Datasets

We curated 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, for property prediction as suggested by Choudhary & 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 bandgap and formation energy, whereas JARVIS-DFT consists of 55,722 materials with 9 different properties.

3.2 Downstream Task Evaluation

To evaluate the effectiveness of CrysGNN, we choose four diverse state of the art algorithms for crystal property prediction, CGCNN Xie & Grossman (2018), GATGNN Louis et al. (2020), CrysXPP Das et al. (2022) and ALIGNN Choudhary & 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 1, 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.

4 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. The pretraining framework can be extended beyond structural graph information in a multi-modal setting to include other important (text and image) information about a crystal which would be our immediate future work.

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.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pp. 1597–1607. PMLR, 2020.

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

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.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Geoffrey Hinton, Oriol Vinyals, Jeff Dean, et al. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2(7), 2015.

Ziniu Hu, Yuxiao Dong, Kuansan Wang, Kai-Wei Chang, and Yizhou Sun. Gpt-gnn: Generative pre-training of graph neural networks. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 1857–1867, 2020.

Dipendra Jha, Kamal Choudhary, Francesca Tavazza, Wei-keng 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.

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.

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.

Jiezhong Qiu, Qibin Chen, Yuxiao Dong, Jing Zhang, Hongxia Yang, Ming Ding, Kuansan Wang, and Jie Tang. Gcc: Graph contrastive coding for graph neural network pre-training. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 1150–1160, 2020.

Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. Fitnets: Hints for thin deep nets. *arXiv preprint arXiv:1412.6550*, 2014.

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.

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

Trieu H Trinh, Minh-Thang Luong, and Quoc V Le. Selfie: Self-supervised pretraining for image embedding. *arXiv preprint arXiv:1906.02940*, 2019.

Aaron Van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv e-prints*, pp. arXiv–1807, 2018.

Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3733–3742, 2018.

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.

Yuning You, Tianlong Chen, Yongduo Sui, Ting Chen, Zhangyang Wang, and Yang Shen. Graph contrastive learning with augmentations. *Advances in Neural Information Processing Systems*, 33: 5812–5823, 2020.