Learning Robust Representation of Crystal Materials for Property Prediction

Kishalay Das
Indian Institute of Technology Kharagpur
Kharagpur, India
kishalaydas@kgpian.iitkgp.ac.in

ABSTRACT

In the past few years, several deep learning techniques have been proposed to enable fast and accurate prediction of different properties for crystal materials, thus facilitating rapid screening of large material search spaces. Particularly, graph neural network (GNN) models have gained prominence due to their capacity to encode graph information in an enriched representation space. Although existing state-of-the-art models predict different material properties with reasonable precision, they suffer from some inherent limitations like *scarcity of labeled data*, *lack of interpretability*, *dependency on domain knowledge*, *lack of Pre-trained model* and *lack of global structural knowledge*. In this context, my research focuses on learning more enriched and robust representations of crystal materials, which not only enhances the accuracy of property prediction but also mitigates the aforementioned limitations.

KEYWORDS

Crystal Representation Learning, Crystal Property Prediction, AI4Science, Graph Pretraining, Multi-modal Learning

ACM Reference Format:
Kishalay Das. 2024. Learning Robust Representation of Crystal Materials for Property Prediction. In . ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 MOTIVATION

Deep learning models have led to significant progress in the field of the chemical, biological, and material science communities, to solve fundamental domain-specific problems. One such fundamental problem in materials science is rapid and accurate prediction of different properties of crystal materials which is imperative for finding new stable crystal materials. This can be done by chemical screening[3] of large material search spaces to find material candidates with desired properties. Historically, Density functional theory (DFT)[15] has been used as an effective tool to estimate several chemical properties however, requires substantial computation costs; hence makes the screening process inefficient. In recent times there has been an ample amount of data-driven works [7, 10–12, 14, 17, 19–21] for predicting crystal properties which are as accurate as DFT but, much faster than it. Particularly, graph neural network (GNN) models have gained prominence due to their capacity to encode complex graph semantic information in an enriched representation space. Existing SOTA GNN models [1, 2, 13, 16, 18, 22] construct multi-edge graphs for a 3D material structure where they create edges between nearby atoms within a pre-specified distance threshold in 3D space and apply GNN model to learn representations of crystal structures that are optimized for downstream property prediction tasks.

Although existing variants of GNN models predict different crystal properties with high precision, they suffer from the following major limitations: **(A) Scarcity of Labeled Data:** Existing models are supervised in nature, possessing large trainable parameters like typical deep neural networks. Hence, a large amount of property-labeled data is needed to train these models, which makes it challenging for many material properties where we don't have enough property-tagged data. **(B) Lack of Interpretability:** Current methods lack interpretability and algorithmic transparency for their results, limiting their utility in material science applications. Therefore it is necessary to explore and provide the reasons behind a prediction for any given property. **(C) Dependency on Domain Knowledge:** The architectural innovations of these models come from incorporating specific domain knowledge into a deep encoding module. However, as different properties expressed by crystal materials 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. **(D) Lack of Pre-trained Graph Model:** While pre-training has been effective in language and vision domains, it remains an open question how to effectively use pre-training on graph datasets like crystals, which will be robust and task agnostic. **(D) Lack of Global Knowledge:** Existing models rely on a single modality of crystal data i.e crystal graph structure and fail to incorporate crucial global periodic structural information, which can aid the property prediction accuracy.

Our research focuses on developing machine learning algorithms to learn more robust and enriched representations for crystal materials, which will enhance the property prediction accuracy and mitigate the aforementioned issues.

2 RESEARCH CONTRIBUTIONS

In this section, we describe three lines of work that deal with learning more robust and enriched crystal representation to improve property prediction.

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

2.1 Transfer Learning-based Unsupervised Framework [5]

In our first work[5], we tackled the data scarcity, and lack of interpretability issues. We leverage a transfer learning-based unsupervised framework to develop an explainable property predictor. It is built upon CrysAE, an auto-encoder-based architecture that is trained with all available (untagged) crystal data. This leads to the deep encoding module capturing all the important structural and chemical information of the constituent atoms of the crystal graph. The learned information is leveraged to build the property predictor, CrysXPP, where the knowledgeable encoder helps to produce a high-quality representation of a candidate crystal. Consequently, the property predictor provides superior performance even when trained with a small amount of property-tagged data. Further, we introduce a feature selector that helps to provide an explanation by highlighting the subset of the atomic features responsible for the manifestation of a property of the given crystal. The node features are first passed through a feature selector which is a trainable weight vector that selects a weighted subset of important node-level features for a given crystal property of interest.

Through extensive analysis of a popular inorganic crystal data set across seven properties, we show that our method can achieve the lowest error compared to other alternative baselines; the improvement is particularly significant when only a small amount of tagged data is available for training. Further, with appropriate case studies, we show that the feature selection module can effectively provide explanations of the importance of different features towards prediction, which are in sync with the domain knowledge.

2.2 Pretrained GNN Model for Crystal Material[6]

In our next work[6], we extend the idea of pre-training further and tackle the issues of domain knowledge dependency and lack of a pre-trained graph model in material science. In this work, we introduce a graph pre-training method that captures (a) the 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, CrysGNN, with the dataset. CrysGNN learns the representation of a crystal graph by initiating both node (atom) and graph (crystal) level losses. 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. Further, 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: $\mathcal{L}_{prop} = \delta\mathcal{L}_{MSE} + (1-\delta)\mathcal{L}_{KD}$ where $\mathcal{L}_{MSE} = (\hat{y}_i - y_i)^2$ denotes the discrepancy between predicted and true property values by $\mathcal{P}_\psi$. 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: $\mathcal{L}_{KD} = \|\mathcal{Z}_i^T - \mathcal{Z}_i^S\|^2$, 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. During property prediction, the pre-trained network is frozen and we backpropagate $\mathcal{L}_{prop}$ through the predictor $\mathcal{P}_\psi$ end to end. With rigorous experimentation across two popular benchmark materials datasets, Material Prjects, and JARVIS-DFT, we show that distilling necessary information from CrysGNN to various property predictors results in substantial performance gains than the vanilla model across all the properties. In specific, the average improvement in CGCNN, CrysXPP, GATGNN and ALIGNN are 16.20%, 12.21%, 8.02%, and 4.19%, respectively.

2.3 Crystal Multi-Modal Representation[4]

In this work[4], we propose to learn a more robust and enriched representation by using multi-modal data i.e graph structure and textual description of materials. One of the major advantages of using the textual description of materials is it provides a diverse set of periodic structural information which is useful to estimate different crystal properties but difficult to incorporate explicitly into a graph structure. We first curate the textual dataset of two popular materials databases (Graph-based), Material Project (MP) and JARVIS-DFT, containing textual descriptions of each material of those databases. We used a popular tool Robocrystallographer [8] to generate descriptions for global crystal structures that include space group number, crystal symmetry, rotational information, component orientations, heterostructure information, etc.

We propose, CrysMMNet, a simple multi-modal framework for crystal materials, which has two components: Graph Encoder and Text Encoder. Given a material, a Graph Encoder uses its graph structure and applies a GNN-based approach to encode the local neighborhood structural information around a node (atom), and subsequently learn graph (crystal) representation. On the contrary, Text Encoder is a transformer-based model, which encodes the global structural knowledge from the textual description of the material and generates a textual representation. Finally, both graph structural and textual representation are fused together to generate a more enriched multimodal representation of materials, which captures both global and local structural knowledge and subsequently improves property prediction accuracy. We use four convolution layers of the graph encoder module and pre-trained MatSciBERT[9] followed by a two-layer neural network (projection layer) as the text encoder module in CrysMMNet. We train it for 1000 epochs using AdamW optimizer with normalized weight decay of 10^{-5} and keep the batch size as 64. We schedule the learning rate according to the one-cycle policy with a maximum learning rate of 0.001. We keep embedding the dimensions of the graph and text encoder as 64 and 256 respectively. We observe that CrysMMNet outperforms all the popular state-of-the-art baselines across ten diverse sets of properties on two popular datasets, Materials Project and JARVIS-DFT.

REFERENCES

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

[2] Kamal Choudhary and Brian DeCost. 2021. Atomistic Line Graph Neural Network for improved materials property predictions. *npj Computational Materials* 7, 1 (2021), 1–8.

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

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

[5] Kishalay Das, Bidisha Samanta, Pawan Goyal, Seung-Cheol Lee, Satadeep Bhattacharjee, and Niloy Ganguly. 2022. CrysXPP: An explainable property predictor for crystalline materials. *npj Computational Materials* 8, 1 (2022), 43.

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

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

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

[9] Tanishq Gupta, Mohd Zaki, N. M. Anoop Krishnan, and Mausam. 2022. MatSciBERT: A Materials Domain Language Model for Text Mining and Information Extraction. *npj Computational Materials* 8, 1 (May 2022), 102. https://doi.org/10.1038/s41524-022-00784-w

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

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

[12] Joohwi Lee, Atsuto Seko, Kazuki Shitara, Keita Nakayama, and Isao Tanaka. 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 (2016), 115104.

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

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

[15] Maylis Orio, Dimitrios A Pantazis, and Frank Neese. 2009. Density functional theory. *Photosynthesis research* 102 (2009), 443–453.

[16] Cheol Woo Park and Chris Wolverton. 2020. Developing an improved crystal graph convolutional neural network framework for accelerated materials discovery. *Physical Review Materials* 4, 6 (Jun 2020). https://doi.org/10.1103/physrevmaterials.4.063801

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

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

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

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

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

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