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

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.

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.
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 pretrained 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 initializing 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 $\mathcal{G}_i$, any SOTA property predictor $\mathcal{P}_y$ and set of property tagged training data $\mathcal{D}_t = \{\mathcal{G}_i, y_i\}$, we aim to find optimal parameter values $\psi^*$ for $\mathcal{P}_y$.
REFERENCES


