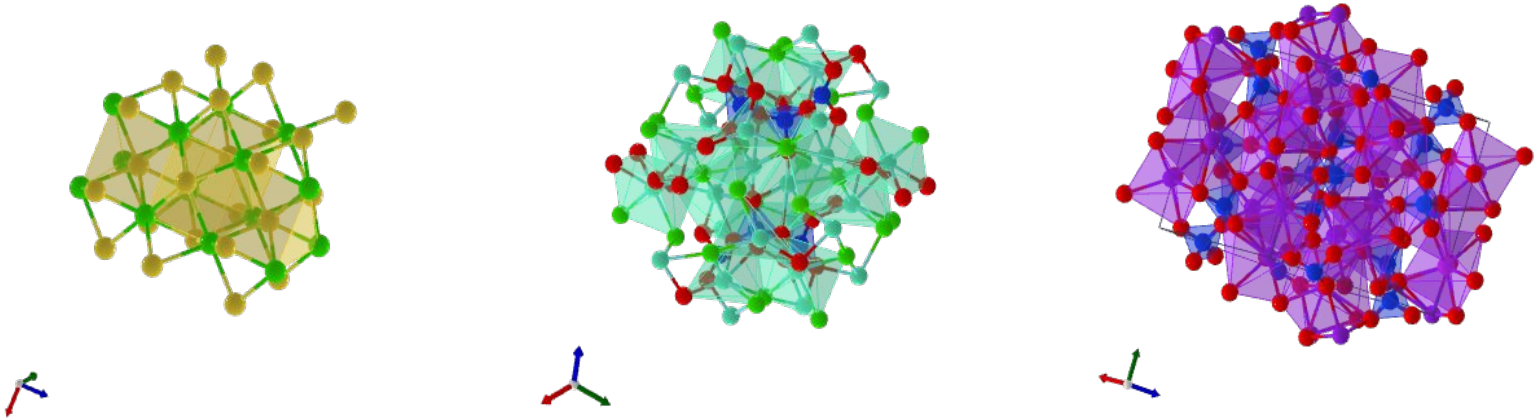
Learning Robust Representation for Crystalline Materials using Graph Neural Networks

Kishalay Das, CSE Dept, IIT-Kharagpur
Ph.D. Supervisors - Pawan Goyal and Niloy Ganguly

Crystalline Material

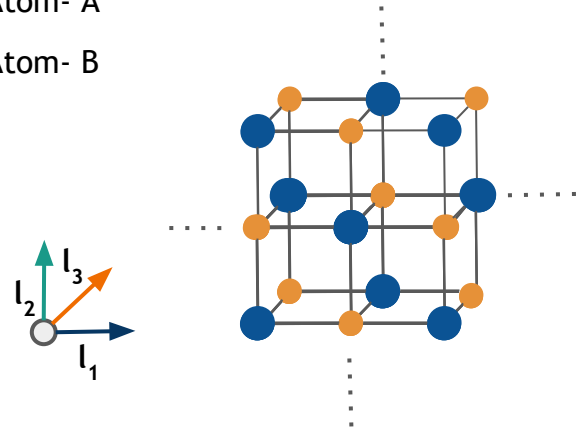


Crystalline materials are typically modeled by a **minimal unit cell** containing all the constituent **atoms in different coordinates**, **repeated** infinite times in **3D space on a regular lattice**, which makes material structures **periodic in nature.**

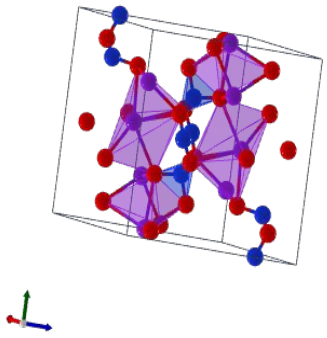
Crystalline Material

Atom- A

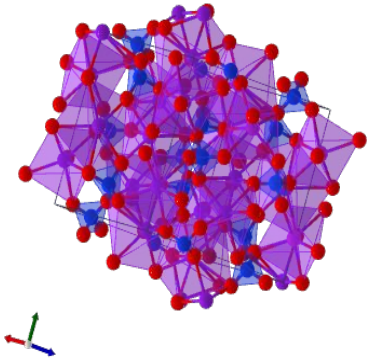
Atom- B



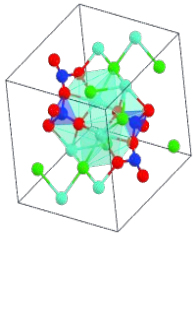
Unit Cell



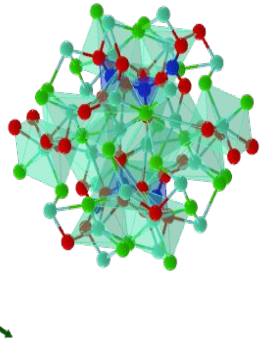
Periodic Structure



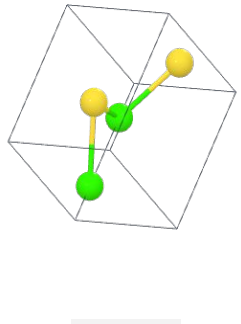
Unit Cell



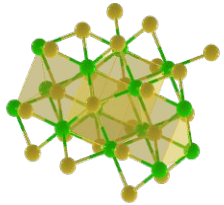
Periodic Structure



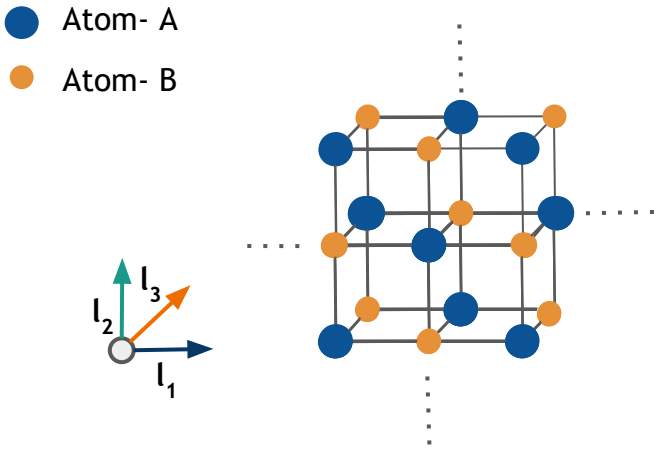
Unit Cell



Periodic Structure



Crystalline Material



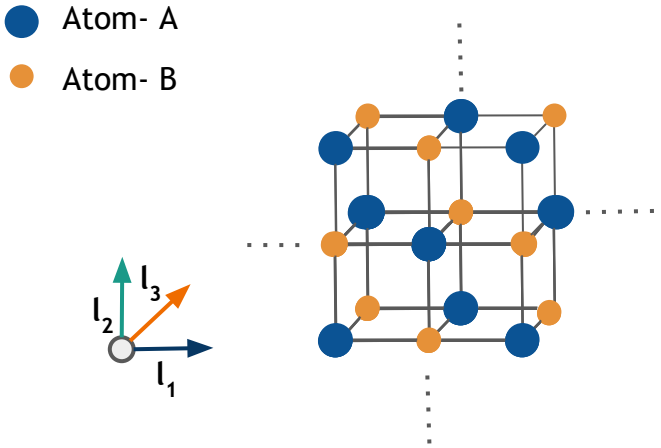
- Crystalline Materials $\mathcal{M} = (\mathcal{C}, \mathcal{X}, \mathcal{L})$
- **Coordinate Matrix** $\mathcal{C} = [\mathbf{c}_1, \mathbf{c}_2, ..., \mathbf{c}_n]^T \in R^{n\times 3}$: atomic coordinate positions, $\mathbf{c}_i \in R^3$ corresponds to cartesian coordinates of i-th atom in the unit cell.
- **Feature Matrix** $\mathcal{X} = [\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n]^T \in R^{n\times d}$: atomic feature set of the material, $\mathbf{x}_i \in R^d$ corresponds to the d-dimensional feature vector of i-th atom.
- **Lattice matrix** $\mathcal{L} = [\mathbf{l}_1, \mathbf{l}_2, \mathbf{l}_3]^T \in R^{3\times 3}$, which describes how a unit cell repeats itself in the 3D space.
- Formally, we can represent the infinite periodic structure of Crystal $\mathcal{M}$ as

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

$$\hat{\mathbf{X}} = \{\hat{\mathbf{x}}_i | \hat{\mathbf{x}}_i = \mathbf{x}_i\}$$

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

Crystalline Material

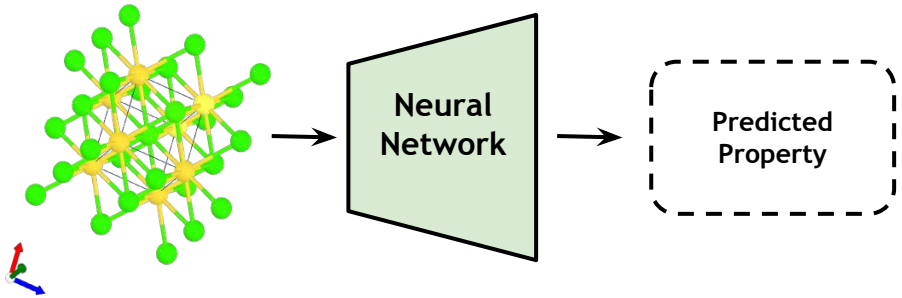


Properties

Property	Unit
Formation_Energy	$eV/(atom)$
Bandgap (OPT)	eV
Formation_Energy	$eV/(atom)$
Bandgap (OPT)	eV
Total_Energy	$eV/(atom)$
Ehull	eV
Bandgap (MBJ)	eV
Bulk Modulus (Kv)	GPa
Shear Modulus (Gv)	GPa
SLME (%)	No unit
Spillage	No unit
ϵ_x (OPT)	No unit
ϵ_y (OPT)	No unit
ϵ_z (OPT)	No unit
ϵ_x (MBJ)	No unit
ϵ_y (MBJ)	No unit
ϵ_z (MBJ)	No unit
n-Seebeck	$\mu V K^{-1}$
n-PF	$\mu W (mK^2)^{-1}$
p-Seebeck	$\mu V K^{-1}$
p-PF	$\mu W (mK^2)^{-1}$

Crystal Property Prediction

- Given a crystal material's 3D structure, predicting different properties is a challenging and important task in material science.
- **Density Functional Theory (DFT)**
  - ↳ **substantial computational costs.**
- **Data driven approaches**
  - ↳ **Accurate as DFT, much faster than it.**
- Majority of the existing approaches, **constructs graphs** by establishing edges only between nearby atoms and use **deep graph neural network (GNN)** to learn crystal structure representation



Limitations of Existing Works

- **Scarcity of Labeled Data**

  Existing models have a large number of trainable parameters, which require a huge amount of tagged training data to learn the models.

- **Lack of Interpretability**

  Existing neural network-based methods hardly provide any explanation for their results, allows little use of them in the field of material science.

- **DFT Error Bias**

  Models are trained using data gathered from the DFT calculations, hence model prediction has a DFT error bias.

- **Dependency on Domain Knowledge**

  Incorporating specific domain knowledge into a deep encoding module.

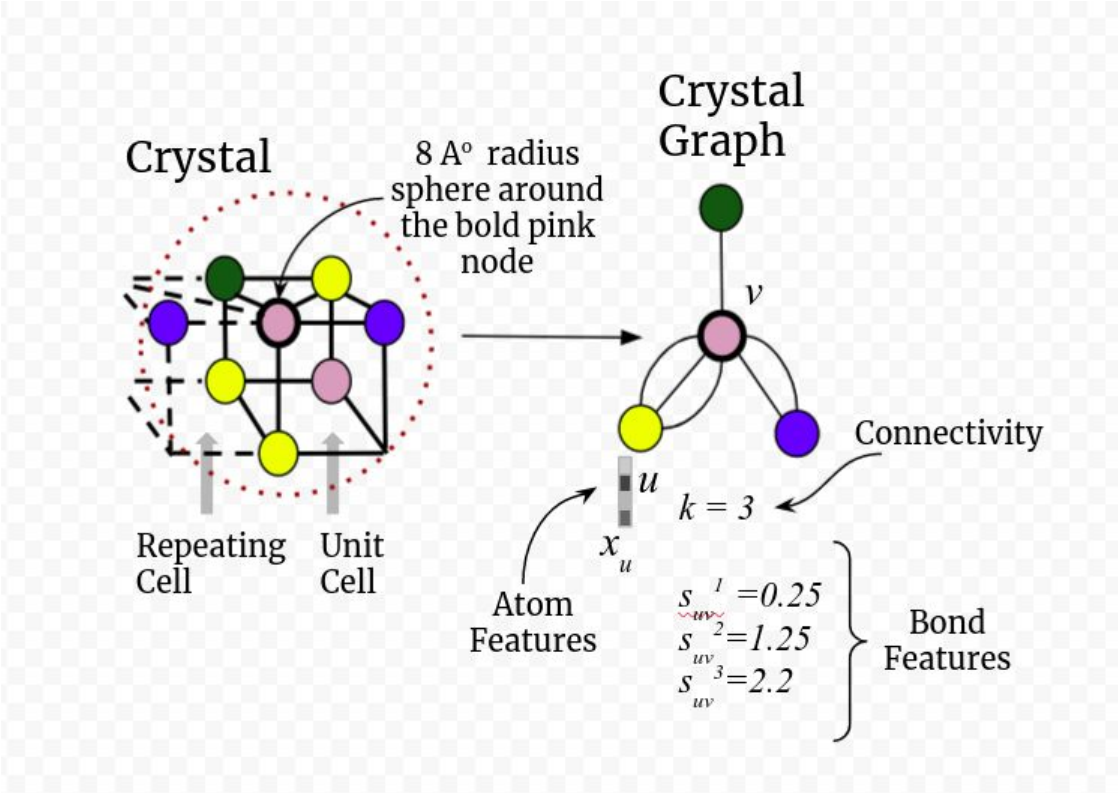
- **Lack of Pre-trained Graph Model**

  It remains an open question how to effectively use pre-training on graph datasets like crystals, which will be robust and task agnostic

Research Question

- Focusing on **developing deep models using Graph Neural Networks (GNN)** to learn more **robust and enriched representations** for crystalline materials, which will **mitigate the existing issues**.
- In specific, we have worked on :
  - leveraging a transfer learning-based unsupervised framework to develop an explainable property predictor **(CrysXPP - NPJ Computational Materials (Nature) Journal, 2022)**
  - Developing a deep pre-trained GNN model using a large curated dataset for crystalline materials. **(CrysGNN - AAAI 2023 [Oral], ML4Materials Workshop at ICLR-2023)**

Multi-Graph Construction of Crystal



Atom Features

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

- Xie, T.; and Grossman, J. C. 2018. Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties. Phys. Rev. Lett.,

CrysXPP: An Explainable Property Predictor for Crystalline Materials

Kishalay Das, Bidisha Samanta, Pawan Goyal, Seung-Cheol Lee, Satadeep Bhattacharjee, Niloy Ganguly

Accepted in **NPJ Computational Materials (Nature)** Journal, 2022

Limitations of Existing Works

- **Scarcity of Labeled Data**

  Existing models have a large number of trainable parameters, which require a huge amount of tagged training data to learn the models.

- **Lack of Interpretability**

  Existing neural network-based methods hardly provide any explanation for their results, allows little use of them in the field of material science.

- **DFT Error Bias**

  Models are trained using data gathered from the DFT calculations, hence model prediction has a DFT error bias.

- **Dependency on Domain Knowledge**

  Incorporating specific domain knowledge into a deep encoding module.

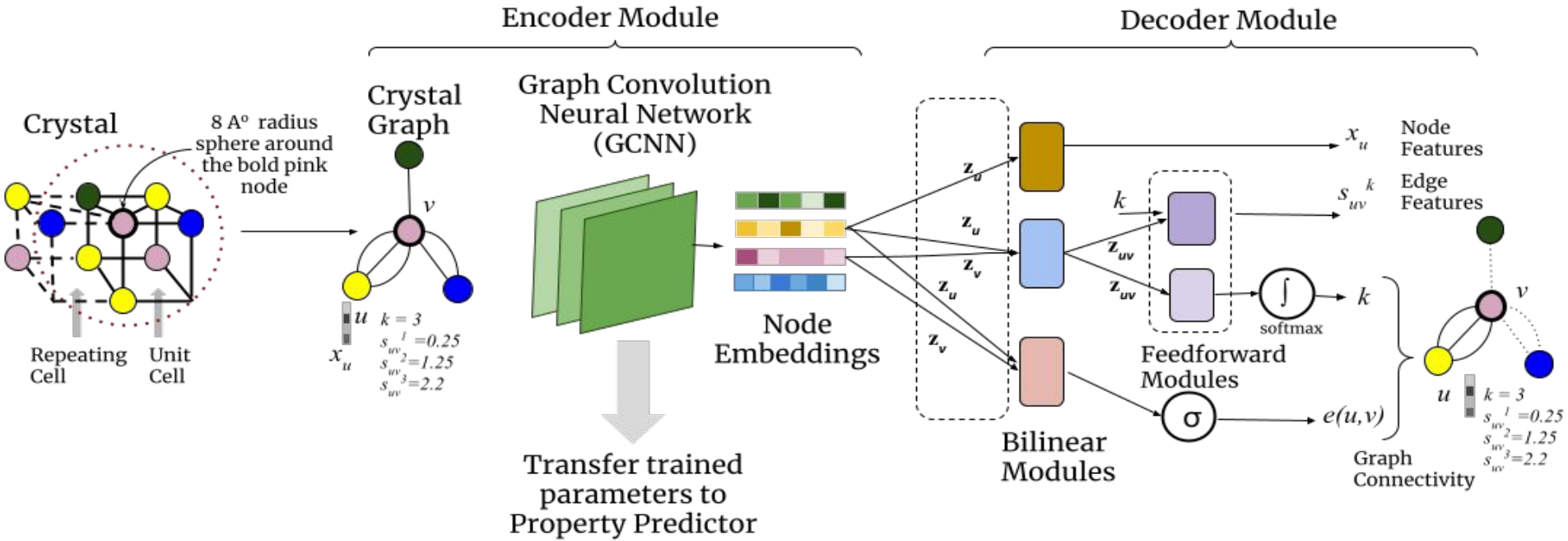
- **Lack of Pre-trained Graph Model**

  It remains an open question how to effectively use pre-training on graph datasets like crystals, which will be robust and task agnostic

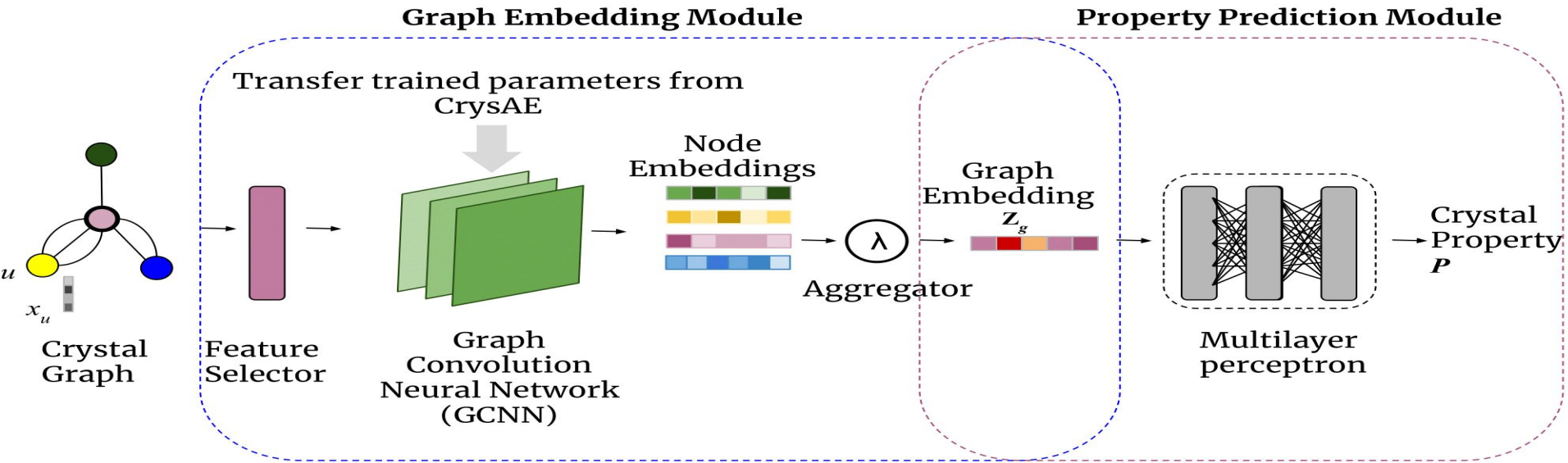
Our Proposed Method

- We propose an **explainable deep property predictor** for crystalline materials which comprised of two modules :
- **CrysAE (Crystal Auto Encoder)** and **CrysXPP (Crystal eXplainable Property Predictor)**
- CrysAE, an **auto-encoder based architecture** which is trained with **all available unlabeled crystal data** (property agnostic), capturing all the **important structural and chemical information** of the constituent atoms (nodes) of the crystal graph.
- This **learned encoding is leveraged** to build up the property predictor, CrysXPP, to which the **knowledge** acquired by the encoder is **transferred** and which is further **trained with a small amount of property-tagged data**, thus largely mitigating the need for having a huge amount of dataset tagged with a specific property.

CrysAE(Crystal Auto Encoder)



CrysXPP(Crystal eXplainable Property Predictor



$$\min_{\zeta,\theta',\psi} \left(\hat{\mathcal{P}} - \mathcal{P}\right)^2 + \lambda_1 * |\zeta|_{L_1}$$

Effectiveness of Property Predictor

Materials Project database (38000 crystalline materials)

Table 2. Summary of the prediction performance (MAE) of different properties trained on 20% data and evaluated on 80% of the data. The best performance is highlighted in bold and second-best with *. We report MAE jointly training most correlated property (average on all property pairs) for MTCGCNN.

	Property	Unit	CGCNN	MTCGCNN	MEGNet	GATGNN	ElemNet	CrysXPP
State properties	Formation energy	eV/atom	0.127	0.112 (0.147)	0.142	0.164	0.098*	**0.086**
	Band gap	eV	0.503	0.497 (0.518)	0.498	0.489*	0.491	**0.467**
	Fermi energy	eV	0.528	0.503* (0.601)	0.533	0.533	0.588	**0.471**
	Magnetic moment	μ_B	1.21	1.16 (1.22)	1.19	1.09	**0.96**	1.03*
Elastic properties	Bulk moduli	log(GPa)	0.09	0.09 (0.09)	0.105	0.088*	0.1057	**0.08**
	Shear moduli	log(GPa)	0.125*	0.120 (0.078)	0.187	0.123	0.148	**0.105**
	Poisson ratio	–	0.04	0.037* (0.039)	0.041	0.039	0.039	**0.035**

Removal of DFT Error bias

- **Setup :**

  We consider a property predictor which has been trained with crystals whose particular property (say Formation Energy, Band Gap) values have been theoretically derived using DFT.

  We then fine tune the parameters with limited amount of experimental data.

- **Formation Energy :** **1500 crystals** whose experimental values of formation energy is known.
- **Band Gap :** we **collect 20 experimental instances from the domain experts**, out of which we **randomly pick 10 instances to fine-tune** the parameters and report the prediction value for the rest

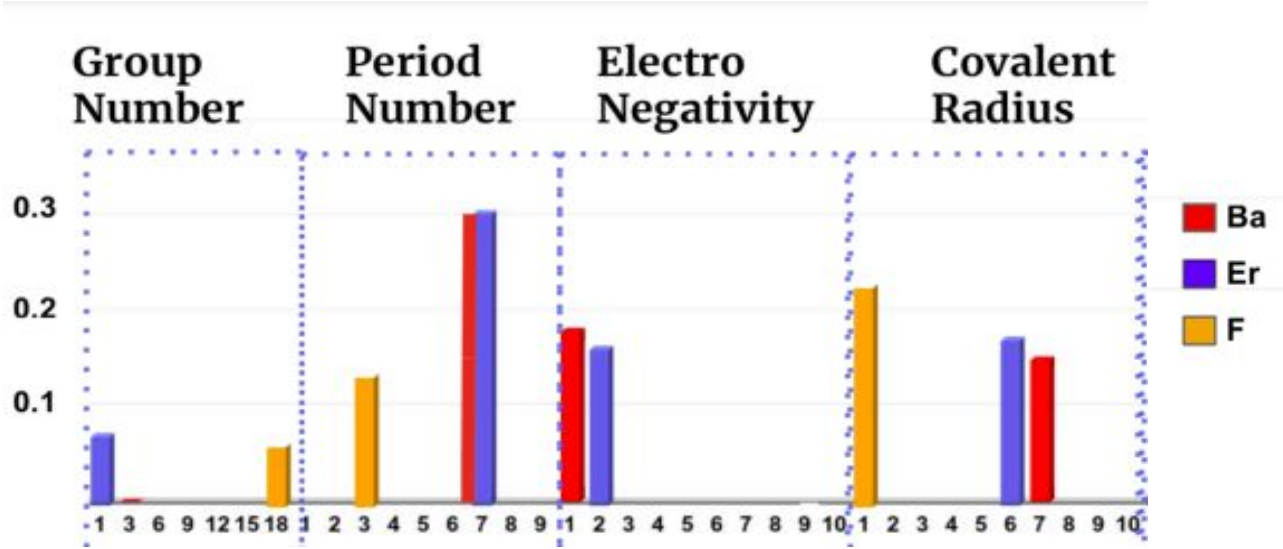
Table 3. MAE of predicting experimental values after fine tuning different methods with different percentages of experimental data for Formation Energy. MAE of the experiment where we replace the experimental data with the same amount of DFT data to train CrysXPP, is provided in the bracket. The closest prediction is marked in bold and second-best with *.

Experiment settings	CGCNN	MTCGCNN	GATGNN	MEGNet	ElemNet	CrysXPP
Train on 20% DFT test on full experimental data	0.24	0.74	0.30	0.28	**0.215**	0.22
Train on 20% DFT, 20% experimental data test on 80% experimental data	0.21	0.24	0.23	0.23	0.16*	**0.15** (0.206)
Train on 80% DFT, 20% experimental data test on 80% experimental data	0.16	0.22	0.19	0.18	0.1344*	**0.1319** (0.195)
Train on 80% DFT, 80% experimental data test on 20% experimental Data	0.12	0.15	0.13	0.125	0.0905*	**0.0892** (0.174)

Materials	Exp	DFT	CrysXPP-Exp	CrysXPP
GaSb	0.72	0.36	0.77	0.9
GaP	2.26	1.69	2.10	1.86
GaAs	1.42	0.18	1.54	1.56
InN	1.97	0.47	1.92	1.85
GaN	3.2	1.73	2.11	1.47
NiO	4.3	2.214	2.45	2.08
Si	1.12	0.85	1.08	0.95
ZnO	3.37	1.05	3.42	2.1
FeO	2.4	0	2.25	1.72
MnO	4	0.20	2.31	1.81

TABLE 8: Experiment (Exp) and predicted value for Band Gap for 10 crystals calculated by DFT and other machine learning models after fine-tuned by experimental data.

Explanation for Formation Energy



- $BaEr_2F_8$ has **Formation Energy -4.41**, indicating **stability** of the materials
- **Period and Group Numbers** provide the information to distinguish each atom.
- **Non-zero difference in Electronegativity** of atoms indicates stability in structure.
- **Covalent Radius** determines the extent of overlap of electron densities of constituents. Higher the radius means weaker the bond. Interesting to note here the trend of weights is the reverse than that of radius itself.

Summary

- In this work, we propose an **explainable property predictor** for crystalline materials, CrysXPP to predict different crystal state and elastic properties with accurate precision **using small amount of property-tagged data**.
- We address the **issue of limited crystal data of a particular property**, using **pretraining - transfer learning paradigm**.
- We further find the encoder knowledge is extremely useful in **de-biasing DFT** error using a meagre instances of experimental results.
- With appropriate case studies, we show that the **explanations** provided by the **feature selection module** are in sync with the domain knowledge.

Paper : https://www.nature.com/articles/s41524-022-00716-8.pdf
Github Repo : https://github.com/kdmsit/crysxpp

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

Kishalay Das, Bidisha Samanta, Pawan Goyal, Seung-Cheol Lee, Satadeep Bhattacharjee, Niloy Ganguly

Accepted in **AAAI-2023** [Oral], **ML4Materials Workshop at ICLR-2023**

Limitations of Existing Works

- **Scarcity of Labeled Data**

  Existing models have a large number of trainable parameters, which require a huge amount of tagged training data to learn the models.

- **Lack of Interpretability**

  Existing neural network-based methods hardly provide any explanation for their results, allows little use of them in the field of material science.

- **DFT Error Bias**

  Models are trained using data gathered from the DFT calculations, hence model prediction has a DFT error bias.

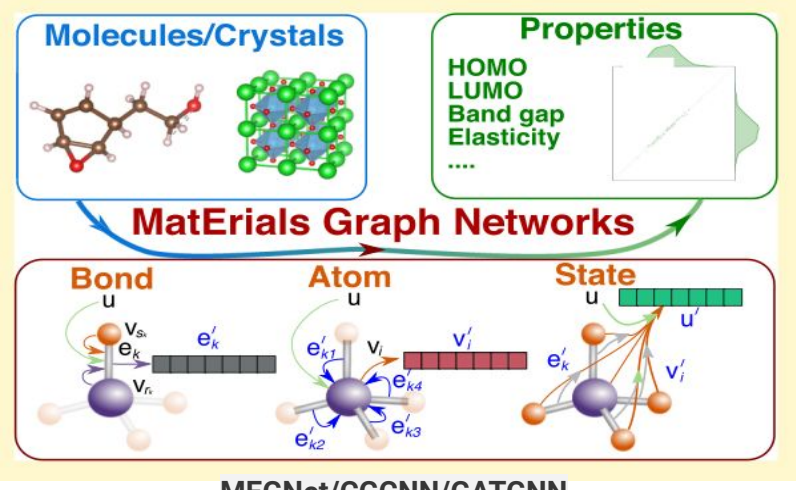
- **Dependency on Domain Knowledge**

  Incorporating specific domain knowledge into a deep encoding module.

- **Lack of Pre-trained Graph Model**

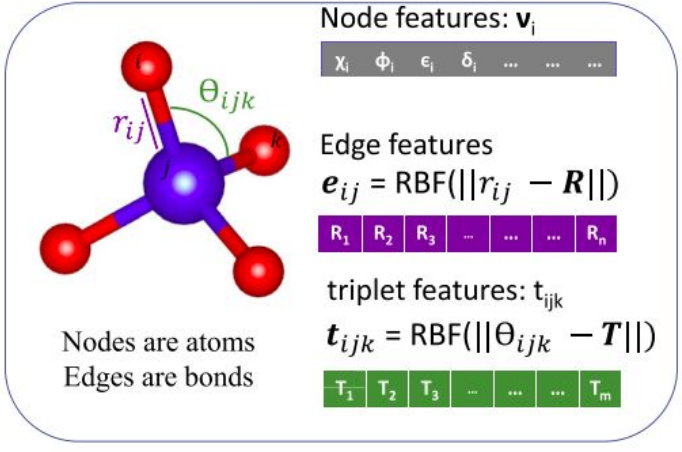
  It remains an open question how to effectively use pre-training on graph datasets like crystals, which will be robust and task agnostic

Limitations of SOTA Models



MEGNet/CGCNN/GATGNN

GNN Based Approaches - *construct graphs by creating edges only between atoms within a* ***pre-specified distance threshold*** *(8 Amstrong)*

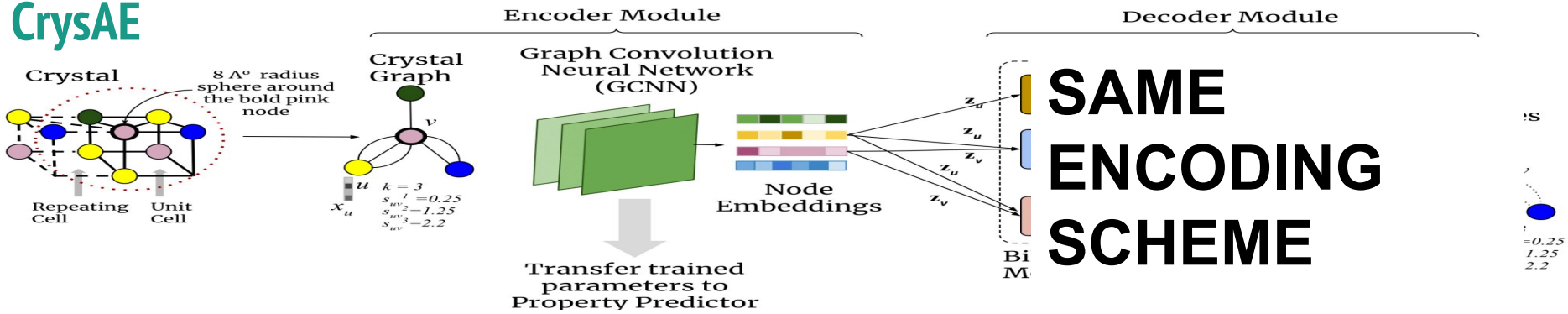


ALIGNN

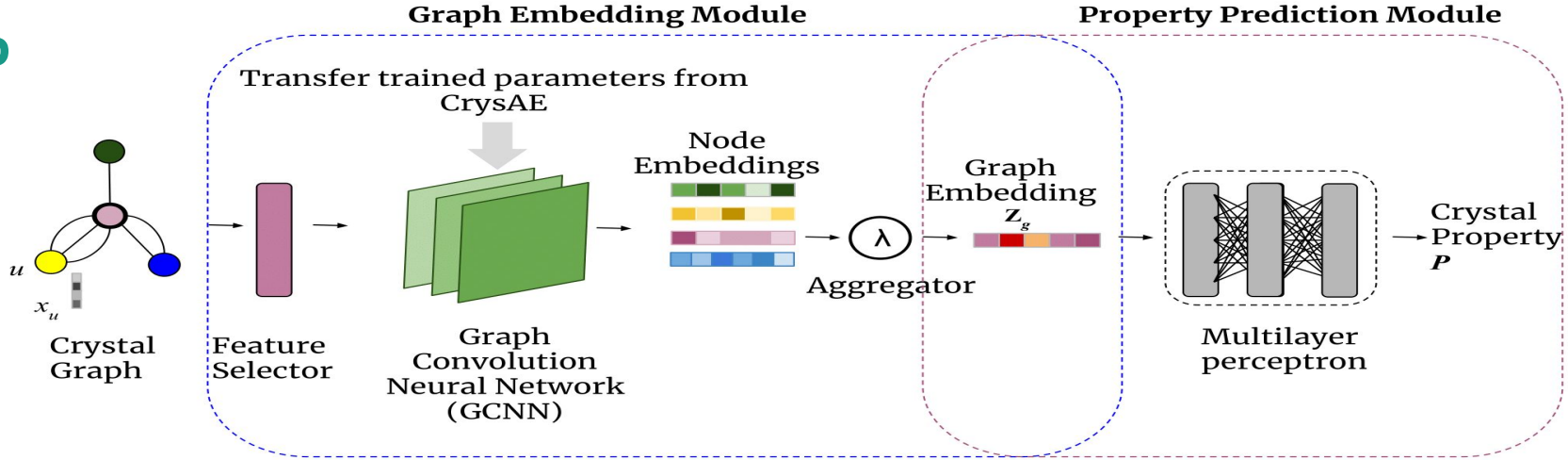
ALIGNN - *Incorporates* ***bond angular information*** *into their encoder module to capture many body interactions between atoms*

Limitations

CrysAE



CrysXPP



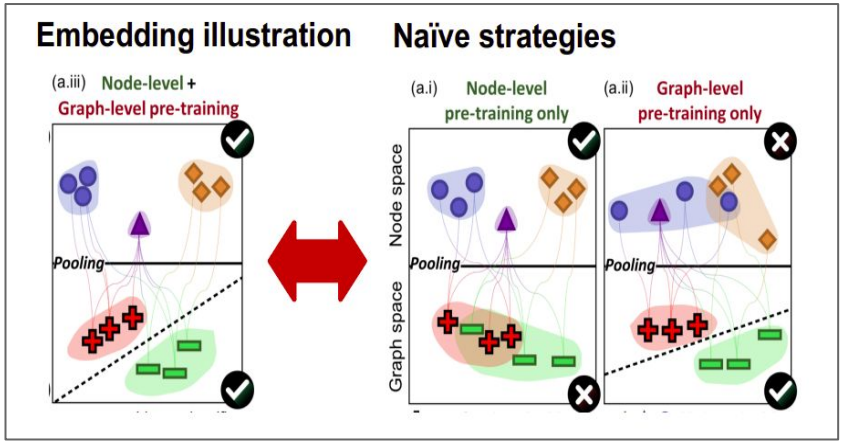
Problem Statement

Can we leverage a **large amount of untagged material structures** to **pretrain a Deep GNN model** which learn the complex hidden features which otherwise are difficult to identify?

Can we **inject** the **pre-trained knowledge into any downstream property predictor**, irrespective of their encoding architecture?

Pretraining GNN

- Prior works focus on molecular and biological dataset, which is **difficult to extend directly to crystalline material.**
- **Structural semantics are different** between molecules and materials.
- For **graph-level pre-training** ⟶ **supervised property prediction** using a huge amount of labelled dataset ⟶ **less effective** in material science where property labeled data is extremely scarce.
- Conventional pre-train fine tuning framework limits knowledge transfer capability of the pre-trained model if the downstream task and dataset is different.

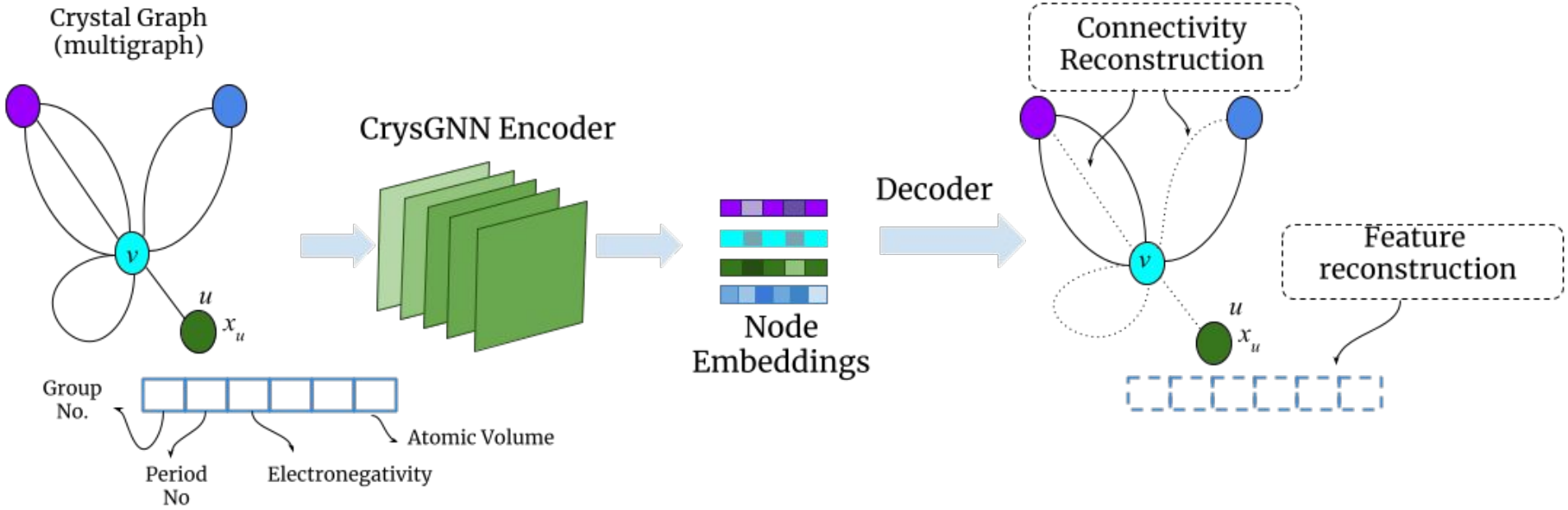


Hu et al. Strategies for Pre-training Graph Neural Networks. (ICLR-2020) : Node-level and graph-level pre-training on GNNs to capture domain specific knowledge about nodes and edges, in addition to global graph-level knowledge. They perform pretraining on large dataset of chemical and biological dataset.

Proposed Methodology

- We developed a **pre-trained GNN model (CrysGNN)** for Crystalline materials, which captures both **local (node level) chemical** and **global (graph level) structural** semantics of crystal graphs.
- We curate a new **large untagged crystal dataset** with **800K crystal graphs** to pretrain CrysGNN.
- We introduce a **self supervised 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 crystal graph data.
- Subsequently we **distill** important **structural and chemical information** of a crystal from the pre-trained CrysGNN model and pass it to the property predictor.
- **Retrofit** the **pre-trained CrysGNN model** into any existing state-of-the-art property predictor, to improve their property prediction performance.

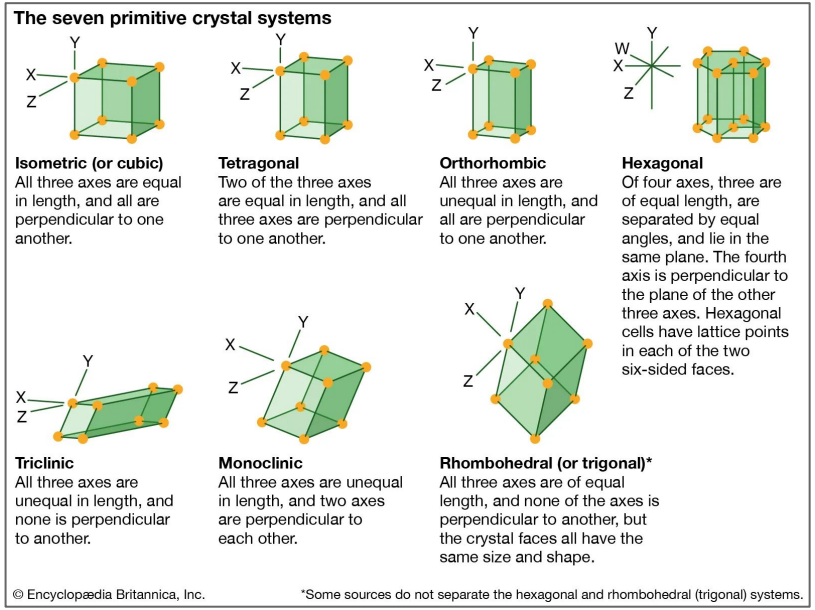
CrysGNN: Node Level Pre-training



CrysGNN: Graph Level Pre-training

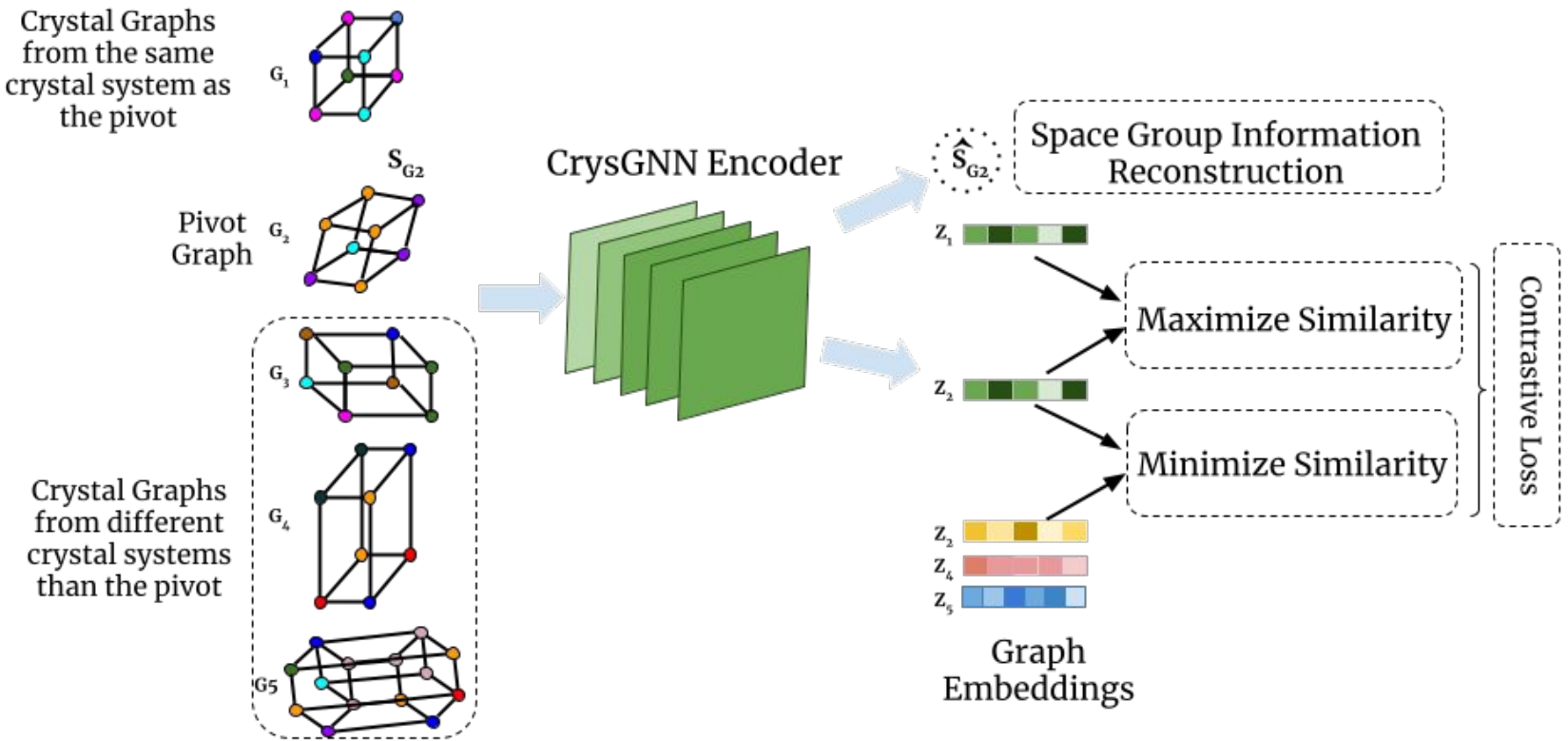
- **Space group of Crystal Structure :**
  - Describe the symmetry of a unit cell of the crystal material.
  - Each crystal has a unique space group number.
  - 230 unique space groups

- **Crystal System:**
  - Space group level information can classify a crystal graph into 7 broad groups of crystal systems.

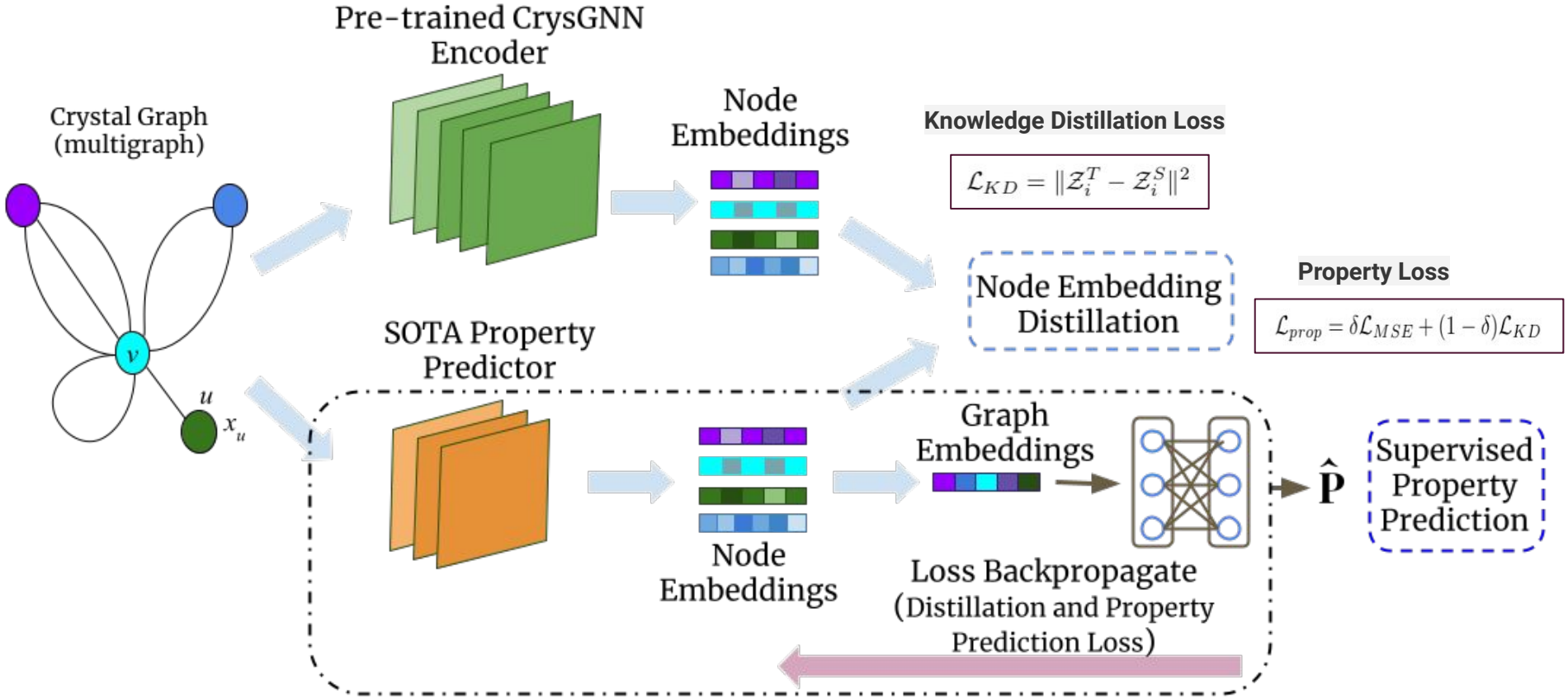


we adopt **supervised** and **contrastive learning** to learn **structural similarities** between graph structures using the **space group and crystal system information** of the materials respectively.

CrysGNN: Graph Level Pre-training



Distillation and Property Prediction



Dataset Details

Table 1: Datasets Details

Task	Datasets	Graph Num.	Structural Info.	Properties Count	Data Type
Pre-training	OQMD	670K	✓	x	DFT Calculated
	Materials Project	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

Downstream Task Evaluation

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 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.

Downstream Task Evaluation

- Distilled version of any state-of- the-art model outperforms the vanilla model across all the properties.
- 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%).
- **Possible reason :** ALIGNN and GATGNN are more complex models that CrysGNN.
- **Potential Improvement :** Incorporating angle-based information or attention mechanism as a part of pre-training framework may improve further.

Comparison with Existing Pre-trained Models.

- Demonstrate the effectiveness of the knowledge distillation method vis-a-vis the conventional fine-tuning approaches.
- We finetune CrysGNN and compare with distilled CGCNN, CrysXPP and pretrain GNN by hu et.al.
- Encoding architecture is same for CrysGNN, CGCNN, and CrysXPP (pretrained-finetuned version of CGCNN)
- Distilled CGCNN outperforms finetuned version of CrysGNN and both the baselines

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 (%)	**4.791**	5.452	5.110	9.853
Spillage	**0.354**	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.

Effectiveness on sparse training dataset.

Property	Train-Val-Test (%)	ALIGNN	ALIGNN (Distilled)	CGCNN	CGCNN (Distilled)	CrysXPP	CrysXPP (Distilled)	GATGNN	GATGNN (Distilled)
Bandgap (MBJ)	20-10-70	0.497	0.485 (2.53)	0.588	0.453* (23.04)	0.598	0.450* (24.82)	0.541	0.521 (3.70)
	40-10-50	0.404	0.395 (2.20)	0.532	0.419* (21.41)	0.496	0.405* (18.40)	0.462	0.448* (2.81)
	60-10-30	0.387	0.380 (1.98)	0.449	0.364 (19.08)	0.435	0.360 (17.36)	0.449	0.439 (2.29)
Bulk Modulus (Kv)	20-10-70	14.70	14.06 (4.35)	16.91	16.26 (3.80)	15.42	14.25* (7.59)	14.80	14.19 (4.12)
	40-10-50	12.47	12.11 (2.89)	14.81	14.46 (2.36)	15.13	14.02* (7.34)	12.98	12.59 (3.00)
	60-10-30	11.23	11.01 (1.96)	14.23	14.05 (1.26)	14.76	13.73 (6.98)	12.01	11.75 (2.16)
Shear Modulus (Gv)	20-10-70	12.71	12.31 (3.15)	13.89	12.50 (10.01)	13.39	12.07* (9.86)	12.83	12.42 (3.20)
	40-10-50	10.98	10.67 (2.82)	12.04	11.54* (4.15)	12.16	11.01* (9.46)	11.43	11.23 (1.75)
	60-10-30	10.24	10.04 (1.95)	11.75	11.31 (3.74)	11.77	10.67 (9.35)	10.65	10.47 (1.69)

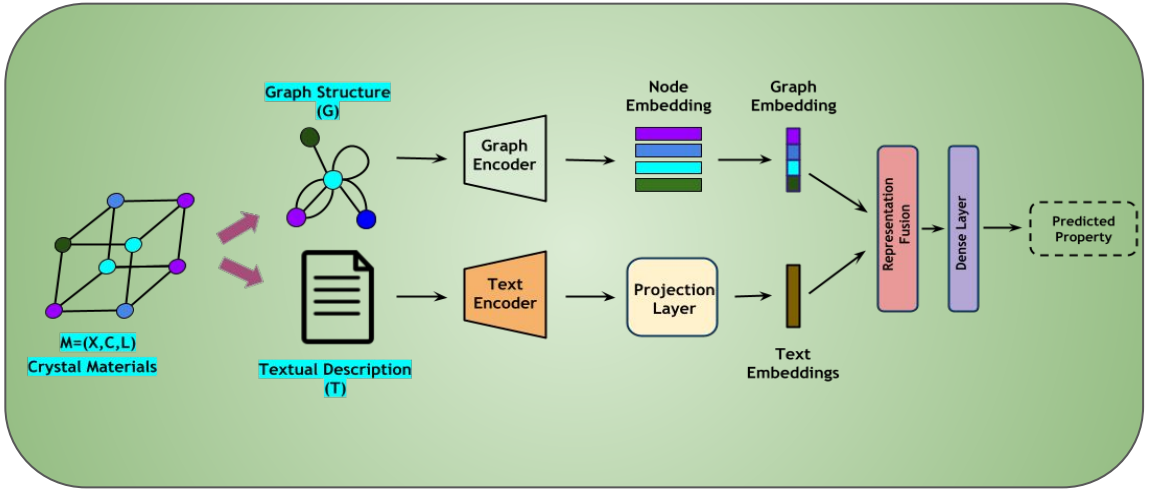
Conclusion

- In this work, we present a novel but simple **pre-trained GNN framework**, CrysGNN, for crystalline materials.
- Captures both **local chemical** and **global structural semantics** of crystal graphs, using node and graph level pre-training respectively
- We curate a new **large untagged crystal dataset** with **800K crystal graphs** to pretrain CrysGNN. We will release the pre-trained model along with the large dataset for the community.
- We **distill important knowledge** from CrysGNN and **inject** it into different state of the art property predictors and **enhance their performance**. We believe this approach can have applications in other domains too.
- Extensive experiments show its superiority over conventional fine-tune models.

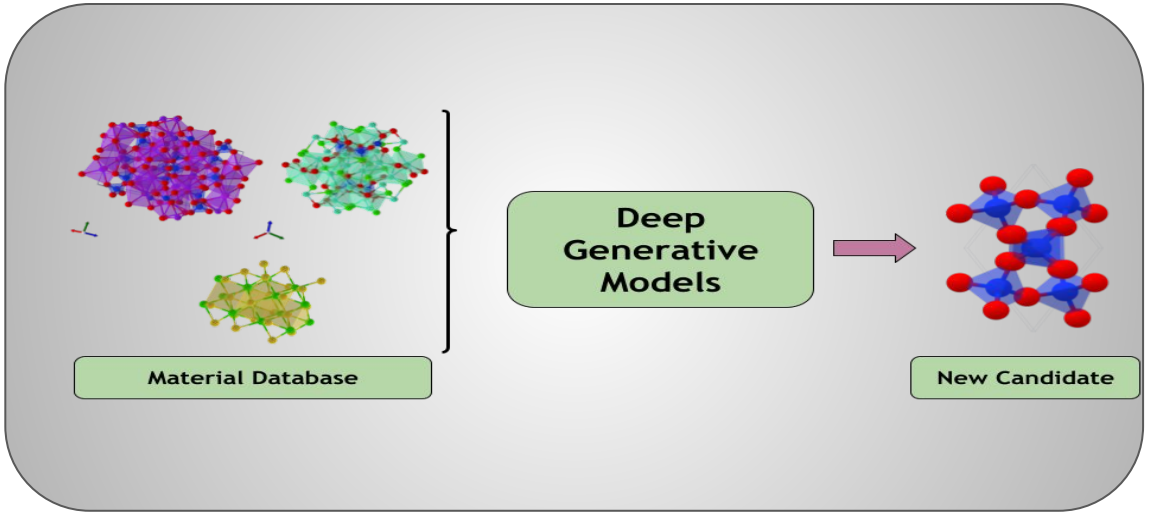
Arxiv : https://arxiv.org/abs/2301.05852
Github Repo for CrysGNN : https://github.com/kdmsit/crysgnn
Github Repo for 800K Dataset : https://github.com/kdmsit/crystal_untagged_800K

Future Work

- Multi-modal Representation.

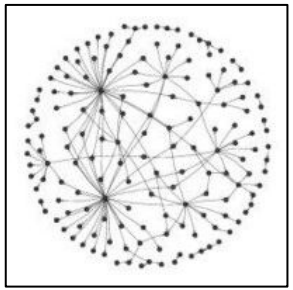


- Discovery of new materials



Thank You for Listening

Any Questions?



Complex Network Research Group (CNeRG) : @cnerg

PMRF
Prime Minister's Research Fellowship