Pretraining Strategies for Structure Agnostic Material Property Prediction

Hongshuo Huang,[†,¶] Rishikesh Magar,[‡,¶] and Amir Barati Farimani[*,‡,†]

[†]*Department of Material Science and Engineering, Carnegie Mellon University, Pittsburgh PA, USA 15213*

[‡]*Department of Mechanical Engineering, Carnegie Mellon University, Pittsburgh PA, USA 15213*

[¶]*Joint First Authorship*

E-mail: barati@cmu.edu

Abstract

In recent years, machine learning (ML), especially graph neural network (GNN) models, have been successfully used for fast and accurate prediction of material properties. However, most ML models rely on relaxed crystal structures to develop descriptors for accurate predictions. Generating these relaxed crystal structures can be expensive and time-consuming, thus requiring an additional processing step for models that rely on them. To address this challenge, structure-agnostic methods have been developed, which use fixed-length descriptors engineered based on human knowledge about the material. However, the fixed-length descriptors are often hand-engineered and require extensive domain knowledge, and generally are not used in the context of learnable models which are known to have a superior performance. Recent advancements have proposed learnable frameworks that can construct representations based on stoichiometry alone, allowing the flexibility of using deep learning frameworks as well as leveraging structure-agnostic learning. In this work, we propose three different

pretraining strategies that can be used to pretrain these structure-agnostic learnable frameworks to further improve the downstream material property prediction performance. We incorporate strategies such as self-supervised learning (SSL), fingerprint learning (FL), and multimodal learning (ML) and demonstrate their efficacy on downstream tasks for the Roost architecture, a popular structure-agnostic framework. Our results show significant improvement in small datasets and data efficiency in the larger datasets, underscoring the potential of our pretrain strategies that that effectively leverage unlabeled data for accurate material property prediction.

Introduction

Machine learning (ML) models have made significant progress in computational material science, both in material property prediction[1–10] and new methods for material property prediction done by Xie et al.[11] The growth of ML models in material science has been fueled by the increasing number of publicly available datasets and improved hardware capabilities.[12–14] Popular ML frameworks take the crystalline structure as input and leverage Graph Neural Networks (GNNs) to construct representations that can be used for property prediction. In these frameworks, the crystal structure information like 3D coordinates is required to construct a graph of the crystalline material.[9,15–18] The general idea is to consider the atoms as the nodes and capture the interactions between them using edges. This structure captures the interactions in the crystalline material, and the models often take optimized structures that are generated via simulations or experiments. Despite the large availability of crystal structure in public repositories such as Materials Project[19] and ICSD,[19] it only represents a fraction of the chemical space of materials. Generating the crystalline structures for all materials in the vast materials space can be a time-consuming process. This has motivated researchers to develop methods that do not require the structure of the material, for crystals that do not have a well defined structure beforehand. These structure agnostic methods can possibly be used for high throughput screening of materials with desired properties. The general approach to develop these structure agnostic models is using fixed length descriptors that encode the chemical composition of the material. These fixed length descriptors can be used to construct a feature vector that captures the material's properties, which can be used to predict its behavior[20–22]Ḣowever, the drawback of this approach is these fixed length descriptors need to be handcrafted and require considerable domain knowledge and expertise. Recently multiple approaches leveraging structure agnostic representations for material property predictions have been developed.[17,23–26] In this work, we focus on the Representation Learning from Stoichiometry (Roost) framework proposed by Goodall et al.[1] The Roost model takes as input the crystal formula and constructs an graph based repre-

sentation to develop a learnable framework. The Roost architecture is able to predict the material properties with a reasonable accuracy using only the stoichiometric data. In this work, we utilize the Roost model and propose three different pretraining strategies to improve the performance of the framework. Our pretraining strategies include 1.) Self Supervised Learning(SSL), 2.) Fingerprint Learning(FL) and 3.) Multimodal Learning(MML). After pretraining the Roost model with the 3 pretraining strategies we observe performance gains in multiple material property prediction tasks(Figure 1). The new generative model based on diffusion[27] is called CDVAE.

For our first strategy, we propose the Self-Supervised Learning approach(SSL) for pretraining the Roost encoder. In recent years, SSL frameworks[28–36] have been successfully utilized in computer vision and natural language processing tasks. The successful application of SSL has spurred many works in molecular machine learning[37–39] and material science.[3,40] Drawing upon the successful strategies of SSL employed in structure-based material property prediction,[3,7] we propose a framework for structure-agnostic SSL using the Roost encoder for generating material representation. The core idea of SSL revolves around pretraining models without the reliance on explicitly labeled datasets by leveraging the intrinsic information present in unlabeled data as the training signal. This framework can especially be advantageous for structure-agnostic material property prediction tasks, where labeled data may be scarce, and complete structural characterization of material is sometimes unavailable. By adopting this approach, we address the challenges associated with limited labeled data and inaccessible full structural information. For FL strategy, we devise a simple methodology of predicting the Magpie fingerprint[20] using the Roost encoder, the core idea is that the pretrained model can learn the information captured by the fingerprint. Using such a strategy allows us build a Roost encoder that can retain the benefits of being a learnable framework and also capture information of a fixed descriptor like Magpie fingerprint. We also introduce a MML strategy in which we leverage the available characterized structure data and predict the embedding generated using a pretrained CGCNN[1] encoder from Crys-

tal Twins Framework.[3] Using such a strategy we are able to learn the structural information using our structure agnostic encoder. By incorporating these three strategies, we successfully enhance the performance of the Roost encoder on downstream tasks within the Matbench suite. Notably, we demonstrate improvements in most material property prediction tasks, highlighting the effectiveness and potential of our proposed pretraining strategies.

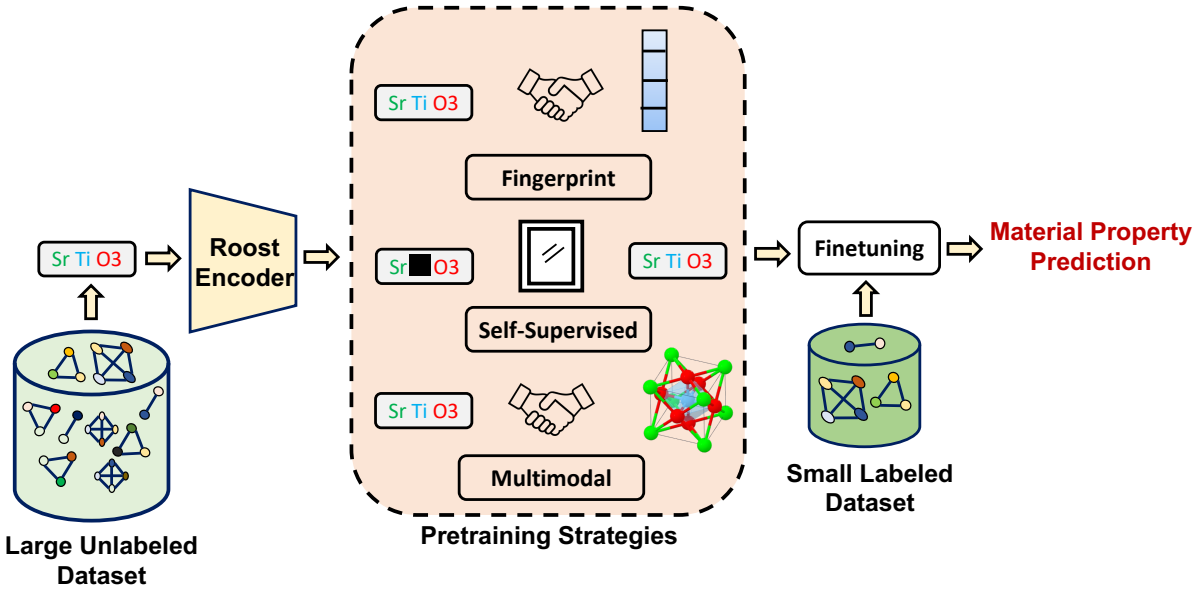


Figure 1: The framework for all the proposed pretraining strategies. We use the Roost encoder to demonstrate the effectiveness of the pretraining strategies for material property prediction tasks. We propose three strategies 1.) Self-Supervised Learning 2.) Fingerprint Learning and 3.) Multimodal Learning. Using these strategies we pretrain the Roost Encoder and finetune the model on different datasets in the Matbench[41] suite. Using such pretraining strategies we are able to demonstrate improvements on downstream tasks.

References

(1) Xie, T.; Grossman, J. C. Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties. *Physical review letters* **2018**, *120*, 145301.

(2) Karamad, M.; Magar, R.; Shi, Y.; Siahrostami, S.; Gates, I. D.; Farimani, A. B. Orbital

graph convolutional neural network for material property prediction. *Physical Review Materials* **2020**, *4*, 093801.

(3) Magar, R.; Wang, Y.; Farimani, A. Crystal twins: self-supervised learning for crystalline material property prediction. npj Comput. *Mater* **2022**, *8*, 231.

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

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

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

(7) Cao, Z.; Magar, R.; Wang, Y.; Farimani, A. B. MOFormer: Self-Supervised Transformer model for Metal-Organic Framework Property Prediction. *arXiv preprint arXiv:2210.14188* **2022**,

(8) Schütt, K. T.; Sauceda, H. E.; Kindermans, P.-J.; Tkatchenko, A.; Müller, K.-R. SchNet–A deep learning architecture for molecules and materials. *The Journal of Chemical Physics* **2018**, *148*, 241722.

(9) Gasteiger, J.; Groß, J.; Günnemann, S. Directional message passing for molecular graphs. *arXiv preprint arXiv:2003.03123* **2020**,

(10) Magar, R.; Farimani, A. B. Learning from mistakes: Sampling strategies to efficiently train machine learning models for material property prediction. *Computational Materials Science* **2023**, *224*, 112167.

(11) Xie, T.; Fu, X.; Ganea, O.-E.; Barzilay, R.; Jaakkola, T. S. Crystal Diffusion Variational Autoencoder for Periodic Material Generation. International Conference on Learning Representations. 2021.

(12) Chen, A.; Zhang, X.; Zhou, Z. Machine learning: accelerating materials development for energy storage and conversion. *InfoMat* **2020**, *2*, 553–576.

(13) Schmidt, J.; Marques, M. R.; Botti, S.; Marques, M. A. Recent advances and applications of machine learning in solid-state materials science. *npj Computational Materials* **2019**, *5*, 1–36.

(14) Choudhary, K.; DeCost, B.; Chen, C.; Jain, A.; Tavazza, F.; Cohn, R.; Park, C. W.; Choudhary, A.; Agrawal, A.; Billinge, S. J.; Holm, E.; Ong, S. P.; Wolverton, C. Recent advances and applications of deep learning methods in materials science. *npj Computational Materials* **2022**, *8*, 59.

(15) Park, C. W.; Wolverton, C. Developing an improved crystal graph convolutional neural network framework for accelerated materials discovery. *Physical Review Materials* **2020**, *4*, 063801.

(16) Yan, K.; Liu, Y.; Lin, Y.; Ji, S. Periodic Graph Transformers for Crystal Material Property Prediction. *arXiv preprint arXiv:2209.11807* **2022**,

(17) Ihalage, A.; Hao, Y. Formula Graph Self-Attention Network for Representation-Domain Independent Materials Discovery. *Advanced Science* **2022**, 2200164.

(18) Gasteiger, J.; Giri, S.; Margraf, J. T.; Günnemann, S. Fast and uncertainty-aware directional message passing for non-equilibrium molecules. **2020**,

(19) Belsky, A.; Hellenbrandt, M.; Karen, V. L.; Luksch, P. New developments in the Inorganic Crystal Structure Database (ICSD): accessibility in support of materials research and design. *Acta Crystallographica Section B: Structural Science* **2002**, *58*, 364–369.

(20) Ward, L.; Agrawal, A.; Choudhary, A.; Wolverton, C. A general-purpose machine learning framework for predicting properties of inorganic materials. *npj Computational Materials* **2016**, *2*, 16028.

(21) Botu, V.; Batra, R.; Chapman, J.; Ramprasad, R. Machine learning force fields: construction, validation, and outlook. *The Journal of Physical Chemistry C* **2017**, *121*, 511–522.

(22) Bartók, A. P.; Kondor, R.; Csányi, G. On representing chemical environments. *Physical Review B* **2013**, *87*, 184115.

(23) Wang, A. Y.-T.; Kauwe, S. K.; Murdock, R. J.; Sparks, T. D. Compositionally restricted attention-based network for materials property predictions. *Npj Computational Materials* **2021**, *7*, 77.

(24) Chen, C.; Ong, S. P. AtomSets as a hierarchical transfer learning framework for small and large materials datasets. *npj Computational Materials* **2021**, *7*, 173.

(25) Goodall, R. E.; Lee, A. A. Predicting materials properties without crystal structure: Deep representation learning from stoichiometry. *Nature Communications* **2020**, *11*, 1–9.

(26) Goodall, R. E.; Parackal, A. S.; Faber, F. A.; Armiento, R.; Lee, A. A. Rapid discovery of stable materials by coordinate-free coarse graining. *Science Advances* **2022**, *8*, eabn4117.

(27) Weng, L. What are diffusion models? *lilianweng.github.io* **2021**,

(28) Chen, T.; Kornblith, S.; Norouzi, M.; Hinton, G. A simple framework for contrastive learning of visual representations. International conference on machine learning. 2020; pp 1597–1607.

(29) Caron, M.; Misra, I.; Mairal, J.; Goyal, P.; Bojanowski, P.; Joulin, A. Unsupervised learning of visual features by contrasting cluster assignments. *arXiv preprint arXiv:2006.09882* **2020**,

(30) Zbontar, J.; Jing, L.; Misra, I.; LeCun, Y.; Deny, S. Barlow twins: Self-supervised learning via redundancy reduction. International Conference on Machine Learning. 2021; pp 12310–12320.

(31) He, K.; Fan, H.; Wu, Y.; Xie, S.; Girshick, R. Momentum contrast for unsupervised visual representation learning. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020; pp 9729–9738.

(32) Grill, J.-B.; Strub, F.; Altché, F.; Tallec, C.; Richemond, P. H.; Buchatskaya, E.; Doersch, C.; Pires, B. A.; Guo, Z. D.; Azar, M. G.; Piot, B.; Kavukcuoglu, K.; Munos, R.; Valko, M. Bootstrap your own latent: A new approach to self-supervised learning. *arXiv preprint arXiv:2006.07733* **2020**,

(33) Chen, X.; He, K. Exploring simple siamese representation learning. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021; pp 15750–15758.

(34) Lan, Z.; Chen, M.; Goodman, S.; Gimpel, K.; Sharma, P.; Soricut, R. ALBERT: A Lite BERT for Self-supervised Learning of Language Representations. International Conference on Learning Representations. 2019.

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

(36) Wu, J.; Wang, X.; Wang, W. Y. Self-Supervised Dialogue Learning. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. 2019; pp 3857–3867.

(37) Wang, Y.; Magar, R.; Liang, C.; Farimani, A. B. Improving Molecular Contrastive Learning via Faulty Negative Mitigation and Decomposed Fragment Contrast. *arXiv preprint arXiv:2202.09346* **2022**,

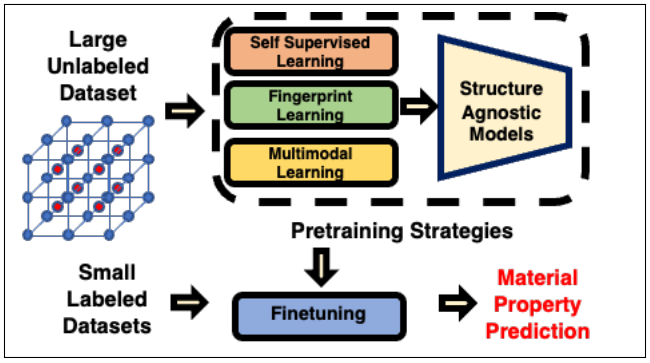
(38) Hu, W.; Liu, B.; Gomes, J.; Zitnik, M.; Liang, P.; Pande, V.; Leskovec, J. Strategies For Pre-training Graph Neural Networks. International Conference on Learning Representations (ICLR). 2020.

(39) Chithrananda, S.; Grand, G.; Ramsundar, B. ChemBERTa: Large-Scale Self-Supervised Pretraining for Molecular Property Prediction. *arXiv preprint arXiv:2010.09885* **2020**,

(40) Suzuki, Y.; Taniai, T.; Saito, K.; Ushiku, Y.; Ono, K. Self-supervised learning of materials concepts from crystal structures via deep neural networks. *Machine Learning: Science and Technology* **2022**, *3*, 045034.

(41) Dunn, A.; Wang, Q.; Ganose, A.; Dopp, D.; Jain, A. Benchmarking materials property prediction methods: the Matbench test set and Automatminer reference algorithm. *npj Computational Materials* **2020**, *6*, 1–10.

TOC Graphic


Large
Unlabeled
Dataset
Self Supervised
Learning
Fingerprint
Learning
Multimodal
Learning
Structure
Agnostic
Models
Pretraining Strategies
Small
Labeled
Datasets
Finetuning
Material
Property
Prediction