XRayPro: A self-supervised multimodal model for MOF application recommendations using PXRD and precursors

Sartaaj Takrim Khan
Chemical Engineering & Applied Chemistry
University of Toronto
Email: sartaaj.khan@mail.utoronto.ca

Seyed Mohamad Moosavi*
Chemical Engineering & Applied Chemistry
University of Toronto
Email: mohamad.moosavi@utoronto.ca

Abstract

In crystal structures, retrieving properties following synthesis is a time-consuming process. As crystal synthesis is often followed by a crystallinity assessment through the calculation of its powder x-ray diffraction (PXRD) pattern, this information (alongside its precursors) can be leveraged to directly predict the properties of these structures. To address this, we developed XRayPro, a model specifically tailored for metal-organic frameworks (MOFs), which can not only directly predict material properties, but also incorporates a recommendation system to suggest new applications - all done with only a PXRD and the MOF precursors. Additionally, self-supervised learning was done against a crystal graph convolutional neural network (CGCNN) to pretrain our multimodal model, leading to a significant improvement in the data efficiency of our model and enhancing its ability to learn chemistry-reliant and quantum-chemical properties. Our multimodal model not only predicts geometric, chemistry-reliant, and quantum-chemical properties, but the recommendation system has also shown potential in discovering new applications for certain MOFs, particularly in carbon capture and methane storage.

1 Introduction

Metal-organic frameworks (MOFs) are pivotal in material discovery because of their extensive surface area, porosity, and tunability. The traditional workflow from MOF synthesis to property analysis involves several intricate steps. Initially, MOFs are characterized using powder X-ray diffraction (PXRD) to assess crystallinity, followed by computational modeling and refinement for structure files such as CIF, which are then cleaned and added to databases like CoRE-MOF 2019 [1]. These procedures are complex and slow down the rapid discovery of MOF applications.

Previous studies have used PXRD for lattice analysis and microstructural characterization. However, PXRD struggles with predicting properties that depend on chemical interactions, such as low-pressure gas uptake [2, 3, 4, 5, 6, 7, 8, 9]. Recent advances like MOFormer and MOFTransformer have partially addressed these issues by incorporating precursors and atom-based embeddings to improve property prediction, although each has limitations regarding the immediacy of data post-synthesis and the scope of property prediction [10, 11].

We propose a multimodal model, XRayPro, that uses both PXRD patterns and molecular precursors (metal nodes and SMILES of linkers) to predict MOF properties and suggest applications directly. This model leverages the strengths of both inputs to provide a comprehensive understanding of the chemistry of MOFs, enhancing prediction accuracy with less computational complexity. XRayPro also incorporates self-supervised learning with CGCNN to enhance its efficiency, particularly in predicting chemistry-dependent properties such as low-pressure carbon dioxide uptake [12].

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

2 Results and Discussion

There are two primary areas of focus: model evaluation on a variety of properties (geometric, chemistry-reliant and quantum-chemical) and the evaluation of the built-in recommendation system to discover new application of MOFs previously synthesized. Furthermore, the data efficiency of the model will be discussed for a chemistry-reliant property - especially at low data regimes.

2.1 Model

As we are working with two different representations (numerical for PXRD and text for SMILES), there are two channels for the model: a convolutional neural network (CNN) and a transformer. The CNN, inspired by the work done by Chitturi et al. (2021) [13], can fully embed the PXRD. To successfully return an embedding of the precursors text string, an encoder and tokenizer were built on top of a transformer model (inspired by Cao et al. (2022))[10, 14]. Further details of the model can be found in the appendix. The embeddings coming out of the CNN and transformer are concatenated and fed into a projector, in which Barlow-Twin [15, 16] self-supervised learning is done between our model and a crystal graph convolutional neural network (CGCNN) [12] in an attempt to learn the chemistry composition and the local environments of the MOF structure better.

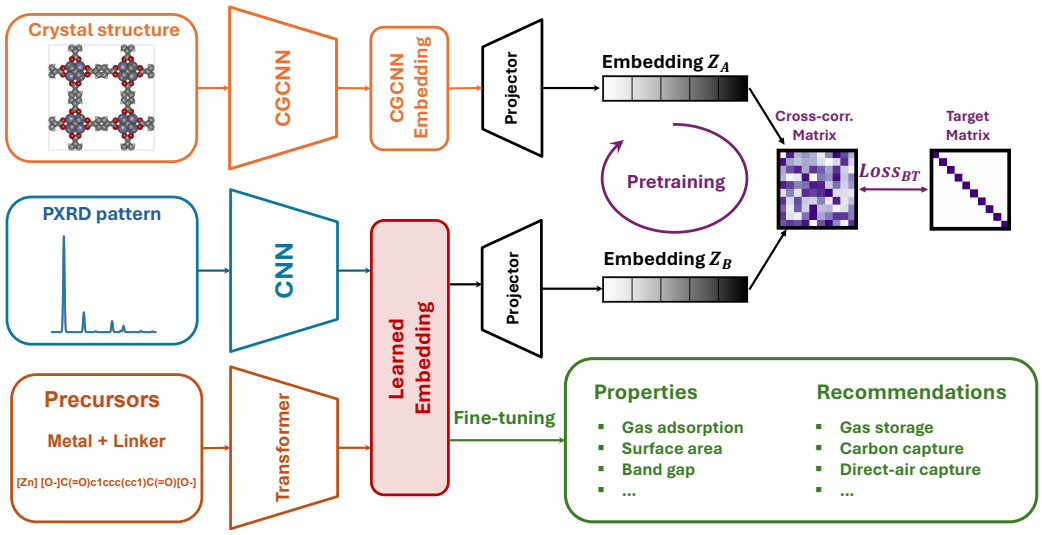

Figure 1: Overview of model architecture, recommendation system, and self-supervised pretraining method described above.

2.2 Regression results

For a comprehensive evaluation of our model, the geometry-reliant properties (gas uptake at high pressure - HP, accessible surface area, crystal density), chemistry-reliant properties (carbon dioxide uptake at low pressure - LP) and quantum-chemical (band gap) were predicted using our model. Data for these properties are extracted from CoRE-2019, QMOF and hMOF databases.[17, 18, 19, 1, 20, 21] Figure 2 showcases the model performance on ranking materials based on these properties (with the reported Spearman Rank Correlation Coefficient – SRCC), alongside comparisons to a descriptor-based model [17], a transformer-based model that accepts precursors only [10] and a crystal graph convolutional neural network (CGCNN).[12]

Our model demonstrates strong performance across both geometric and chemistry-related properties. Notably, it outperforms the CGCNN and models only accepting precursors for geometric property predictions and performs similarly for chemistry-reliant property predictions. This improvement stems from incorporating PXRD data, which captures the MOF's global structure, unlike CGCNN's local focus or precursor-based chemistry insights. Our model also approaches the descriptor-based

model's performance on properties such as hydrogen capacity and xenon uptake at HP, though the descriptor model excels in surface area and density due to using these descriptors as inputs. For chemistry-driven and quantum-chemical properties, our model remains competitive with other approaches.

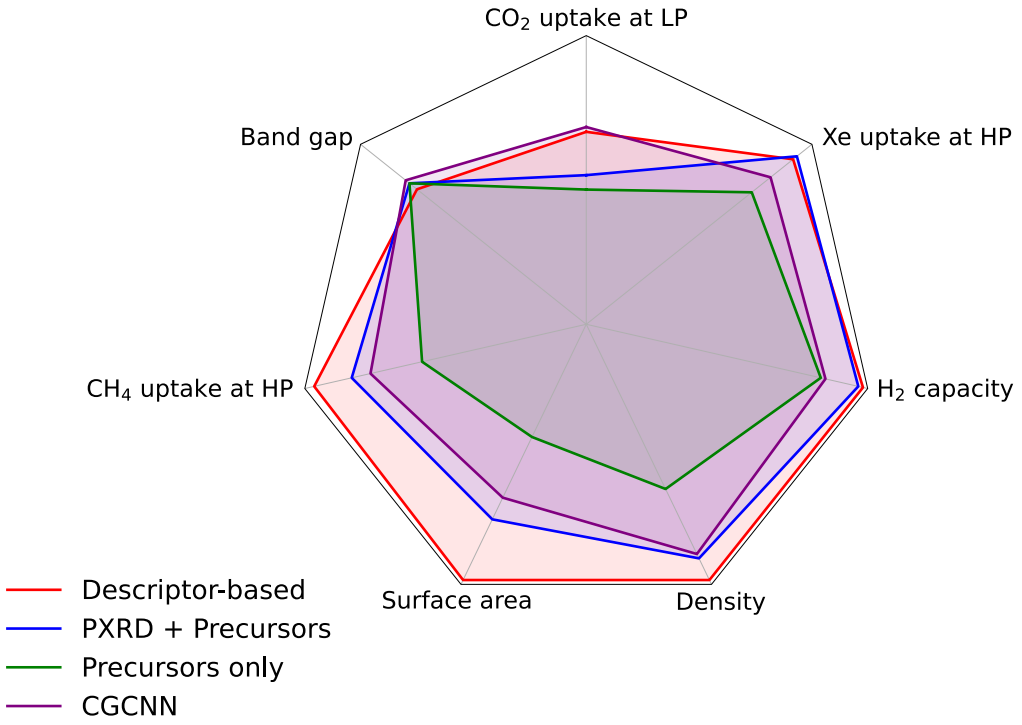

Figure 2: Summary of regression results (for our model, XRayPro, that accepts PXRD + precursors) and comparisons made to a descriptor-based ML model, a state-of-the-art transformer model accepting only precursors and a crystal graph convolutional neural network (CGCNN).[17, 10, 22, 12] This was evaluated across geometric, chemistry-reliant and quantum-chemical properties across multiple databases such as CoRE-2019, QMOF and hMOF.[1, 19, 21, 17] These values are all Spearman Rank Correlation Coefficient (SRCC).

2.3 Limitations

There are aspects to our work that need to be addressed. While the chemistry-reliant property predictions for BW20K, ARABG and QMOF were acceptable, predicting this for CoRE-2019 was proven to be challenging. While the reason for this has not been confirmed, one possible reason can be due to the large diversity of metal chemistry - CoRE has underrepresented metals such as Sr which have very different electrochemical properties such as a larger atomic radius and weaker affinity for carbon dioxide molecules in comparison to a metal node that is represented well in the database like Zn. Furthermore, while the precursors give some information about the metal and organic chemistry, it does not give a good amount of information about the local environment of the MOF - which was somewhat mitigated through pretraining against a CGCNN. However, that is outside the scope of the work.

3 Conclusion

Our multimodal model is able to predict geometry, chemistry-reliant and quantum chemical properties with just a PXRD pattern and the precursors. Through pretraining via the self-supervised pipeline against a CGCNN, the model also has a deeper understanding of the local environment of the MOF, which is crucial for the chemistry-reliant properties that the PXRD and precursors are unable to

showcase sufficiently alone. The key takeaway is that this model enables researchers to bypass the labor-intensive steps of MOF synthesis and property calculations by directly predicting properties and generating comprehensive recommendations using only a PXRD and its precursors. This advancement promises to accelerate both material discovery and material application discovery, offering a cost-effective and efficient tool for the field.

References

[1] Yongchul G. Chung et al. "Advances, Updates, and Analytics for the Computation-Ready, Experimental Metal–Organic Framework Database: CoRE MOF 2019". In: *Journal of Chemical Engineering Data* 64.12 (2019), pp. 5985–5998. DOI: 10.1021/acs.jced.9b00835.

[2] Vasile-Adrian Surdu and Romuald Győrgy. "X-ray Diffraction Data Analysis by Machine Learning Methods—A Review". In: *Applied Sciences* 13.17 (2023). ISSN: 2076-3417. DOI: 10.3390/app13179992. URL: https://www.mdpi.com/2076-3417/13/17/9992.

[3] Felipe Oviedo et al. "Fast and interpretable classification of small X-ray diffraction datasets using data augmentation and deep neural networks". In: *npj Computational Materials* 5 (2019), p. 60. DOI: 10.1038/s41524-019-0196-x. URL: https://www.nature.com/articles/s41524-019-0196-x.

[4] Keishu Utimula et al. "Machine-Learning Clustering Technique Applied to Powder X-Ray Diffraction Patterns to Distinguish Compositions of ThMn12-Type Alloys". In: *Advanced Theory and Simulations* 3.7 (2020), p. 2000039. DOI: 10.1002/adts.202000039. URL: https://onlinelibrary.wiley.com/doi/full/10.1002/adts.202000039.

[5] A Boulle and A Debelle. "Convolutional neural network analysis of x-ray diffraction data: strain profile retrieval in ion beam modified materials". In: *Machine Learning Science and Technology* 4.2 (2023). DOI: 10.1088/2632-2153/acab4c. URL: https://iopscience.iop.org/article/10.1088/2632-2153/acab4c/meta.

[6] Urko Petralanda et al. "Smart Nanomaterials: Lessons from Nature and Directions to Reach the Next Level". In: *Advanced Materials* 36.15 (2023), p. 2203879. DOI: https://doi.org/10.1002/adma.202203879. URL: https://onlinelibrary.wiley.com/doi/full/10.1002/adma.202203879.

[7] Kajetan A. Arkus et al. "Exploring the potential of framework topologies for guest separation using a machine learning approach". In: *Inorganic Chemistry Frontiers* 8 (2021), pp. 1463–1474. DOI: 10.1039/D0QI01513J. URL: https://pubs.rsc.org/en/content/articlehtml/2021/qi/d0qi01513j.

[8] Hongyu Wang et al. "A Machine Learning Study of MOF Stability in Aqueous Solutions". In: *Scientific Reports* 10.1 (2020), p. 19239. DOI: 10.1038/s41598-020-77474-4. URL: https://www.nature.com/articles/s41598-020-77474-4.

[9] Yosuke Suzuki et al. "Symmetry prediction and knowledge discovery from X-ray diffraction patterns using an interpretable machine learning approach". In: *Scientific Reports* 10 (2020), p. 21790. DOI: 10.1038/s41598-020-77474-4. URL: https://www.nature.com/articles/s41598-020-77474-4.

[10] Zhonglin Cao et al. *MOFormer: Self-Supervised Transformer Model for Metal–Organic Framework Property Prediction*. https://pubs.acs.org/doi/10.1021/jacs.2c11420. [Accessed 14-08-2024]. 2023.

[11] Yeonghun Kang et al. "A multi-modal pre-training transformer for universal transfer learning in metal–organic frameworks". In: *Nature Machine Intelligence* 5 (2023), pp. 309–318.

[12] Tian Xie and Jeffrey C Grossman. "Crystal Graph Convolutional Neural Networks for an Accurate and Interpretable Prediction of Material Properties". In: *arXiv preprint arXiv:1710.10324* (2017). URL: https://arxiv.org/abs/1710.10324.

[13] Sathya R. Chitturi et al. "Automated prediction of lattice parameters from X-ray powder diffraction patterns". In: *Journal of Applied Crystallography* 54.6 (2021), pp. 1799–1810. DOI: 10.1107/S1600576721010840. URL: https://journals.iucr.org/j/issues/2021/06/00/vb5020/index.html.

[14] Ashish Vaswani et al. "Attention is All You Need". In: *arXiv preprint arXiv:1706.03762* (2017). URL: https://arxiv.org/abs/1706.03762.

[15] Jure Zbontar et al. "Barlow Twins: Self-Supervised Learning via Redundancy Reduction". In: *arXiv preprint arXiv:2103.03230* (2021).

[16] Rishikesh Magar, Yuyang Wang, and Amir Barati Farimani. “Crystal twins: self-supervised learning for crystalline material property prediction”. In: *npj Computational Materials* 8.1 (2022), p. 231.

[17] Seyed Mohamad Moosavi et al. “Understanding the diversity of the metal-organic framework ecosystem”. In: *Nature Communications* 11 (2020), p. 4068. DOI: 10.1038/s41467-020-17755-8.

[18] N. Scott Bobbitt et al. “Title of the Article”. In: *Journal of Chemical Engineering Data* 68.2 (2023). DOI: 10.1021/acs.jced.2c00583.

[19] Andrew S. Rosen et al. “Machine learning the quantum-chemical properties of metal–organic frameworks for accelerated materials discovery”. In: *Matter* 4.5 (2021), pp. 1578–1597. DOI: 10.1016/j.matt.2021.02.017. URL: https://www.cell.com/matter/fulltext/S2590-2385(21)00070-9.

[20] Peter G. Boyd and Tom K. Woo. “A generalized method for constructing hypothetical nanoporous materials of any net topology from graph theory”. In: *CrystEngComm* 18 (2016), pp. 3777–3792. DOI: 10.1039/C6CE00407E. URL: https://pubs.rsc.org/en/content/articlehtml/2016/ce/c6ce00407e.

[21] Christopher E Wilmer et al. “Large-scale screening of hypothetical metal–organic frameworks”. In: *Nature Chemistry* 4 (2012), pp. 83–89. DOI: 10.1038/nchem.1192.

[22] Benjamin J. Bucior et al. “Identification Schemes for Metal–Organic Frameworks To Enable Rapid Search and Cheminformatics Analysis”. In: *Crystal Growth Design* 19.11 (2019), pp. 6682–6697. DOI: 10.1021/acs.cgd.9b01050. URL: https://pubs.acs.org/doi/10.1021/acs.cgd.9b01050.

[23] Peter G Boyd et al. “Data-driven design of metal–organic frameworks for wet flue gas CO$_2$ capture”. In: *Nature* 576.7786 (2019), pp. 253–256. DOI: 10.1038/s41586-019-1798-7.

[24] Shyue Ping Ong et al. “Python Materials Genomics (pymatgen): A Robust, Open-Source Python Library for Materials Analysis”. In: *Computational Materials Science* 68 (2013), pp. 314–319. DOI: 10.1016/j.commatsci.2012.10.028.

[25] Marc De Graef and Michael E. McHenry. *Structure of Materials: An Introduction to Crystallography, Diffraction, and Symmetry*. Cambridge, UK: Cambridge University Press, 2007. ISBN: 9780521651516.