Generative model for enhancing reticular material discovery

Theo Jaffrelot Inizan[a], Aaron Kaplan[b], Sherry Yang[c], Yen-hsu Lin[d], Mona Abdelgaid[a], Jian Yin[d], Zhiling Zheng[d], Saber Mirzaei[d], Ali H. Alawadhi[d], Ekin D. Cubuk[c], Christian Borgs[a], Jennifer T. Chayes[a], Kristin A. Persson[b], Omar M. Yaghi[d]

[a] *Bakar Institute of Digital Materials for the Planet, College of Computing, Data Science, and Society, University of California, Berkeley, CA, USA.*

[b] *Molecular Foundry, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA*

[c] *Google DeepMind, USA*

[d] *Department of Chemistry, University of California, Berkeley, CA, USA*

* Presenting author

1. Introduction

The discovery of synthesizable materials is a one of the main challenge in chemistry and condensed matter physics. While computational models such as generative models have enabled the generation of hundreds of thousands of energetically stable inorganic materials,[1, 2] only one have been synthesized.[3] Metal-Organic Frameworks (MOFs) [4] are no exception and present an even greater challenge due to their large combinatorial chemistry space, arising from the combination of organic linkers, inorganic nodes and topologies, making their synthesis highly complex. Due to their high-porosity and tunable building blocks, MOFs have found many applications such as direct air carbon capture.[5] In this context, generative AI models can play a pivotal role in accelerating the discovery of experimentally accessible MOFs.

In this work we developed an AI-driven framework that combine several generative models, quantum mechanical computations, Monte Carlo simulation as well as synthesizability-feasibility rules to generate experimental MOFs. We validated our methodology by high throughput synthesis, successfully synthesizing a set of purely AI-generated MOFs, the largest number of synthesized materials obtained by generative models to date.

2. Substantial Section

2.1 Novelty and Methodology

We developed a framework based on a combination of generative models, a fine-tuned large language model that generate input to a chemical formula-conditioned diffusion model. The diffusion model takes inspiration of models used image generation and was tuned to tackle large MOF structures with more than 256 atoms. The model was trained on computational MOF databases as well as all experimentally synthesized MOF, incorporating the latest structures curated from the CDCC database.

The generated structures are then processed through a multi-stage workflow integrating multi-level density functional theory (DFT) computations, machine learning interatomic potential,[6] Monte Carlo simulation and expert-designed synthesizable rules to filter out the most relevant MOF structures. The final structure are then geometry optimized at a very high level DFT and key stability-related properties such as formation energy are computed. This approach resulted in the generation of hundred of thousands of MOFs, far exceeding the number of MOF historically synthesized, and a similar number of organic linkers never reported before that escaped prior human intuition. Finally a key point in our method was to includes experimental constraints, proposed by human-expert, like human-in-the-loop, ensuring relevance to real-world synthesis. Finally, we validated our methodology by successfully synthesizing several of them.

2.2 Comparison to Related Work

Our diffusion model architecture represente a MOF structure as point clouds allowing to target large systems such as MOFs, In contrast, other generative models designed for inorganic crystal generation such as UniMat[2] and MatterGen [3] cannot be directly applied to MOFs due to their size.

Several works have developed diffusion model for MOF but none have implemented all-atom diffusion. For example MOFDiff [7] employs a coarse-grained approach, by representing the MOF into organic linkers and inorganic SBUs, another work focused on generating the organic linkers alone.[8]

In any case, our approach surpasses prior efforts, such as those in [1, 2, 3], by bridging the gap between computational prediction and experimental synthesis, by synthesizing MOF. Unlike "energy above the hull" metrics commonly used in inorganic materials, our workflow accounts for synthesis-constraint and multiple synthesizable-checking rules, enabling the curation and filtering of only experimentally relevant structures.

2.3 Results and Validation

The generated database includes over a hundred of thousands of MOFs and their corresponding organic linkers. After filtering, we ended-up with 80,000 novel and synthesizable organic linkers. Their synthesizability was assessed through the combination of several crietria, through the Allchemy software, scoring functions, machine learning-predicted synthesizability and expert evaluations from experimentalists. From these organic linkers, we used high-throughput synthesis and robotic automation to synthesize several MOFs.

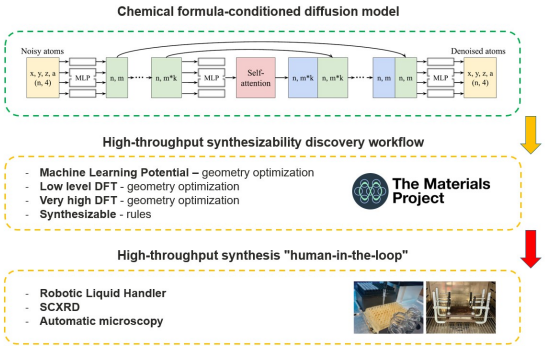

Fig. 1: Overview of the workflow. At the top the architecture of the chemical formula-conditioned diffusion model. In the middle the workflow developed within the Material Project infrastructure.[9] Bottom, high-throughput experimental synthesis.

3. Conclusion and Outlook

Our generative model and workflow significantly accelerate the discovery of synthesizable MOF. Beyond the computational generation of hundred of thousands of MOFs, the synthesis of several AI-designed MOFs show the promise of the use of generative models combined with human-expert rules into actual experimental discovery, accelerating the lenghtly trial-and-error process of current chemistry while giving idea to synthetic chemist for generating new compounds. This approach paves also the way for the development of new diffusion model architecture that could generate closely-to-experiment MOF structures.

Future directions include property-conditioned generation, as done in MatterGen[3], and the inrporation of additional experimental constraints to further enhance synthesizability. Our work highlights the transformative potential of generative AI in chemistry, bridging AI innovation and experimental validation.

References

[1] Amil Merchant, Simon Batzner, Samuel S. Schoenholz, Muratahan Aykol, Gowoon Cheon, and Ekin Dogus Cubuk. Scaling deep learning for materials discovery. *Nature*, 624(7990):80–85, December 2023.

[2] Sherry Yang, KwangHwan Cho, Amil Merchant, Pieter Abbeel, Dale Schuurmans, Igor Mordatch, and Ekin Dogus Cubuk. Scalable Diffusion for Materials Generation, June 2024. arXiv:2311.09235 [cs].

[3] Claudio Zeni, Robert Pinsler, Daniel Zügner, Andrew Fowler, Matthew Horton, Xiang Fu, Zilong Wang, Aliaksandra Shysheya, Jonathan Crabbé, Shoko Ueda, Roberto Sordillo, Lixin Sun, Jake Smith, Bichlien Nguyen, Hannes Schulz, Sarah Lewis, Chin-Wei Huang, Ziheng Lu, Yichi Zhou, Han Yang, Hongxia Hao, Jielan Li, Chunlei Yang, Wenjie Li, Ryota Tomioka, and Tian Xie. A generative model for inorganic materials design. *Nature*, pages 1–3, January 2025.

[4] Omar M. Yaghi, Michael O'Keeffe, Nathan W. Ockwig, Hee K. Chae, Mohamed Eddaoudi, and Jaheon Kim. Reticular synthesis and the design of new materials. *Nature*, 423(6941):705–714, June 2003. Number: 6941 Publisher: Nature Publishing Group.

[5] Zihui Zhou, Tianqiong Ma, Heyang Zhang, Saumil Chheda, Haozhe Li, Kaiyu Wang, Sebastian Ehrling, Raynald Giovine, Chuanshuai Li, Ali H. Alawadhi, Marwan M. Abduljawad, Majed O. Alawad, Laura Gagliardi, Joachim Sauer, and Omar M. Yaghi. Carbon dioxide capture from open air using covalent organic frameworks. *Nature*, 635(8037):96–101, November 2024. Publisher: Nature Publishing Group.

[6] Ilyes Batatia, Dávid Péter Kovács, Gregor N. C. Simm, Christoph Ortner, and Gábor Csányi. MACE: Higher Order Equivariant Message Passing Neural Networks for Fast and Accurate Force Fields, January 2023. arXiv:2206.07697 [stat].

[7] Xiang Fu, Tian Xie, Andrew S. Rosen, Tommi Jaakkola, and Jake Smith. MOFDiff: Coarse-grained Diffusion for Metal-Organic Framework Design, October 2023. arXiv:2310.10732 [physics].

[8] Hyun Park, Xiaoli Yan, Ruijie Zhu, Eliu A. Huerta, Santanu Chaudhuri, Donny Cooper, Ian Foster, and Emad Tajkhorshid. A generative artificial intelligence framework based on a molecular diffusion model for the design of metal-organic frameworks for carbon capture. *Communications Chemistry*, 7(1):1–18, February 2024.

[9] Anubhav Jain, Shyue Ping Ong, Geoffroy Hautier, Wei Chen, William Davidson Richards, Stephen Dacek, Shreyas Cholia, Dan Gunter, David Skinner, Gerbrand Ceder, and Kristin A. Persson. Commentary: The Materials Project: A materials genome approach to accelerating materials innovation. *APL Materials*, 1(1):011002, July 2013.