

Multi-modal conditional diffusion model using signed distance functions for metal-organic frameworks generation

Corresponding Author: Professor Jihan Kim

This file contains all reviewer reports in order by version, followed by all author rebuttals in order by version.

Version 0:

Reviewer comments:

Reviewer #1

(Remarks to the Author)

The paper introduces a generative model for metal organic frameworks (MOFs). The proposed MOFFUSION method combines several interesting technical ideas: (1) signed distance functions (SDFs) for representing MOF structures; (2) VQVAE for encoding and decoding of SDFs, and a latent diffusion model in the VQVAE latent space; (3) a variety of conditional generation capabilities/tasks through latent-space cross attention; (4) a decoder that decodes the topology and building block identities, from which a MOF assembly algorithm such as PORMAKE can be used to obtain the atomic MOF structures. The model demonstrates strong performance in generating diverse MOFs and generation under a variety of numeric, categorical, and some basic text conditions. This reviewer believes the technical ideas and the demonstrated capabilities presented in this paper are valuable to the community and recommends acceptance of this paper to Nature Communications.

The overall workflow is considerably complicated, but each module is well-motivated:

- SDFs as MOF representation is a novel contribution to this reviewer's knowledge, and is a clever idea to encode the key structural information of a MOF, without using the atomic representations of MOFs which are hard to generate very well directly, especially with complicated topology/building blocks. An assembly algorithm for recovering the MOF atomic structures from the SDF offload the difficult task (generating atom types/coordinates) to the more stable algorithm.

- The downside of an assembly algorithm like PORMAKE is perhaps its lack of account for the structural information of a MOF. The SDF representation completes well this drawback of assembly algorithms. The various conditions can be incorporated well with the cross-attention scheme.

- The authors construct a new database for the training and validation of their model. The new database seems structurally diverse, and could also be of use to future works. The conditional generation results are comprehensive and convincing. The supplementary information contains satisfying details of the proposed method.

- It would be great if the authors could include some discussion on the limitation of the method presented — e.g., when does the proposed method fail; what is the gap between these computationally designed MOF and making MOFs in the real world.

Questions:

- The embedding of the MOF SDF should ideally be periodicity-aware and translational invariant. It seems the VQ-VAE architecture is not explained in details in the paper/appendix. Can the authors discuss on whether the auto encoder respect symmetry properties of MOF SDFs and elaborate on the architecture in the paper?

(Remarks on code availability)

The codebase seems satisfactory.

Reviewer #2

(Remarks to the Author)

I read with interest your article. It combines a variety of ML methods, from diffusion modeling, to 3D structure prediction, to NLP, and text conditioning.

You present a nice set of exciting results. Please read the suggestions below to ensure your article is suitable for publication.

1. The framing of your work is poorly done. I could only understand that you validated the hydrogen capacity of your MOFs with GCMC until Section 5.2 of the Supplementary information. You need to include a couple of paragraphs of the introduction, with a clear, easy to follow figure, where you clearly state the aims and workflow of your work, from data production with PORMAKE and validation with LAMMPS, to assembly of MOFs, and characterization with GCMC. Else, you leave the reader with too many gaps and guesses about what you are doing.
2. Training data. You produced your training data with PORMAKE. A bit risky. You mentioned that you validated these structures with LAMMPS. How did you do that, what metrics did you use to ensure that the structures are stable and chemically consistent? How long were the LAMMPS simulations to ensure stability? Why would you do that if we have this massive dataset: <https://open-dac.github.io/>?
3. SDFs. Nice application. Did you check that the interatomic distance of your AI predicted MOFs adhere to experimental values provided by databases such as OCChemDB? There is no indication of such valuation in your article.
4. AI models. Really basic models, ResNet, U-Net. A bit underwhelming, it is a weak part of your work.
5. Open source code. Well done!
6. Unconditional generation result. I am a bit concerned that the generated structures are very similar to the training dataset. This indicates that your model is not capable of generalizing to new structures.
7. Multi-target conditioning. You demonstrated this feature using several properties. Can you more clearly correlate this capability to the design of high performing MOFs for hydrogen working capacity? What's the optimal metal node, pore size, geometry, etc., combination?
8. Text conditioning. If you are already using text conditioning, can you use this capability to screen your set of predicted MOFs to identify the optimal region of parameter space for the production of high performing MOFs?
9. Please, please do not push critical information about your work to supplementary materials. The reader needs to know what you do from data creation to validation, and the numerical methods you have used for such screening and validation tests.

Once you address these concerns, I am happy to review the paper again.

(Remarks on code availability)

Yes, the article is available at <https://github.com/parkjunkil/MOFFUSION>

It provides step by step guidance to install and use their code. It is easy to follow and use.

I was able to follow their instructions to install the code, and run it.

Excellent work!

Reviewer #3

(Remarks to the Author)

Park and coworkers present a MOFFUSION, a generative algorithm for MOFs. The approach has multiple innovations including the unconditional and conditional modes of the algorithm. The conditional can also have multiple conditions which can be numeric, categorical, or text data. The work also highlights the possibility of pore crafting: selecting empty regions of a structure and using that a specification to generate new ones. Overall, the tools has great potential for future studies looking for generative support of porous crystalline materials, not just MOFs. Some questions should be addressed prior to publication:

1. In many of the results, the authors show the resulting distributions of a target property requested. The resulting distributions can be very different depending on the value of the property. Some of the distributions are very sharp while others are very broad. Authors should expand on that behavior.
2. The authors present a lot of generation results, however, only the SI has any mention of how many are novel, unique, and/or valid. I think the main manuscript should have discussions on these observations. For example, does the multi-conditional generation result in any new structures outside of the training set? Similarly, how was novelty, uniqueness, and validity determined? How do the generated structures compare to those of the training set? Do the generated structures

need to be energy minimized or, in the cases where the generated structures are similar to those in the training set, how different are they? How do the unit cell parameters compare, how do the adsorption isotherms compare, how do simulated PXRDs compare?

3. In many instances the authors refer to the "accuracy" of the generation but it is unclear what is meant. It seems that each task would have its own measure of accuracy but it is unclear what the ground truth being compared is. For example, when requesting a target void fraction, what is the accuracy of the final generated structures?

(Remarks on code availability)

Appropriate code was provided to reproduce results and has enough instruction to install and use.

Reviewer #4

(Remarks to the Author)

Overview

Park et al. report a new approach for the "inverse design" of metal-organic frameworks.

The inverse design of materials, particularly MOFs, is of wide interest, and a range of commercial players are now also investigating this topic.

The authors use diffusion-based models and a classifier that returns the corresponding topologies and building blocks.

Major comments

- In spirit, the work bears similarities to approaches such as ZeoGAN from some of the same authors --- i.e., to use some form of voxelized representation of the structure. This tends to make it challenging to introduce chemistry (would, e.g., require the addition of many channels, which increases the dimensionality of the problem).

In the work at hand, the authors address this by adding a classifier that predicts topologies and building blocks based on SDFs. However, this leaves a few open questions:

- With a fixed set of building blocks and topologies, the system is limited to a fixed, enumerable set of output structures. In this case, is it fair to compare the model to models that can generate structures who can not be built from a predefined list of building blocks?

- What is the advantage of having the diffusion model instead of just directly predicting topology and building blocks? That is, what is the benefit of the "indirection" via the SDF?

Minor comments

- Given the set of building blocks is fixed

- I think a comparison to a GA (e.g. the one reported by the same authors) would make a lot of sense

- Is the comparison in terms of the diversity metrics (Tab 1) fair given that some of the other techniques go beyond a fixed set of building blocks

- The limitation of models being trained on different datasets should not only be in the SI, but also in the main text

- "the methodology of directly coordinating the atoms (or grains) yet presents challenges creating an intact structure due to the high structural complexity of MOFs."

- I am not sure that the study contains empirical proof for this claim

- The resolution of 32^3 will most likely not be enough for phenomena that requires exact atomic resolution (chirality, functional groups, etc. - as the low accuracies for edge predictions show). However, increasing the dimensionality will likely lead to computational challenges. It probably makes sense to more clearly acknowledge this limitation.

- While the pore-crafting technique is nice, it would be great to see experimental quantitative evidence that it systematically works

- Other relevant metrics for Tab 1 include uniqueness and distance to the training data

Code and reproducibility

The code is openly available and looks well structured and documented.

Limitations of expertise

I've been building machine learning models for MOFs. I do not have hands-on experience building diffusion models for MOFs. However, I am experienced in doing so for other materials.

(Remarks on code availability)

Version 1:

Reviewer comments:

Reviewer #1

(Remarks to the Author)

I thank the authors for replying to my comments and questions. I have read the updated draft and comments from other reviewers. My concerns are addressed and I recommend acceptance.

(Remarks on code availability)

the code release seems satisfactory.

Reviewer #2

(Remarks to the Author)

Thank you for addressing my comments and suggestions. Your manuscript is ready for publication. Great work!

(Remarks on code availability)

They follow best practices for code development and sharing. The code includes a demo, and step-by-step guidance to install and run training and inference software.

Reviewer #3

(Remarks to the Author)

All my comments were properly addressed.

(Remarks on code availability)

Reviewer #4

(Remarks to the Author)

Park et al.'s revised version of the manuscript clarified some points. However, some of my concerns remain, and I feel that addressing them would strengthen the manuscript.

I'm still very happy to see novel inverse design techniques for MOFs - however, I feel that it would help the manuscript to more clearly ablate what parts of the new workflow lead to which performance gains and to be even more explicit about the limitations.

- Text conditioned inverse design has been done before. Here a recent work (on Arxiv after the paper at hand) on using LLM-based agents for MOF linkers <https://arxiv.org/abs/2410.03963> but those LLM agents for materials design have been around. Also related, but earlier <https://chemrxiv.org/engage/api-gateway/chemrxiv/assets/orp/resource/item/65fbf03366c1381729744c2d/original/post-pretraining-large-language-model-enabled-reverse-design-of-mo-fs-for-hydrogen-storage.pdf>

For molecules, such an approach is relatively common by now (e.g. <https://openreview.net/pdf?id=FdUloEgBSE>)

- That is, the argument on "text conditioning" being novel can perhaps benefit from a bit more context

- "the first application of 3D modeling techniques for the generation of porous materials"

I am not sure if this is not a bit too broad. The authors themselves used different 3D techniques (e.g., energy grids) for the modeling of porous materials, such as ZeoDiff.

- Regarding my comment on going via the SDF: I was asking for evidence of the benefits of going via the SDF. I can see that it can perhaps enable things like "pore crafting" - but if I condition on a given pore size and window diameters it can be almost the same. And the question that is still open to me is if there is any quantitative benefit of going then via the SDF instead of directly predicting the relevant building blocks (i.e., directly go from the multi-modal conditioning inputs to the building blocks)

- Why am I so persistent on this? The use of SDFs is portrayed as one of the main advances in this work ("The high validity of MOFFUSION stems from its novel architecture and the benefits of SDFs, which provide a deep understanding of the geometric nature of MOFs."), but there is little direct evidence of their benefit. Yes, we see that the overall pipeline works, but there is no direct evidence that going via the SDFs is important for that. In addition, you add another indirection, and the model from SDF to building blocks also only has limited accuracy (80%, according to the authors). That is, if one can avoid multiplying errors that might lead to increased overall performance. In addition, the SDF will also lose some "chemical resolution" which might be essential for important applications like direct air capture (where not only the pore shape is important but also the chemical environment, Fig. S6d might be such a case). While I can conceptually see that the system generates MOFs and that they also fulfill some target characteristics the article gives, at this point, little signal on whether going via the SDFs is a more fruitful approach than going via energy grids, reinforcement learning on the building blocks (<https://pubs.rsc.org/en/content/articlelanding/2024/dd/d4dd00010b>) coarse-grained diffusion, etc. Distilling a clearer message on a point like this might make the article also more impactful in the long run.

- Another point I had been pedantic on in my first review is that MOFFUSION relies on a set of building blocks:

- For pore-crafting: How well would a lookup work (e.g. based on some zeo++ statistics, or persistent homology)

- How much do we gain for all the other properties compared to having a machine learning model for the working capacity or some other target property and simply predicting this property for all MOFs. I do not necessarily ask for new experiments on all of that -- but I feel that it currently is not clear to the reader that those enumeration-based approaches might also be an option and that they might even perform better in some instances.

(Remarks on code availability)

Version 2:

Reviewer comments:

Reviewer #4

(Remarks to the Author)

The authors addressed my remaining concerns and the manuscript now provides more useful context.

(Remarks on code availability)

Open Access This Peer Review File is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

In cases where reviewers are anonymous, credit should be given to 'Anonymous Referee' and the source.

The images or other third party material in this Peer Review File are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

To view a copy of this license, visit <https://creativecommons.org/licenses/by/4.0/>

Dear Editor,

We greatly appreciate the feedback as the comments made by the reviewers helped us improve the manuscript. Below are point-by-point responses to the comments. Thank you.

Sincerely,

Jihan Kim

In this point-by-point response to the Referees' comments:

1. Black color - comments from the Referees
2. Blue color - response to the comments
3. Green color - added in the manuscript with page number provided
(The page number is based on word file)

Reviewer #1

1. It would be great if the authors could include some discussion on the limitation of the method presented — e.g., when does the proposed method fail; what is the gap between these computationally designed MOF and making MOFs in the real world.

We appreciate the valuable comment provided by the reviewer. MOFFUSION was trained with hypothetical MOFs, which may create a gap between the generated structures and those experimentally reported. This gap is a primary challenge that the field of materials design using generative models is focusing on. In the discussion section of our original manuscript, we mentioned future directions for accounting for the synthesizability of materials. By utilizing MOFFUSION's versatile conditioning capabilities, there is potential to guide the generation process to selectively generate structures with high synthesizability. However, as the reviewer noted, the gap between the hypothetical structures and experimental structures is a clear limitation in most materials generation approaches using generative models, and we have added a new sentence to clarify this point. We appreciate the reviewer for the helpful comment.

(Page 26) A primary challenge in materials design using generative models is the gap between hypothetical structures and those exist in the real world. Guiding the generation process toward the structures likely to be synthesized is considered a practical solution, but assessing the synthesizability of structures in a virtual space is a complex problem.

2. The embedding of the MOF SDF should ideally be periodicity-aware and translational invariant. It seems the VQ-VAE architecture is not explained in details in the paper/appendix. Can the authors discuss on whether the auto encoder respect symmetry properties of MOF SDFs and elaborate on the architecture in the paper?

We appreciate the insightful feedback from the reviewer. We agree with the reviewer's opinion that our original manuscript lacks sufficient description of periodicity-awareness and invariance during data processing. Periodicity is a crucial factor in handling crystalline materials, and the MOFFUSION architecture was carefully designed to be periodicity-aware. For the data type, as described in section S1-1 of the Supplementary Information, we first proposed the concept of periodic SDF, which reflects the pore morphology of crystals while considering their periodic nature. Additionally, the convolutional layers within our VQ-VAE and U-Net architecture use periodic padding (also called 'circular padding'), which accounts for periodicity during processing. However, the explanation of periodic padding was not covered in our original manuscript. On the other hand, like many other generations tasks using SDFs, translational invariance was not explicitly considered in MOFFUSION. Although MOFFUSION achieved satisfactory performance without translational invariant, this can be addressed further using data augmentation techniques as described in Kim et al.[1] We agree that these aspects will be informative for our readers, so we have added a phrase explaining the above discussion. Additionally, we have added further details about the source from which the VQ-VAE architecture used in MOFFUSION was adopted. We greatly appreciate the reviewer's help in developing our manuscript.

[1] Kim, B., Lee, S. & Kim, J. Inverse design of porous materials using artificial neural networks. *Science advances* **6**, eaax9324 (2020)

(Page 9) Notably, within the convolutional layers of the VQ-VAE architecture, periodic padding (or circular padding) was used to account for the periodicity of MOFs during the processing.

(Page 26) Furthermore, explicit consideration of translational and/or rotational invariance could be an option worth considering for enhancing the overall performance of the architecture.

(Page 6, Supplementary Information) The architecture of VQ-VAE was adopted from Aaron et al., who first proposed the architecture.

(Page 9, Supplementary Information) As with the VQ-VAE architecture, circular padding option was used for convolutional layers to account for periodicity of crystalline materials during data processing.

Reviewer #2

1. The framing of your work is poorly done. I could only understand that you validated the hydrogen capacity of your MOFs with GCMC until Section 5.2 of the Supplementary information. You need to include a couple of paragraphs of the introduction, with a clear, easy to follow figure, where you clearly state the aims and workflow of your work, from data production with PORMAKE and validation with LAMMPS, to assembly of MOFs, and characterization with GCMC. Else, you leave the reader with too many gaps and guesses about what you are doing.

We appreciate the insightful comment provided by the reviewer. We agree with the reviewer that some parts of the organization of our original manuscript is not reader-friendly, and readers (especially who lack sufficient domain knowledge) may find it difficult to grasp the overall flow. Especially, as noted by the reviewer, the introduction section didn't provide enough information about the overall organization of our work. Therefore, we have added a more detailed explanation regarding the overall picture, from data preparation to the characterization of the generated structures, in the introduction section. We anticipate that this additional information will greatly help readers understand the flow of the subsequent Results section. We appreciate the reviewer for helping us in improving our manuscript.

(Page 4) A database composed of 247,742 hypothetical MOF structures was used for the training, which were generated using PORMAKE (a MOF construction software) and further geometry optimized using LAMMPS software.

(Page 5) As a representative numeric data, MOFFUSION was conditioned on hydrogen working capacity and its capability to generate structures with the desired hydrogen capacity was validated through Grand Canonical Monte Carlo (GCMC) simulations. MOFFUSION also demonstrated outstanding performance in handling categorical data, achieving high accuracy in generating MOFs with targeted geometric topologies. Furthermore, its capability to condition with text data was validated by generating MOFs with desired characteristics based on input prompts. Lastly, MOFFUSION's ability to account for multiple target properties during conditional generation was confirmed, and its inherent capability of generating MOFs with user-tailored pore shapes was further demonstrated.

2. Training data. You produced your training data with PORMAKE. A bit risky. You mentioned that you validated these structures with LAMMPS. How did you do that, what metrics did you use to ensure that the structures are stable and chemically consistent? How long were the LAMMPS simulations to ensure stability? Why would you do that if we have this massive dataset: <https://open-dac.github.io/>?

We appreciate the valuable feedback from the reviewer. The training dataset for MOFFUSION was prepared using PORMAKE software, an actively used tool for generating hypothetical MOFs. However, PORMAKE doesn't guarantee that the generated structures will exist in their energetically most stable state, and occasionally, a small portion of them suffers from excessive steric hindrance. Therefore, subsequent geometry optimization is often performed for the generated structures using molecular dynamics simulations.

In the case of MOFFUSION, for the preparation of training dataset and for the further processing of the generated structures, we performed geometry optimization as follow; energy minimization, relaxation of simulation box, and energy minimization again. Energy minimization was conducted with stopping tolerance for energy (E_{tol}) of $1e-6$ and stopping tolerance for force (F_{tol}) of $1e-8$ (both are unitless). This setting cannot be said to be highly strict, but it was a reasonable choice in handling large number of structures. Notably, this is a same setting with the precedent work [1], which generated a large dataset of the hypothetical MOFs and trained a property prediction model with high accuracy.

The validity of the generated structures was assessed using the following metrics. First, the structure must be able to be generated using PORMAKE. This includes: (i) the connection points of building blocks should be compatible with the topology, and (ii) building blocks need to be placed to a given site within a certain root mean square distance. (More details on the PORMAKE generation scheme can be found in [2].) Second, for the generated structures, geometry optimization is conducted and those structures that are chemically non-feasible or suffer from excessive steric hindrance (which might lead to either failure in force field assignment or simulation convergence) are filtered out.

Utilizing a DFT-optimized structures, such as those in openDAC, is also a promising approach for training generative models. Since the structures in the training set are accurately optimized, further geometry optimization process through LAMMPS can be omitted. One concern is the size of the dataset, where the size of the DFT-based database might be small to train generative models effectively. (To our knowledge, openDAC comprises approximately 8,000 MOFs from

CoRE MOF, where 247,742 MOFs were used for the training of MOFFUSION.) However, the existence of the DFT-optimized structures is promising. One approach might be to pre-train a model with a large amount of hypothetical data and then fine-tune it with DFT-optimized structures for the generation of more realistic MOFs.

We strongly agree with the reviewer's opinion that our original manuscript lacked detailed descriptions of the LAMMPS optimization step and validity metrics. We have added more information in our revised manuscript. We greatly appreciate the reviewer's thoughtful feedback.

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

[2] Lee *et al.* Computational Screening of Trillions of Metal-Organic Frameworks for High-Performance Methane Storage. *ACS Appl. Mater. Interfaces* 13, 2-, 23647-23654 (2021)

(Page 11) The structure was deemed valid if it could be generated using PORMAKE (which accounts for building block compatibility and steric hindrance), and if it successfully underwent geometry optimization without failure.

(Page 18, Supplementary Information) Structure validity is a metric widely used for assessing the quality of the generated structures through generative models. As described in the main text, the validity of the structures generated through MOFFUSION was assessed using the following metric. First, the structure must be able to be generated using PORMAKE. This includes: (i) the connection points of building blocks should be compatible with the topology, and (ii) building blocks need to be placed to a given site within a certain root mean square distance. (More details on the PORMAKE generation scheme can be found in Lee *et al.*) Second, for the generated structures, geometry optimization is conducted to filter out those structures that are chemically non-feasible or suffer from excessive steric hindrance (which might lead to either failure in force field assignment or convergence). It is our opinion that our metric is more stringent compared to some precedent works, which considered the structure valid only if there was no atom overlap.

(Page 28, Supplementary Information)

Section S5-3. Details on Geometry Optimization

The geometry optimization of the generated structures and those comprising training dataset was conducted using LAMMPS software as follow; energy minimization, relaxation of simulation box, and energy minimization again. Energy minimization was conducted with stopping tolerance for energy (E_{tol}) of $1e-6$ and stopping tolerance for force (F_{tol}) of $1e-8$ (both are unitless). During the relaxation of simulation box, the dimensions of the simulation cell was gradually adjusted to achieve an isotropic pressure of 0.0 with a maximum change in the cell size of 0.001 per step.

3. SDFs. Nice application. Did you check that the interatomic distance of your AI predicted MOFs adhere to experimental values provided by databases such as OChemDB? There is no indication of such valuation in your article.

We appreciate the valuable feedback from the reviewer. In evaluating the MOFFUSION generated structures, we initially compared their similarity only to the structures comprising the trainset (Table 1), without comparing them to experimental structures in other public databases. We have now compared the generated structures from MOFFUSION with experimental crystal structures in OChemDB. Specifically, among various interatomic distances, we focused on metal-oxygen bond distances, which are considered one of the most important bond types within MOFs. We examined Zn, Cu and Co atoms, which are among the most common metals in MOFs. As shown in the Figure below, the distributions of all three bond types in MOFFUSION generated structures exhibited similar profile to those in OChemDB, validating the reliability of the output structures and the entire workflow. Since this result would be informative for readers, we have added this discussion to our revised manuscript. We appreciate the reviewer for the helpful comment in improving our manuscript.

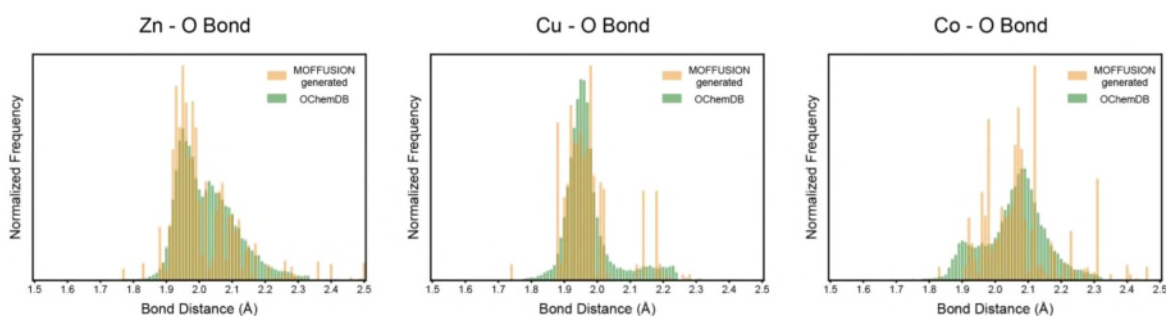


Fig. A. Distribution of metal-oxygen bond distances in Open Chemistry Database (OCChemDB) and MOFFUSION generated structures. The distributions of metal-oxygen bond distances for Zn, Cu, and Co atoms in structures obtained from 10,000 generation attempts by MOFFUSION (orange) are compared with the structures from the OCChemDB database (green).

(Page 12) In addition, the interatomic distances of the generated structures were compared with the experimental structures comprising the Open Chemistry Database (OCChemDB). The metal-oxygen bond distances in the generated structures from MOFFUSION showed similar distributions to those of the structures from OCChemDB, which further validates the reliability of the generated structure and the entire workflow (Fig. S9).

(Page 16, Supplementary Information) Figure S9.

4. AI models. Really basic models, ResNet, U-Net. A bit underwhelming, it is a weak part of your work.

We appreciate the insightful feedback offered by the reviewer. As noted by the reviewer, the components used in the MOFFUSION architecture are relatively simple and basic models. We employed a ResNet within the MOF Constructor, which is a fundamental model widely used in various tasks, especially for image classification. The U-Net, which is integrated into the diffusion part of MOFFUSION, is a crucial and popular model for image processing that has undergone numerous modifications and extensions.

We would like to highlight that our primary focus during the development of our model was to address the issue of limited flexibility regarding data modalities, while relative less attention was given to novelty in the model architecture. We aimed to enable the conditioning with various data modalities during the generation of MOFs, and achieved satisfactory performance with relatively simple but representative components (e.g., ResNet, U-Net). However, replacing these components with more up-to-date models (e.g., Vision Transformers, Diffusion Transformers) is anticipated to enhance the overall performance of the model, and represents a promising future direction. We appreciate the valuable comment provided by the reviewer.

5. Open source code. Well done!

We appreciate the reviewer's praise.

6. Unconditional generation result. I am a bit concerned that the generated structures are very similar to the training dataset. This indicates that your model is not capable of generalizing to new structures.

The point raised by the reviewer is one that should not be overlooked. Generation of data points similar to those in the given dataset is a key object of generative models. The high similarity between the generated structures from MOFFUSION and the structures in the training dataset, as shown in Table 1 and Figure S8, illustrates the outstanding generation performance of MOFFUSION. However, this also might raise concerns about whether the model can generate new structures that were not seen during training.

In this context, we analyzed the uniqueness and novelty of the generated structures in addition to their structure validity. Uniqueness is a metric that indicates how non-redundant the generated structures are, while novelty measures the percentage of generated structures not included in the training data. Among the generated structures, 79.9 % were *valid and unique*, while 58.4 % were *valid, unique and novel*. This means that, among the valid and unique structures, roughly 3 out of 4 were entirely new structures unseen during training. This analysis validates MOFFUSION's capability to generate new structures. This discussion was originally included in the Supplementary Information, but we agree with the reviewer that it is crucial for validating the model, and we have added it to the main-text. We deeply appreciate the reviewer in improving our manuscript.

(Page 12) The uniqueness and novelty of structures are additional metrics often considered for evaluating generative models. Uniqueness indicates how non-redundant the generated structures are, while novelty measures the percentage of generated structures not included in the training data. For the structures generated using MOFFUSION, 79.9 % were found to be valid and unique, while 58.4 % were valid, unique, and novel. This validates MOFFUSION's capability to generate new structures unseen during the training. However, uniqueness and novelty were not compared with other models, as their respective papers did not report these

specific values, except for MOFDIFF who reported that its proportion of valid, novel, and unique structures was 30.0 %.

7. Multi-target conditioning. You demonstrated this feature using several properties. Can you more clearly correlate this capability to the design of high performing MOFs for hydrogen working capacity? What's the optimal metal node, pore size, geometry, etc., combination?

This is a valuable point raised by the reviewer, which merits further discussion. Generation of the samples with multiple target properties is expected to offer various potential applications, with the chemistry-aware design of high performing MOFs being one of the promising applications, as the reviewer mentioned. Figure B shows the correlation between void fraction and surface area (the same combination as in Figure 5A) with the hydrogen capacity of MOFs. From the figure, it can be identified that the region around a surface area of 5000 m²/g and a void fraction 0.75 is an optimal region for achieving high hydrogen working capacity. With this insight, multi-conditional generation with the target values set at (SA, VF) = (5000, 0.75) would likely lead to the generation of structures with high hydrogen capacity. Instead of direct conditioning with a single target, researchers might able to gain a deeper understanding and chemical intuition from this approach. Additionally, further investigation into multi-target conditioning and exploring its additional application is a promising direction for future work. We appreciate the insightful comment from the reviewer.

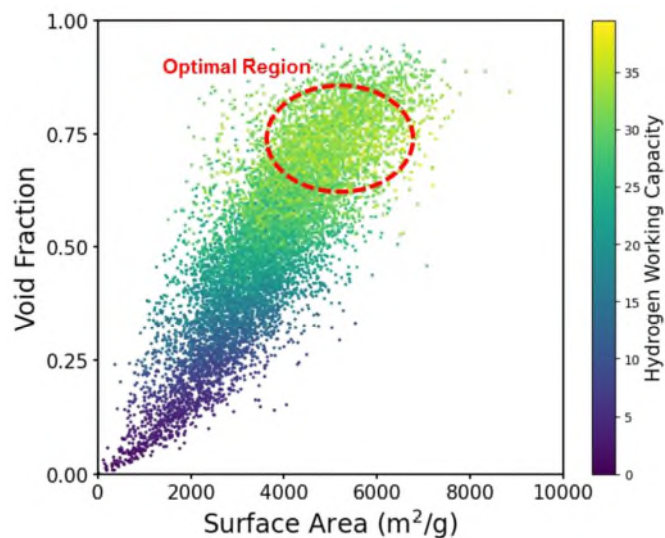


Figure B. Distribution of hydrogen working capacity against void fraction and surface area. The distribution was drawn from randomly selected 10,000 structures from the training set. Hydrogen working capacity was calculated as a pressure swing between 5 bar and 100 bar at 77 K.

8. Text conditioning. If you are already using text conditioning, can you use this capability to screen your set of predicted MOFs to identify the optimal region of parameter space for the production of high performing MOFs?

We appreciate the insightful comment provided by the reviewer. In our original manuscript, we only discussed the generation of samples with specific target properties (e.g., void fraction, topology in Figure 4b), but the description of generating high performing MOFs through text conditioning was lacking. We believe there are several viable solutions for this. The most straightforward approach would be to directly condition the generation process with a sufficiently high target value, using a prompt such as “a void fraction of 0.90”. Alternatively, the user could prepare the text conditioning model with a different training set formatted to include performance levels, such as [very low, low, high, very high]. This would allow for the generation of structures with prompts like “MOFs with very high void fraction”. A complex but comprehensive solution would involve integrating a large language model, such as ChatGPT, into the workflow. This would enable the model to automatically understand the relationship between terms like ‘low’ and ‘high’ and their corresponding values. We mentioned the utilization of large language models as a future direction in our original manuscript, and we have now added specific details on the possibilities for more flexible conditioning. We appreciate the reviewer’s contribution to improving our manuscript.

(Page 19) Otherwise, in contrast to our use of the frozen SciBERT model, fine-tuning the language model specifically for the materials domain could also be a promising method for improving the overall performance and enabling more versatile text conditioning.

9. Please, please do not push critical information about your work to supplementary materials. The reader needs to know what you do from data creation to validation, and the numerical methods you have used for such screening and validation tests.

We deeply agree with the feedback from the reviewer. After receiving the feedback, we reviewed our manuscript once again and identified too much information was allocated in the Supplementary Information, including some crucial details that are important for the overall flow of the work. Therefore, we have moved much of the information that was originally included in the Supplementary Information to the main text. This includes the description of the metrics for evaluating the generated structures, information about the geometry optimization process, and the discussion of the resolution of the SDFs. Additionally, as mentioned in comment #1, we have supplemented our introduction section to help reader follow the overall flow of the manuscript more easily. We are grateful for the thoughtful comments provided by the reviewer, as they have greatly helped us improve our manuscript.

(Page 4) A database composed of 247,742 hypothetical MOF structures was used for the training, which were generated using PORMAKE (a MOF construction software) and further geometry optimized using LAMMPS software.

(Page 5) As a representative numeric data, MOFFUSION was conditioned on hydrogen working capacity and its capability to generate structures with the desired hydrogen capacity was validated through Grand Canonical Monte Carlo (GCMC) simulations. MOFFUSION also demonstrated outstanding performance in handling categorical data, achieving high accuracy in generating MOFs with targeted geometric topologies. Furthermore, its capability to condition with text data was validated by generating MOFs with desired characteristics based on input prompts. Lastly, MOFFUSION's ability to account for multiple target properties during conditional generation was confirmed, and its inherent capability of generating MOFs with user-tailored pore shapes was further demonstrated.

(Page 8) It is worth noting that during the generation process of PORMAKE, we limited the maximum cell length of the generated structures to 60 Å. Therefore, the length of one grid lattice is no longer than 2 Å, ensuring the most chemical bonds are allocated to more than one grid. However, a finer grid will lead to more detailed pore morphology, which will enhance the overall performance of the workflow.

(Page 12) The uniqueness and novelty of structures are additional metrics often considered for evaluating generative models. Uniqueness indicates how non-redundant the generated

structures are, while novelty measures the percentage of generated structures not included in the training data. For the structures generated using MOFFUSION, 79.9 % were found to be valid and unique, while 58.4 % were valid, unique, and novel. This validates MOFFUSION's capability to generate new structures unseen during the training. However, uniqueness and novelty were not compared with other models, as their respective papers did not report these specific values, except for MOFDIFF who reported that its proportion of valid, novel, and unique structures was 30.0 %.

(Page 12) The structure was deemed valid if it could be generated using PORMAKE (which accounts for building block compatibility and steric hindrance), and if it successfully underwent geometry optimization without failure.

Reviewer #3

1. In many of the results, the authors show the resulting distributions of a target property requested. The resulting distributions can be very different depending on the value of the property. Some of the distributions are very sharp while others are very broad. Authors should expand on that behavior.

We appreciate the insightful comment from the reviewer. This is an important point, which is closely related to the nature of generative models. Generative models generally perform well with interpolation problems, while they face relatively more difficulties with extrapolation. In other words, when there are many data points near a target property value, the model performs well in conditionally generating materials with the desired property, resulting in a sharp distribution. However, if there are not enough data points near the target value, conditional generation still works but tends to produce a broad distribution. Specifically, the distribution often shifts toward regions where data points are more abundant. This can be observed in the results of the conditional generation for hydrogen working capacity, as described below. The target WC value of 35 g/L showed a sharp distribution, while the target value of 5 g/L showed a relatively broad distribution. We have added an entirely new paragraph in our manuscript about the shifts in data distributions and how they affect the broadness. We appreciate the reviewer for helping us in improving our manuscript.

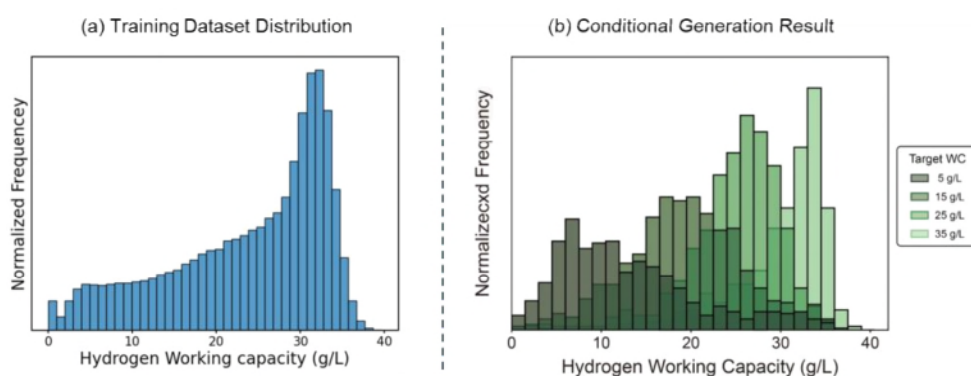


Figure C. (a) Distribution of hydrogen working capacity against of the structures comprising the training set. (b) Conditional generation result from MOFFUSION. Target values for conditional generation were 5, 15, 25, and 35 g/L. Hydrogen working capacity was calculated as a pressure swing between 5 bar and 100 bar at 77 K.

(Page 25) One observation made throughout the workflow is that, as with other generative models, MOFFUSION performs well with interpolation problems but faces relatively more difficulties with extrapolation. In other words, when there are many data points near a target property value, the model performs well in conditionally generating materials with the desired property, resulting in a sharp distribution in the target property of the generated samples. However, if there are not enough data points near the target value, conditional generation still works but tends to produce a relatively broader distribution. Specifically, the distribution often shifts toward regions where data points are more abundant, and this further emphasizes the necessity of preparing a well-distributed training dataset.

2. The authors present a lot of generation results, however, only the SI has any mention of how many are novel, unique, and/or valid. I think the main manuscript should have discussions on these observations. For example, does the multi-conditional generation result in any new structures outside of the training set? Similarly, how was novelty, uniqueness, and validity determined? How do the generated structures compare to those of the training set? Do the generated structures need to be energy minimized or, in the cases where the generated structures are similar to those in the training set, how different are they? How do the unit cell parameters compare, how do the adsorption isotherms compare, how do simulated PXRDs compare?

We deeply agree with the feedback from the reviewer. After receiving the feedback, we reviewed our manuscript once again and identified much crucial information, including the evaluation metrics for generated structures, was allocated to the Supplementary Information.

The validity of the generated structures was assessed using the following metrics. First, the structure must be able to be generated using PORMAKE. This includes: (i) the connection points of building blocks should be compatible with the topology, and (ii) building blocks need to be placed to a given site within a certain root mean square distance. (More details on the PORMAKE generation scheme can be found in [1].) Second, for the generated structures, geometry optimization is conducted and those structures that are chemically non-feasible or suffer from excessive steric hindrance (which might lead to either failure in force field assignment or convergence) are filtered out.

Uniqueness is a metric that indicates how non-redundant the generated structures are, while novelty measures the percentage of generated structures not included in the training data. Among the generated structures, 79.9 % were *valid and unique*, while 58.4 % were *valid, unique and novel*. This means that, among the valid and unique structures, roughly 3 out of 4 were entirely new structures, validating MOFFUSION's capability to generate new structures outside of the training set.

As shown in Table 1 and Figure S8, the generated structures exhibited property distributions highly similar to those of the structures comprising the training set. This confirms that MOFFUSION effectively learned the material space provided by the training set and successfully generated new structures. However, since both the training set structures and those generated through MOFFUSION were generated using the same software, PORMAKE, and both underwent the same energy minimization procedure, it is unlikely that two structures would be chemically identical but have only minor differences in their lattice or atomic coordinates (which we understand as 'similar' structures mentioned by the reviewer).

We agree with the reviewer that the discussion regarding evaluation metrics and the presence of the generated structures outside of the training set is crucial. Therefore, we have added this discussion to our revised manuscript. We deeply appreciate the reviewer in improving our manuscript.

[1] Lee *et al.* Computational Screening of Trillions of Metal-Organic Frameworks for High-Performance Methane Storage. *ACS Appl. Mater. Interfaces* 13, 2-, 23647-23654 (2021)

(Page 12) The uniqueness and novelty of structures are additional metrics often considered for evaluating generative models. Uniqueness indicates how non-redundant the generated structures are, while novelty measures the percentage of generated structures not included in the training data. For the structures generated using MOFFUSION, 79.9 % were found to be valid and unique, while 58.4 % were valid, unique, and novel. This validates MOFFUSION's capability to generate new structures unseen during the training. However, uniqueness and novelty were not compared with other models, as their respective papers did not report these specific values, except for MOFDIFF who reported that its proportion of valid, novel, and unique structures was 30.0 %.

(Page 12) The structure was deemed valid if it could be generated using PORMAKE (which accounts for building block compatibility and steric hindrance), and if it successfully underwent geometry optimization without failure.

(Page 18, Supplementary Information) Structure validity is a metric widely used for assessing the quality of the generated structures through generative models. As described in the main text, the validity of the structures generated through MOFFUSION was assessed using the following metric. First, the structure must be able to be generated using PORMAKE. This includes: (i) the connection points of building blocks should be compatible with the topology, and (ii) building blocks need to be placed to a given site within a certain root mean square distance. (More details on the PORMAKE generation scheme can be found in Lee et al.) Second, for the generated structures, geometry optimization is conducted to filter out those structures that are chemically non-feasible or suffer from excessive steric hindrance (which might lead to either failure in force field assignment or convergence). It is our opinion that our metric is more stringent compared to some precedent works, which considered the structure valid only if there was no atom overlap.

3. In many instances the authors refer to the "accuracy" of the generation but it is unclear what is meant. It seems that each task would have its own measure of accuracy but it is unclear what the ground truth being compared is. For example, when requesting a target void fraction, what is the accuracy of the final generated structures?

In our original manuscript, we tested various conditional generation approaches with diverse modalities and presented the results. For the conditioning of numeric target values (e.g., hydrogen working capacity), we validated the results by showing the distribution of the target property of the generated samples. As reviewer noted, defining accuracy can be challenging in this context, so we qualitatively compared how well the distribution aligns with the respective target property values. On the other hand, in the case of categorical values (e.g., topology), correctness is more straightforward. If we set a specific topology (e.g., pcu) as the target, the generated structures are correct if they exhibit pcu topology; otherwise, they are incorrect. Therefore, we can calculate accuracy for these cases by determining the percentage of samples

that correctly match the intended target. To summarize, accuracy was quantitatively assessed only for categorical values, where the accuracy calculation is straightforward.

We appreciate the insightful and thoughtful comments provided by the reviewer throughout the revision process. We are grateful for the reviewer's contribution to enhancing the clarity and overall quality of our manuscript.

Reviewer #4

Major Comments

1. With a fixed set of building blocks and topologies, the system is limited to a fixed, enumerable set of output structures. In this case, is it fair to compare the model to models that can generate structures who can not be built from a predefined list of building blocks?

We appreciate the insightful feedback provided by the reviewer. From a broader perspective, approaches for the generation of MOFs can be classified into two categories: placing the building blocks into the topology (top-down) or generating the structures atom by atom (bottom-up). In the methods introduced in the manuscript, MOFFUSION and SMVAE can be considered top-down methodologies, while MOFDiff adopts a bottom-up approach. These methods demonstrate distinct strengths and weaknesses. For instance, MOFFUSION and SMVAE exhibit high structure validity, but their chemical space is constrained. Conversely, MOFDiff can explore a theoretically unlimited chemical space, though the validity of the generated structures was unsatisfactory.

We strongly agree with the reviewer's comment that a direct comparison between these methods may not be entirely fair given their different approaches of generating structures. Specifically, the comparison of database diversity can be misleading, as dataset diversity may affect the workflow differently depending on the approach used. Furthermore, since MOFDiff is the only study in MOF generation research that adopts a bottom-up approach, it is problematic to generalize its low validity as a limitation of the entire bottom-up method. These points will be further discussed in the minor comments.

We wish to emphasize that, the primary goal of MOFFUSION was to apply conditioning with diverse data modalities, rather than to propose a novel architecture that outperforms the baseline models. Although a minimal comparison with precedent models is unavoidable, in our revised manuscript, we have limited or rephrased implications of directly comparing methods that stem from different approaches using the same criteria. (Detailed corrections made in our revised manuscript are demonstrated in the responses to subsequent minor comments.) This perspective provided by the reviewer should not be overlooked, and we are very grateful to the reviewer for helping us improve the parts of our manuscript that may have been misleading and ambiguous.

2. What is the advantage of having the diffusion model instead of just directly predicting topology and building blocks? That is, what is the benefit of the "indirection" via the SDF?

We appreciate the reviewer's valuable comment, which is crucial in the overall flow of the research. In the context of materials generation, consideration of the geometric shape and surface morphology has recently received huge attention. In practice, recent studies on molecules and proteins have shown that utilizing information on 3D shapes for the representation learning leads to distinct performance improvements [1], and research directly utilizing 3D shapes in materials generation process is actively being explored [2-4].

Furthermore, the importance of recognizing 3D shapes becomes even more pronounced in porous materials. What differentiates porous materials from others is the presence of pores, and for materials like MOFs with high porosity, their chemical/physical properties are closely related to the shape of the pore structures. For instance, in predicting the properties of MOFs, methodologies considering their 3D morphologies have demonstrated state-of-the-art performance [5,6] compared to those relying solely on types of topology and building blocks. This underscores the importance of explicitly considering the 3D pore structure in understanding the complex nature of MOFs, rather than relying only on simple combinations of building blocks.

In the case of MOFFUSION, it delicately represents the shape of MOF pore structures through SDF and, based on this representation, generates new structures with desired properties using a diffusion model. By employing SDF, MOFFUSION is able to gain a deeper understanding of the pore structure of MOFs compared to methods that use only combinations of building blocks. Therefore, similar to previous research where consideration of 3D structure of materials led to performance improvements, the understanding of 3D structures through SDF would have significantly contributed to the exceptional conditioning performance of MOFFUSION.

In addition, by utilizing SDFs, MOFFUSION offer capabilities that are unattainable with simple representation of build blocks. By applying direct abstraction on 3D pore structure, the generation of the structures with desired pore morphology is feasible. This aspect is most vividly revealed in the pore crafting section of the manuscript, though various other conditional generation approaches for the desired pore structure can also be explored. Such flexibility makes the MOF design process more versatile, potentially allowing for the design of MOFs

with strong binding site or optimal channels for separation. This is crucial for experimentalists as well given that upon trying to design MOF structures, they envision pore shapes, channels, and topologies to optimize for the performance. These achievements are not viable with representation of simple building block combinations, and we believe this approach holds unlimited potential with high research value.

- [1] Lee *et al.*, Pre-training Sequence, Structure, and Surface Features for Comprehensive Protein Representation Learning. *The Twelfth International Conference on Learning Representations* (2023)
- [2] Lutz *et al.*, Top-down design of protein architectures with reinforcement learning. *Science* 380, 266-273 (2023)
- [3] Song *et al.*, SurfPro: Functional Protein Design Based on Continuous Surface. *arXiv preprint arXiv:2405.06693* (2024).
- [4] Wu *et al.*, SurfBind: Surface Distance aided Geometric Deep Learning for Binding Conformations. *Deep Learning on Graphs: Method and Applications (DLG-AAAI'23)* (2023)
- [5] Kang *et al.*, A multi-modal pre-training transformer for universal transfer learning in metal-organic frameworks. *Nature Machine Intelligence* 5, 309-318 (2023)
- [6] Bucior *et al.*, Energy-based descriptors to rapidly predict hydrogen storage in metal-organic frameworks. *Mol. Syst. Des. Eng.* 4, 162-174 (2023)

Minor Comments

1. Given the set of building blocks is fixed, I think a comparison to a GA (e.g. the one reported by the same authors) would make a lot of sense.

We appreciate the valuable comment from the reviewer. The genetic algorithm (GA) is one of the traditional methods investigated for MOF generation. As cycles proceeds through operations such as selection, crossover, and mutation, chromosomes with high fitness score (i.e., MOFs with target property) are gradually obtained. GA and MOFFUSION share similarities, as both utilize a fixed set of building blocks for the generation of MOFs.

One aspect worth discussing in the comparison of two methods is their speed. The training of MOFFUSION may take considerable time (e.g., several days); however, in the inference phase, the user can generate a large number of MOFs (e.g., 10,000) in just a *few minutes*. In contrast, GA necessarily requires numerous cycles for the property of the generated MOFs to converge. In practice, in the workflow proposed in [1], it took roughly *one week* to generate MOFs with the desired property. Furthermore, once a MOFFUSION model is trained, it can be used to generate MOFs for any target values (e.g., 5, 15, 25, and 35 g/L), whereas GA requires entirely

new cycles for each target value. Therefore, from the user's perspective, MOFFUSION would exhibit much higher efficiency in generating structures.

Another aspect is the versatility of conditioning, which was mainly emphasized in the original manuscript. The latent diffusion architecture and attention algorithm allow MOFFUSION to condition on diverse modalities, including categorical data or text. However, GA has primarily been investigated with numeric target values, and novel applications, such as text conditioning, are not feasible with GA.

We agree with the reviewer's feedback that a comparison with GA would be informative for the readers, so we have added this discussion in our revised manuscript. We greatly appreciate the reviewer's helpful comment.

[1] Park *et al.*, Computational Design of Metal-Organic Frameworks with Unprecedented High Hydrogen Working Capacity and High Synthesizability. *Chem. Mater.* 35, 1, 9-16 (2023)

(Page 15) A comparison of MOFFUSION with the genetic algorithm would be valuable, as the latter has been investigated for a similar objective (i.e., the generation of MOFs with desired hydrogen capacity). The genetic algorithm and MOFFUSION share similarities, as both utilize a fixed set of building block for the generation of MOFs. Although the training of MOFFUSION may take considerable time (e.g., several days), in the inference phase, the user can generate a large number of structures (e.g., 10,000) in a few minutes. In contrast, genetic algorithm necessarily requires numerous cycles for the property of the generated MOFs to converge, which took roughly one week in the workflow from Park et al. Furthermore, the genetic algorithm requires entirely new cycles for each different target value (e.g., 5, 15, 25, and 35 g/L), whereas MOFFUSION does not require re-training, making MOFFUSION a more efficient option.

2. Is the comparison in terms of the diversity metrics (Tab 1) fair given that some of the other techniques go beyond a fixed set of building blocks?

We appreciate the insightful feedback from the reviewer. As addressed in the main comments, a direct comparison between methods stemming from different approaches may be not entirely

fair. This is especially true when considering dataset diversity. In a bottom-up approach, structures with building blocks outside of the training set can be obtained theoretically. However, in a top-down approach, only building blocks within the training set can be used, meaning that the diversity of the dataset directly determines the search space for MOFs. In this context, top-down approaches could fundamentally require a more diverse dataset compared to bottom-up approaches. Although the degree of diversity of the components is significantly different and MOFDiff does not mention observations of building component outside of the training set (e.g., new topology) to our knowledge, we decided to remove or limit implications that dataset diversity affects the generation performance of the models. This will reduce ambiguity in the manuscript and help readers understand the information more objectively. We appreciate the review for the meaningful comment which improved our manuscript.

~~(Page 11) Notably, while other databases expand their diversity by mainly focusing on organic linkers, the database used for MOFFUSION has a significantly greater variety in geometric topologies and types of metal nodes. Since generative models learn probability distributions from their training datasets, the diversity of the dataset is a crucial factor directly correlating to the diversity of the output structures. In that sense, compared to other methodologies, the high diversity of building components within the database used for MOFFUSION's training enables the generation of more architecturally diverse structures. (deleted)~~

~~(Page 11) ... spanning a wide diversity of MOF building components, especially in geometric topologies and types of metal nodes (Table 1).~~

~~(Table 1) Bolded numbers were removed from dataset diversity section, to reduce the sense of comparison.~~

3. The limitation of models being trained on different datasets should not only be in the SI, but also in the main text.

We have mentioned the limitation of comparing models trained on different datasets both in the main text and in the SI. The portion of the main text addressing this imitation is shown

below. We may have confused the reviewer by stating redundant information in both the main text and the SI.

(Page 11, extracted) Unlike Perov-5 for perovskites, or Carbon-24 for carbon materials, there exist no publicly shared dataset for MOF generation tasks. All precedent research used their own datasets for training and it is unfeasible to re-train every model with a shared dataset due to the differences in their detailed methodologies. This presents limitations in thoroughly comparing the performance between different models. However, a comparison was conducted as fairly as possible, by focusing solely on the output structures obtained from 10,000 generation attempts for each trained model.

4. "the methodology of directly coordinating the atoms (or grains) yet presents challenges creating an intact structure due to the high structural complexity of MOFs." I am not sure that the study contains empirical proof for this claim.

This is a crucial point brought up by the reviewer. What we originally meant by the phrase reviewer noted was that, given the relatively low structure validity exhibited by MOFDiff compared to other methods, bottom-up generation of MOFs would be more challenging, as MOFs particularly demonstrate significant chemical complexity compared other simple materials. However, since MOFDiff is the only study in MOF generation research that adopts a bottom-up approach, it could be misleading to generalize its low validity as a limitation of the entire bottom-up method. More fundamentally, as addressed in the main comments, a direct comparison between methods stemming from different approaches may be not entirely fair. Therefore, we decided to remove the phrases, as it could be misleading and less objective. We deeply appreciate the reviewer in improving the clarity and objectivity of our manuscript.

(Page 12) ~~Notably, relatively high validity observed in MOFFUSION and SMVAE compared to MOFDiff implies that, the methodology of directly coordinating the atoms (or grains) yet presents challenges in creating an intact structure due to the high structural complexity of MOFs.~~ (deleted)

5. The resolution of 32^3 will most likely not be enough for phenomena that requires exact atomic resolution (chirality, functional groups, etc. - as the low accuracies for edge predictions show). However, increasing the dimensionality will likely lead to computational challenges. It probably makes sense to more clearly acknowledge this limitation.

The point raised by the reviewer is crucial and should not be overlooked. The resolution of the SDF is a topic worth further discussion. As the reviewer pointed out, the resolution of the SDF involves a trade-off between accuracy and computational cost. As resolution increases, more detailed pore morphology can be captured, but it requires more memory and time. In our setting, we used a resolution of 32^3 to balance these two aspects. It is worth noting that, during the generation process of PORMAKE, we limited the maximum cell length of the generated structures to 60 Å. Therefore, the length of one grid lattice is no longer than 2 Å, ensuring the most chemical bonds are allocated to more than one grid. We briefly addressed this discussion on our original manuscript (in *SDF representations for periodic crystals* of the Results section), but we have increased the portion of this part, as it provides valuable information for our readers. We are grateful for the helpful comment from the reviewer.

(Page 8) It is worth noting that during the generation process of PORMAKE, we limited the maximum cell length of the generated structures to 60 Å. Therefore, the length of one grid lattice is no longer than 2 Å, ensuring the most chemical bonds are allocated to more than one grid. However, a finer grid will lead to more detailed pore morphology, which will enhance the overall performance of the workflow.

6. While the pore-crafting technique is nice, it would be great to see experimental quantitative evidence that it systematically works.

We appreciate the valuable comment from the reviewer. Pore crafting is a concept that is introduced for the first time in this work. Similar to the image interpolation or shape completion tasks, the user can guide the de-noising process into a way that the designated region of SDFs remains empty, leading the corresponding portion of the resulting MOFs to stay void. To provide quantitative evidence for pore crafting, we examined what percentage of the generated structures with a request for a center pocket actually exhibited a void at the center as desired.

(The presence of the center pocket was assessed by checking the fractional coordinates of the atoms to determine whether any atom(s) had x, y, z coordinates within the range of [0.4 ~ 0.6]. If no atoms were within this range, the structure was considered to have a void at the center.) Among the generated structures, 64.2 % exhibited a center pocket which is much higher compared to the 35.0 % observed in unconditionally generated structures without pore crafting. This supports the efficacy of pore crafting in generating structures with the desired pore morphology.

In our original manuscript, we included only a qualitative sanity check for the pore crafting. Pore crafting is a concept we devised during the development of MOFFUSION, whose primary goal was to address the issue of limited data modalities during conditioning. Though it was not our original focus, pore crafting has a high potential and may be valuable in future materials generation, which is why we believe it is worth mentioning and have allocated a small section to it in our manuscript. One of the future research directions for us is to place special emphasis on pore crafting and conduct further in-depth analysis in subsequent work. We appreciate the reviewer's insightful comment that prompted deep reflection on our research direction.

7. Other relevant metrics for Tab 1 include uniqueness and distance to the training data

We appreciate the valuable feedback provided by the reviewer. Generation of data points similar to those in the given dataset is a key object of generative models. In Table 1, property statistics measures the distance of the generated structures to those comprising training data. For pore limiting diameter, surface area, and density, the earth mover's distance between generated structures and the training set was compared to assess the models' capability to generate structures similar to the training set.

In addition to structure validity, the uniqueness and novelty of the generated structures were analyzed. Uniqueness is a metric that indicates how non-redundant the generated structures are, while novelty measures the percentage of generated structures not included in the training data. Among the generated structures, 79.9 % were *valid and unique*, while 58.4 % were *valid, unique and novel*. This means that, among the valid and unique structures, roughly 3 out 4 were entirely new structures, validating MOFFUSION's capability to generate unseen structures. We have added this discussion to the main-text.

We deeply appreciate the insightful feedback provided by the reviewer throughout the revision process. These comments made us reflect on our study, reconsider its meaning and significance. We would like to express our sincere gratitude to the reviewer for the contribution to improving our manuscript.

(Page 12) The uniqueness and novelty of structures are additional metrics often considered for evaluating generative models. Uniqueness indicates how non-redundant the generated structures are, while novelty measures the percentage of generated structures not included in the training data. For the structures generated using MOFFUSION, 79.9 % were found to be valid and unique, while 58.4 % were valid, unique, and novel. This validates MOFFUSION's capability to generate new structures unseen during the training. However, uniqueness and novelty were not compared with other models, as their respective papers did not report these specific values, except for MOFDIFF who reported that its proportion of valid, novel, and unique structures was 30.0 %.

Dear Editor,

We greatly appreciate the feedback as the comments made by the reviewers helped us improve the manuscript. Below are point-by-point responses to the comments. Thank you.

Sincerely,

Jihan Kim

In this point-by-point response to the Referees' comments:

1. Black color - comments from the Referees
2. Blue color - response to the comments
3. Green color - added in the manuscript with page number provided

(The page number is based on word file)

Reviewer #4

1. Text conditioned inverse design has been done before. Here a recent work (on Arxiv after the paper at hand) on using LLM-based agents for MOF linkers (*Ansari et al. arxiv, 2024*) but those LLM agents for materials design have been around. Also related, but earlier (*Zhimeng et al., ChemRxiv. 2024*). For molecules, such an approach is relatively common by now (e.g. *Luo et al., arxiv, 2024*). That is, the argument on "text conditioning" being novel can perhaps benefit from a bit more context

We appreciate the comment from the reviewer. As mentioned, since text conditioning for generative models has been reported in small molecules before, the claim that this work is one of the first attempts at text conditioning for materials design may be a bit broad. However, this is still the first generative model focused on the text-conditioned generation of porous materials, and more broadly, it is also the first attempt for crystalline materials as well, to the best of our knowledge. In our revised manuscript, we have rephrased our statement to narrow its scope and specify the domain that this work addresses (i.e., materials → porous materials).

LLM agents are emerging as promising tools for materials design. Although their approach is somewhat different from ours (conditioning a diffusion model using text), both can be regarded as forms of text-based materials design. In our original manuscript, we noted that LLMs are actively integrated into materials science. We have now added the works suggested by the reviewer as references to this statement, further informing users about the introduction of LLMs to various materials, including MOFs. We appreciate the reviewer for the helpful comment.

(Page 18) We examined MOFFUSION for the conditional generation based on text data, as one of the first attempts at text-based conditional generation of porous materials.

[Ref. 70] Luo, Y., Li, S., Liu, Z., Wu, J., Yang, Z., He, X., Wang, X. and Tian, Q. Text-guided Diffusion Model for 3D Molecule Generation. *arXiv preprint arXiv:2410.03803* (2024).

[Ref. 77] Ansari, M., Watchorn, J., Brown, C. E. & Brown, J. S. dZiner: Rational Inverse Design of Materials with AI Agents. *arXiv preprint arXiv:2410.03963* (2024).

[Ref. 78] Liu, Z., Su, Y., Guo, Y., Lin, J., Wang, S., Song, Z., Xi, Z., Gao, H., Shi, L. & Wang, G. Post-Pretraining Large Language Model Enabled Reverse Design of MOFs for Hydrogen Storage. *ChemRxiv* (2024).

2. “the first application of 3D modeling techniques for the generation of porous materials”

I am not sure if this is not a bit too broad. The authors themselves used different 3D techniques (e.g., energy grids) for the modeling of porous materials, such as ZeoDiff.

We used the phrase “the first application of 3D modeling techniques” because SDF is widely used for the visualization of 3D objects and for representing assets in virtual space by explicitly describing their 3D surfaces. However, as the reviewer mentioned, there exist other representations used for modeling porous materials that implicitly convey 3D morphology. The energy grid is a representative example, as it fundamentally represents the interactions between probe gases and the atoms comprising materials, thereby conveying the 3D shape of voids in an indirect way. To the best of our knowledge, the implementation of energy grids (or any other 3D shape-related representations) for generating porous materials has only been conducted by our group, particularly for zeolites (ZeoGAN¹ and ZeoDiff²). We agree with the reviewer that our claim could be seen as a bit broad; therefore, we speculated that, in MOFFUSION, the 3D

modeling techniques were first implemented for the generation of MOFs (not for all porous materials). We appreciate the reviewer for improving the clarity of our manuscript.

[1] Kim *et al.* Inverse design of porous materials using artificial neural networks. *Science advances* **6**, eaax9324 (2020).

[2] Park *et al.* Inverse design of porous materials: a diffusion model approach. *Journal of Materials Chemistry A*, **12**(11), 6507-6514 (2024).

(Page 4) We applied a signed distance function (SDF) for the input representation, marking the first application of 3D modeling techniques for the generation of MOFs.

3. Regarding my comment on going via the SDF: I was asking for evidence of the benefits of going via the SDF. I can see that it can perhaps enable things like "pore crafting" - but if I condition on a given pore size and window diameters it can be almost the same. And the question that is still open to me is if there is any quantitative benefit of going then via the SDF instead of directly predicting the relevant building blocks (i.e., directly go from the multi-modal conditioning inputs to the building blocks)

Why am I so persistent on this? The use of SDFs is portrayed as one of the main advances in this work ("The high validity of MOFFUSION stems from its novel architecture and the benefits of SDFs, which provide a deep understanding of the geometric nature of MOFs."), but there is little direct evidence of their benefit. Yes, we see that the overall pipeline works, but there is no direct evidence that going via the SDFs is important for that. In addition, you add another indirection, and the model from SDF to building blocks also only has limited accuracy (80%, according to the authors). That is, if one can avoid multiplying errors that might lead to increased overall performance. In addition, the SDF will also lose some "chemical resolution" which might be essential for important applications like direct air capture (where not only the pore shape is important but also the chemical environment, Fig. S6d might be such a case). While I can conceptually see that the system generates MOFs and that they also fulfill some target characteristics the article gives, at this point, little signal on whether going via the SDFs is a more fruitful approach than going via energy grids, reinforcement learning on the building blocks (Park *et al.*, *Digital Discovery*, 2024) coarse-grained diffusion, etc. Distilling a clearer message on a point like this might make the article also more impactful in the long run.

We appreciate the insightful comment from the reviewer. Firstly, we would like to highlight the concept of pore crafting. Pore crafting is a novel inverse design approach for porous materials that offers significant potential unattainable through other methods that do not utilize SDFs. With pore crafting, users can specify the desired pore structures and also their positions within the unit cell without restrictions. This represents a key distinction from simply conditioning on geometry parameters, such as window size, and notably, the latter is less likely to work as the desired pore structure becomes complex. This will be further discussed in our response to comment #5.

To provide further quantitative evidence on pore crafting, we compared its performance in generating MOFs with center pockets (which is one of the simplest target pore structures) against other viable methods. In obtaining structures with pocket at the center, it is not straightforward to determine which set of geometric parameters should be used for conditioning. One method worth comparing could be the genetic algorithm, which we have also discussed in detail in the first round of revision. We implemented the genetic algorithm to generate MOFs with center pocket, using the same workflow applied for designing MOFs with high hydrogen storage in Ref [1]. We first trained a machine learning model to predict the presence of center pockets within the structure and used this as a cost function for the genetic algorithm. As shown in Figure A, while the genetic algorithm succeeded in generating the desired structures, it couldn't surpass the performance of pore crafting. The details on genetic algorithm is summarized at the end of this response. This suggests that pore crafting is not only a novel approach but also could be a practical method for designing porous materials.

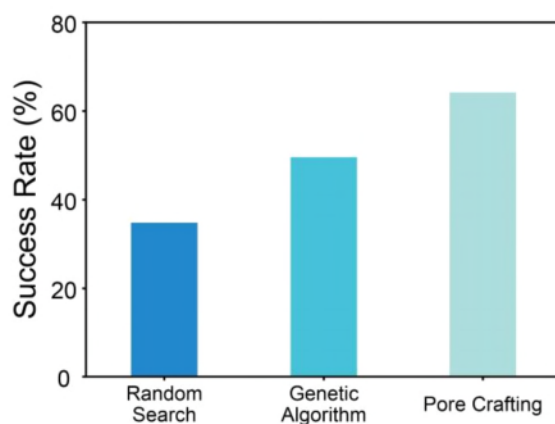


Fig. A. Comparison of pore crafting and the genetic algorithm in generating structures with center pockets. Random search refers to the percentage of structures exhibiting center pockets among 50,000 randomly sampled structures. For both the genetic algorithm and pore crafting, the percentage of structures exhibiting center pockets (as desired) among all generated structures was compared.

Secondly, to provide additional evidence of the advantage of going through SDFs, we examined the performance of the workflow after excluding the SDF phase, allowing the diffusion model to generate building blocks directly. This serves as an ablation study to assess the necessity of SDFs within the MOFFUSION workflow. We utilized the D3PM² architecture, one of the most widely used diffusion models specifically designed for discrete data. The model is trained to learn the valid building block compositions (e.g., “tbo + N15 + N330 + E17”) from the training set, and to generate new compositions based on that. We referred this workflow (without SDFs) as Building Block Diffusion. The exactly same training data was used with MOFFUSION, and the chemical space searchable for both workflows are identical. We assessed the validity, uniqueness, and novelty of 10,000 generated structures and confirmed that the original MOFFUSION workflow exhibits higher generation performance, as shown in Figure B. We interpret this result to indicate that representing structures using SDFs as 32x32x32 continuous float data offers greater flexibility and a deeper understanding compared to representing them as four discrete entities, thereby enhancing the generation performance of the diffusion model. We agree with the reviewer that we provided the limited evidence on the advantage of using SDFs, and we have added this discussion to the manuscript.

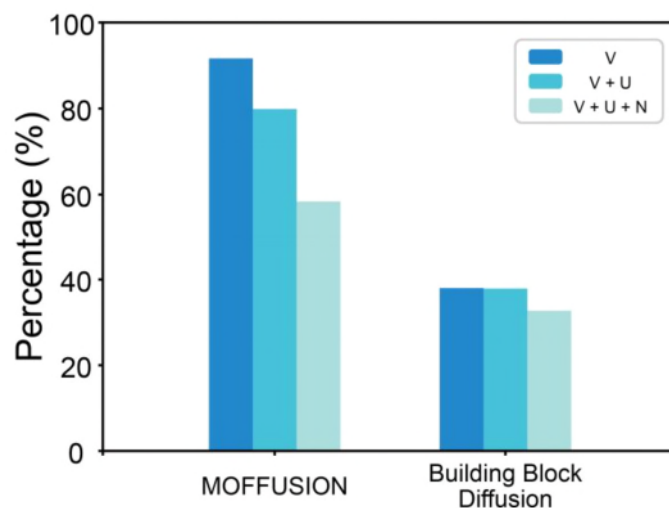


Fig. B. Comparison of 10,000 generated structures obtained from MOFFUSION and Building Block Diffusion, respectively. “V” refers to the ratio of valid structures, “V+U” indicates valid and unique structures, and “V+U+N” represents valid, unique and novel structures among the generated outputs. Further descriptions of the evaluation metrics can be found in the main text.

Lastly, we acknowledge the point that SDFs could lose some chemical resolution, which we emphasized in our original manuscript by dedicating an entire paragraph in the discussion

section to this topic. We noted that MOFFUSION focuses more on physical shapes rather than on chemistry aspects, which could be particularly important for specific applications. In this regard, we mentioned in our manuscript that incorporating an energy grid as an additional channel could enhance the model’s understanding of chemistry. While it is not feasible to retrain a different version of MOFFUSION with the extra energy grid channel at this revision stage, we believe this represents a promising direction for future work which could expand the model’s range of applications. We appreciate the insightful feedback from the reviewer, which significantly helped us improve our manuscript.

[Details on Genetic Algorithm]

For machine learning prediction, the final layer of MOF-net³ was modified to fit logistic regression for classification tasks. The model returns the likelihood that a structure possesses a center pocket as a float in the range [0,1]. As shown in Figure C, the trained model exhibited satisfactory accuracy, with high true positive and true negative ratios. This trained model served as the cost function for the multispecies genetic algorithm, following the same approach used in Ref [1]. The population size was set to 1,000 for each topology, and a total of 100 generations were conducted to ensure the convergence of the algorithm. Chromosome pairs were selected using tournament selection with a probability of 0.9. For 90 % of the selected pairs, uniform crossover was performed, while a random chromosome was passed to the next generation (without crossover) for the remaining 10 %. Random mutation was also applied with a probability of 0.2. For a fair comparison, the genetic algorithm was designed to target exactly same chemical space as MOFFUSION.

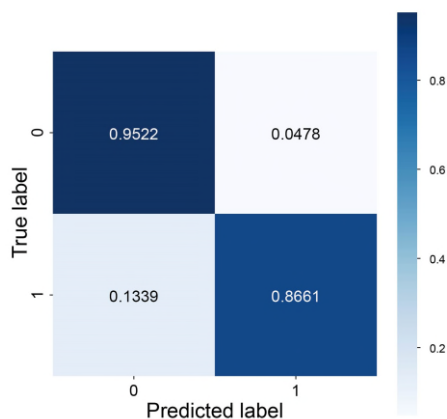


Fig. C. Precision of the trained classification model for predicting the existence of center pockets. Label 1 refers to samples with center pockets, while Label 0 refers to those without center pockets.

[Details on Building Block Diffusion]

As MOF building blocks are discrete objects, the D3PM², a denoising diffusion probabilistic model developed for the generation of discrete data, was used for Building Block Diffusion. The model is designed to generate four categorical data (each corresponding to topology, metal node, organic node, and edge) simultaneously to suit our purpose. Similar to DiT⁴, the architecture is based on transformers, and the number of model parameters was set

to 790M, which is comparable to that of the original MOFFUSION (690M). Building Block Diffusion and MOFFUSION share the exact same search space, and the same training set used in MOFFUSION was employed. A total of 300 epochs was used for training, which was confirmed to be long enough for the training loss to converge. The best checkpoint model was used for the generation with regard to the validation loss.

- [1] Park *et al.*, Computational Design of Metal-Organic Frameworks with Unprecedented High Hydrogen Working Capacity and High Synthesizability. *Chem. Mater.* 35, 1, 9-16 (2023)
- [2] Austin *et al.*, Structured denoising diffusion models in discrete state-spaces. *Advances in Neural Information Processing Systems*, 34, 17981-17993 (2021)
- [3] Lee *et al.* Computational screening of trillions of metal–organic frameworks for high-performance methane storage. *ACS Applied Materials & Interfaces*, 13.20, 23647-23654 (2021)
- [4] Peebles *et al.* "Scalable diffusion models with transformers." *Proceedings of the IEEE/CVF International Conference on Computer Vision*. (2023)

(Page 13) Furthermore, an ablation study on the SDFs was performed to assess the significance of representing structures as SDFs, as shown in the Supplementary Information Section S3-4. The comparison with a diffusion model that directly predicts the building block entities highlights that representing structures using SDFs played a significant role in the overall performance of MOFFUSION.

(Page 20, Supplementary Information) Section 3-4. Ablation study of the SDFs
(Page 21, Supplementary Information) Fig. S10

4. Another point I had been pedantic on in my first review is that MOFFUSION relies on a set of building blocks:

The model's reliance on a set of building blocks was a focal topic of discussion during the first revision round. We agreed with the reviewer's comments that a direct comparison of methodologies based on different approaches (i.e., top-down vs bottom-up) can be misleading, and we revised our manuscript by actively incorporating this feedback. We eliminated/minimized the nuances of comparing different methods and focused on highlighting the performance of MOFFUSION itself and its ability to condition on diverse data modalities. This was a crucial point that should not be overlooked, and the comment provided by the reviewer significantly helped clarify the ambiguity in our original manuscript.

5. For pore-crafting: How well would a lookup work (e.g. based on some zeo++ statistics, or persistent homology)

As we understood, the term ‘lookup’ might refer to filtering the MOF database using geometric descriptors (e.g., those obtainable from zeo++) to search for structures with the desired pore morphology. If the desired pore structure is simple, this lookup could yield similar outcomes to pore crafting. However, representing desired pore structures with only a few geometric descriptors is highly limited. In particular, recently discovered MOFs with novel applications often feature unique and intrinsic pore structures. For example, the exceptional separation behavior observed in Zr-fum-fcu-MOF originates from its asymmetric pore windows, which selectively allow nitrogen gas to penetrate while blocking other natural gas constituents (Figure A).¹ It would be extremely difficult to represent such complex pore structures with just a few geometric descriptors. In contrast, the pore crafting method presented in this work allows users to freely express the desired appearance of the pores and even their positions within the unit cell, offering significant potential to achieve high-level inverse design without constraints on the complexity of the desired pore morphology.

More directly, we conducted a genetic algorithm to compare with pore crafting, as we responded to the comment #3. Although we targeted the center pocket, which is one of the simplest pore morphologies to consider, pore crafting exhibited decent performance compared to the genetic algorithm. While the genetic algorithm might not correspond to the lookup method mentioned by the reviewer, this indicates that pore crafting is not only a novel methodology but can also be practically applied to design new porous materials.

Another point we want to mention is related to computational cost. Since pore crafting follows a blended diffusion methodology, once the unconditional MOFFUSION model is prepared, no further training is required (regardless of the desired pore morphology requested by users). In contrast, lookup methods necessitate different screening attempts for each target pore shape, making pore crafting significantly more advantageous in terms of computational cost. Although we are still in the early stages, we believe that pore crafting has significant potential compared to existing methodologies, and we hope that sharing this concept with readers will contribute to the advancement of the field.

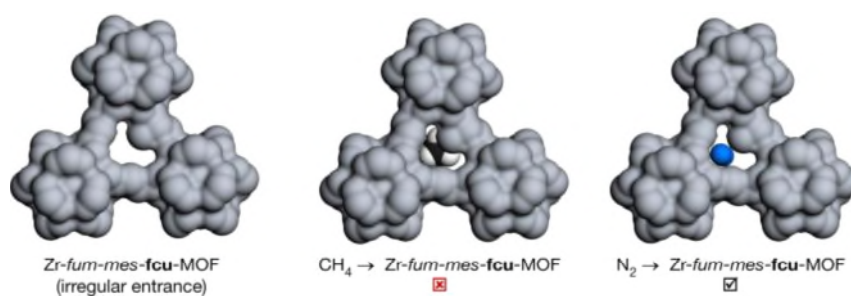


Fig. A. Illustrations of the irregular pore window of Zr-fum-mes-fcu-MOF. Reprinted figure with permission from ref [1]. Copyright 2022 Springer Nature.

[1] Zhou, Sheng, *et al.*, Asymmetric pore windows in MOF membranes for natural gas valorization., *Nature*, 606.7915, 706-712 (2022)

6. How much do we gain for all the other properties compared to having a machine learning model for the working capacity or some other target property and simply predicting this property for all MOFs. I do not necessarily ask for new experiments on all of that -- but I feel that it currently is not clear to the reader that those enumeration-based approaches might also be an option and that they might even perform better in some instances.

We appreciate the comment from the reviewer. From the statement, “having a machine learning model for the working capacity or some other target property and simply predicting this property for all MOFs”, we understand that the reviewer appears to be discussing a comparison with high-throughput screening methods. Screening methods are the most straightforward way to discover materials with desired properties, and they have been widely applied in the field of MOFs, yielding meaningful results. As the reviewer mentioned, screening method can be the most effective approach in certain circumstances, especially when the number of candidate structures is manageable (e.g., when considering the specific metal types or topologies).

However, screening-based methodologies have a drawback because, as the size of the database increases, the required computational costs become prohibitively high. This issue becomes even more pronounced in the case of MOFs, where the number of virtual structures is known to reach into the trillions.¹ While machine learning predictions are much faster than conventional computational simulation-based evaluations, they are still not fast enough to cover such a vast search space. Therefore, few optimization methods have been explored for

more efficient searches, and recently, generative models have been introduced as alternatives. We outlined the flow of changes in these methodologies in the introduction section of the original manuscript, and a comparison of our workflow with the genetic algorithm (which could be regarded as an enumeration-based method from a broad perspective) can be found in Page 15 of the main text.

We deeply appreciate the insightful feedback throughout the revision process, which stems from the reviewer's thorough understanding. We would like to express our sincere gratitude to the reviewer for the contribution to improving our manuscript.

[1] Lee *et al.* Computational screening of trillions of metal–organic frameworks for high-performance methane storage. *ACS Applied Materials & Interfaces*, 13.20, 23647-23654 (2021)