Symmetry is the key to the design of reticular frameworks

Andrea Darù,[*,†] John S. Anderson,[†] Davide M. Proserpio,[‡] and Laura Gagliardi[*,†,¶]

†*Department of Chemistry, University of Chicago, Chicago, Illinois 60637, United States*

‡*Dipartimento di Chimica, Università degli studi di Milano, Via Golgi 19, 20133 Milano, Italy*

¶*Pritzker School of Molecular Engineering, University of Chicago, Chicago, Illinois 60637, United States*

E-mail: adaru@uchicago.edu; lgagliardi@uchicago.edu

Abstract

De novo prediction of reticular framework structures is a challenging task for chemists and materials scientists. Herein, a computational workflow that predicts a list of possible reticular frameworks based on only the connectivity and symmetry of node and linker building blocks is presented. This list is ranked based on the occurrence of topologies in known structures, thus providing a manageable number of structures that can be optimized using density functional theory, and inform future experiments. This workflow is broadly applicable, correctly predicts known reticular materials, and furthermore identifies heretofore unknown phases for some systems. This workflow is available online at `https://rationaldesign.pythonanywhere.com/`.

Introduction

The computational generation and prediction of reticular frameworks, such as metal-organic frameworks (MOFs) and covalent organic frameworks (COFs), is a highly active field in chemistry and materials science.[1–5] Although efforts have focused on leveraging neural networks, machine learning methods,[6–8] and high-throughput techniques to generate extensive datasets of structures,[9–13] significant manual effort is still required.[14] Furthermore, while many approaches can accurately predict known MOF structures that are related to existing databases, they have comparatively limited utility with fundamentally new frameworks. This is because generative automated approaches, based on neural networks, face challenges such as uncertain synthesizability and difficulties in accurately assessing connectivity and topology which lead to a low rate of correctly predicted structures. The high computational cost and still limited success in the MOF/COF research field of these methods highlight the need for more effective approaches that integrate chemistry principles alongside machine learning models, rather than solely relying on training off of existing data.[8,15–19]

On a positive perspective, the high demand of data has led to the generation of databases containing either curated experimental structures refined via density functional theory (DFT), and/or computationally generated hypothetical structures (i.e. Materials Project,[9] CSD MOF Collection,[20,21] ARC database,[12] MOFX-DB database,[10] CoRE-MOF database,[11] ReDD-COFFEE database,[13] TopCryst database[22]). Such collections of structures and their properties are providing valuable data to base exploratory research and to evolve the field. However, a workflow which predicts candidate structures based on fundamental principles of frameworks components would both liberate predictions from known databases and also enable the prediction of fundamentally new structures.[23–26]

These challenges inspired us to investigate a new approach, where the generation of novel frameworks is based on the fundamental feature of the building blocks: their symmetry. We then envisioned comparing possible structural solutions with how often their symmetry features appear in existing databases to narrow down likely candidate structures. In this way

the crystallographic information associated with a reticular framework is generated only for a small amount of potential structures; to allow for DFT calculations on this reduced set. While traditional machine learning models are trained on existing datasets, this approach focuses on high-quality de novo predictions by leveraging the structural properties of linkers and nodes to design new MOFs, without depending on pre-existing patterns. Machine learning algorithms can infer symmetry from known MOF structures, but this often fails for novel MOFs due to limited training data. In contrast, symmetry remains universally applicable, making it a more reliable principle for predicting new MOFs. This workflow aims to create a foundation for a suite of computational tools that ultimately can deliver broad predictivity without high computational costs.

Results and Discussions

Workflow

The workflow is designed to make the generation of frameworks easier and more *rational* by following symmetry and structural rules. It starts with the identification of the node and the linker of choice, followed by the creation of a list of potential topologies that can be accessible using those building blocks. The workflow is shown in Figure 1.

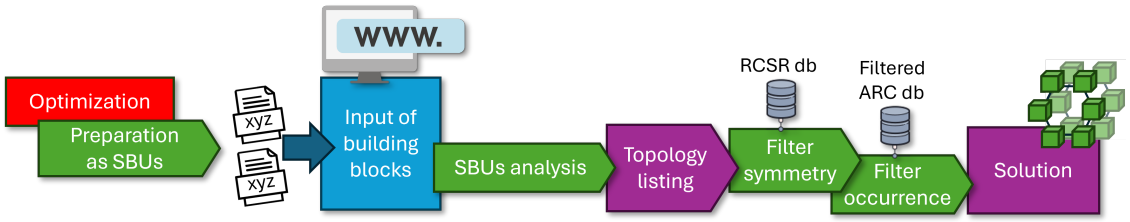

Figure 1: Workflow schematics including the current working modules.

A general description is provided herein, which is subsequently expanded with examples of MOF and COF structures taken from the literature. The node and linker structures chosen by the user are inputted as secondary building units (SBUs) in XYZ file format. To

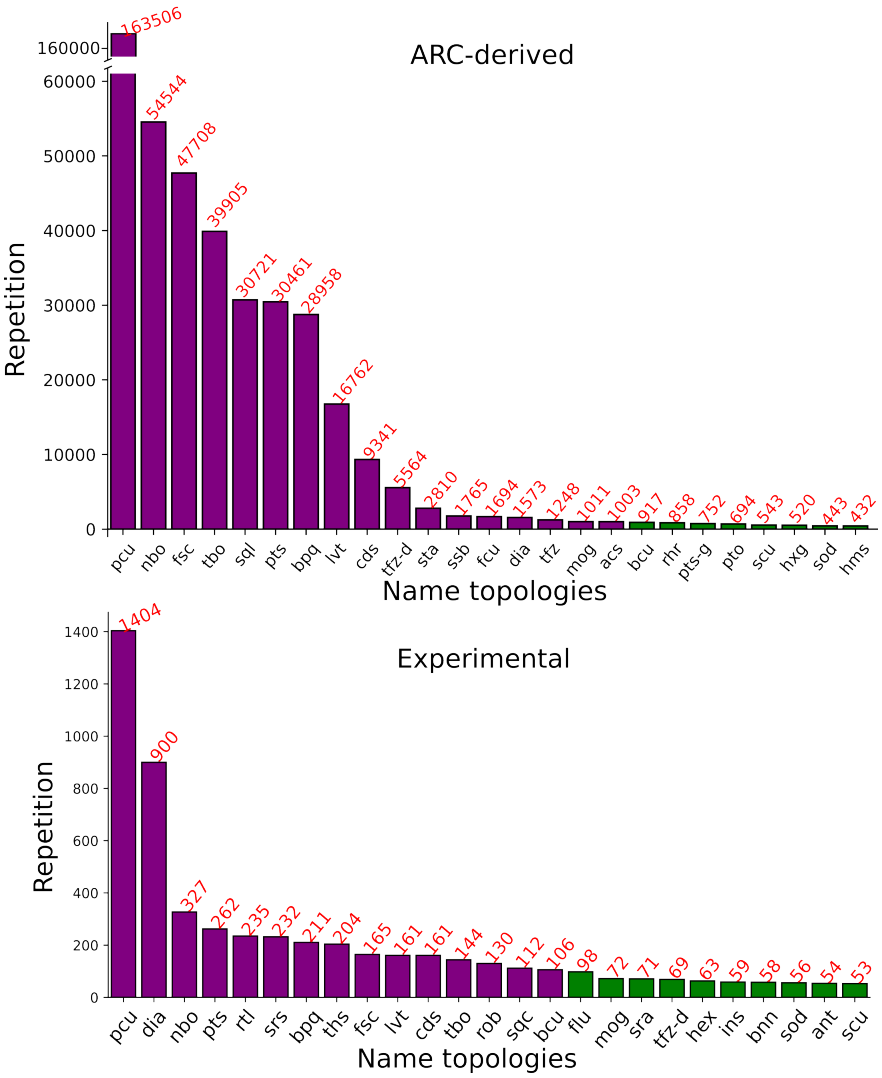

Figure 2: Ranking of topologies from ARC-derived (top), and experimental (bottom) datasets. The top plot shows the 25 most frequent topologies from the ARC-derived dataset with the 17 most frequent repetitions of at least 1000 times highlighted in purple. The bottom plot shows the 25 most frequent topologies from the experimental dataset with the 15 most frequent repetitions of at least 100 times highlighted in purple

make such SBUs the user builds, on their computer, the building blocks and optimizes them preferably at the DFT level, or otherwise they can be taken from curated SBU structures present in public databases. Each SBU has the ligating atoms capped with ghost atoms (X). Therefore, the building blocks have to be modified accordingly, by the user, to set the position of the new connection points substituting some atoms with X. (See Supporting Information for full description).

Using the number of connections (X atoms) in the SBUs, a list of potential topologies is provided. At this stage, state of the art high-throughput techniques optimize and rank the structures based on their relative energies (or free energies), typically using force-field methods. However, force-field optimized structures are less accurate than those optimized with DFT. The key innovation in this approach is that, rather than optimizing all generated structures, only a selection of them is chosen based on the symmetry of the building blocks, and topological occurrence data from previously synthesized structures in the literature. This reduces the list to a more manageable set of potential structures that can be accurately optimized at the DFT level. In contrast, standard high-throughput techniques consider every structure that fits the connectivity, and neural networks are limited to data of previously existing reticular frameworks.

The workflow algorithm determines the symmetry group of the building blocks and uses it to refine the initial list which is based only on connectivity numbers. To perform this selection the data from the Reticular Chemistry Structure Resource database (RCSR) is used.[27] The RCSR database provides valuable topology information extracted from synthesized frameworks. In the present workflow, these pieces of information are used in the other way around to build frameworks based on building blocks and their structural match with a known topology. Indeed, every framework topology entry in RCSR provides, among others, details on the number of nodes and linkers present in the framework, and their symmetry.

The symmetrization module is a fundamental piece of this workflow. In the current implementation of the code the symmetry group of each SBU is considered, together with its related higher and lower symmetry groups, for up to three symmetry groups per SBU. This is based on the fact that when the building blocks that form a framework are assembled, they can be slightly stretched, compressed, or twisted, and therefore their symmetry can increase or decrease compared to the isolated building blocks.

To further refine the selection process, an additional criterion is employed: the frequency of each topology from synthesized MOFs; this is obtained by analyzing and cleaning the ARC

database (See Supporting Information).[12] This ARC-derived dataset contains the topology name of each structure and its frequency of occurrence. In addition a fully experimental dataset was created with the frequency of occurrence of topologies from CoRE MOF[11] and CSD databases[20,21] (See Supporting Information). Each entry of the final topology list from the workflow is ranked accordingly to the ARC-derived dataset primarily, with insights from the experimental dataset.

Analyzing the ARC-derived dataset, 61 topologies appear with at least 100 repetitions (**pcu** – **cdz**). Among these, 17 are repeated at least 1000 times (**pcu** – **acs** highlighted in purple) Figure 2. The experimental dataset provides a different topology order of frequencies of occurrence: 15 topologies appear at least 100 times and only one more than 1000 times (**pcu**) Figure 2. Overall, **pcu** is always the most common topology, together with **nbo**, **pts**, **bpq**, **lvt**, **cds**, and **tbo** which are on the high end of the occurrence frequency number. The ARC dataset has been chosen even though it collects *hypothetical* MOF structures to get a wide collection of underlying net topologies. Our concept aims at leveraging stable computational calculated structures which can still be porous after optimization without amorphization, beyond simple proof of their existence. The experimental dataset is used to support the actual existence of such topologies and to weight for the experimental frequency of occurrence. Both datasets are still unbalanced towards common topologies and are thus prone to be biased always towards the most frequent topology; improving the predictive power for less common structures is an ongoing area of research for this project.

The workflow has seven steps. The first two are based on user preferences while the others are performed automatically:

1. Geometry optimization of node and linker building blocks at any computational level, preferably DFT

2. Definition of connectivity of the building blocks to create the final SBU (see Supporting Information)

3. Detection of symmetry group of each SBU

4. Listing of potential framework topologies based on the connectivity number

5. Exclusion of items from 4 based on symmetry of SBUs

6. Ranking items in 5 based on occurrence from the literature

7. Output of topologies names

Validation

To validate the workflow, a series of MOF and COF structures have been selected from the literature and blindly rebuilt starting from their components. This series includes UIO-66,[28] MOF-808,[29] PCN-6',[30,31] NU-50,[32] MOF-841,[29], MOF-801,[29], MOF-802,[29], MOF-812;[29] and COFs such as LZU-306[33] and LZU-111[34,35]. We showcase here MOF-841, PCN-6', NU-50, and LZU-111, while the other examples are presented in the SI.

MOF-841[29] is formed by the node $Zr_6O_4(OH)_4(-CO_2)_8$ which has 8 connections (CN = 8), and the linker H_4MTB (4,4',4",4"'-methanetetrayltetrabenzoic acid) which has 4 connections (CN = 4) (Figure 3). Considering solely the connectivity number, there are 15 potential topologies: **crd**, **crs-d**, **csq**, **fla**, **flu**, **hcp-d**, **jus**, **ken**, **nin-d**, **scu**, **sqc**, **xax**, **xbi**, **xbm**, **xbt**. However, 15 topological frameworks can be a lot to be explored with DFT. Additionally, it is likely that some of these frameworks cannot be constructed based on the symmetry of the node and the linker, D_{4h} and T_d, respectively in Schönflies notation. Following the concept that only certain symmetry groups belong to a space symmetry, among the initial 15 topologies we identify those with the space symmetry that incorporates both the node and linker symmetries. None are found in this case. The symmetry refinement step of the workflow also considers the higher and lower symmetry groups associated with both the node and linker. This adjustment allows the node to be considered also as D_4 and O_h, and the linker as T and O_h. Every pair combination is inspected to see if it matches one of

the topologies present in the initial list. The final result provides only **flu** as a topology with node symmetry O_h and linker symmetry T_h, which can now be used to build the MOF structure, in agreement with the experimental topology of MOF-841[29]. This is then followed by DFT geometry optimization to obtain a final reliable structure. (See Computational Details section in Supporting Information) For this example, the experimental PXRD pattern of MOF-841 is in good agreement with the PXRD pattern computed for the DFT optimized structures following this workflow (Figure 3).[29]

The second example is PCN-6'[30,31], which is formed by the node $Cu_2(-CO_2)_4(H_2O)_2$ with 4 connections (CN = 4), and the linker TATB (4,4',4"-s-triazine-2,4,6-triyltribenzoate) with 3 connections (CN = 3) (Figure 4). Based on the connectivity numbers of each building block, there are 72 possible topologies (See Supporting Information). Exploring all of them would be computationally prohibitive, even with classical simulations. The symmetry groups of the DFT optimized node and linker are D_{4h} and D_{3h}, respectively (See Supporting Information). This pair of symmetry groups does not belong to any of the above topologies, thus both symmetries can be augmented and reduced providing the node with the additional symmetries D_4 (lower), and T_d (higer), and the linker with C_{3h} (lower), D_{6h} (higher). Coupling these symmetry groups and search for their existing combination in the previous list provides a refined list of potential topologies that can be ranked based on the occurrences in the ARC-derived, and experimental datasets: **tbo** (node: D_4, linker: C_{3h}. 39905 and 144 occurrences), **pto** (node: D_4, linker: C_{3h}. 694 and 38 occurrences), **bor** (node: D_4, linker: D_{3h}. 9 and 2 occurrences). This analysis provides **tbo** as most probable, followed by **pto**, and **bor**; this result is in agreement between the two datasts and most importantly with the literature, where **tbo** was the assigned topology.[30,31] Possible PCN-6' MOFs with these four topologies were optimized using DFT (See Computational Details). The **tbo** and **pto** topologies differ in electronic energy only by 0.4 kcal/mol, suggesting the possibility that the topology might change with some synthetic modifications (i.e. solvent, temperature, pressure, etc.).

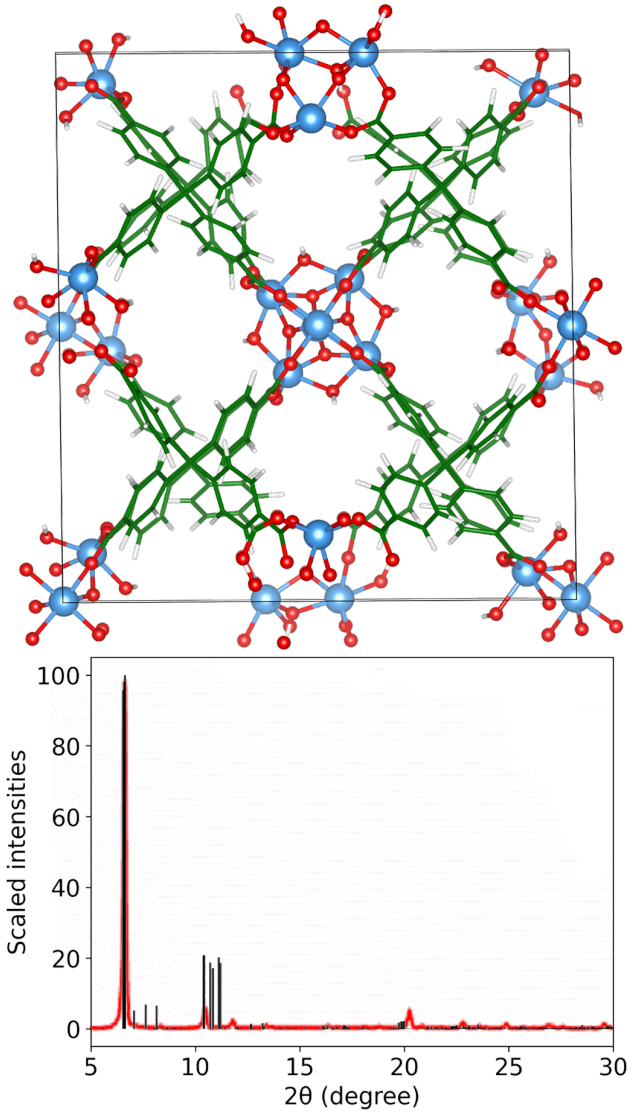

Figure 3: MOF-841 in topology **flu** optimized at DFT level with PBE-D3BJ functional.[36,37] Experimental PXRD pattern (red line) taken from previously published data from reference[29]. Computational PXRD pattern calculated in this work (black lines). Color-code: Zr blue, O red, C green, H white.

Next, we considered MOF NU-50[32] with In(III) node (from $InCl_3$) and H_4TBAPy linker (1,3,6,8-tetrakis(p-benzoic acid)pyrene), each with a connectivity number of 4 which provides 241 possibilities (See Supporting Information for full list). The symmetry groups of node and linker are D_{2d} and D_2 respectively, giving no potential topologies. However, by symmetry relaxation, symmetry groups D_{4h} (higher), and D_2 (lower) are available for the node, and D_{2h} (higher) and C_2 (lower) for the linker. This provides, after pairwise combination, the

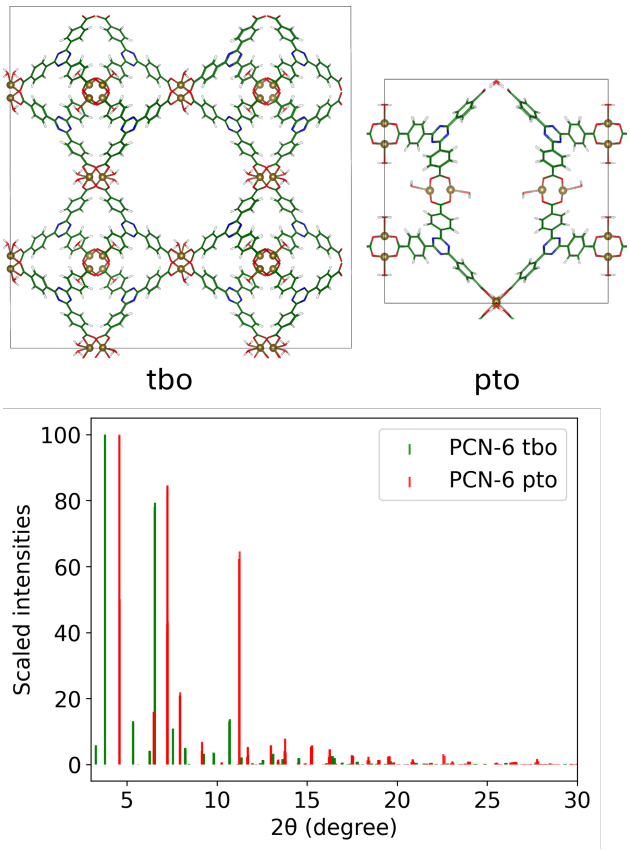

Figure 4: PCN-6' structures with topologies **tbo** and **pto** optimized at DFT level with PBE-D3BJ functional[36, 37] (top); and related computational PXRD pattern (bottom). Colorcode: Cu gold, N blue, O red, C green, H white.

topologies (ordered by frequency of occurrence): **pts** (D_{2d} and D_{2h}, 30461 and 262), **ssb** (D_2 and D_2, 1765 and 13), **pth** (D_2 and D_2, 297 and 9), and **ssa** (D_2 and D_2, 108 and 3). According to this ranking, **pts** is the most likely topology for this MOF (from both datasets) which is in agreement with experiment. This also shows the importance of considering higher and lower symmetries for the node and linker. We note that other topologies from this ranking could be used to generate related NU-50 MOFs ; **ssb**, **pth**, and **ssa**. MOF structures with these four topologies were DFT optimized with the PBE-D3BJ[36,37] functional (see SI for full Computational Details). To balance the framework charges one Na^+ per node has been added in its proximity before full geometry optimization. The **pts** and **pth** structures have similar electronic energies, with the latter 2.0 kcal/mol lower (Figure 5). **ssb** is about 12.8 kcal/mol higher than **pts** (See Supporting Information). These results provide

valuable insights, particularly regarding the experimental observations that the framework exhibits structural deformation after thermal activation and solvent removal, leading to pore contraction that does not revert to the pristine structure even after reactivation by submersion in water.[32] We propose that this behavior may be related to a thermally accessible **pth** morphology which could potentially be present post-activation. The probe-occupable available volume (POAV) of these two structures (**pts**, and **pth**) measured using Zeo++[38] fortifies this suggestion. **pts**-NU-50 POAV is 2.5 cm^3/g with the largest included sphere of 13.8 Å, while **pth**-NU-50 shows some shrinkage with POAV is 2.1 cm^3/g with largest included sphere of 11.0 Å. In addition, the computed PXRD for both **pts** and **pth** MOFs (Figure 5) shows a trend consistent with experimental findings.[32] The green PXRD pattern, both in the experimental and computational panel, corresponds to the **pts**-NU-50 structure and the red PXRD pattern to the activated structure, computationally associated with **pth**. The highest experimental peak shifts from 8° to 10°; the computed PXRD peak shifts from 6° in **pts**-NU-50 to 7° in **pth**-NU-50, mirroring the experimental trend. Therefore, computations suggest that the **pth** topology is a reasonable assignment for the activated NU-50 structure.

The final example is COF LZU-111[34,35] with topology **lon-b** and a 3-fold interpenetration. **lon-b** is a rare topology due to the use of two nodes rather than the more common **lon** (38 repetitions in our experimental database) which contains only one node. In this case the workflow predicts **pts** as the most favorable topology. Calculating the suggested topology reveals a discrepancy when comparing the PXRD patterns. **pts** has the most intense peak at 4° and a pair of peaks at 5° with a lower intensity, instead **lon-b** shows the first and less intense peak positions at 5° followed by the most intense peak positions at around 6°. In Figure 6 both PXRD patterns from calculations are shown together with the related COF structures. The COF LZU-111 case also shows the need to implement the interpenetration feature which is very common in COF structures. In fact, COF LZU-111 has been manually made also at non- and 2- fold and structurally optimized via DFT (See Supporting Information). From the optimized structures the related PXRD patterns have then been plotted

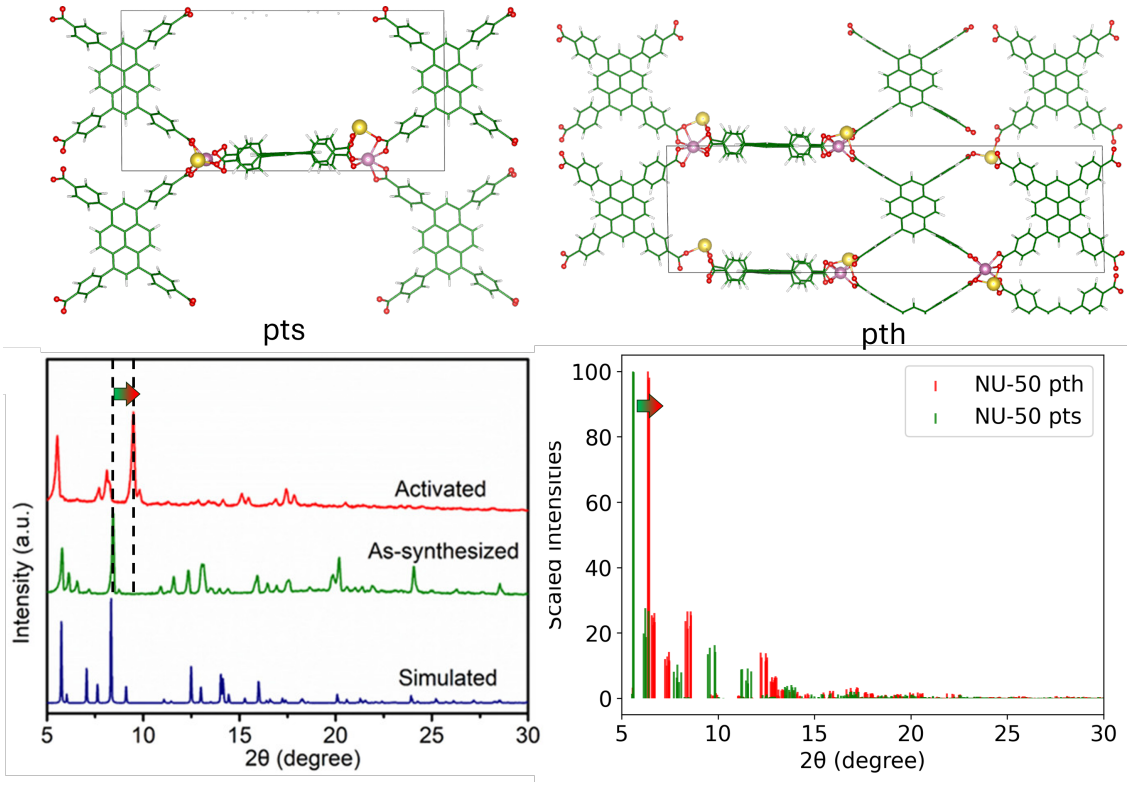

Figure 5: NU-50 structures with topologies **pts** and **pth** optimized at DFT level with PBE-D3BJ functional[36, 37] (top); related computational PXRD pattern (bottom right) showing the peak shift with potential relation to post-activated structure deformation; and experimental PXRD pattern (bottom left) from reference [32] added for comparison. Colorcode: In pink, Na yellow, O red, C green, H white.

showing the 3-fold structure to be quite different than both the 2-fold and the none-fold (See Supporting Information). Therefore, even though the workflow would have had guessed the correct topology, the missing two folds prevent it from generating the correct structure.

More examples are described in the Supporting information in the same fashion as the case studies described here.

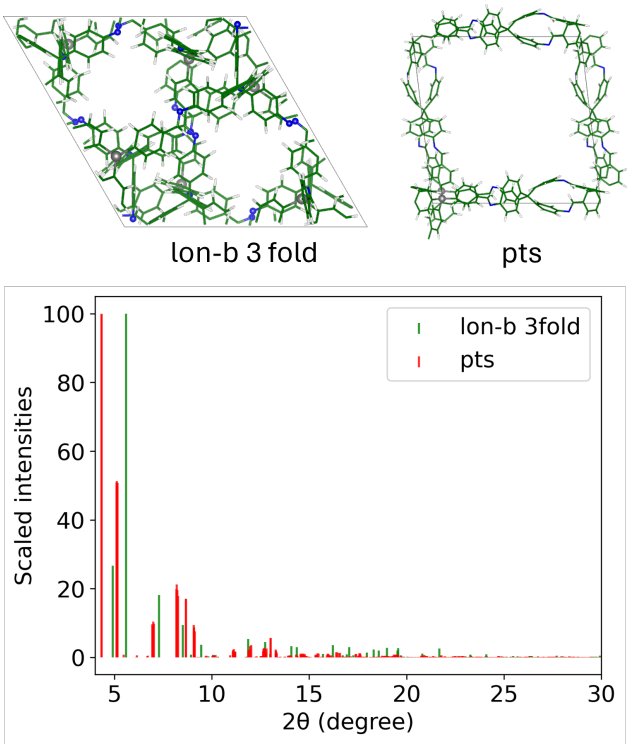

Figure 6: COF LZU-111 structures with topologies **lon-b** 3-folded interpenetrated, and **pts** optimized at DFT level with r^{2}SCAN+rVV10 functional[39, 40] (top), and related computational PXRD pattern (bottom) showing for comparison. Colorcode: Si gray, N blue, O red, C green, H white.

Conclusion

We presented a comprehensive and unique workflow for the generation of MOF or COF structures from building blocks. The implementation of this workflow is available at `https://rationaldesign.pythonanywhere.com`. The workflow has been tested on molecular frameworks already reported in the literature, confirming its reliability. Additionally, it has provided insights into alternative topologies that can be formed through post-synthetic processes. Finally, one case where the workflow fails is described, to highlight the necessity of more balanced structural databases as well as the need to implement a module that considers interpenetration, a common feature of molecular frameworks. This workflow will be extended to 2D structures and rod-MOFs. The method presented is a powerful tool for the prediction of new MOF or COF structures, particularly those featuring novel building blocks

that have not been previously explored.

Acknowledgement

This research was supported by the Catalyst Design for Decarbonization Center an Energy Frontier Research Center funded by the U.S. Department of Energy, Office of Science, Basic Energy Sciences under Award No. DE-SC0023383. We acknowledge the University of Chicago's Research Computing Center for providing the resources to carry out the computational structural optimizations presented in this work. DMP thanks the MUR for the grant PRIN2020 "Nature Inspired Crystal Engineering (NICE)

References

(1) Butler, K. T.; Davies, D. W.; Cartwright, H.; Isayev, O.; Walsh, A. Machine learning for molecular and materials science. *Nature* (**2018**), 547–555.

(2) Sanchez-Lengeling, B.; Aspuru-Guzik, A. Inverse molecular design using machine learning: Generative models for matter engineering. *Science* (**2018**), 360–365.

(3) Sparks, T. D.; Kauwe, S. K.; Parry, M. E.; Tehrani, A. M.; Brgoch, J. Machine learning for structural materials. *Annu. Rev. Mater. Res.* (**2020**), 27–48.

(4) Freund, R.; Zaremba, O.; Arnauts, G.; Ameloot, R.; Skorupskii, G.; Dincă, M.; Bavykina, A.; Gascon, J.; Ejsmont, A.; Goscianska, J., et al. The current status of MOF and COF applications. *Angew. Chem. Int. Ed.* (**2021**), 23975–24001.

(5) Schmidt, J.; Marques, M. R.; Botti, S.; Marques, M. A. Recent advances and applications of machine learning in solid-state materials science. *Npj Comput. Mater.* (**2019**), 83.

(6) Zeni, C.; Pinsler, R.; Zügner, D.; Fowler, A.; Horton, M.; Fu, X.; Shysheya, S.; Crabbé, J.; Sun, L.; Smith, J.; Nguyen, B.; Schulz, H.; Lewis, S.; Huang, C.-W.; Lu, Z.; Zhou, Y.; Yang, H.; Hao, H.; Li, J.; Tomioka, R.; Xie, T. MatterGen: a generative model for inorganic materials design, 2024.

(7) Merchant, A.; Batzner, S.; Schoenholz, S. S.; Aykol, M.; Cheon, G.; Cubuk, E. D. Scaling deep learning for materials discovery. *Nature* (**2023**), DOI: 10.1038/s41586-023-06735-9.

(8) Luo, Y.; Bag, S.; Zaremba, O.; Cierpka, A.; Andreo, J.; Wuttke, S.; Friederich, P.; Tsotsalas, M. MOF synthesis prediction enabled by automatic data mining and machine learning. *Angew. Chem. Int. Ed.* (**2022**), e202200242.

(9) Jain, A.; Ong, S. P.; Hautier, G.; Chen, W.; Richards, W. D.; Dacek, S.; Cholia, S.; Gunter, D.; Skinner, D.; Ceder, G., et al. Commentary: The Mater. Project: A materials genome approach to accelerating materials innovation. *APL mater.* (**2013**).

(10) Bobbitt, N. S.; Shi, K.; Bucior, B. J.; Chen, H.; Tracy-Amoroso, N.; Li, Z.; Sun, Y.; Merlin, J. H.; Siepmann, J. I.; Siderius, D. W.; Snurr, R. Q. MOFX-DB: An Online Database of Computational Adsorption Data for Nanoporous Mater. *J. Chem. Eng. Data* (**2023**), 483–498.

(11) Chung, Y. G.; Haldoupis, E.; Bucior, B. J.; Haranczyk, M.; Lee, S.; Zhang, H.; Vogiatzis, K. D.; Milisavljevic, M.; Ling, S.; Camp, J. S.; Slater, B.; Siepmann, J. I.; Sholl, D. S.; Snurr, R. Q. Advances, Updates, and Analytics for the Computation-Ready, Experimental Metal–Organic Framework Database: CoRE MOF 2019. *J. Chem. Eng. Data* (**2019**), 5985–5998.

(12) Burner, J.; Luo, J.; White, A.; Mirmiran, A.; Kwon, O.; Boyd, P. G.; Maley, S.; Gibaldi, M.; Simrod, S.; Ogden, V.; Woo, T. K. ARC–MOF: A Diverse Database of Metal-Organic Frameworks with DFT-Derived Partial Atomic Charges and Descriptors for Mach. Learn. *Chem. Mater.* (**2023**), 900–916.

(13) De Vos, J. S.; Borgmans, S.; Van Der Voort, P.; Rogge, S. M.; Van Speybroeck, V. ReDD-COFFEE: a ready-to-use database of covalent organic framework structures and accurate force fields to enable high-throughput screenings. *J. Mater. Chem. A* (**2023**), 7468–7487.

(14) Batatia, I.; Benner, P.; Chiang, Y.; Elena, A. M.; Kovács, D. P.; Riebesell, J.; Advincula, X. R.; Asta, M.; Avaylon, M.; Baldwin, W. J.; Berger, F.; Bernstein, N.; Bhowmik, A.; Blau, S. M.; Cărare, V.; Darby, J. P.; De, S.; Pia, F. D.; Deringer, V. L.; Elijošius, R.; El-Machachi, Z.; Falcioni, F.; Fako, E.; Ferrari, A. C.; Genreith-Schriever, A.; George, J.; Goodall, R. E. A.; Grey, C. P.; Grigorev, P.; Han, S.; Handley, W.; Heenen, H. H.; Hermansson, K.; Holm, C.; Jaafar, J.; Hofmann, S.; Jakob, K. S.; Jung, H.; Kapil, V.; Kaplan, A. D.; Karimitari, N.; Kermode, J. R.; Kroupa, N.; Kullgren, J.; Kuner, M. C.; Kuryla, D.; Liepuoniute, G.; Margraf, J. T.; Magdău, I.-B.; Michaelides, A.; Moore, J. H.; Naik, A. A.; Niblett, S. P.; Norwood, S. W.; O'Neill, N.; Ortner, C.; Persson, K. A.; Reuter, K.; Rosen, A. S.; Schaaf, L. L.; Schran, C.; Shi, B. X.; Sivonxay, E.; Stenczel, T. K.; Svahn, V.; Sutton, C.; Swinburne, T. D.; Tilly, J.; van der Oord, C.; Varga-Umbrich, E.; Vegge, T.; Vondrák, M.; Wang, Y.; Witt, W. C.; Zills, F.; Csányi, G. A foundation model for atomistic materials chemistry, 2024.

(15) Huang, G.; Guo, Y.; Chen, Y.; Nie, Z. Application of Mach. Learn. in Material Synthesis and Property Prediction. *Mater.* (**2023**), DOI: 10.3390/ma16175977.

(16) Lin, J.; Liu, Z.; Guo, Y.; Wang, S.; Tao, Z.; Xue, X.; Li, R.; Feng, S.; Wang, L.; Liu, J.; Gao, H.; Wang, G.; Su, Y. Machine learning accelerates the investigation of targeted MOFs: Performance prediction, rational design and intelligent synthesis. *Nano Today* (**2023**), 101802.

(17) Yang, P.; Zhang, H.; Lai, X.; Wang, K.; Yang, Q.; Yu, D. Accelerating the Selection of Covalent Organic Frameworks with Automated Mach. Learn. *ACS Omega* (**2021**), 17149–17161.

(18) Kang, Y.; Kim, J. ChatMOF: an artificial intelligence system for predicting and generating metal-organic frameworks using large language models. *Nat. Commun.* (**2024**), 4705.

(19) Fu, X.; Xie, T.; Rosen, A. S.; Jaakkola, T. S.; Smith, J. A. In *The Twelfth International Conference on Learning Representations*, 2024.

(20) Moghadam, P. Z.; Li, A.; Wiggin, S. B.; Tao, A.; Maloney, A. G. P.; Wood, P. A.; Ward, S. C.; Fairen-Jimenez, D. Development of a Cambridge Structural Database Subset: A Collection of Metal–Organic Frameworks for Past, Present, and Future. *Chemistry of Materials* (**2017**), 2618–2625.

(21) Li, A.; Perez, R. B.; Wiggin, S.; Ward, S. C.; Wood, P. A.; Fairen-Jimenez, D. The launch of a freely accessible MOF CIF collection from the CSD. *Matter* (**2021**), 1105–1106.

(22) Alexander P. Shevchenko Aleksandr A. Shabalin, I. Y. K.; Blatov, V. A. Topological representations of crystal structures: generation, analysis and implementation in the TopCryst system. https://topcryst.com/. *STAM Methods* (**2022**), 250–265.

(23) Islamov, M.; Babaei, H.; Anderson, R.; Sezginel, K. B.; Long, J. R.; McGaughey, A. J.; Gomez-Gualdron, D. A.; Wilmer, C. E. High-throughput screening of hypothetical metal-organic frameworks for thermal conductivity. *Npj Comput. Mater.* (**2023**), 11.

(24) Altintas, C.; Erucar, I.; Keskin, S. High-Throughput Computational Screening of the Metal Organic Framework Database for CH4/H2 Separations. *ACS Appl. Mater. Interfaces.* (**2018**), 3668–3679.

(25) Majumdar, S.; Moosavi, S. M.; Jablonka, K. M.; Ongari, D.; Smit, B. Diversifying Databases of Metal Organic Frameworks for High-Throughput Computational Screening. *ACS Appl. Mater. Interfaces.* (**2021**), 61004–61014.

(26) Rosen, A. S.; Fung, V.; Huck, P.; O'Donnell, C. T.; Horton, M. K.; Truhlar, D. G.; Persson, K. A.; Notestein, J. M.; Snurr, R. Q. High-throughput predictions of metal–organic framework electronic properties: theoretical challenges, graph neural networks, and data exploration. *Npj Comput. Mater.* (**2022**), 1–10.

(27) O'Keeffe, M.; Peskov, M. A.; Ramsden, S. J.; Yaghi, O. M. The Reticular Chemistry Structure Resource (RCSR) Database of, and Symbols for, Crystal Nets. *Acc. Chem. Res* (**2008**), 1782–1789.

(28) Cavka, J. H.; Jakobsen, S.; Olsbye, U.; Guillou, N.; Lamberti, C.; Bordiga, S.; Lillerud, K. P. A new zirconium inorganic building brick forming metal organic frameworks with exceptional stability. *J. Ame. Chem. Soc.* (**2008**), 13850–13851.

(29) Furukawa, H.; Gándara, F.; Zhang, Y.-B.; Jiang, J.; Queen, W. L.; Hudson, M. R.; Yaghi, O. M. Water Adsorption in Porous Metal–Organic Frameworks and Related Mater. *J. Am. Chem. Soc.* (**2014**), 4369–4381.

(30) Sun, D.; Ma, S.; Ke, Y.; Collins, D. J.; Zhou, H.-C. An Interweaving MOF with High Hydrogen Uptake. *J. Am. Chem. Soc.* (**2006**), 3896–3897.

(31) Ma, S.; Sun, D.; Ambrogio, M.; Fillinger, J. A.; Parkin, S.; Zhou, H.-C. Framework-Catenation Isomerism in Metal-Organic Frameworks and Its Impact on Hydrogen Uptake. *J. Am. Chem. Soc.* (**2007**), 1858–1859.

(32) Chen, Y.; Zhang, X.; Chen, H.; Drout, R. J.; Chen, Z.; Mian, M. R.; Maldonado, R. R.; Ma, K.; Wang, X.; Xia, Q.; Li, Z.; Islamoglu, T.; Snurr, R. Q.; Farha, O. K. Tuning the Atrazine Binding Sites in an Indium-Based Flexible Metal–Organic Framework. *ACS Appl. Mater. Interfaces.* (**2020**), 44762–44768.

(33) Liang, L.; Qiu, Y.; Wang, W. D.; Han, J.; Luo, Y.; Yu, W.; Yin, G.-L.; Wang, Z.-P.; Zhang, L.; Ni, J.; Niu, J.; Sun, J.; Ma, T.; Wang, W. Non-Interpenetrated Single-Crystal Covalent Organic Frameworks. *Angew. Chem. Int. Ed.* (**2020**), 17991–17995.

(34) Ma, T.; Wei, L.; Liang, L.; Yin, S.; Xu, L.; Niu, J.; Xue, H.; Wang, X.; Sun, J.; Zhang, Y.-B., et al. Diverse crystal size effects in covalent organic frameworks. *Nat. Commun.* (**2020**), 6128.

(35) Ma, T.; Kapustin, E. A.; Yin, S. X.; Liang, L.; Zhou, Z.; Niu, J.; Li, L.-H.; Wang, Y.; Su, J.; Li, J., et al. Single-crystal x-ray diffraction structures of covalent organic frameworks. *Science* (**2018**), 48–52.

(36) Perdew, J. P.; Burke, K.; Ernzerhof, M. Generalized gradient approximation made simple. *Phys. Rev. Lett.* (**1996**), 3865.

(37) Grimme, S.; Ehrlich, S.; Goerigk, L. Effect of the damping function in dispersion corrected density functional theory. *J. Comput. Chem.* (**2011**), 1456–1465.

(38) Willems, T. F.; Rycroft, C. H.; Kazi, M.; Meza, J. C.; Haranczyk, M. Algorithms and tools for high-throughput geometry-based analysis of crystalline porous materials. *Microporous Mesoporous Mater.* (**2012**), 134–141.

(39) Furness, J. W.; Kaplan, A. D.; Ning, J.; Perdew, J. P.; Sun, J. Accurate and numerically efficient r2SCAN meta-generalized gradient approximation. *J. Phys. Chem. Lett.* (**2020**), 8208–8215.

(40) Ning, J.; Kothakonda, M.; Furness, J. W.; Kaplan, A. D.; Ehlert, S.; Brandenburg, J. G.; Perdew, J. P.; Sun, J. Workhorse minimally empirical dispersion-corrected density functional with tests for weakly bound systems: r^2SCAN + rVV10. *Phys. Rev. B.* (7 **2022**), 075422.