Diversifying databases of metal organic frameworks for high-throughput computational screening

Sauradeep Majumdar, Seyed Mohamad Moosavi, Kevin Maik Jablonka,
Daniele Ongari, and Berend Smit*

Laboratory of Molecular Simulation, Institut des Sciences et Ingénierie Chimiques, École Polytechnique Fédérale de Lausanne (EPFL), Rue de l'Industrie 17, CH-1951 Sion, Valais, Switzerland

E-mail: berend.smit@epfl.ch

Abstract

By combining metal nodes and organic linkers, an infinite number of metal organic frameworks (MOFs) can be designed in silico. When making new databases of such hypothetical MOFs, we need to assure that they not only contribute towards the growth of the count of structures but also add different chemistry to existing databases. In this study, we designed a database of $\sim 20,000$ hypothetical MOFs which are diverse in terms of their chemical design space—metal nodes, organic linkers, functional groups and pore geometries. Using Machine Learning techniques, we visualized and quantified the diversity of these structures. We find that on adding the structures of our database, the overall diversity metrics of hypothetical databases improve, especially in terms of the chemistry of metal nodes. We then assessed the usefulness of diverse structures by evaluating their performance, using grand-canonical Monte Carlo simulations, in two important environmental applications—post combustion carbon capture and hydrogen

storage. We find that many of these structures perform better than widely used benchmark materials such as Zeolite-13X (for post combustion carbon capture) and MOF-5 (for hydrogen storage).

Keywords: MOFs, Molecular Simulations, Machine Learning, Diversity, Carbon Capture, Hydrogen Storage

1 Introduction

Metal Organic Frameworks (MOFs) have been an exciting class of crystalline nanoporous materials since their discovery about two decades ago. By combining metal nodes and organic linkers, one can, in principle, make an infinite number of MOFs.¹ Over 100,000 MOFs have already been currently synthesized.²⁻⁴ Due to characteristics like high surface area, large pore volume, and wide range of pores sized from micro to mesoscale, MOFs have found applications in several areas like gas storage,⁵ catalysis⁶ and nondistillative separations⁷⁻⁹.

At present, of the millions of possible MOF structures, in practice we can only synthesize a small fraction of all possible structures. This is because synthesizing a new MOF, and then characterizing and testing it, could take many months.¹⁰ Therefore, computational researchers have been building databases of hypothetical MOF structures for high-throughput screening purposes. The idea stems from the fact that efficient computational algorithms can help in generating MOF structures and evaluating them for different applications, in a less expensive and faster way. Ranking these hypothetical MOFs based on specific material properties then helps in identifying the most promising materials for a specific application.^{4,11} Experimentally, we can then focus our efforts on synthesizing only the promising materials. Thus, along with the 100,000 experimental structures, there are also millions of hypothetical MOFs, which have been generated computationally.^{10,12-15}

One of the earliest hypothetical MOF databases was generated by Wilmer et al.,¹⁴ which consisted of a database of around 137,000 MOFs constructed using a ‘Tinkertoy’ algorithm.

This database of hypothetical MOFs was used in various screening studies for gas storage and separations and some of those hypothetical MOFs have been experimentally synthesized.¹⁶ This approach of MOF construction however had a limitation. These 137,000 MOFs only sample from six topologies, with most of them having a **pcu** topology. What this algorithm did was it sequentially connected the molecular building blocks (SBUs) until a period crystal was formed. Or, in other words, it used a bottom-up approach for generating MOFs.

Along with the building blocks, topologies also play an important role in MOF performance. The net of a MOF, also called a topology, represents the underlying connectivity of the metal nodes and organic linkers. Gomez Gualdron et al.¹⁷ showed in their computational study of Zr-MOFs for volumetric methane storage how Zr-MOFs based on **ftw** topology outperform Zr-MOFs based on **scu** and **csq** topologies, even if the same organic linkers are used for all the three topologies. Subsequent algorithms to generate MOFs explore the different topologies using a top-down MOF construction algorithms. One such topology based MOF construction algorithm has been developed by Boyd and Woo^{12,18} and is called ToBasCCo. This study generated around 325,000 MOFs which were screened for post combustion carbon capture and two of those structures, namely Al-PMOF and Al-PyrMOF have also been synthesized. Another topology based MOF construction algorithm was developed by Gomez-Gualdron et al.¹³ and is called ToBaCCo. Hypothetical MOFs generated using ToBaCCo have been used for screening applications in hydrogen storage, methane storage and xenon-krypton separation. Some structures from this database were also synthesized and tested.¹⁹ Recently, a study by Lee et al.¹⁵ presented an algorithm to explore a MOF space of over 100 trillion materials, which was used to find the most optimal structures for methane storage.

At present there are thus several databases with millions of hypothetical MOF structures in total. The algorithms underlying these databases have been focused on enumerating as many possible structures for a given topology, metal node, organic linker and functional group. The end result is that we have now reached such a large number of structures that it is practically impossible to screen all structures for a possible application. In addition, as we

have an infinite number of possible structures, this is a fundamental problem we cannot solve with faster computers. It is therefore important to take a different approach and carefully select a representative set of diverse structures as a starting point for a screening study. And subsequently use as a strategy to only add novel structures if in our set of most diverse materials some materials are missing.

In this respect, a detailed analysis of the diversity of the Computation-Ready Experimental Metal-Organic Framework (CoRE MOF) database,³ which represents the synthesized MOFs from the Cambridge Structural Database (CSD),²⁰ and hypothetical databases was performed by Moosavi et al.²¹ Descriptors were built to capture features of a MOF, such as pore geometry, metal chemistry, linker chemistry, and functional groups, which combine to form the chemical design space for a MOF chemist. The chemical diversity of a MOF was then expressed in terms of these features. Moosavi et al.²¹ concluded that with respect to pore geometry, linker chemistry and functional groups, the hypothetical databases seem to sufficiently well covered. However, with respect to metal chemistry, the hypothetical databases turn out to be less explored. The variety of metal chemistry in hypothetical databases was found to be surprisingly low, when compared to those in the experimental databases.²¹ Hence, there are many MOF structures corresponding to these missing metal nodes, lying in the less explored regions of the material space. It is therefore important that we are able to generate such structures to study their properties. By harvesting these metal nodes using the algorithms developed in this respect,^{15,21,22} one could generate such structures.

In this study, we thus designed a database of $\sim 20,000$ hypothetical MOFs, keeping in mind their chemical diversity in terms of—pore geometry, metal chemistry, linker chemistry and functional groups. We focused on improving the diversity of metal nodes in hypothetical MOF databases, by harvesting metal nodes from experimental structures. Diversity of metal nodes can be important for important environmental applications like carbon capture.²¹ We are interested in carbon capture and storage because it is considered to be one of the most

promising and viable technologies to address the rising CO₂ emissions in the atmosphere.^{23,24} In this study, we have specifically looked into post-combustion carbon capture. Another application we have looked here into, is the storage of hydrogen, a promising vehicular fuel.^{25,26} Promising structures found from this screening study could then be added to the list of already available top-performing structures and the resultant list of structures would thus be chemically more diverse. This would then also help to choose from a wider range of structures and try synthesizing them.

2 Methods

2.1 Building Blocks Selection and Structure Generation

Moosavi et al.²¹ developed a methodology to mine metal nodes from experimental MOF databases. These are some of the metal nodes which are not commonly used for structure generation in hypothetical MOF databases. Thus, in this work, we focused on some of these metal nodes, as a proof of concept to validate our argument of improving metal diversity. Figure 1 shows some of these metal nodes. In total, 14 metal nodes have been used and they are all listed in the Supporting Information. We have chosen metal nodes consisting of different metals such as nickel, zinc, cadmium, copper, manganese, cobalt, and lead. Additionally, we have included metal nodes with different connectivity, ranging from 4-connected nodes up to 12-connected nodes, as well as different coordination geometries such as triangular, tetranuclear, square-planar and others. We selected organic linkers from the libraries reported in the hypothetical databases in literature.^{9,12} We have also used several functional groups to decorate these organic linkers (see Supporting Information). The topologies were selected from the Reticular Chemistry Structure Resource Database (RCSR)²⁷ (see Supporting Information for the list of all topologies used in this study). These building blocks were then used to build the hypothetical MOFs using the ToBaCCo (Topologically Based Crystal Constructor)^{9,13} algorithm.

Figure 1: Some metal nodes used in this study to generate hypothetical MOFs. The metal type, connection type and the metal node names as we used in this study are provide below each node. These metal nodes are shown to highlight different metals, connectivities and geometries used in this study. The entire list of metal nodes used in this study is provided in the Supporting Information.

2.2 Structure Optimization and Charge Generation

The hypothetical MOFs of our database were optimized using the Universal Force Field (UFF).²⁸ The optimization of the structures was performed using LAMMPS,²⁹ the input and data files for which were generated using lammps_ interface.³⁰ The EQeq (extended charge equilibration) method¹⁶ was used to generate the partial charges of the framework atoms of the hypothetical MOFs designed in this work.³¹ Additional details on the structure optimization and charge generation process are provided in the Supporting Information.

2.3 Diversity Analysis

To analyze the diversity of MOF databases, we used a set of descriptors to quantify the similarity of MOF structures. Since both pore geometry and material chemistry are important in gas separation applications, we need descriptors for both aspects. Several material

descriptors have been developed to characterize different aspects of the similarity of MOF materials.^{32–35} We used classic geometric characteristics, such as the largest included sphere, surface area, density, and pore volume to describe the pore geometry. These descriptors were computed using Zeo++.^{36,37} We described the chemistry of MOF structures using revised-autocorrelations (RACs). RACs are the product or difference of atomic heuristics, e.g., Pauling electronegativity, connectivity, and covalent radii, computed on a molecular or crystal graph.³⁸ While RACs were initially introduced for machine-learning open shell transition metal complex properties,^{38–40} they were recently adapted to MOF chemistry²¹ and shown to be successful in capturing structure-property relationships for gas adsorption²¹ and photoelectronic properties (e.g., color) of MOFs.⁴¹ In this approach, the MOF structure is described with three groups of features, describing the metal centres, organic linkers, and the functional groups. In total, 156 RACs descriptors were computed using the molSimplify package^{39,42} to describe the chemistry of a MOF structure.

We computed variety, balance, and disparity to assess the diversity of the material databases. The diversity metrics were calculated for each aspect of the material chemistry, that includes chemistry of metals, linkers, functional groups, and the pore geometry. The chemical and geometric descriptors construct high-dimensional feature spaces. We first split these high-dimensional spaces into 1000 bins using k -means clustering method. In this approach, the structures were assigned to their closest centroid. Then, the three diversity metrics were computed using this binning. The variety is computed as the percentage of all the bins sampled by a given database. We used Pielou’s evenness,⁴³ which measures how even the structures are distributed among the sampled bins, as a measure of the balance. To compute disparity, we computed the covered area of the concave hull by a database in the map of the first two principal components. We normalized this number with the area of all databases together. The covered area was computed using the Shapely package with circumference to area ratio cutoff of 1.⁴⁴ A detailed description of the material descriptors and diversity analysis can be found in the previous work.²¹

2.4 Property Calculation

The pore limiting diameter and blocking spheres for each MOF were calculated using Zeo++.^{36,37} For blocking spheres, we considered spherical probes with diameters of 3.05 Å for CO₂ (oxygen’s sigma in TraPPE), 3.31 Å for N₂ (nitrogen’s sigma in TraPPE) and 2.96 Å for H₂ (hydrogen’s sigma in the Buch force field⁴⁵) respectively. The force field parameters for the framework atoms were extracted from UFF.²⁸ CO₂ and N₂ molecules were described by the TraPPE force field⁴⁶ and H₂ was described by the Buch force field⁴⁵ with the Feynman Hibbs correction⁴⁷ (see Supporting Information for the full list of parameters used). The gas-framework interactions were modeled using Lennard Jones potential, truncated at 12 Å (for CO₂ and N₂) and 12.8 Å (for H₂), with tail corrections.⁴⁸ The Lennard Jones interactions between dissimilar atoms were approximated using Lorentz-Berthelot rules.⁴⁹ The Coulombic electrostatic interactions were computed using Ewald summation. The gas adsorption calculations were performed in RASPA.⁵⁰ Grand-canonical Monte Carlo (GCMC) simulations were used to compute the gas uptake of the MOFs. Each calculation consisted of 10,000 initialization cycles followed by 10,000 equilibration cycles. The pure component CO₂ uptakes were calculated at 1 bar and 298 K. We also calculated the uptakes of CO₂ and N₂ for a binary mixture of CO₂ and N₂ in the ratio of 0.15:0.85. For the binary mixture, we considered the flue gas to be adsorbed at 1 bar and 298 K, and regenerated at 0.1 bar and 363 K. These conditions have been used in several studies for post combustion CO₂ capture.^{12,51–53} The H₂ uptakes were calculated at 100 bar and 77 K. These conditions have been used in several studies for hydrogen storage.^{9,25}

3 Results and Discussion

3.1 Diversity Analysis

Using the workflow as described in the Methods section, a database of $\sim 20,000$ MOFs were generated. We took the combination of all synthesized MOF structures and hypothetical MOF structures as the current chemical space of MOFs. This chemical space has been described using the high dimensional pore geometry and chemistry feature vectors, and we have thus made a projection of it on two dimensions to visualize which regions of the material design space our hypothetical MOFs are covering. For experimental structures, we have considered the CoRE-2019 database.³ For hypothetical structures, we have considered the database developed by Anderson et al.,⁵⁴ the ToBaCCo database,¹³ a diverse subset of 20,000 structures from the database developed by Boyd and Woo.¹² These databases were used for the analysis in the work by Moosavi et al.²¹

Figure 2 shows a dimensionality reduction visualization of all hypothetical MOF databases, including the database we have developed in this study, when overlaid on the total set of all experimental and hypothetical MOF databases. The distributions of the databases are shown with respect to their pore geometry, metal chemistry, linker chemistry and functional groups. For pore geometry, linker chemistry and functional groups, the hypothetical databases are covering and sampling the design space well. For metal chemistry, we find that the sampling of the design space has improved on including the structures from this study, when compared to the previous distribution reported in Moosavi et al.²¹ This overall improvement of the diversity in metal chemistry has been quantified below (Table 1).

Figure 3 shows which regions of the material design space we have specifically contributed to, through database of this study. As discussed in the Methods section, for organic linkers, functional groups and topologies, we have selected them from the respective libraries reported in literature. But for metal nodes, we have tried to focus on the ones which have not been commonly used in the other hypothetical databases mentioned in this study. Thus, if we

look at the metal chemistry map, we find that our metal nodes are complementing different regions of the space as compared to the previously used metal nodes in the other hypothetical databases.²¹ And when we combine all these metal nodes used in all the hypothetical databases together, we get the map of metal chemistry as shown in Figure 2.

Figure 2: Visualization of the material design space. The t-Distributed Stochastic Neighbour Embedding (t-SNE)⁵⁵ method was used to project the pore geometry, metal chemistry, linker chemistry, and functional groups descriptor spaces to 2D maps. The t-SNE method preserves pairwise distances, ensuring similar structures are mapped close to each other in two dimensions. (See principal component analysis (PCA) figures in Supporting Information for the global similarities.) Only descriptors up to the second coordination shell were included for metal chemistry to emphasize the local metal chemistry environment. In each panel, the structures from all the hypothetical databases considered in this study were colored and overlaid on the entire known design space represented in gray.

Figure 3: Visualization of the material design space. The t-SNE method was used to project the pore geometry, metal chemistry, linker chemistry, and functional groups descriptor spaces to 2D maps. Only descriptors up to the second coordination shell were included for metal chemistry to emphasize the local metal chemistry environment. In each panel, the structures from the hypothetical database developed in this study were colored and overlaid on the entire known design space represented in gray.

We have quantified the diversity of the databases in terms of their variety, balance and disparity (Table 1). The variety of a database indicates how many distinct types of structures exist in our database. Balance indicates how even the distribution of structures are. Disparity of a database reflects how much dissimilar or distinct the structures of our database are. A high disparity would thereby indicate that we have structures from far apart points in the material design space. We thus calculated these metrics for the hypothetical databases in two scenarios: before adding the database of this study and after adding the database of this study.

For the geometric features of the hypothetical databases, we find a slight increase in the variety and disparity and a slight decrease in the balance. For the metal center features, we

see that on adding the structures from the database of this study, the variety of structures have improved. The balance of the structures decreases slightly, and the disparity of the structures also improves. This gives us an indication that the overall diversity of structures with respect to the metal chemistry has improved upon adding the structures from the database of this study. For the linker chemistry, as like the geometric features, we see a slight increase in the variety and disparity, and a decrease in the balance. And for functional groups, we see an improvement in all the three diversity metrics.

Table 1: Diversity metrics for the different features of hypothetical databases. We first splitted the high-dimensional spaces into 1000 bins using k-means clustering method. Variety measures the percentage of all the bins sampled by a given database. Balance measures the evenness of the distribution of the structures among the sampled bins. And disparity the spread of the sampled bins. We normalized these number with the area of all databases together.

Feature	Hypothetical databases	Variety	Balance	Disparity
Geometric	Excluding this study	0.977	0.849	0.874
	Including this study	0.988	0.775	0.933
Metal Center	Excluding this study	0.068	0.334	0.078
	Including this study	0.107	0.296	0.104
Linker Chemistry	Excluding this study	0.648	0.617	0.737
	Including this study	0.684	0.446	0.798
Functional Group	Excluding this study	0.722	0.213	0.782
	Including this study	0.851	0.323	0.834

Now that we have designed a diverse set of hypothetical MOF structures, our next aim was to see if we could use them in some practical applications.

3.2 Post-Combustion Carbon Capture

Figure 4 shows the distribution of the hypothetical MOFs of the current study for the uptake of pure CO₂ at 1 bar and 298 K. A reference line has been drawn to denote the pure CO₂ uptake of Zeolite-13X, which is often used as a benchmark CO₂ adsorbent.¹² From

the distribution, we find that there are many structures which perform as well as Zeolite-13X—pure CO₂ uptake of $\sim 5 \text{ mmol g}^{-1}$,^{51,56} and there are also many structures—around 800—which surpass the performance of Zeolite-13X.

Figure 4: Results from the computational screening of $\sim 20,000$ MOFs for post-combustion carbon capture (pure CO₂ adsorption at 1 bar and 298 K). This plot shows the distribution of the pure CO₂ uptake of the MOFs. The blue reference line denotes the pure CO₂ uptake of Zeolite-13X.

We further investigated the structures for their performance in separating CO₂ from flue gas. For this, we considered a binary mixture of CO₂ and N₂. We then calculated the CO₂ working capacity and CO₂/N₂ selectivity of these hypothetical structures. Figure 5 shows how the structures perform. Again, we find that there are many structures—around 250—which surpass the performance of Zeolite-13X at dry conditions—CO₂ working capacity greater than $\sim 2 \text{ mmol g}^{-1}$ and CO₂/N₂ selectivity greater than ~ 50 .^{12,57}

Figure 5: Results from the computational screening of $\sim 20,000$ MOFs for post-combustion carbon capture (15:85 CO_2/N_2 mixture with adsorption at 1 bar and 298 K, and regeneration at 0.1 bar and 363 K). This plot shows the CO_2 working capacity versus CO_2/N_2 selectivity of the MOFs. The color coding represents the number of MOFs according to the color bar on the right. The blue reference lines denote the CO_2 working capacity and CO_2/N_2 selectivity of Zeolite-13X.

Figure 6: Pure CO_2 uptake distribution for MOFs of different metal nodes at 1 bar and 298 K. a. Distribution of MOFs of node mn1 (left) and the structure of node mn1 (right). b. Distribution of MOFs of node mn3 (left) and the structure of node mn3 (right). c. Distribution of structures of node mn2 (left) and the structure of node mn2 (right). d. Distribution of structures of node mn12 (left) and the structure of node mn12 (right). The number of bins for the structures of metal node mn2 are less than the other nodes because fewer structures were generated with this node mn2 (as it was one of the last metal nodes to be added in our database). The blue reference lines denote the pure CO_2 uptake of Zeolite-13X.

Based on diversity analysis, Moosavi et al.²¹ concluded that for CO₂ adsorption at low pressures, metal chemistry as a factor cannot be ignored. To illustrate the importance of using different metal nodes in our database and to have a look at how some of the metal nodes in our study performed, we plotted the pure CO₂ uptake at 1 bar and 298 K for some of the metal nodes used in this study. For example, Figure 6 shows the distribution of the pure CO₂ uptake for the metal node mn1—a Ni based metal node, mn3—a Cu based metal node, mn2—a Zn based metal node and mn12—a Ni based metal node. Nodes mn1, mn2, and mn3 have similar tetranuclear metal clusters. But in mn3, Cu forms a 5-connected cluster—Cu₄(μ₃—OH)₂(COO)₅,⁵⁸ and in mn1 and mn2, Ni and Zn form a 6-connected cluster—Ni₄(μ₃—OH)₂(COO)₆ and Zn₄(μ₃—OH)₂(COO)₆, respectively.^{59,60} Many of the structures of this study containing nodes mn1 and mn2 have pure CO₂ uptakes above 5 mmol g⁻¹ while hardly any of the structures containing node mn12 have pure CO₂ uptakes above 5 mmol g⁻¹. This shows that even if we have metal nodes of similar geometry, the type of metal can impact the formation of a node with different connectivities and different CO₂ adsorption characteristics of the MOF.

If we then compare the distribution of the nodes mn1 and mn12, both metal nodes are made of Ni but have different geometries. In mn1, Ni forms a tetranuclear cluster while in mn12, Ni forms a 6-connected triangular cluster, Ni₃(μ₃—OH)₂(COO)₆.⁶¹ And the structures made of both of these metal nodes perform quite well in their CO₂ uptakes. This shows how the same metal can form two nodes of different geometries and that the structures of both the nodes could be promising for CO₂ capture. Capturing these variations in the metal nodes is important because when we include all these metal nodes, we get the final distribution as shown in Figure 4, which is very different from the individual distributions. And the presence of these different metal nodes help in obtaining many high performing structures. This would also help to choose from a wider range of metal nodes while synthesizing new MOFs for carbon capture.

We find that most of the top performing structures contain the metal nodes mn1 (Ni

based)— $\sim 40\%$ of the top performing structures, mn12 (Ni based)— $\sim 30\%$, mn13 (Co based)— $\sim 8\%$, and mn2 (Zn based)— $\sim 5\%$. For the linkers, we find the top performing structures to have simple 2-coordinated linkers like benzene dicarboxylic acids (BDC) to more complicated 3-coordinated linkers like benzene-1,3,5-tricarboxylic acid and triazine to further complicated 6-coordinated linkers like bicyclooctanes and a combination of different linkers. Figure 7 shows the structure of one of the high performing MOFs for post-combustion carbon capture.

Figure 7: Structure of a top performing MOF for post-combustion carbon capture—ddmof_559—metal node mn1 + organic edge oe31 + topology **snk**

3.3 Hydrogen Storage

Figure 8 shows a plot of the gravimetric uptake versus the volumetric uptake of H_2 in our structures at 100 bar and 77 K. This plot shows a volcano type relationship between the two type of uptakes, as observed in previous studies.^{9,25} A reference line has been drawn to denote the gravimetric H_2 uptake— $9.20 \text{ wt}\%$ ²⁵ and volumetric H_2 uptake— 52.64 g L^{-1} ²⁵ in MOF-5, a widely used benchmark material for H_2 storage selected by the Hydrogen Storage Engineering Centre of Excellence (HSECoE).^{25,62,63} We find that many structures from our database have a gravimetric uptake higher than that of MOF-5. An ideal H_2 adsorbent should

however exhibit a balance between high gravimetric uptake and high volumetric uptake.²⁵ This is because the volumetric uptake of the H₂ storage system has a greater impact on the driving range of fuel cell vehicles (FCVs) than the gravimetric uptake.^{25,62–65} And in this respect, we find around 50 MOFs from our database which outperform MOF-5. Figure 9 shows the structure of one such promising hypothetical MOF of this study for hydrogen storage.

Figure 8: Results from the computational screening of $\sim 20,000$ MOFs for hydrogen storage (pure H₂ adsorption at 100 bar and 77 K). The colour coding represents the number of MOFs according to the colour bar on the right. The blue reference lines denote the volumetric H₂ uptake and gravimetric H₂ uptake of MOF-5.

Figure 9: Structure of a top performing MOF for hydrogen storage—ddmof_6749—metal node mn13 + organic node on1 + organic edges oe33, oe68 + **tfz-d** topology.

Here, we find that almost all of the top performing structures have a **tfz-d** topology. This is also shown in Figure 10. Metal nodes which form structures in these topologies are thereby more prevalent in these top 50 structures. In this study, these are mainly 8 connected metal nodes like mn13 (Co based)—~85% of the top-performing structures and mn4 (Cd based)—~15% of the top-performing structures. The exploration of different topologies with these metal nodes led to generating MOFs with different pore geometries which finally led to some of these structures to be promising for hydrogen storage. This again highlights the importance of having a diverse database. Here, we did not pre-bias the structures for a particular application. We tried to make a diverse set of structures, which covers different aspects of MOF chemistry. Some features of these structures play an important role in one application and some features in other applications. And therefore, some of these structures turn out to be good for one application and some for other. In this case of hydrogen storage, it is the topology of a MOF which is playing a more important role in forming a top-performing storage than metal chemistry.

Figure 10: Results from the computational screening of $\sim 20,000$ MOFs for hydrogen storage (pure H_2 adsorption at 100 bar and 77 K). The structures with the **tfz-d** topology are highlighted in blue and all the remaining structures are in grey

4 Conclusion

In this work, we have designed a database of $\sim 20,000$ hypothetical MOFs with the aim to increase the chemical diversity of the existing databases. We show that adding the structures of our database improves the overall diversity metrics of hypothetical databases, especially in terms of metal chemistry.

To highlight the usefulness of these diverse structures, we evaluated their performance for two important environmental applications—post-combustion carbon capture and hydrogen storage. In the case of post-combustion carbon capture, we find that many of these structures outperform Zeolite-13X, a widely used benchmark material for carbon capture, in terms of their pure CO_2 uptake—around 800 structures, CO_2 working capacity and CO_2/N_2 selectivity—around 250 structures. For hydrogen storage, we find around 50 structures

which outperform MOF-5, a widely used benchmark material for hydrogen storage, in terms of their balance between gravimetric uptake of H_2 and volumetric uptake of H_2 . For post-combustion carbon capture, we find that including different metal nodes help in obtaining high performing structures. In the case of hydrogen storage, we find that it is the topology of the MOF which plays the more dominant role for a structure to be high-performing. The promising structures found in this study could be added to the existing list of promising structures in literature and this would provide us with a more diverse range of materials to choose from, for synthesizing.

Through this work, we thus show that on starting with a relatively small but diverse set of materials one could still obtain interesting materials for different applications. This would help us to locate the interesting regions of the material space. To avoid brute-force screening of an infinite number of possible MOFs, as a next step, one could then explore around these interesting regions using active learning, Bayesian optimization,^{66,67} or generative models.^{68,69}

Supporting Information Available

Additional details for structure generation—list of metal nodes, organic linkers, functional groups and topologies used to generate the hypothetical MOFs in this study; additional details for structure optimization and charge generation calculations; force field parameters used for grand-canonical Monte Carlo simulations; additional figures for diversity analysis; additional figures for hydrogen storage.

Acknowledgments

SM would like to acknowledge Peter G. Boyd for his help and guidance in the structure generation process. This research was supported by the European Research Council (ERC) under the European Union’s Horizon 2020 research and innovation programme

(grant agreement 666983, MaGic), by the NCCR-MARVEL, funded by the Swiss National Science Foundation, and by the PrISMa Project (299659), funded through the ACT Programme (Accelerating CCS Technologies, Horizon 2020 Project 294766). Financial contributions from the Department for Business, Energy & Industrial Strategy (BEIS) together with extra funding from the NERC and EPSRC Research Councils, United Kingdom, the Research Council of Norway (RCN), the Swiss Federal Office of Energy (SFOE), and the U.S. Department of Energy are gratefully acknowledged. Additional financial support from TOTAL and Equinor is also gratefully acknowledged. All of the MOF structures developed in this study and their properties are available on the Materials Cloud platform⁷⁰ at <https://doi.org/10.24435/materialscloud:yn-de>. The structures and properties of the experimental and hypothetical MOFs used for diversity analysis were collected from the Materials Cloud platform at <https://doi.org/10.24435/materialscloud:3y-gr>.

References

- (1) Eddaoudi, M.; Kim, J.; Rosi, N.; Vodak, D.; Wachter, J.; O’Keeffe, M.; Yaghi, O. M. Systematic Design of Pore Size and Functionality in Isoreticular MOFs and Their Application in Methane Storage. *Science* **2002**, *295*, 469–472.
- (2) 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. *Chem. Mater.* **2017**, *29*, 2618–2625.
- (3) Chung, Y.; Haldoupis, E.; Bucior, B.; Haranczyk, M.; Lee, S.; Zhang, H.; Vogiatzis, K.; Milisavljevic, M.; Ling, S.; Camp, J.; Slater, B.; Siepmann, J.; Sholl, D.; Snurr, R. Advances, Updates, and Analytics for the Computation-Ready, Experimental Metal-Organic Framework Database: CoRE MOF 2019. *J. Chem. Eng.* **2019**, *64*, 5985–5998.
- (4) Ongari, D.; Talirz, L.; Smit, B. Too Many Materials and Too Many Applications: An Experimental Problem Waiting for a Computational Solution. *ACS Cent. Sci.* **2020**, *6*, 1890–1900.
- (5) Sculley, J.; Yuan, D.; Zhou, H.-C. The current status of hydrogen storage in metal–organic frameworks—updated. *Energy Environ. Sci.* **2011**, *4*, 2721–2735.
- (6) Farha, O. K.; Shultz, A. M.; Sarjeant, A. A.; Nguyen, S. T.; Hupp, J. T. Active-Site-Accessible, Porphyrinic Metal Organic Framework. *J. Am. Chem. Soc.* **2011**, *133*, 27.
- (7) Simon, C. M.; Mercado, R.; Schnell, S. K.; Smit, B.; Haranczyk, M. What Are the Best Materials To Separate a Xenon/Krypton Mixture? *Chem. Mater.* **2015**, *27*, 4459–4475.
- (8) Sikora, B. J.; Wilmer, C. E.; Greenfield, M. L.; Snurr, R. Q. Thermodynamic analysis of Xe/Kr selectivity in over 137000 hypothetical metal–organic frameworks. *Chem. Sci.* **2012**, *3*, 2217–2223.

- (9) Colón, Y. J.; Gómez-Gualdrón, D. A.; Snurr, R. Q. Topologically Guided, Automated Construction of Metal-Organic Frameworks and Their Evaluation for Energy-Related Applications. *Cryst. Growth Des.* **2017**, *17*, 5801–5810.
- (10) Boyd, P. G.; Lee, Y.; Smit, B. Computational development of the nanoporous materials genome. *Nat. Rev. Mater.* **2017**, *2*, 1–15.
- (11) Moosavi, S. M.; Jablonka, K. M.; Smit, B. The Role of Machine Learning in the Understanding and Design of Materials. *J. Am. Chem. Soc.* **2020**, *142*, 20273–20287.
- (12) Boyd, P. G., et al. Data-driven design of metal–organic frameworks for wet flue gas CO₂ capture. *Nature* **2019**, *576*, 253–256.
- (13) Evaluating topologically diverse metal-organic frameworks for cryo-adsorbed hydrogen storage †. *Energy Environ. Sci.* **2016**, *9*, 3279–3289.
- (14) Wilmer, C. E.; Leaf, M.; Lee, C. Y.; Farha, O. K.; Hauser, B. G.; Hupp, J. T.; Snurr, R. Q. Large-scale screening of hypothetical metal-organic frameworks. *Nat. Chem.* **2012**, *4*.
- (15) Lee, S.; Kim, B.; Cho, H.; Lee, H.; Lee, S. Y.; Cho, E. S.; Kim, J. Computational Screening of Trillions of Metal–Organic Frameworks for High-Performance Methane Storage. *ACS Appl. Mater. Interfaces* **2021**, *13*, 23647–23654.
- (16) Wilmer, C. E.; Kim, K. C.; Snurr, R. Q. An Extended Charge Equilibration Method. *J. Phys. Chem. Lett.* **2012**, *3*.
- (17) Gomez-Gualdrón, D.; Gutov, O.; Krungleviciute, V.; Borah, B.; Mondloch, J.; Hupp, J.; Yildirim, T.; Farha, O.; Snurr, R. Computational design of metal-organic frameworks based on stable zirconium building units for storage and delivery of methane. *Chem. Mater.* **2014**, *26*, 5632–5639.

- (18) Boyd, P. G.; Woo, T. K. A generalized method for constructing hypothetical nanoporous materials of any net topology from graph theory. *CrystEngComm.* **2016**, *18*, 3777–3792.
- (19) Gómez-Gualdrón, D. A.; Colón, Y. J.; Zhang, X.; Wang, T. C.; Chen, Y.-S.; Hupp, J. T.; Yildirim, T.; Farha, O. K.; Zhang, J.; Snurr, R. Q. Evaluating topologically diverse metal–organic frameworks for cryo-adsorbed hydrogen storage. *Energy Environ. Sci.* **2016**, *9*, 3279–3289.
- (20) Groom, C. R.; Bruno, I. J.; Lightfoot, M. P.; Ward, S. C. The Cambridge Structural Database. *Acta. Crystallogra. B* **2016**, *72*, 171–179.
- (21) Moosavi, S. M.; Nandy, A.; Jablonka, K. M.; Ongari, D.; Janet, J. P.; Boyd, P. G.; Lee, Y.; Smit, B.; Kulik, H. J. Understanding the diversity of the metal-organic framework ecosystem. *Nat. Commun.* **2020**, *11*, 1–10.
- (22) Bucior, B. J.; Rosen, A. S.; Haranczyk, M.; Yao, Z.; Ziebel, M. E.; Farha, O. K.; Hupp, J. T.; Siepmann, J. I.; Aspuru-Guzik, A.; Snurr, R. Q. Identification schemes for metal–organic frameworks to enable rapid search and cheminformatics analysis. *Cryst. Growth Des.* **2019**, *19*, 6682–6697.
- (23) Bui, M., et al. Carbon capture and storage (CCS): the way forward. *Energy Environ. Sci.* **2018**, *11*, 1062–1176.
- (24) Smit, B.; Reimer, J. A.; Oldenburg, C. M.; Bourg, I. C. *Introduction to carbon capture and sequestration*; World Scientific, 2014; Vol. 1.
- (25) Ahmed, A.; Liu, Y.; Purewal, J.; Tran, L. D.; Wong-Foy, A. G.; Veenstra, M.; Matzger, A. J.; Siegel, D. J. Balancing gravimetric and volumetric hydrogen density in MOFs. *Energy Environ. Sci.* **2017**, *10*, 2459–2471.
- (26) Thornton, A. W., et al. Materials genome in action: identifying the performance limits of physical hydrogen storage. *Chem. Mater.* **2017**, *29*, 2844–2854.

- (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**, *41*, 1782–1789.
- (28) Rappé, A. K.; Casewit, C. J.; Colwell, K.; Goddard III, W. A.; Skiff, W. M. UFF, a full periodic table force field for molecular mechanics and molecular dynamics simulations. *J. Am. Chem. Soc.* **1992**, *114*, 10024–10035.
- (29) Plimpton, S. Fast Parallel Algorithms for Short-Range Molecular Dynamics. *J. Comput. Phys.* **1995**, *117*, 1–19.
- (30) Boyd, P. G.; Mohamad Moosavi, S.; Witman, M.; Smit, B. Force-Field Prediction of Materials Properties in Metal-Organic Frameworks. *J. Phys. Chem. Lett.* **2017**, *8*, 357–363.
- (31) Ongari, D.; Boyd, P. G.; Kadioglu, O.; Mace, A. K.; Keskin, S.; Smit, B. Evaluating charge equilibration methods to generate electrostatic fields in nanoporous materials. *J. Chem. Theory Comput.* **2018**, *15*, 382–401.
- (32) Jablonka, K. M.; Ongari, D.; Moosavi, S. M.; Smit, B. Big-data science in porous materials: materials genomics and machine learning. *Chem. Rev.* **2020**, *120*, 8066–8129.
- (33) Fernandez, M.; Boyd, P. G.; Daff, T. D.; Aghaji, M. Z.; Woo, T. K. Rapid and accurate machine learning recognition of high performing metal organic frameworks for CO₂ capture. *J. Phys. Chem. Lett.* **2014**, *5*, 3056–3060.
- (34) Lee, Y.; Barthel, S. D.; Dłotko, P.; Moosavi, S. M.; Hess, K.; Smit, B. Quantifying similarity of pore-geometry in nanoporous materials. *Nat. Commun.* **2017**, *8*, 1–8.
- (35) Moosavi, S. M.; Xu, H.; Chen, L.; Cooper, A. I.; Smit, B. Geometric landscapes for

- material discovery within energy–structure–function maps. *Chem. Sci.* **2020**, *11*, 5423–5433.
- (36) Ongari, D.; Boyd, P. G.; Barthel, S.; Witman, M.; Haranczyk, M.; Smit, B. Accurate characterization of the pore volume in microporous crystalline materials. *Langmuir* **2017**, *33*, 14529–14538.
- (37) 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**, *149*, 134–141.
- (38) Janet, J. P.; Kulik, H. J. Resolving transition metal chemical space: Feature selection for machine learning and structure–property relationships. *J. Phys. Chem. A* **2017**, *121*, 8939–8954.
- (39) Nandy, A.; Duan, C.; Janet, J. P.; Gugler, S.; Kulik, H. J. Strategies and Software for Machine Learning Accelerated Discovery in Transition Metal Chemistry. *Ind. Eng. Chem. Res.* **2018**, *57*, 13973–13986.
- (40) Nandy, A.; Duan, C.; Taylor, M. G.; Liu, F.; Steeves, A. H.; Kulik, H. J. Computational Discovery of Transition-metal Complexes: From High-throughput Screening to Machine Learning. *Chem. Rev.* **2021**,
- (41) Jablonka, K. M.; Moosavi, S. M.; Asgari, M.; Ireland, C.; Patiny, L.; Smit, B. A data-driven perspective on the colours of metal–organic frameworks. *Chem. Sci.* **2021**, *12*, 3587–3598.
- (42) Ioannidis, E. I.; Gani, T. Z. H.; Kulik, H. J. molSimplify: A toolkit for automating discovery in inorganic chemistry. *J. Comput. Chem.* **2016**, *37*, 2106–2117.
- (43) Pielou, E. C. The measurement of diversity in different types of biological collections. *J. Theor. Biol.* **1966**, *13*, 131–144.

- (44) Shapely: manipulation and analysis of geometric objects. <https://github.com/Toblerity/Shapely>, using March 4, 2021 release throughout this study.
- (45) Buch, V. Path integral simulations of mixed para-D2 and ortho-D2 clusters: The orientational effects. *J. Chem. Phys.* **1994**, *100*, 7610–7629.
- (46) Potoff, J. J.; Siepmann, J. I. Vapor–liquid equilibria of mixtures containing alkanes, carbon dioxide, and nitrogen. *AIChE J.* **2001**, *47*, 1676–1682.
- (47) Feynman, R. P.; Hibbs, A. R.; Styer, D. F. *Quantum mechanics and path integrals*; Courier Corporation, 2010.
- (48) Jablonka, K. M.; Ongari, D.; Smit, B. Applicability of tail corrections in the molecular simulations of porous materials. *J. Chem. Theory Comput.* **2019**, *15*, 5635–5641.
- (49) Lorentz, H. A. About the application of the virial theorem in the kinetic theory of gases. *Ann. Phys.* **1881**, *248*, 127–136.
- (50) Dubbeldam, D.; Calero, S.; Ellis, D. E.; Snurr, R. Q. RASPA: molecular simulation software for adsorption and diffusion in flexible nanoporous materials. *Mol. Simul.* **2016**, *42*, 81–101.
- (51) Hu, Z.; Wang, Y.; Shah, B. B.; Zhao, D. CO2 capture in metal–organic framework adsorbents: an engineering perspective. *Adv. Sustain. Syst.* **2019**, *3*, 1800080.
- (52) Wang, L.; Liu, Z.; Li, P.; Wang, J.; Yu, J. CO2 capture from flue gas by two successive VPSA units using 13XAPG. *Adsorption* **2012**, *18*, 445–459.
- (53) Tlili, N.; Grévillet, G.; Vallières, C. Carbon dioxide capture and recovery by means of TSA and/or VSA. *Int. J. Greenh. Gas Control.* **2009**, *3*, 519–527.
- (54) Anderson, R.; Rodgers, J.; Argueta, E.; Biong, A.; Gómez-Gualdrón, D. A. Role of pore chemistry and topology in the CO2 capture capabilities of MOFs: from molecular simulation to machine learning. *Chem. Mater.* **2018**, *30*, 6325–6337.

- (55) Van der Maaten, L.; Hinton, G. Visualizing data using t-SNE. *J. Mach. Learn. Res.* **2008**, *9*, 2509–2605.
- (56) Cavenati, S.; Grande, C. A.; Rodrigues, A. E. Layered pressure swing adsorption for methane recovery from CH₄/CO₂/N₂ streams. *Adsorption* **2005**, *11*, 549–554.
- (57) Ho, M. T.; Allinson, G. W.; Wiley, D. E. Reducing the cost of CO₂ capture from flue gases using pressure swing adsorption. *Ind. Eng. Chem. Res.* **2008**, *47*, 4883–4890.
- (58) Wang, Y.-X.; Li, Z.-Y.; Zhu, J.; Wang, X.-Q.; Xu, L.-L.; Lachgar, A.; Zhang, J.-J. Four (5, 5)-connected three-dimensional metal organic materials based on pentacarboxylate ligand: Synthesis, structures and characterization. *CrystEngComm.* **2013**, *15*, 6395–6402.
- (59) Liu, F.; Zhang, L.; Wang, R.; Sun, J.; Yang, J.; Chen, Z.; Wang, X.; Sun, D. Five MOFs with different topologies based on anthracene functionalized tetracarboxylic acid: syntheses, structures, and properties. *CrystEngComm.* **2014**, *16*, 2917–2928.
- (60) Li, X.-J.; Jiang, F.-L.; Wu, M.-Y.; Chen, L.; Qian, J.-J.; Zhou, K.; Yuan, D.-Q.; Hong, M.-C. Construction of Two Microporous Metal–Organic Frameworks with flu and pyr Topologies Based on Zn₄ (μ_3 -OH)₂ (CO₂)₆ and Zn₆ (μ_6 -O)(CO₂)₆ Secondary Building Units. *Inorg. Chem.* **2014**, *53*, 1032–1038.
- (61) Liu, S.-J.; Han, S.-D.; Chang, Z.; Bu, X.-H. Cluster-and chain-based magnetic MOFs derived from 3d metal ions and 1, 3, 5-benzenetricarboxylate. *New J. Chem.* **2016**, *40*, 2680–2686.
- (62) Veenstra, M., et al. Ford/BASF/UM activities in support of the hydrogen storage engineering center of excellence. 2015.
- (63) Purewal, J.; Liu, D.; Sudik, A.; Veenstra, M.; Yang, J.; Maurer, S.; Müller, U.;

- Siegel, D. J. Improved Hydrogen Storage and Thermal Conductivity in High-Density MOF-5 Composites. *J. Phys. Chem. C* **2012**, *116*, 20199–20212.
- (64) Mason, J. A.; Veenstra, M.; Long, J. R. Evaluating metal–organic frameworks for natural gas storage. *Chem. Sci.* **2014**, *5*, 32–51.
- (65) *DOE Technical Targets for Onboard Hydrogen Storage for Light-Duty Vehicles*; 2017.
- (66) Jablonka, K. M.; Jothiappan, G. M.; Wang, S.; Smit, B.; Yoo, B. Bias free multiobjective active learning for materials design and discovery. *Nat. Commun.* **2021**, *12*, 1–10.
- (67) Deshwal, A.; Simon, C.; Doppa, J. R. Bayesian optimization of nanoporous materials. **2021**, <https://doi.org/10.33774/chemrxiv-2021-4624n-v2>.
- (68) Yao, Z.; Sánchez-Lengeling, B.; Bobbitt, N. S.; Bucior, B. J.; Kumar, S. G. H.; Collins, S. P.; Burns, T.; Woo, T. K.; Farha, O. K.; Snurr, R. Q.; Aspuru-Guzik, A. Inverse design of nanoporous crystalline reticular materials with deep generative models. *Nat. Mach. Intell.* **2021**, *3*, 76–86.
- (69) Kim, B.; Lee, S.; Kim, J. Inverse design of porous materials using artificial neural networks. *Sci. Adv.* **2020**, *6*, eaax9324.
- (70) Talirz, L., et al. Materials Cloud, a platform for open computational science. *Sci. Data* **2020**, *7*, 1–12.