

# The Rise of Generative AI for Metal–Organic Framework Design and Synthesis

Chenru Duan,<sup>1,2,\*</sup> Aditya Nandy,<sup>1,3,\*</sup> Shyam Chand Pal,<sup>4,5</sup> Xin Yang,<sup>6</sup> Wenhao Gao,<sup>7</sup> Yuanqi Du,<sup>8</sup> Hendrik Kraß,<sup>9</sup> Yeonghun Kang,<sup>10,11</sup> Varinia Bernales,<sup>10,12</sup> Zuyang Ye,<sup>13</sup> Tristan Pyle,<sup>14</sup> Ray Yang,<sup>4,5</sup> Zeqi Gu,<sup>8</sup> Philippe Schwaller,<sup>15</sup> Shengqian Ma,<sup>14</sup> Shijing Sun,<sup>13</sup> Alán Aspuru-Guzik,<sup>9,10,11,12,16,17,18,19</sup> Seyed Mohamad Moosavi,<sup>9</sup> Robert Wexler,<sup>4,5</sup> Zhiling Zheng<sup>1,4,5\*</sup>

<sup>1</sup>These authors contributed equally to this work

<sup>2</sup>Deep Principle, Inc., Cambridge, MA, 02139, United States

<sup>3</sup>Department of Chemical and Biomolecular Engineering, University of California, Los Angeles, Los Angeles, CA 90095, United States

<sup>4</sup>Department of Chemistry, Washington University, St. Louis, MO 63130, United States

<sup>5</sup>Institute of Materials Science & Engineering, Washington University, St. Louis, Missouri 63130, United States

<sup>6</sup>David H. Koch Institute for Integrative Cancer Research, Massachusetts Institute of Technology, Cambridge, MA 02139, United States

<sup>7</sup>Department of Chemical Engineering, Massachusetts Institute of Technology, Cambridge, MA 02139, United States

<sup>8</sup>Department of Computer Science, Cornell University, Ithaca, NY 14853, United States

<sup>9</sup>Department of Chemical Engineering and Applied Chemistry, University of Toronto, Toronto, ON M5S 3E5, Canada

<sup>10</sup>Department of Chemistry, University of Toronto, Toronto, ON, M5S 3H6, Canada

<sup>11</sup>Vector Institute for Artificial Intelligence, Toronto, ON M5G 0C6, Canada

<sup>12</sup>Acceleration Consortium, University of Toronto, Toronto, ON M5G 1X6, Canada

<sup>13</sup>Department of Mechanical Engineering, University of Washington, Seattle, WA 98195, United States

<sup>14</sup>Department of Chemistry, University of North Texas, Denton, Texas 76201, United States

<sup>15</sup>Laboratory of Artificial Chemical Intelligence, École Polytechnique Fédérale de Lausanne, Lausanne 1015, Switzerland

<sup>16</sup>Department of Computer Science, University of Toronto, Toronto, ON M5S 1A1, Canada

<sup>17</sup>Department of Materials Science & Engineering, University of Toronto, 184 College St., Toronto, ON M5S 3E4, Canada

<sup>18</sup>Senior Fellow, Canadian Institute for Advanced Research, 661 University Ave., Toronto, ON M5G 1M1, Canada

<sup>19</sup>NVIDIA, 431 King St W #6th, Toronto, ON M5V 1K4, Canada

\*E-mail: z.z@wustl.edu; duanchenru@gmail.com; aditya.nandy@ucla.edu

## Abstract

Advances in generative artificial intelligence are transforming how metal–organic frameworks (MOFs) are designed and discovered. This Perspective introduces the shift from laborious enumeration of MOF candidates to generative approaches that can autonomously propose and synthesize in the laboratory new porous reticular structures on demand. We outline the progress of employing deep learning models, such as variational autoencoders, diffusion models, and large language model-based agents, that are fueled by the growing amount of available data from the MOF community and suggest novel crystalline materials designs. These generative tools can be combined with high-throughput computational screening and even automated experiments to form accelerated, closed-loop discovery pipelines. The result is a new paradigm for reticular chemistry in which AI algorithms more efficiently direct the search for high-performance MOF materials for clean air and energy applications. Finally, we highlight remaining challenges such as synthetic feasibility, dataset diversity, and the need for further integration of domain knowledge.

## 1. Introduction

Reticular synthesis of metal–organic frameworks (MOFs) epitomizes a materials design paradigm where modular inorganic nodes and organic linkers are linked into crystalline networks to engineer a periodic structure at the atomic level.<sup>1,2</sup> Since the concept of reticular chemistry was proposed after the synthesis of the first MOFs<sup>3,4</sup>, the field has exploded with more than 100,000 distinct MOF structures synthesized to date,<sup>5,6</sup> leading to a wide range of programmable materials with novel properties. Notably, the global MOF market continues to expand, with commercial demand projected to reach several hundred tonnes annually, and multiple companies have already deployed MOFs at industrial scale as solid sorbents.<sup>7,8</sup> However, This immense structural diversity is both a triumph and a challenge: researchers can, in principle, create an almost boundless number of MOFs by swapping building blocks and topologies, yet identifying the optimal material for a given application is very similar to finding a needle in an infinitely large haystack due to the intractable combinatorics of MOF chemical space. Early efforts tackled this search by systematic enumeration.<sup>9,10</sup> While such template-based approaches provided invaluable databases for gas storage and separation studies, they inevitably explored only a tiny fraction of the vast “MOF universe.” Indeed, it is estimated that millions of MOF structures have been predicted *in silico* beyond those made experimentally.<sup>11</sup> Societal needs like carbon capture,<sup>12,13</sup> clean water,<sup>6,14</sup> clean energy,<sup>15–17</sup> and catalysis<sup>18,19</sup> may be met by reticular materials with exceptional performance that lie in uncharted regions of chemical space.<sup>2,20</sup> Thus, the following question arises: *how can we intelligently navigate this exponentially large design space to discover MOFs with transformative properties for desired applications?*

The recent revolution in deep generative artificial intelligence (GenAI) models, ranging from variational autoencoders (VAEs)<sup>21</sup> to generative adversarial networks (GANs)<sup>22</sup>, diffusion models<sup>23</sup> and language models<sup>24</sup>, offers an unprecedented opportunity to “dream up” novel reticular materials *in silico* and subsequently realize them in the laboratory.<sup>2,25–30</sup> Broadly speaking, GenAI models are machine learning models that learn the underlying patterns of data to create new, similar data.<sup>31</sup> While traditional

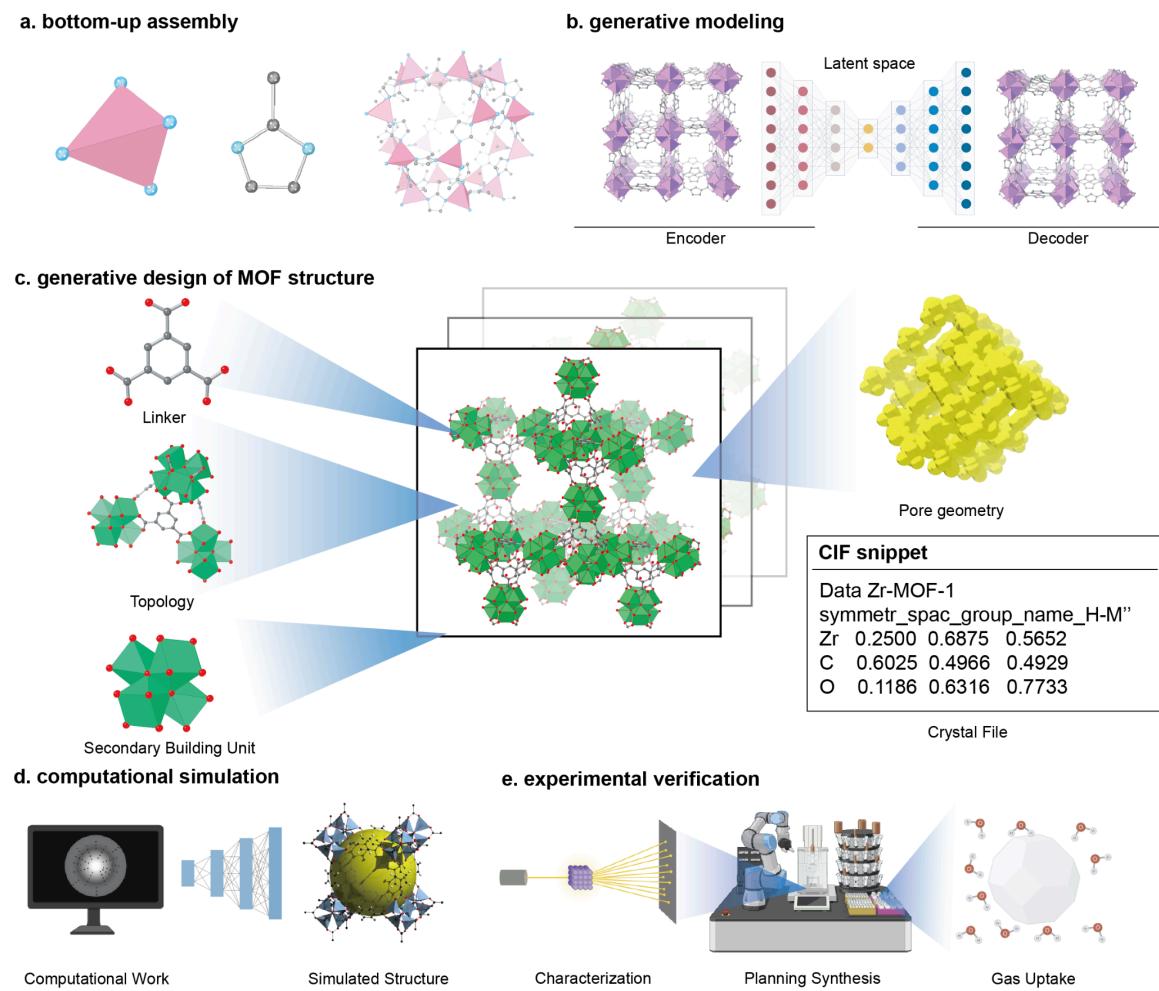
trial-and-error approaches struggle to keep up with the sheer combinatorial possibilities offered by reticular chemistry, generative models offer a way to unlock creativity at the atomic scale, proposing candidates that meet multiple design objectives simultaneously (e.g. high gas uptake and selectivity/stability / reduced cost).<sup>2,11,28,29,32–36</sup> In this Perspective, we explore how generative AI is reshaping MOF discovery, melding computational prediction with experimental realization (**Figure 1**). We first discuss the evolution of generative techniques for crystalline materials and how they were adapted to MOFs. We then highlight specific strategies, which range from fragment-based assembly to 3D diffusion, that address the unique challenges of MOF generation. Next, we examine computational workflows and platforms that integrate generative models with high-throughput screening and even autonomous labs, closing the loop between virtual and real-world discovery. A key focus is to demonstrate how machine imagination can propose MOFs beyond human intuition, how these proposals are vetted through physics-based simulations and experiments, and how a new generation of AI tools and foundation models is accelerating the entire workflow of materials discovery.

## 2. From Enumerative Design to Generative Modeling

In the early 2010s, the dominant computer-assisted approach to discovering new MOFs was enumerative design: researchers built hypothetical MOFs by stitching together known molecular components in all allowable ways. For example, an early study by Wilmer *et al.* assembled MOFs from libraries of secondary building units (SBUs) and linkers according to known nets, producing a “hypothetical MOF” database that could be screened for high methane storage capacity.<sup>9</sup> Such databases yielded valuable insights as top candidates from screening pre-existing building blocks and network architectures were later synthesized and confirmed as adsorbents (**Figure 1**).<sup>37</sup> Efforts were subsequently taken by Colón *et al.* to diversify the enumerated topologies.<sup>38</sup> Such efforts on topology screening have demonstrated success in identifying MOFs with high CO<sub>2</sub> adsorption capacity and concomitant selectivity in a flue gas stream.<sup>39</sup> The next leap was to accelerate the diversity and efficiency of building blocks assembly and potentially remove those topological limitations in order to let an algorithm imagine entirely new MOFs not confined to prior templates. This required a fundamentally different approach: generative models that learn the underlying rules of assembly from data and then extrapolate to create novel structures.

Generative modeling of crystalline materials is a non-trivial problem. Unlike molecules (which can be described by finite graphs) or images (fixed-size pixel grids), a MOF is an infinite periodic structure with long-range order and hundreds to thousands of atoms in a unit cell (**Figure 1b**). Successes in generative AI for materials occurred in similar systems such as zeolites, a class of inorganic porous crystals with analogies to MOFs. In 2020, Kim *et al.* reported an approach to generate zeolite structures using a generative adversarial network, dubbed “ZeoGAN”.<sup>40</sup> They represented zeolites on a 3D grid to capture the pore structure of the crystal and trained a GAN to produce new frameworks with targeted methane adsorption energies. This was one of the first demonstrations that a neural network could learn the “construction rules” of a crystal family and inverse-design materials with user-specified properties. However, extending such methods to MOFs is far more challenging: MOFs contain multiple components (metal oxyhydroxy clusters

and organic linkers), a wider variety of chemistry than aluminosilicate composition of zeolites, and typically larger unit cells. In fact, experimentally synthesized MOFs already contain building blocks derived from over 60 elements, and computationally simulated MOFs can include over 100 atom types, making naive atom-by-atom generation schemes computationally intractable (**Figure 1c**).<sup>29,41,42</sup> Generative models for MOFs therefore required leveraging their modular nature that serves as the core principle of reticular chemistry to simplify the representation and learn insights from both the computational and experimental workflows (**Figure 1d, 1e**).



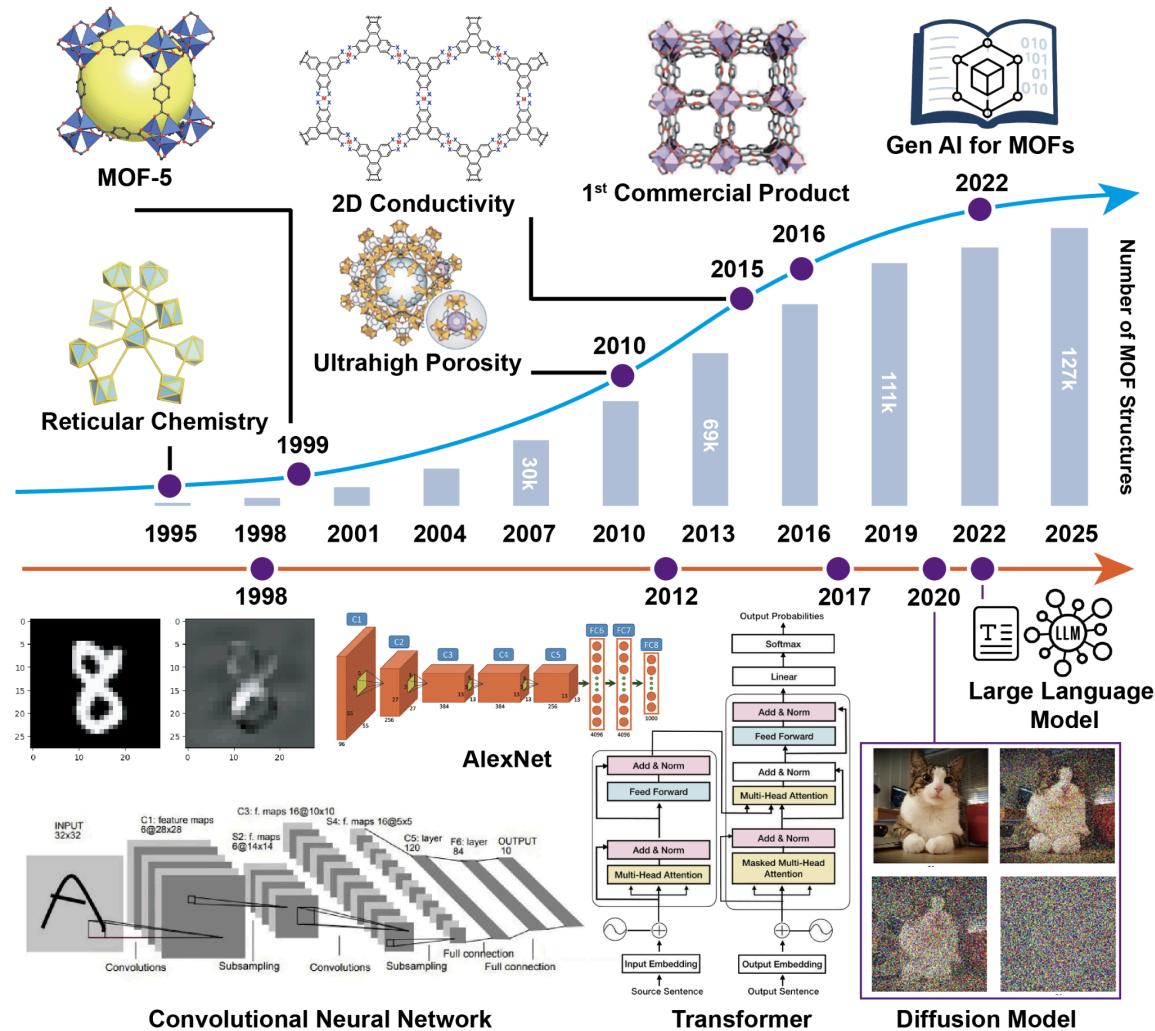
**Figure 1 | Overview of design and characterization principles of MOF.** a) Bottom-up assembly approach starting with the choices of SBUs, linkers, and stoichiometry. b) Generative approach by encoding the MOF structures in a latent space then decoding a latent vector to a MOF structure. c) Generative design of MOF structure files by selectively generating SBUs, linkers, and topologies, with optional conditions on, for example, pore geometry. d) Generated scripts from LLMs to use computational tools for simulating structures and properties of newly constructed new MOFs. e) Generated robotic actions from LLMs for experimental synthesis and measurement for designed new MOFs.

### 3. From 2D Linkers to 3D frameworks design

Early MOF discovery relied heavily on fragment-based 1D or 2D chemical representation databases and human intuition due to the reticular nature of MOFs (**Figure 2**). MOF linkers, when connectivity was considered without 3D geometry, were often specified by IUPAC names or SMILES<sup>43</sup> strings, and researchers searched large databases (e.g. PubChem<sup>44,45</sup>) for compatible organic molecules to serve as linkers for known nodes with different coordination numbers. Thus, conventional approaches for SMILES generation that have been developed for organic chemistry can be directly applied, enabling systematic *in silico* screening of MOF linker fragments and the MOF they can build up for targeted properties, for example, proposing analogues of known structures by utilizing new linkers.<sup>46,47</sup> Translating a 2D blueprint into a 3D crystal structure, however, is a non-trivial step in MOF chemical space due to the complexity of topological nets that connect SBUs and linkers and the periodic nature of crystalline materials. Many MOF structures significantly change upon 3D optimization, with some frameworks collapsing during the solvent removal or some breathing MOFs adjusting their pores volumes.<sup>48–50</sup> On the other hand, these 2D-based workflows often rely on known topologies. Recent advances in generative AI now enable direct design of 3D structures, yielding fully formed MOF crystal structures complete with realistic geometries and spatial constraints without requiring manual assembly or downstream optimization.<sup>51</sup> By learning from known structures and conditioning on desired properties, GenAI models have potentials of accelerating the discovery of novel frameworks with high validity rates and tailor pore environments to specific applications. Together, these models offer unprecedented flexibility for generating 3D MOF structures, reducing reliance on hand-crafted databases and enabling property-driven exploration of chemical space.

From the different components of MOFs that can be altered (i.e. topological net, inorganic node, and organic linker), the linker offers the chemical design space (**Figure 3a**). In contrast, the nuclearity and connectivity of inorganic nodes are largely dictated by reaction conditions and are therefore difficult to control directly. Leveraging this asymmetry, Park *et al.* fine-tuned the DiffLinker (a diffusion model for 3D organic molecule generation) to design organic linkers predicted to boost CO<sub>2</sub> uptake, then inserted them into a prototypical MOF architecture by fixing the metal nodes (e.g., Cu paddlewheels and Zn tetramers) and constraining the topology to the *pcu* net, which is a primitive 3-D cubic network frequently seen in MOF crystal structures.<sup>52</sup> This modular strategy, named GHP-MOFassemle, where GHP stands for generative high-performance, targeted carbon-capture materials without requiring generation of the entire framework at once. A set of 540 molecular fragments that frequently appeared in high-performing MOFs for CO<sub>2</sub> capture was used to allow the model to assemble over 12,000 new linker molecules, which were then used to assemble about 120,000 hypothetical MOFs. While these MOFs have yet to be synthesized and tested experimentally, the study provided clear proof-of-concept that generative AI can rapidly explore the MOF search space and pinpoint designs that would have been unlikely to emerge from human imagination or brute-force enumeration.

Beyond linker-centered design, 3D structures of MOFs can also be directly generated (**Table 1**). Yao *et al.* introduced SmVAE (supramolecular variational autoencoder), which can encode MOFs into latent vectors (**Figure 3b**), through deconstructing them into metal nodes, multiconnected organic nodes, ditopic linkers, and topological nets, and decode them back to full frameworks.<sup>53</sup> About 61.5% of random latent samples are chemically valid, as verified by building-block compatibility and steric-hindrance checks, meaning a majority of the materials generated by the SmVAE were chemically plausible. In addition, an auxiliary predictor was trained to steer generation in the latent space toward generating MOFs with high CO<sub>2</sub>-uptake.



**Figure 2 | Parallel evolution of metal–organic frameworks (top) and artificial intelligence (bottom) from 1995 to 2025.** Stacked bars (right axis) show the cumulative number of MOF crystal structures (1D, 2D, and 3D) deposited in the Cambridge Structural Database (CSD) as of July 2025, where values are given in thousands (K). The upper timeline marks representative innovations in experimental MOF research,<sup>4,54–57</sup> while the lower timeline traces key breakthroughs in machine learning that have inspired new computational approaches to scientific discovery.<sup>23,24,58–61</sup>

With the popularization of diffusion models, Fu *et al.* developed MOFDiff, a coarse-grained diffusion model that constructs MOFs *de novo* by sequentially placing SBU and linker fragments together as rigid 3D objects.<sup>62</sup> A post-processing step aligned these fragments into a 3D periodic crystal, avoiding topological biases and producing approximately 30% novel, chemically valid frameworks. More recently, a model called MOFFUSION introduced a novel way to encode a MOF for generative tasks by using signed distance functions to represent the pore structure of the framework.<sup>63</sup> By training a latent diffusion model on signed distance representations of thousands of MOFs, MOFFUSION learned to generate new signed distances that correspond to plausible MOF structures. These signed distance functions were then decoded into actual MOF atomic structures via a reconstruction step, resulting in high validity of its generated structures. In addition, classifier-free guidance enables conditional generation on numeric (e.g., surface area), categorical (metal, topological net), or textual descriptors, yielding property-conditioned MOF designs.

With joint diffusion that encoded the distribution of multiple building blocks, Duan *et al.* developed building-block-aware (BBA) MOF diffusion,<sup>64</sup> employing an object-aware SE(3)-equivariant diffusion model that learns 3D, all-atom representations of individual building blocks while explicitly encoding crystallographic nets. BBA MOF Diffusion overcame the size constraints of periodic materials generation and readily sampled MOFs of 1000 atoms with high geometric validity, novelty, and diversity (**Figure 3b**). A top-ranked [Zn(1,4-TDC)(EtOH)<sub>2</sub>], where TDC<sup>2-</sup> = thiophenedicarboxylate, was synthesized and confirmed by powder X-ray diffraction (PXRD), thermogravimetric analysis (TGA), and N<sub>2</sub> sorption, demonstrating the practical utility of BBA-MOF diffusion for the design of high-performance MOFs. Building upon these advances, MOFFLOW-2 introduced a flexible two-stage framework that combines autoregressive SMILES-based generation of novel building blocks with a flow-matching model that predicts translations, rotations, and torsions for 3D assembly.<sup>65,66</sup> This enabled the design of chemically diverse MOFs with flexible linkers and accurate structures, overcoming the limitations of fixed libraries and rigid conformations.

**Table 1. Comparison of Key Features and Distinctions of the Major Generative MOF Models.**

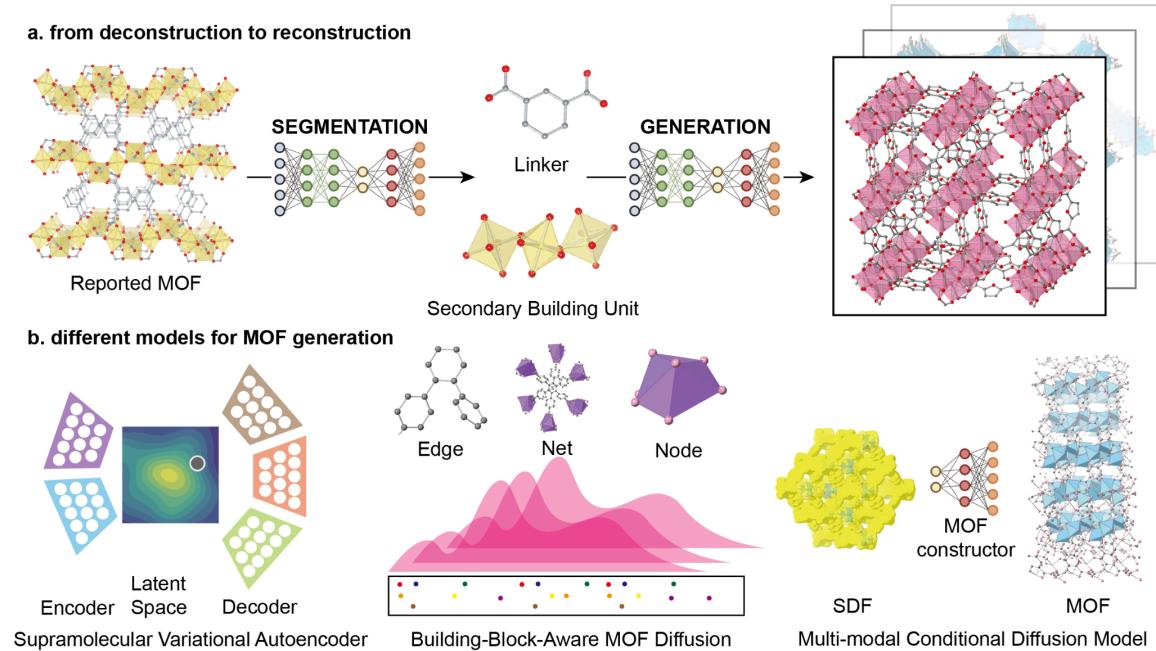
Model	Architecture	Representation	I/O	Target Properties
<b>SmVAE</b>	VAE (2-module encoder-decoder)	Topology + building blocks (discrete)	Input: topology + node/linker IDs; Output: same format.	CO <sub>2</sub> capture (capacity/selectivity)
<b>GHP-MOF assemble</b>	Diffusion (for linker gen) + rule-based assembly	SMILES linkers; fixed metal nodes & pcu net	Input: fragments from high-performing MOFs; Output: novel linker SMILES → assembled MOF.	CO <sub>2</sub> uptake (low-pressure); HPC screening for max capacity

<b>MOFDiff</b>	SE(3)-Equivariant Diffusion	Coarse-grained 3D graph (nodes = building blocks)	Input: noise; Output: block types + 3D coords.	CO <sub>2</sub> working capacity
<b>MOFFUSION</b>	Latent diffusion (VQ-VAE + 3D U-Net)	3D volumetric grid	Input: Signed distance function grid; Output: new MOF signed distance function → atom structure. Conditioning: numeric, categorical, or text prompts	H <sub>2</sub> storage
<b>BBA-Diffusion</b>	Object-aware SE(3)-Equivariant diffusion	All-atom building blocks + net connectivity	Input: noise + topology; Output: full MOF structure in that net. Topology conditioning built-in	High-scoring candidates found via heuristics
<b>MOFFlow2</b>	Autoregressive transformer + Flow-matching	3D building blocks	Input: SMILES of building blocks; Output: transition, rotation, torsion, lattice	Unconditional generation
<b>ChatMOF</b>	Agentic LLM + genetic algorithm	Text ↔ MOF (via internal tools)	Input: natural language; Output: predicted MOF (formula/CIF). Conditioning via prompt	H <sub>2</sub> Uptake
<b>MOFTransformer</b>	Multi-modal transformer encoder	Atom-based graph + methane-probe energy-grid patches	Input: MOF structure; Output: predicted properties via fine-tuned model	Gas uptake, diffusion, band gap, stability metrics

#### 4. From Native Transformers to Large Language Models

In addition to VAEs, GANs, and diffusion models, first introduced in 2017, a different type of architecture called transformers has emerged as a powerful tool for learning patterns in structured domains by using self-attention to capture contextual relationships in sequence data in the field of natural language processing (NLP).<sup>52</sup> Chemistry itself can be viewed as a language since atoms and building blocks follow a grammar making it well-suited to transformer-based modeling.<sup>67</sup> Soon after the development of landmark models like BERT<sup>68</sup> and generative pre-trained transformers (GPT)<sup>63</sup> which attained unprecedented performance in language generation tasks, the successes quickly inspired adaptations in chemistry and materials science, where sequential or graph-based representations of molecules could similarly benefit from attention mechanisms.<sup>24,69</sup> Transformer-based networks have been applied to molecular property prediction and reaction modeling, often outperforming traditional descriptors.<sup>70</sup> Naturally, transformer architectures have also been leveraged for structure–property predictions in porous crystals. A MOFormer model that encodes MOF components as text strings (e.g. “MOFid”<sup>10</sup> identifiers) achieved rapid property predictions without requiring 3D structures by coupling with Crystal Graph Convolutional Neural Networks (CGCNN).<sup>71</sup> Likewise, a multi-modal MOFTransformer pre-trained on over a million hypothetical

MOFs can be fine-tuned to predict diverse properties with improved accuracy over previous machine learning (ML) models,<sup>72</sup> such as gas adsorption isotherms, diffusion coefficients, and band gaps. These models not only accelerated high-throughput screening of MOFs but can also provided some interpretability (e.g. by analyzing attention weights to identify important structural features), though understanding of learned representations remains an active area of research, indicating the promise of transformers as generalizable, data-driven predictors for MOF chemistry.<sup>73,74</sup>



**Figure 3 | Generative design strategies for metal–organic frameworks.** a) Deconstruction-to-reconstruction workflow where reported MOF structures MOF are segmented into its organic linker and inorganic secondary building unit fragments and a generative module subsequently modifies and recombines these fragments to propose new framework topologies and atomic structures. b) Representative deep-learning architectures used for MOF generation. From left to right: a supramolecular variational autoencoder that encodes complete frameworks into a continuous latent space and decodes them back; a building-block-aware diffusion model that samples edges, nets and nodes to assemble MOFs stochastically; and a multi-modal conditional diffusion model that generates signed-distance-function (SDF) volumes which are subsequently converted into full atomic MOF structures.

Beyond prediction tasks, general-purpose large language models (LLMs) have demonstrated remarkable generative capabilities that are now being repurposed for chemical and materials innovation. In their native domain of NLP, transformer-based LLMs can compose fluent passages of text or even generate computer code, illustrating their capacity for open-ended synthesis of information. This generative power has been harnessed in chemistry for tasks like *de novo* molecular design, reaction pathway generation, synthesizability prediction, and even crystal structure proposal.<sup>75–78</sup> For instance, language models fine-tuned on chemical corpora can auto-complete SMILES

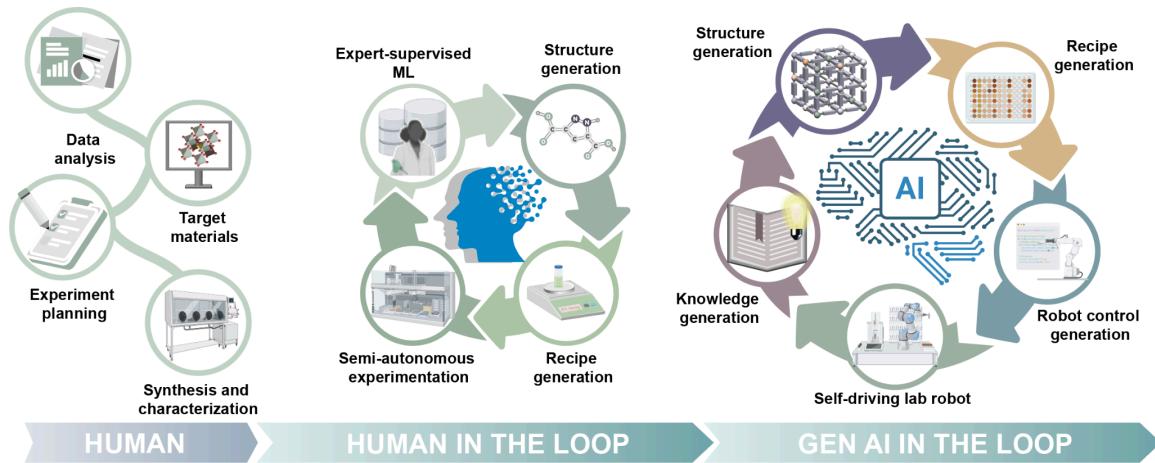
strings to propose novel compounds or suggest multi-step synthetic routes, effectively treating molecular design as a language-generation task.<sup>79</sup> Similarly, language models fine-tuned on SMILES string and IUPAC names of existing reported MOF linkers demonstrate the ability to propose new and valid organic linkers based on human instruction.<sup>80</sup> Furthermore, coupled with evolutionary algorithms, the internal chemistry and materials knowledge and reasoning capability of LLMs were exploited for functional organic molecules,<sup>76</sup> transition metal complexes,<sup>81</sup> and bulk materials design,<sup>77</sup> performing on par or better than conventional genetic algorithms and Bayesian optimization.<sup>82</sup> More recently, a text-guided GenAI model, Chemeleon, conditions a diffusion-based crystal generator on textual descriptions, allowing it to propose novel crystal structures informed by both chemistry language and 3D structural data.<sup>77</sup> Together, these research findings suggest that LLMs can serve not only as natural language interfaces, image creators, or coding assistance popularized for commercial use, but also as creative engines in MOF research for generating new MOF building block combinations, suggesting synthesis plans, and integrating domain knowledge on-the-fly to accelerate the discovery of next-generation MOFs.

## 5. Human-in-the-Loop Workflows for Gen AI-Driven MOF Discovery

Generative modeling provides an unprecedented way of proposing new candidate materials, potentially unlocking chemical spaces unimaginable by enumerative library screening. Generative models alone, however, are only the beginning rather than the end of a materials discovery pipeline (**Figure 4**). Once obtaining new MOF structures, we need to answer crucial questions: *Are these structures physically plausible and synthesizable? Do they have the properties we want? How do we actually make them in the lab?* Therefore, building end-to-end MOF discovery workflows that couple generative models with simulation, optimization, and robotic experimentation are desirable for reticular materials generation with targeted properties.<sup>2,25,83,84</sup>

Traditional MOF discovery relies on manual workflow, where researchers search and interpret literature, plan experiments, and conduct synthesis and characterization.<sup>2,25,26</sup> These approaches, while effective, are limited by human cognitive biases and throughput constraints, making them inadequate for exploring the vast combinatorial MOF chemical space.<sup>85</sup> In contrast, interactive GenAI platforms now let chemists exploit data-driven literature mining without bespoke coding.<sup>86,87</sup> Some early examples of interactive GenAI for MOF research include the ChatGPT Chemistry Assistant<sup>88</sup>, Paragraph2MOFInfo<sup>89</sup>, LLM-NERRE<sup>90</sup>, Eunomia,<sup>91</sup> and L2M3<sup>92</sup>, which helps extract MOF data from the literature, allowing experimentalists to query papers and synthesis procedures conversationally via natural language. By unifying literature-derived insights with computational chemistry data and linking them to crystal structures, LLM-based workflows such as MOF-ChemUnity<sup>93</sup> enable literature-informed AI assistants for scalable, multisource data mining. In the meantime, multimodal generative models with vision capabilities, such as GPT-4V have also shown promise in helping labeling and extracting MOF characterization data from literature figures.<sup>94</sup> On the other hand, agentic-AI is a promising approach to augment LLMs with proper tools for task execution, such as data mining<sup>91</sup> and generation. For example, ChatMOF<sup>95</sup> developed by Kang and Kim demonstrated a tool-augmented system where an LLM agent marshals

tools for database search, property prediction, and structure generation, responding to plain-language requests, such as “*Can you suggest a MOF structure with a pore size around 1.2 nm and high methane storage capacity?*” The agent interfaces with provided generative modules, ranging from genetic algorithms to diffusion models, to propose novel MOF structures.



**Figure 4 | Evolution of design and synthesis practices for MOF innovations.** Schematic illustration of the progressive integration of generative AI into close-looped MOF discovery. The transition from human-led decision-making (left and middle) to AI-driven orchestration is enabled through structure, recipe, instruction, and knowledge generation modules (right).

Beyond data mining and inferencing, GenAI agents have the potential to integrate with enhanced-sampling simulations to form a self-optimizing, closed-loop pipeline for MOF discovery. An autonomous agent could coordinate multi-fidelity strategies (e.g., nested sampling<sup>96</sup>, metadynamics<sup>97</sup>, thermodynamic maps<sup>98</sup>) to act as a physics-informed oracle. Inexpensive coarse evaluations could outline the free-energy landscape, while high-resolution profiles would refine priority frameworks to capture features like sorption-induced phase transitions<sup>99</sup> and gate-opening events<sup>100</sup> that influence usable capacity. This approach mirrors multifidelity Bayesian optimization<sup>101</sup>, filtering candidates efficiently and reserving costly evaluations for top performers. Besides, it is envisioned that thermodynamic observables ( $\Delta H_{\text{ads}}$ ,  $\Delta S_{\text{ads}}$ ) and phase boundaries be introduced to suggest optimal linker-node-topology combinations that aim to maximize gas uptake, selectivity, and structural robustness under realistic operating conditions. Leading candidates could be synthesized by automated platforms and validated through high-throughput characterization, feeding data back into the workflow.

Closer to the experimental end, Yan *et al.* developed MOFA<sup>102</sup>, a workflow that couples a generative linker model (MOFLinker) with automated assembly, simulation-based screening, and on-line retraining, forming a self-improving loop for MOF discovery. They found MOF candidates whose predicted CO<sub>2</sub> uptakes (GCMC at 0.1 bar, 300 K) ranked in the top decile of a 137,652 hypothetical MOF benchmark. Despite these advances, experimental verification remains predominantly manual, with researchers synthesizing and characterizing a limited subset of AI-generated candidates constrained

by synthetic feasibility and available resources.<sup>103–105</sup> Moreover, even successfully synthesized MOFs often require post-synthetic functionalization to achieve optimal performance. Those modifications are difficult to model *in silico* and represent an additional layer of complexity in translating AI-generated designs to functional materials. To address these limitations, echoing developments in LLM-driven organic synthesis (e.g., ChemCrow<sup>106</sup> and Co-Scientist<sup>107</sup> ), MOF research is beginning to incorporate LLM-copilot workflows such as the ChatGPT Research Group<sup>108</sup> and Chemical Robotic Explorer<sup>109</sup> for iterative reaction optimization. These systems reveal that the application of GenAI models extends beyond merely structure generation as they can also generate scripts for synthetic protocols and robotic execution. Recently, Inizan *et al.*<sup>51</sup> developed an agentic AI system to propose and validate synthesizable MOFs. Here, an LLM is first used to propose novel MOF compositions, followed by a diffusion model to generate crystal structures, and quantum chemistry calculations to optimize and filter candidates. Five AI-predicted MOFs were experimentally synthesized with synthesis conditions systematically explored using a robotic high-throughput platform.

As laboratory automation continues to evolve, it is envisioned that GenAI-in-the-loop workflows will be increasingly viable. In such systems, AI agents autonomously generate candidate structures, design synthesis protocols, interface with robotic platforms, and iteratively refine outputs based on experimental feedback. This paradigm represents a shift from static and manual workflows towards adaptive and closed-loop pipelines that can significantly accelerate MOF discovery.

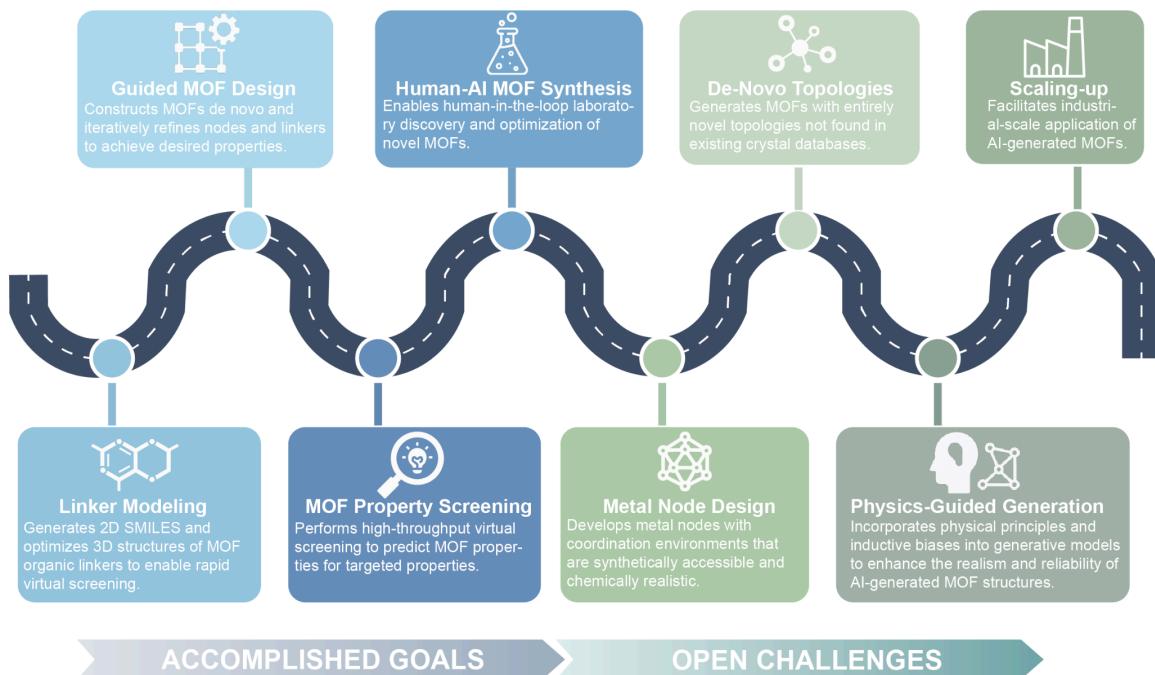
## 6. Outlook

Generative AI is poised to fundamentally reshape how we discover MOFs, transforming the process from a largely manual, intuition-driven craft into a data-rich, AI-accelerated science. To draw an analogy, reticular chemistry gave us the chemical language and grammar to construct MOFs deliberately, rather than by serendipity. Now, generative AI is like a creative agent fluent in that language, brainstorming new sentences (MOF structures) that are grammatically correct (chemically valid) but novel and inventive. The collaboration between humans and machines can amplify innovation: humans set the objectives and constraints inspired by application needs, and AI explores the possibilities at a speed and breadth that is close to what an entire team of human experts can achieve in years.

To fully realize the potential of GenAI, several challenges still remain ahead (**Figure 5**). First, validation bottlenecks will arise as AI generates more candidates than we can reasonably test experimentally. Manually synthesizing and evaluating a predicted MOF can take months, and might result in a failure due to inaccurate scoring functions. This is where improved screening models and autonomous experimentation must step in. Second, experimental constraints pose significant integration challenges. MOF discovery often requires new experimental protocols or customized hardware, limiting the applicability of existing robotic platforms. Developing flexible and modular self-driving laboratories capable of accommodating diverse synthesis routes is therefore critical. In parallel, GenAI must also evolve to generate robotic workflows tailored to the constraints of individual systems. Third, diversity vs. realism will potentially be a perpetual trade-off: pushing a model too hard to explore out-of-distribution MOFs may result in the

model outputting unphysical structures, but constraining it too tightly will lead to recapitulation of known MOFs. This also raises the issue of hallucination, where generative models may confidently create structures or synthetic pathways that are chemically implausible. Addressing these issues requires integrating domain knowledge and human-in-the-loop strategies to penalize unrealistic generations. Future work on techniques like reinforcement learning with human chemist feedback could allow MOF researchers to iteratively refine generative models by giving feedback on small subsets of outputs. Community standards, including unified benchmarking metrics that reflect practical real-world benefits, and sharing of models will also be important. Indeed, many groups have released their code and even pretrained models.<sup>62,63,72,95</sup> Democratizing the toolsets and making them easily accessible (e.g. through language model interfaces) will enable researchers, especially synthetic chemists, who are not AI specialists, to apply generative design to their specific MOF problems.

While most generative models leverage repeatable building blocks to enable scaling to hundreds of atoms, limiting the available substructures to known motifs in the training data hinders the discovery of truly novel structures, including previously unknown linkers, metal nodes, and topologies. Models operating directly on atoms, such as All-Atom Diffusion Transformers<sup>62</sup>, are promising for small molecules and crystals, but lag behind building block-based approaches in terms of validity for MOF generation. An important future direction of research will be combining the flexibility of all-atom methods with the coarse-grained nature of building block-based models, which could furthermore incorporate transfer learning to leverage knowledge about crystals and molecules.



**Figure 5 | Accomplished goals (blue) and open challenges (green) towards realization of generative-AI designed MOFs.**

It is also worthwhile noting that another critical limitation lies in the quality of available structural data. While hypothetical MOF datasets such as hMOF<sup>9,110</sup>, ToBaCCo<sup>38,111,112</sup>, and other *in silico* libraries provide valuable inputs for training generative models, they often lack experimental validation. By contrast, databases like CoRE MOF<sup>113–115</sup> and QMOF<sup>111,112</sup> are rooted in experimental data but may not cover MOFs that are harder to crystalize and can suffer from potential bias of structural false negatives. With this data limitation, how generative models can be designed to generate novel metal nodes and topology is an open challenge. In addition to previously raised concerns about geometric and charge errors,<sup>116</sup> recent work found over 40% of widely used “computation-ready” MOFs contain flawed metal oxidation states,<sup>117</sup> undermining their reliability for DFT-level modeling and generative workflows. These findings underscore the importance of structure validation and repair pipelines that are not just for simulations, but also for ensuring generative models are trained on physically meaningful data. Moreover, MOF chemical data is inherently multimodal and derived from diverse sources. While current efforts in generative AI focus primarily on crystal structure generation, incorporating additional data modalities such as X-ray diffraction patterns could strengthen links between computational modeling and synthesis.<sup>118</sup> Realizing this vision will require holistic datasets that integrate computational results, experimental measurements, and human-curated insights.<sup>93</sup>

One exciting prospect is the integration of physics-based priors into generative models. Imagine a generative model that inherently respects force-field energetics or DFT stabilities, which could be approached by combining energy-based modeling with deep learning, so that the model preferentially generates structures in low-energy configurations.<sup>119,120</sup> Some initial work using score-based generative models with physical constraints is appearing in molecular domains;<sup>121,122</sup> for crystals, the idea of crystal symmetry and energy could be baked into the model’s training objectives. Community datasets that combine atomic coordinates<sup>112</sup> with comprehensive energetics, including enthalpic and entropic contributions, could support transfer-learning workflows for adsorption<sup>74</sup> and catalysis<sup>123</sup>, thereby facilitating each discovery cycle and enriching the collective knowledge base.

Furthermore, this knowledge can also be incorporated post-training. Reinforcement learning-based fine-tuning with physics-based reward models and inference-time control using reward guidance are two promising directions. As available data will be run out for training generative models, inference-time scaling can be key to further improving model performance. As generative models produce candidates that push into unknown chemistry regimes, close collaboration with experimentalists will be essential to identify which suggestions are truly synthesizable. Furthermore, experimentalists will be able to verify the properties of the proposed new materials once obtained. Looking forward, this paradigm shift invites new thinking on how success in GenAI-guided MOF discovery should be defined, whether through the accurate synthesis of predicted structures, the generation of experimentally testable hypotheses, or the continual refinement of foundation models with new knowledge. Establishing robust success metrics will be essential for benchmarking progress and guiding future trajectory of MOF discovery in the GenAI era.

In reflecting on the journey from early initial trial-and-error MOF designs to today's AI-driven discovery, one is reminded of how data-driven insight can complement fundamental knowledge. Reticular chemistry gave us principles like "the topology can be decoupled from the components" and "infinite nets are achievable via strong bonds."<sup>2,124</sup> Generative AI doesn't change those principles; in fact, it leverages them.<sup>25,26</sup> Reticular materials provided an ideal playground for GenAI, as it is a domain with modular chemistry and a wealth of data, and AI, in turn, is giving back to MOF science by revealing how vast and varied that playground really is. As the field progresses, we anticipate seeing AI-designed MOFs setting new records: perhaps an ultraporous MOF that captures and catalytically breaks down methane from agricultural or industrial sources, a water harvesting MOF with high capacity and fast kinetics providing clean water solution for semi-desert areas, or a CO<sub>2</sub>-capturing MOF that works efficiently in humid air and makes direct air capture (DAC) economically viable. In fact, in recent years companies like BASF and Svante are already advancing newly discovered, bench-developed MOFs toward industrial production.<sup>8</sup> In conclusion, the MOF of the future might very well be conceived by an AI, but it will be the human scientists who guide the AI, synthesize the creation, and ultimately implement it in solving real-world problems. In that sense, generative AI is not replacing the reticular chemist. In fact, it is empowering them, much like a powerful new synthesis platform or characterization tool would. The reticular revolution has entered the digital age, and the prospects for innovation have never been more exciting.

### Competing interests

The authors declare no competing interests.

### Acknowledgements

Z.Z. acknowledge the support provided by the EQT Foundation Breakthrough Science Grants. A.N. gratefully acknowledges support from the Eric and Wendy Schmidt AI in Science Postdoctoral Fellowship, a Schmidt Sciences, LLC program. Z.Y. and S.S. acknowledge support by the National Science Foundation through University of Washington Materials Research Science and Engineering Center, DMR-2308979. A.A.-G. thanks Anders G. Frøseth for his generous support. A.A.-G. and Y.K. acknowledge the generous support of Natural Resources Canada and the Canada 150 Research Chairs program. A.A.-G. and V.B. are supported by the University of Toronto's Acceleration Consortium, which receives funding from the CFREF-2022-00042 Canada First Research Excellence Fund. Y.K. was supported by the CIFAR AI Safety Catalyst Award (Catalyst Fund Project #CF26-AI-001).

### References

1. Furukawa, H., Cordova, K. E., O'Keeffe, M. & Yaghi, O. M. The Chemistry and Applications of Metal-Organic Frameworks. *Science* **341**, 1230444 (2013).
2. Lyu, H., Ji, Z., Wuttke, S. & Yaghi, O. M. Digital Reticular Chemistry. *Chem* **6**, 2219–2241 (2020).
3. Yaghi, O. M. *et al.* Reticular synthesis and the design of new materials. *Nature* **423**, 705–714 (2003).
4. Yaghi, O. M., Li, G. & Li, H. Selective binding and removal of guests in a

- microporous metal–organic framework. *Nature* **378**, 703–706 (1995).
- 5. Moghadam, P. Z. *et al.* Targeted classification of metal–organic frameworks in the Cambridge structural database (CSD). *Chem. Sci.* **11**, 8373–8387 (2020).
  - 6. Moosavi, S. M. *et al.* Understanding the diversity of the metal–organic framework ecosystem. *Nat. Commun.* **11**, 4068 (2020).
  - 7. Wright, A. M., Kapelewski, M. T., Marx, S., Farha, O. K. & Morris, W. Transitioning metal–organic frameworks from the laboratory to market through applied research. *Nat. Mater.* **24**, 178–187 (2025).
  - 8. A market for metal–organic frameworks. *Nat. Mater.* **24**, 157–157 (2025).
  - 9. Wilmer, C. E. *et al.* Large-scale screening of hypothetical metal–organic frameworks. *Nat. Chem.* **4**, 83–89 (2012).
  - 10. Glasby, L. T. *et al.* DigiMOF: A Database of Metal–Organic Framework Synthesis Information Generated via Text Mining. *Chem. Mater.* **35**, 4510–4524 (2023).
  - 11. Park, H., Majumdar, S., Zhang, X., Kim, J. & Smit, B. Inverse design of metal–organic frameworks for direct air capture of CO<sub>2</sub> via deep reinforcement learning. *Digit. Discov.* **3**, 728–741 (2024).
  - 12. Sumida, K. *et al.* Carbon Dioxide Capture in Metal–Organic Frameworks. *Chem. Rev.* **112**, 724–781 (2012).
  - 13. Li, J.-R. *et al.* Carbon dioxide capture-related gas adsorption and separation in metal–organic frameworks. *Coord. Chem. Rev.* **255**, 1791–1823 (2011).
  - 14. Song, W., Zheng, Z., Alawadhi, A. H. & Yaghi, O. M. MOF water harvester produces water from Death Valley desert air in ambient sunlight. *Nat. Water* **1**, 626–634 (2023).
  - 15. Murray, L. J., Dincă, M. & Long, J. R. Hydrogen storage in metal–organic frameworks. *Chem. Soc. Rev.* **38**, 1294–1314 (2009).
  - 16. Rosi, N. L. *et al.* Hydrogen Storage in Microporous Metal–Organic Frameworks. *Science* **300**, 1127–1129 (2003).
  - 17. A. Mason, J., Veenstra, M. & R. Long, J. Evaluating metal–organic frameworks for natural gas storage. *Chem. Sci.* **5**, 32–51 (2014).
  - 18. Yang, P., Yang, C., Wu, Z. & Tang, Z. Evolving metal–organic frameworks for highly active oxygen evolution. *Matter* **8**, (2025).
  - 19. Bavykina, A. *et al.* Metal–Organic Frameworks in Heterogeneous Catalysis: Recent Progress, New Trends, and Future Perspectives. *Chem. Rev.* **120**, 8468–8535 (2020).
  - 20. Freund, R. *et al.* The Current Status of MOF and COF Applications. *Angew. Chem. Int. Ed.* **60**, 23975–24001 (2021).
  - 21. Kingma, D. P. & Welling, M. Auto-Encoding Variational Bayes. (2013).
  - 22. Goodfellow, I. J. *et al.* Generative adversarial nets. *Adv. Neural Inf. Process. Syst.* **27**, (2014).
  - 23. Ho, J., Jain, A. & Abbeel, P. Denoising Diffusion Probabilistic Models. in *Advances in Neural Information Processing Systems* vol. 33 6840–6851 (Curran Associates, Inc., 2020).
  - 24. Vaswani, A. *et al.* Attention is All you Need. in *Advances in Neural Information Processing Systems* vol. 30 (Curran Associates, Inc., 2017).
  - 25. Zheng, Z. *et al.* Large language models for reticular chemistry. *Nat. Rev. Mater.* 1–13 (2025) doi:10.1038/s41578-025-00772-8.
  - 26. Ozcan, A. *et al.* Artificial Intelligence Paradigms for Next-Generation Metal–Organic

- Framework Research. *J. Am. Chem. Soc.* (2025) doi:10.1021/jacs.5c08214.
- 27. Xie, E., Wang, X., Siepmann, J. I., Chen, H. & Snurr, R. Q. Generative AI for Design of Nanoporous Materials: Review and Future Prospects. *Digit. Discov.* (2025) doi:10.1039/D5DD00221D.
  - 28. Han, Z. *et al.* Development of the design and synthesis of metal–organic frameworks (MOFs) – from large scale attempts, functional oriented modifications, to artificial intelligence (AI) predictions. *Chem. Soc. Rev.* **54**, 367–395 (2025).
  - 29. Park, J., Kim, H., Kang, Y., Lim, Y. & Kim, J. From Data to Discovery: Recent Trends of Machine Learning in Metal–Organic Frameworks. *JACS Au* **4**, 3727–3743 (2024).
  - 30. MacKnight, R. *et al.* Rethinking chemical research in the age of large language models. *Nat. Comput. Sci.* 1–12 (2025) doi:10.1038/s43588-025-00811-y.
  - 31. Moosavi, S. M., Jablonka, K. M. & Smit, B. The Role of Machine Learning in the Understanding and Design of Materials. *J. Am. Chem. Soc.* **142**, 20273–20287 (2020).
  - 32. Nandy, A., Duan, C. & Kulik, H. J. Using Machine Learning and Data Mining to Leverage Community Knowledge for the Engineering of Stable Metal–Organic Frameworks. *J. Am. Chem. Soc.* **143**, 17535–17547 (2021).
  - 33. Nandy, A. *et al.* MOFSimplify, machine learning models with extracted stability data of three thousand metal–organic frameworks. *Sci. Data* **9**, 74 (2022).
  - 34. Jablonka, K. M., Rosen, A. S., Krishnapriyan, A. S. & Smit, B. An Ecosystem for Digital Reticular Chemistry. *ACS Cent. Sci.* **9**, 563–581 (2023).
  - 35. Moosavi, S. M. *et al.* A data-science approach to predict the heat capacity of nanoporous materials. *Nat. Mater.* **21**, 1419–1425 (2022).
  - 36. Nandy, A. *et al.* A database of ultrastable MOFs reassembled from stable fragments with machine learning models. *Matter* **6**, 1585–1603 (2023).
  - 37. Boyd, P. G. & Woo, T. K. A generalized method for constructing hypothetical nanoporous materials of any net topology from graph theory. *CrystEngComm* **18**, 3777–3792 (2016).
  - 38. 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.* **17**, 5801–5810 (2017).
  - 39. Boyd, P. G. *et al.* Data-driven design of metal–organic frameworks for wet flue gas CO<sub>2</sub> capture. *Nature* **576**, 253–256 (2019).
  - 40. Kim, B., Lee, S. & Kim, J. Inverse design of porous materials using artificial neural networks. *Sci. Adv.* **6**, eaax9324 (2020).
  - 41. Rappe, A. K., Casewit, C. J., Colwell, K. S., Goddard, W. A. I. & Skiff, W. M. UFF, a full periodic table force field for molecular mechanics and molecular dynamics simulations. *J. Am. Chem. Soc.* **114**, 10024–10035 (1992).
  - 42. Griffin, S. L. & Champness, N. R. A periodic table of metal-organic frameworks. *Coord. Chem. Rev.* **414**, 213295 (2020).
  - 43. Weininger, D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci.* **28**, 31–36 (1988).
  - 44. Wang, Y. *et al.* PubChem: a public information system for analyzing bioactivities of small molecules. *Nucleic Acids Res.* **37**, W623–W633 (2009).
  - 45. Kim, S. *et al.* PubChem Substance and Compound databases. *Nucleic Acids Res.* **44**,

- D1202-1213 (2016).
46. Manion, C. *et al.* Automated design of flexible linkers. *Dalton Trans.* **45**, 4338–4345 (2016).
  47. Witman, M. *et al.* In silico design and screening of hypothetical MOF-74 analogs and their experimental synthesis. *Chem. Sci.* **7**, 6263–6272 (2016).
  48. Zhang, X. *et al.* A historical overview of the activation and porosity of metal–organic frameworks. *Chem. Soc. Rev.* **49**, 7406–7427 (2020).
  49. Serre, C. *et al.* Very Large Breathing Effect in the First Nanoporous Chromium(III)-Based Solids: MIL-53 or CrIII(OH)·{O<sub>2</sub>C—C<sub>6</sub>H<sub>4</sub>—CO<sub>2</sub>}·{HO<sub>2</sub>C—C<sub>6</sub>H<sub>4</sub>—CO<sub>2</sub>H}<sub>x</sub>·H<sub>2</sub>O<sub>y</sub>. *J. Am. Chem. Soc.* **124**, 13519–13526 (2002).
  50. Alhamami, M., Doan, H. & Cheng, C.-H. A Review on Breathing Behaviors of Metal–Organic-Frameworks (MOFs) for Gas Adsorption. *Materials* **7**, 3198–3250 (2014).
  51. Inizan, T. J. *et al.* System of Agentic AI for the Discovery of Metal–Organic Frameworks. Preprint at <https://doi.org/10.48550/arXiv.2504.14110> (2025).
  52. Pyzer-Knapp, E. O. *et al.* Foundation models for materials discovery – current state and future directions. *Npj Comput. Mater.* **11**, 1–10 (2025).
  53. Yao, Z. *et al.* Inverse design of nanoporous crystalline reticular materials with deep generative models. *Nat. Mach. Intell.* **3**, 76–86 (2021).
  54. Li, H., Eddaoudi, M., O’Keeffe, M. & Yaghi, O. M. Design and synthesis of an exceptionally stable and highly porous metal-organic framework. *Nature* **402**, 276–279 (1999).
  55. Furukawa, H. *et al.* Ultrahigh Porosity in Metal–Organic Frameworks. *Science* **329**, 424–428 (2010).
  56. Campbell, M. G., Sheberla, D., Liu, S. F., Swager, T. M. & Dincă, M. Cu<sub>3</sub>(hexaiminotriphenylene)2: An Electrically Conductive 2D Metal–Organic Framework for Chemiresistive Sensing. *Angew. Chem. Int. Ed.* **54**, 4349–4352 (2015).
  57. Frameworks for commercial success. *Nat. Chem.* **8**, 987–987 (2016).
  58. Lecun, Y., Bottou, L., Bengio, Y. & Haffner, P. Gradient-based learning applied to document recognition. *Proc. IEEE* **86**, 2278–2324 (1998).
  59. Hinton, G. E., Osindero, S. & Teh, Y.-W. A Fast Learning Algorithm for Deep Belief Nets. *Neural Comput.* **18**, 1527–1554 (2006).
  60. Krizhevsky, A., Sutskever, I. & Hinton, G. E. ImageNet Classification with Deep Convolutional Neural Networks. in *Advances in Neural Information Processing Systems* vol. 25 (Curran Associates, Inc., 2012).
  61. OpenAI *et al.* GPT-4 Technical Report. Preprint at <https://doi.org/10.48550/arXiv.2303.08774> (2024).
  62. Fu, X., Xie, T., Rosen, A. S., Jaakkola, T. & Smith, J. MOFDiff: Coarse-grained Diffusion for Metal–Organic Framework Design. Preprint at <https://doi.org/10.48550/arXiv.2310.10732> (2023).
  63. Radford, A., Narasimhan, K., Salimans, T. & Sutskever, I. Improving Language Understanding by Generative Pre-Training.
  64. Duan, C. *et al.* Building-Block Aware Generative Modeling for 3D Crystals of Metal Organic Frameworks. Preprint at <https://doi.org/10.48550/arXiv.2505.08531> (2025).

65. Kim, N., Kim, S., Kim, M., Park, J. & Ahn, S. MOFFlow: Flow Matching for Structure Prediction of Metal–Organic Frameworks. Preprint at <https://doi.org/10.48550/arXiv.2410.17270> (2025).
66. Kim, N., Kim, S. & Ahn, S. Flexible MOF Generation with Torsion-Aware Flow Matching. Preprint at <https://doi.org/10.48550/arXiv.2505.17914> (2025).
67. Schwaller, P., Hoover, B., Reymond, J.-L., Strobel, H. & Laino, T. Extraction of organic chemistry grammar from unsupervised learning of chemical reactions. *Sci. Adv.* **7**, eabe4166 (2021).
68. Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Preprint at <https://doi.org/10.48550/arXiv.1810.04805> (2019).
69. Dao, T., Fu, D. Y., Ermon, S., Rudra, A. & Ré, C. FlashAttention: Fast and Memory-Efficient Exact Attention with IO-Awareness. Preprint at <https://doi.org/10.48550/arXiv.2205.14135> (2022).
70. Schwaller, P. *et al.* Molecular Transformer: A Model for Uncertainty-Calibrated Chemical Reaction Prediction. *ACS Cent. Sci.* **5**, 1572–1583 (2019).
71. Cao, Z., Magar, R., Wang, Y. & Barati Farimani, A. MOFormer: Self-Supervised Transformer Model for Metal–Organic Framework Property Prediction. *J. Am. Chem. Soc.* **145**, 2958–2967 (2023).
72. Kang, Y., Park, H., Smit, B. & Kim, J. A multi-modal pre-training transformer for universal transfer learning in metal–organic frameworks. *Nat. Mach. Intell.* **5**, 309–318 (2023).
73. Pan, J. Transfer learning for metal–organic frameworks. *Nat. Comput. Sci.* **3**, 280–280 (2023).
74. J. Bucior, B. *et al.* Energy-based descriptors to rapidly predict hydrogen storage in metal–organic frameworks. *Mol. Syst. Des. Eng.* **4**, 162–174 (2019).
75. Wang, H. *et al.* LLM-Augmented Chemical Synthesis and Design Decision Programs. in (2025).
76. Wang, H. *et al.* Efficient Evolutionary Search Over Chemical Space with Large Language Models. Preprint at <https://doi.org/10.48550/arXiv.2406.16976> (2025).
77. Gan, J. *et al.* Large Language Models Are Innate Crystal Structure Generators. Preprint at <https://doi.org/10.48550/arXiv.2502.20933> (2025).
78. Kim, S., Jung, Y. & Schrier, J. Large Language Models for Inorganic Synthesis Predictions. *J. Am. Chem. Soc.* **146**, 19654–19659 (2024).
79. Jablonka, K. M., Schwaller, P., Ortega-Guerrero, A. & Smit, B. Leveraging large language models for predictive chemistry. *Nat. Mach. Intell.* **6**, 161–169 (2024).
80. Zheng, Z. *et al.* Shaping the Water-Harvesting Behavior of Metal–Organic Frameworks Aided by Fine-Tuned GPT Models. *J. Am. Chem. Soc.* **145**, 28284–28295 (2023).
81. Lu, J. *et al.* Generative Design of Functional Metal Complexes Utilizing the Internal Knowledge of Large Language Models. Preprint at <https://doi.org/10.48550/arXiv.2410.18136> (2024).
82. Rajabi-Kochi, M., Mahboubi, N., Gill, A. P. S. & Moosavi, S. M. Adaptive representation of molecules and materials in Bayesian optimization. *Chem. Sci.* **16**, 5464–5474 (2025).
83. Moosavi, S. M. *et al.* Capturing chemical intuition in synthesis of metal-organic

- frameworks. *Nat. Commun.* **10**, 539 (2019).
- 84. Zhao, Y., Zhao, Y., Wang, J. & Wang, Z. Artificial Intelligence Meets Laboratory Automation in Discovery and Synthesis of Metal–Organic Frameworks: A Review. *Ind. Eng. Chem. Res.* **64**, 4637–4668 (2025).
  - 85. Nandy, A. From pages to patterns: Towards extracting catalytic knowledge from structure and text for transition-metal complexes and metal-organic frameworks. *J. Catal.* **448**, 116174 (2025).
  - 86. White, A. D. The future of chemistry is language. *Nat. Rev. Chem.* **7**, 457–458 (2023).
  - 87. Lála, J. *et al.* PaperQA: Retrieval-Augmented Generative Agent for Scientific Research. Preprint at <http://arxiv.org/abs/2312.07559> (2023).
  - 88. Zheng, Z., Zhang, O., Borgs, C., Chayes, J. T. & Yaghi, O. M. ChatGPT Chemistry Assistant for text mining and the prediction of MOF synthesis. *J. Am. Chem. Soc.* **145**, 18048–18062 (2023).
  - 89. Zhang, W. *et al.* Fine-tuning large language models for chemical text mining. *Chem. Sci.* **15**, 10600–10611 (2024).
  - 90. Dagdelen, J. *et al.* Structured information extraction from scientific text with large language models. *Nat. Commun.* **15**, 1418 (2024).
  - 91. Ansari, M. & Moosavi, S. M. Agent-based learning of materials datasets from the scientific literature. *Digit. Discov.* **3**, 2607–2617 (2024).
  - 92. Kang, Y. *et al.* Harnessing Large Language Models to Collect and Analyze Metal–Organic Framework Property Data Set. *J. Am. Chem. Soc.* **147**, 3943–3958 (2025).
  - 93. Pruyne, T., Aswad, A., Khan, S. T., Black, R. & Moosavi, S. M. MOF-ChemUnity: Unifying metal-organic framework data using large language models. *ChemRxiv* (2025) doi:<https://doi.org/10.26434/chemrxiv-2025-8h5q7>.
  - 94. Zheng, Z. *et al.* Image and data mining in reticular chemistry powered by GPT-4V. *Digit. Discov.* **3**, 491–501 (2024).
  - 95. Kang, Y. & Kim, J. ChatMOF: an artificial intelligence system for predicting and generating metal-organic frameworks using large language models. *Nat. Commun.* **15**, 4705 (2024).
  - 96. Yang, M., Pártay, L. B. & Wexler, R. B. Surface phase diagrams from nested sampling. *Phys. Chem. Chem. Phys.* **26**, 13862–13874 (2024).
  - 97. Mu, A. U. *et al.* Structure–Activity Relationships in Ether-Functionalized Solid-State Metal–Organic Framework Electrolytes. *Chem. Mater.* **37**, 2783–2794 (2025).
  - 98. Herron, L., Mondal, K., Schneekloth, J. S. & Tiwary, P. Inferring phase transitions and critical exponents from limited observations with thermodynamic maps. *Proc. Natl. Acad. Sci.* **121**, e2321971121 (2024).
  - 99. Bette, S. *et al.* CO<sub>2</sub> and H<sub>2</sub>O Sorption Induced Bulk-Phase Changes of CALF-20 Captured Using In Situ Laboratory X-ray Powder Diffraction. *J. Am. Chem. Soc.* **147**, 25662–25671 (2025).
  - 100. Lee, G., Choi, D. & Oh, M. Activating the Gate-Opening of a Metal–Organic Framework and Maximizing Its Adsorption Capacity. *J. Am. Chem. Soc.* **147**, 12811–12820 (2025).
  - 101. Gantzler, N., Deshwal, A., Doppa, J. R. & Simon, C. M. Multi-fidelity Bayesian optimization of covalent organic frameworks for xenon/krypton separations. *Digit.*

- Discov.* **2**, 1937–1956 (2023).
- 102. MOFA: Discovering Materials for Carbon Capture with a GenAI- and Simulation-Based Workflow. <https://arxiv.org/html/2501.10651v1>.
  - 103. Lin, J. *et al.* Conversational Large-Language-Model Artificial Intelligence Agent for Accelerated Synthesis of Metal–Organic Frameworks Catalysts in Olefin Hydrogenation. *ACS Nano* **19**, 23840–23858 (2025).
  - 104. Liu, Z. *et al.* Leveraging Post-Pretrained LLMs for Inverse Engineering High-Capacity Hydrogen-Storage Metal-Organic Frameworks: From Virtual Structures to Synthesized Materials. Preprint at <https://doi.org/10.26434/chemrxiv-2024-xz77q-v2> (2025).
  - 105. Zheng, Z. *et al.* A GPT-4 Reticular Chemist for Guiding MOF Discovery. *Angew Chem Int Ed* **62**, e202311983 (2023).
  - 106. M. Bran, A. *et al.* Augmenting large language models with chemistry tools. *Nat. Mach. Intell.* **6**, 525–535 (2024).
  - 107. Boiko, D. A., MacKnight, R., Kline, B. & Gomes, G. Autonomous chemical research with large language models. *Nature* **624**, 570–578 (2023).
  - 108. Zheng, Z. *et al.* ChatGPT Research Group for Optimizing the Crystallinity of MOFs and COFs. *ACS Cent. Sci.* **9**, 2161–2170 (2023).
  - 109. Huang, L. *et al.* Natural-Language-Interfaced Robotic Synthesis for AI-Copilot-Assisted Exploration of Inorganic Materials. *J. Am. Chem. Soc.* **147**, 23014–23025 (2025).
  - 110. Chung, Y. G. *et al.* In silico discovery of metal-organic frameworks for precombustion CO<sub>2</sub> capture using a genetic algorithm. *Sci. Adv.* **2**, e1600909 (2016).
  - 111. Rosen, A. S. *et al.* High-throughput predictions of metal–organic framework electronic properties: theoretical challenges, graph neural networks, and data exploration. *Npj Comput. Mater.* **8**, 112 (2022).
  - 112. Rosen, A. S. *et al.* Machine learning the quantum-chemical properties of metal–organic frameworks for accelerated materials discovery. *Matter* **4**, 1578–1597 (2021).
  - 113. Zhao, G. *et al.* CoRE MOF DB: A curated experimental metal-organic framework database with machine-learned properties for integrated material-process screening. *Matter* **8**, (2025).
  - 114. Chung, Y. G. *et al.* Computation-Ready, Experimental Metal–Organic Frameworks: A Tool To Enable High-Throughput Screening of Nanoporous Crystals. *Chem. Mater.* **26**, 6185–6192 (2014).
  - 115. Chung, Y. G. *et al.* Advances, Updates, and Analytics for the Computation-Ready, Experimental Metal–Organic Framework Database: CoRE MOF 2019. *J. Chem. Eng. Data* **64**, 5985–5998 (2019).
  - 116. Jin, X., Jablonka, K. M., Moubarak, E., Li, Y. & Smit, B. MOFChecker: a package for validating and correcting metal–organic framework (MOF) structures. *Digit. Discov.* **4**, 1560–1569 (2025).
  - 117. White, A. J., Gibaldi, M., Burner, J., Mayo, R. A. & Woo, T. K. High Structural Error Rates in “Computation-Ready” MOF Databases Discovered by Checking Metal Oxidation States. *J. Am. Chem. Soc.* **147**, 17579–17583 (2025).
  - 118. Khan, S. T. & Moosavi, S. M. Connecting metal-organic framework synthesis to applications using multimodal machine learning. *Nat. Commun.* **16**, 5642 (2025).

119. Kraß, H., Huang, J. & Moosavi, S. M. MOFSimBench: Evaluating Universal Machine Learning Interatomic Potentials In Metal--Organic Framework Molecular Modeling. Preprint at <https://doi.org/10.48550/arXiv.2507.11806> (2025).
120. Rubungo, A. N., Fajardo-Rojas, F., Gómez-Gualdrón, D. & Dieng, A. B. Highly Accurate and Fast Prediction of MOF Free Energy Via Machine Learning. Preprint at <https://doi.org/10.26434/chemrxiv-2025-93xmj> (2025).
121. Liu, X. & Wu, L. Learning diffusion bridges on constrained domains. in *international conference on learning representations (ICLR)* (2023).
122. Lou, A. & Ermon, S. Reflected Diffusion Models. in *Proceedings of the 40th International Conference on Machine Learning* 22675–22701 (PMLR, 2023).
123. Moghadam, P. Z. *et al.* Structure-Mechanical Stability Relations of Metal-Organic Frameworks via Machine Learning. *Matter* **1**, 219–234 (2019).
124. Jiang, H., Alezi, D. & Eddaoudi, M. A reticular chemistry guide for the design of periodic solids. *Nat. Rev. Mater.* **6**, 466–487 (2021).