Digital Discovery

ROYAL SOCIETY OF CHEMISTRY

REVIEW

View Article Online
View Journal | View Issue



Cite this: *Digital Discovery*, 2025, **4**, 2336

# Generative AI for design of nanoporous materials: review and future prospects

Evan Xie,[a] Xijun Wang,[*b] J. Ilja Siepmann,[c] Haoyuan Chen[d] and Randall Q. Snurr[*b]

Generative artificial intelligence (AI) is emerging as a powerful tool for advancing the design of nanoporous materials such as metal–organic frameworks, covalent–organic frameworks, and zeolites. These materials have potential application in important areas such as carbon capture, catalysis, gas storage, chemical separation, and drug delivery due to their modular, tunable structures, and their performance in these areas depends on precise control over their structure, chemical functionalities, and properties. Herein, we provide a review of generative AI algorithms that are emerging as powerful tools for the design of nanoporous materials, namely generative adversarial networks, variational autoencoders, diffusion models, genetic algorithms, reinforcement learning, and large language models. Some models are particularly good at generating diverse and high-quality designs, while others excel at exploring large design spaces or optimizing materials with desired properties. Certain algorithms also allow for efficient transitions between different designs, and some offer versatility in generating materials based on textual input. We discuss the advantages, limitations, and applications of these algorithms in porous material design and emphasize the future potential of integrating AI with experimental workflows to accelerate the development and validation of AI-generated materials.

Received 22nd May 2025
Accepted 11th July 2025

DOI: 10.1039/d5dd00221d

rsc.li/digitaldiscovery



## 1. Introduction

Artificial intelligence (AI) is revolutionizing material design and discovery, especially through the use of new generative AI models. Traditional methods for material discovery often involve a trial-and-error process, extensively sampling the material space to search for those that meet the desired properties.[1,2] This approach is not only time-consuming but also resource-intensive, requiring substantial investments in laboratory equipment, materials, and human time. More recently, high-throughput computational screening of materials has emerged as a way to more quickly find top-performing materials for a given application.[3–5] The properties or performance of the materials may be predicted using methods such as electronic structure calculations (especially density functional theory), molecular simulations (*e.g.*, Monte Carlo (MC) or molecular dynamics (MD) simulations), or other methods (*e.g.*, geometric analysis). Since the cost of such calculations can quickly become prohibitive, machine learning (ML) algorithms, such as decision trees,[6] random forest,[7] and XG boost,[8] can be used instead to predict the properties of candidate materials at much lower cost, but usually also reduced accuracy. These ML models are often trained on computational data and, once trained, allow vast chemical spaces to be rapidly explored. Training a ML model in this way is referred to as "supervised" learning, and the goal is to create a surrogate model that can predict the properties of a candidate material more quickly than, say, a MC simulation. In contrast, generative models suggest new candidate materials, where the suggested materials have specific targeted properties. This capability significantly accelerates the material discovery process by identifying promising candidates early in the research cycle, allowing researchers to focus more detailed simulations or experiments on the most promising candidates, significantly reducing the time and cost in material discovery.

Nanoporous materials,[9] such as activated carbons and zeolites, are important in a variety of important processes, including adsorptive separations and heterogeneous catalysis. Zeolites are crystalline framework materials made from interconnected rings of silicon (or other atoms in tetrahedral sites) and oxygen atoms. They are widely used in petroleum refining, air separation, and other separations.[10] Activated carbons, by contrast, are amorphous materials with a high surface area and tunable porosity, commonly employed in gas purification, water treatment, and energy storage applications.[11] In the past 25 years, several classes of new nanoporous materials have emerged in which the materials are synthesized from well-

[a]*Deerfield Academy, 7 Boyden Lane, Deerfield, Massachusetts 01342, USA*
[b]*Department of Chemical and Biological Engineering, Northwestern University, Evanston, Illinois 60208, USA. E-mail: snurr@northwestern.edu; wangxijun1016@gmail.com*
[c]*Department of Chemistry and Chemical Theory Center, University of Minnesota, 207 Pleasant Street SE, Minneapolis, Minnesota 55455, USA*
[d]*Department of Chemistry, Southern Methodist University, Dallas, Texas 75275, USA*

defined building blocks. For example, metal–organic frameworks (MOFs) are synthesized from metal nodes and organic "linkers" that connect the metal nodes. Covalent–organic frameworks (COFs) are constructed from organic molecules linked together by strong covalent bonds. Due to the building-block synthesis approach, a wide variety of MOFs and COFs can be synthesized, and it is possible to tune properties such as their porosity,[12–14] surface area,[15,16] and topology.[17–19] These attributes make them ideal candidates for various applications contributing to clean energy solutions and environmental sustainability. For example, nanoporous materials are being developed for storage of hydrogen[20] and methane[21] and for carbon dioxide capture[22,23] and other molecular separations.[24–26] In catalysis, metal atoms in MOF nodes or decorated in these frameworks can serve as active sites to catalyze various chemical reactions, including hydrogenation, oligomerization, and electron donor–acceptor reactions.[27–30] Additionally, the porous structures of these frameworks allow for the loading of drugs into their cavities.[31] By modifying the pore sizes, topology, and surface chemistry, the release rate of the encapsulated drugs can be finely tuned, ensuring sustained and controlled drug delivery over time.[32–35]

The immense application potential of nanoporous materials has motivated tremendous efforts to accelerate their discovery using ML. These efforts have successfully predicted gas adsorption,[36–38] catalytic,[39,40] thermal,[41] and electronic properties[42,43] for various families of nanoporous materials.[44–47] However, ML in this field relies on large, labeled datasets for model training. Acquiring such datasets can be challenging and resource-intensive due to the inherent complexity of porous materials, especially when considering that performance metrics may require predictions at a range of temperatures, pressures, and adsorbate compositions. Additionally, traditional ML models struggle with generalizing beyond the data they are trained on, making it difficult to efficiently explore the vast chemical space or generate new materials with targeted properties. In contrast, generative models have shown great promise in mitigating these challenges, either by rapidly generating a large number of new materials beyond the training data for further screening or by purposefully designing new materials with desired properties. This enables more efficient exploration of the vast material space with reduced sampling requirements and thereby facilitates material design, where desired properties directly guide the generation of suitable material structures.[48] This approach is particularly compelling for porous frameworks, given their modular nature, which allows for precise tuning of building blocks to achieve targeted properties.

The remainder of this review is organized as follows. First, we present six generative AI approaches that have shown potential in the design of porous materials. Next, we examine key practical considerations, including data requirements, user-friendliness, and the scalability of these AI approaches. Then, we discuss the challenges and opportunities in applying generative AI to porous material design. We conclude with a summary of key findings and a perspective on the future of generative AI in nanoporous materials design.

## 2. Generative AI approaches for design of porous materials

In this section, we provide an overview and illustrative examples of six generative AI approaches that have demonstrated potential in designing nanoporous materials (Fig. 1): generative adversarial networks (GANs), variational autoencoders (VAEs), diffusion models (DMs), genetic algorithms (GAs), reinforcement learning (RL), and large language models (LLMs). Each of these approaches offers distinct solutions to the challenge of porous material design, allowing researchers to generate new structures and explore the vast chemical space in ways previously unattainable with traditional methods. We highlight their advantages, limitations, and specific case studies that demonstrate their impact in the discovery and optimization of

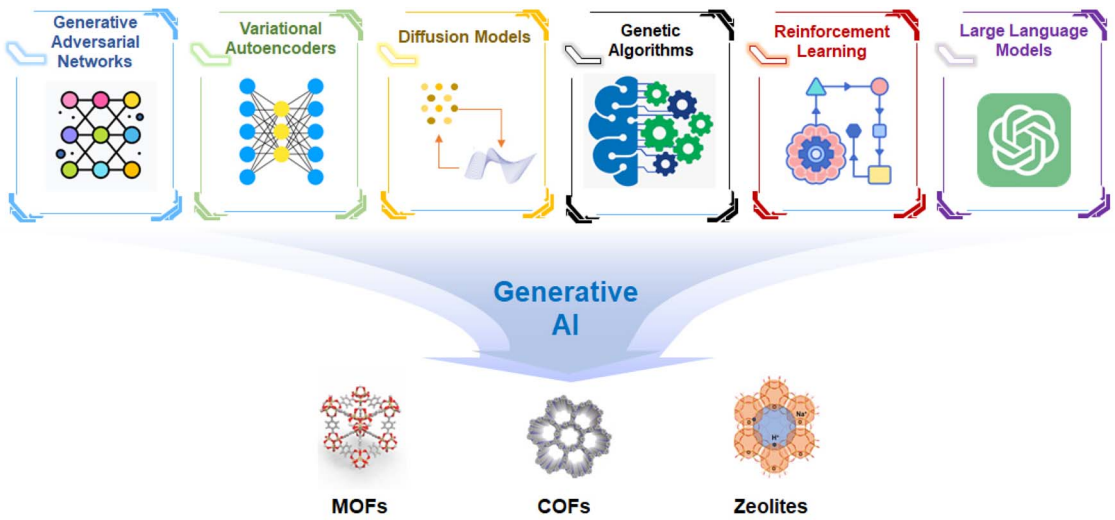


Fig. 1 Schematic illustration of generative AI applied to nanoporous material design.

Table 1 Summary of reviewed studies on generative AI for porous material design. The table highlights key aspects of each study, including the system studied, application, generative AI method, training dataset size, features used for AI, performance metrics, validation methods, key findings, and limitations or remarks

| Ref. no. | Year | Generative AI method | System studied | Application | Training dataset size | Features used for AI | Performance metrics | Validation methods | Key findings | Limitations/ remarks |
|---|---|---|---|---|---|---|---|---|---|---|
| Ref. 55 | 2020 | ZeoGAN (WGAN-GP variant) | Pure silica zeolites | Methane adsorption | 31 713 zeolites | Energy grids (methane potential energy), material grids (Si and O positions) | Methane heat of adsorption: 18–22 kJ $mol^{-1}$ | Molecular simulations (classical); comparison with IZA/PCOD databases | Demonstrated design of zeolites with specific methane adsorption properties; generated 121 new crystalline materials | Limited to pure silica zeolites; requires significant computational resources and cleanup steps for connectivity |
| Ref. 72 | 2021 | Supramolecular variational encoder (SmVAE) | MOFs | Separation of carbon dioxide from natural gas | 45 000 MOFs with property data, ~2 million MOFs without property data | Representation of MOFs in RFcode (composed of edges, vertices, topologies) | $CO_2$ uptake, $CH_4$ uptake, $CO_2/CH_4$ selectivity for natural gas separation; $CO_2$ uptake, $N_2$ uptake, $CO_2/N_2$ selectivity for flue gas separation | Comparison of top performing MOFs with well-known MOFs and zeolites reported in previous literature | Demonstrated effectiveness of automated design process of MOFs using SmVAE; identified top-performing MOF with $CO_2$ capacity of 7.55 mol $kg^{-1}$ and a $CO_2/CH_4$ selectivity of 16.0 | Hard to compare performance of top materials with literature, since experimental measurements are done at different conditions |
| Ref. 69 | 2023 | Cage-VAE | Porous organic cages (POCs) | General application | 1.2 million structures (after data augmentation) | Tri-topic precursor (BB1) skeletons, di-topic precursor (BB2) skeletons, reaction type | Validity, novelty, uniqueness, precursor validity, number of reaction sites, symmetry | MD simulations for stability validation; PCA analysis of latent space; manual inspection for shape-persistency | Successfully generated novel shape-persistent POCs with the $Tri^4Di^6$ topology using latent space traversal | Limitations in predicting shape persistence accurately; limitations in exploring diverse reaction types without model adjustments |
| Ref. 82 | 2024 | DiffLinker (diffusion model) | MOFs | $CO_2$ capture | 78 238 MOFs from the hMOF dataset; 12 305 linkers after filtering | Molecular fragments from MOF linkers in the hMOF dataset | $CO_2$ adsorption capacity threshold (high performing if > 2 mmol $g^{-1}$ at 0.1 bar); validity, synthesizability (SAscore, SCscore), uniqueness, internal diversity for MOF linker evaluation | Interatomic distance check; pre-simulation check; structural validation using MD simulations; property validation using GCMC simulations | Identified six AI-generated MOFs with $CO_2$ adsorption capacities >2 mmol $g^{-1}$ at 0.1 bar, outperforming 96.9% of MOFs in the hMOF dataset; combines generative modeling, AI prediction, and | Generated linkers occasionally failed valency checks, requiring additional filtering; steps become more computationally intensive |

Table 1 (*Contd.*)

| Ref. no. | Year | Generative AI method | System studied | Application | Training dataset size | Features used for AI | Performance metrics | Validation methods | Key findings | Limitations/ remarks |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | | | molecular simulations to screen 120 000 MOFs in under 12 hours using distributed computing | |
| Ref. 80 | 2024 | ZeoDiff (based on DDPM workflow) | Pure silica zeolites | Methane adsorption | 63 246 structures from the IZA and PCOD databases; structures were randomly split into equal training and validation sets of 31 713 each | Three-dimensional grids composed of energy, silicon, and oxygen channels | Structural validity; geometric uniqueness (102 geometrically unique structures) Chemical properties: void fraction (target values 0.05, 0.1, 0.15, 0.2, 0.25), Henry coefficient, heat of adsorption (15, 20, 25 kJ $mol^{-1}$) | Post-processing for correct Si/O ratio and accurate connectivity; chemical property distribution analysis; comparison with ZeoGAN model; test of conditional generation with user-desired properties | Generates valid zeolite structures 2000 times more effectively than GAN; successfully generated novel zeolite structures, including those with user-desired properties | Model efficiency is limited by the slow sampling speed of diffusion models; challenges with generating Henry coefficient-optimized structures; applicability to other porous materials might require increased data dimensionality |
| Ref. 107 | 2016 | Genetic algorithm | MOFs | Precombustion carbon capture | 51 163 MOFs from WLLFHS database | Chromosome representation of MOFs (using 6 integers for 6 features) | $CO_2$ working capacity; $CO_2/H_2$ selectivity; adsorbent performance score (APS) | Test of GA robustness by identifying hMOFs with highest gravimetric and volumetric surface areas and methane working capacity; experimental synthesis and testing of top-performing MOFs; GCMC simulations; performance comparison of newly generated MOFs with previously identified MOFs | Identified a list of 50 top-performing MOFs; synthesized MOF NOTT-101/OEt, which achieved a $CO_2$ working capacity of 3.8 mol $kg^{-1}$ (highest ever under studied conditions at the time) and $CO_2/H_2$ selectivity of 60; GA reduced computational time by over 99% compared to brute-force search | Limited identification of high-performing MOFs through structure–property relationships from GA; future work could extend the GA to more complex applications and larger databases |

Table 1 (*Contd.*)

| Ref. no. | Year | Generative AI method | System studied | Application | Training dataset size | Features used for AI | Performance metrics | Validation methods | Key findings | Limitations/ remarks |
|---|---|---|---|---|---|---|---|---|---|---|
| Ref. 81 | 2025 | MOFFUSION – composed of vector quantized-variational autoencoder (VQ-VAE), diffusion model, and a MOF constructor | MOFs | General application | Dataset of 247 742 hypothetical MOFs (hMOFs)<br>Total database contained 605 different topologies, 432 metal nodes, 51 organic nodes, and 220 organic edges | Signed distance function (SDF) for MOF representation | Hydrogen working capacity (WC) – target values: 5, 15, 25, and 35 g $L^{-1}$<br>Largest cavity diameter (LCD) – target values: 5, 15, 25, and 35 Å<br>Void fraction: 0.6<br>Surface area: 5000 $m^2$ $g^{-1}$ | MOF generation comparison with previous models SMVAE and MOFDiff<br>Validation of hydrogen WC through GCMC simulations | Demonstrated the effectiveness of using SDF for MOF representation. MOFFUSION showed a structural validity of 81.7%, outperforming SMVAE and MOFDiff models. Shows MOFFUSION's ability to process diverse input data formats | MOFFUSION faces challenges with extrapolation; struggles to generate structures with the desired target property when there is limited data |
| Ref. 109 | 2021 | Multispecies genetic algorithm with fitness approximation (MSGA-FA), combined with artificial neural network for property prediction (MOF-NET) | MOFs | Methane storage | Over 100 trillion hypothetical MOFs | Topology, building block information (consists of edge building blocks and node building blocks) | Methane working capacity (high performing working capacity > 180 $cm^3$/$cm^3$) | GCMC simulations through in-house GPU code and RASPA software | Successfully identified 964 MOFs with methane working capacities exceeding 200 $cm^3$/$cm^3$, with 96 of them surpassing the existing world record of 208 $cm^3$/$cm^3$<br>Demonstrated the ability of a systematic approach (evolutionary algorithm + ML + MOF constructor) for efficient screening of MOFs | Computational resources required for iterative GA cycles; further exploration needed to understand correlation between building blocks and MOF performance |
| Ref. 110 | 2021 | Genetic algorithm (GA) combined with machine learning model (MOF-NET) and a flexible cost function | MOFs | Xenon/krypton (Xe/Kr) separation from used nuclear fuel | 245 618 MOFs were screened | Used PORMAKE to generate hypothetical MOFs. Consists of node building block (NBB), edge building block (EBB), topology | Xe/Kr selectivity; xenon and krypton Henry coefficients | Molecular simulations to see the impact of framework flexibility; RASPA simulations; polymorphic simulations | Discovered two viable MOFs with record-breaking Xe/Kr selectivity; demonstrated their model can also incorporate fine-tuned targeting of user-desired properties | High computational cost due to iterative GA cycles; prediction capability of machine learning model decreases with higher selectivity values |

Table 1 (*Contd.*)

| Ref. no. | Year | Generative AI method | System studied | Application | Training dataset size | Features used for AI | Performance metrics | Validation methods | Key findings | Limitations/ remarks |
|---|---|---|---|---|---|---|---|---|---|---|
| Ref. 111 | 2016 | MOF functionalization GA (MOFF-GA) | MOFs | Postcombustion $CO_2$ capture | 1.64 trillion structures | Chromosome representation of MOFs using parent MOF and functional group code (FGC) | $CO_2$ uptake capacity (>3 mmol $g^{-1}$ at 0.15 atm and 298 K); surface area; parasitic energy | Validation of MOFF-GA on a set of 48 experimentally characterized MOFs; evaluation of $CO_2$ uptake with GCMC simulations | Discovered an average of 3.7-fold increase in $CO_2$ uptake for 141 optimized MOFs; demonstrated effectiveness in finding top-performing structures with minimal sampling | Some structures may be difficult or impossible to synthesize |
| Ref. 122 | 2024 | Deep reinforcement learning | MOFs | Direct air capture (DAC) of $CO_2$ | 646 907 MOFs used for generator pre-training; 33 000 MOFs used for predictor $CO_2$ heat of adsorption training; 24 000 MOFs used for predictor $CO_2$/$H_2O$ selectivity training | Combination of organic linkers (using SELFIES representation), metal clusters, and topologies | $CO_2$ heat of adsorption (>30 kJ $mol^{-1}$); $CO_2$/$H_2O$ selectivity (>1); validity, scaffold, and uniqueness of generated MOFs Note: integrated two predictive models, one optimizing $CO_2$ heat of adsorption, the other $CO_2$/$H_2O$ selectivity | Molecular simulations for generated MOF validation; structural feasibility tests through synthetic accessibility score and topological RMSD | Successfully designed structures with high $CO_2$ affinity (heat of adsorption > 40 kJ $mol^{-1}$) and $CO_2$/$H_2O$ selectivity (>1); revealed distinctive features in top-performing structures | Relies on large training dataset, which requires a tradeoff between computational cost and predictive accuracy; limited experimental validation of results |
| Ref. 142 | 2023 | ChatGPT-based workflow with ChemPrompt engineering | MOFs | Use of LLMs as chemical research assistant through text mining and data analysis | 228 MOF peer-reviewed papers (to extract 26 257 distinct synthesis parameters pertaining to ~800 MOFs) | 18 248 individual text segments from 228 research articles; each text segment was converted into 1536-dimensional text embedding | Precision (>95%), recall (>90%), F1 scores (>92%) for text mining; accuracy (87%) and F1 score (92%) in determining MOF crystalline state based on synthesis conditions | Manual verification of results; use of training/test sets for model predictability; comparison of predicted crystalline states with experimental results | Introduces an AI-driven workflow using ChatGPT to efficiently mine, analyze, and present MOF synthesis data; successfully predicts MOF experimental crystallization outcomes; introduces a data-driven MOF chatbot | Difficulties in accurately determining volumes/ concentration of chemicals; limited by factors such as token count and paragraph segmentation |
| Ref. 139 | 2023 | GPT-4-based reticular chemist | MOFs | Guided discovery and synthesis of MOFs | Not applicable | Leverages features like MOF structures, synthesis parameters, properties, and literature data to guide prompt | Accuracy, validity, precision of the GPT-4 answer/ suggestions | Experimental validation (NMR, XRD, *etc.*) | Demonstrates that iterative human-AI collaboration can accelerate material discovery and optimization. Successfully | Performance is reliant on human feedback for learning; challenges with advanced analytical tasks, |

Table 1 (*Contd.*)

| Ref. no. | Year | Generative AI method | System studied | Application | Training dataset size | Features used for AI | Performance metrics | Validation methods | Key findings | Limitations/ remarks |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | engineering and in-context learning for GPT-4 | | | discovered and synthesized four new isoreticular MOFs (MOF-521 variants) | such as detailed topological analysis of MOF structures, are beyond GPT-4's capabilities |
| Ref. 140 | 2024 | GPT-based ChatMOF system (GPT-4, GPT-3.5-turbo, and GPT-3.5-turbo-16k) | MOFs | Search, prediction, and generation of MOFs with user-desired properties | MOFs from CoRE MOF and QMOF databases | ChatMOF uses 4 categories of tools. Searcher: uses MOFs from CoRE MOF and QMOF. Predictor: Utilizes MOFTransformer with features like bonds, atoms, surface area, and topology. Generator: applies a genetic algorithm based on topology and building blocks. Utilities: Leverages LangChain for file search, internet search, calculations, *etc.* | Accuracy: 96.9% (search) 95.7% (prediction), 87.5% (generation tasks); RASPA simulations for generated structures | Computational simulations; manual verification of results; accuracy analysis | Demonstrated the versatility of LLMs in predicting, generating, and searching for MOF structures based on user input | Constrained by token and computational limits in LLMs; scarcity of specialized data; need for experimental validation of the generated MOFs |

nanoporous frameworks. A comprehensive overview of the reviewed research studies is provided in Table 1. The table includes the systems studied, target applications, the generative AI methods used, dataset sizes, challenges addressed, performance metrics, validation approaches, and notable findings, aiming to highlight how these methods advance the field of porous material design.

### 2.1 Generative adversarial networks (GANs)

GANs have been extensively applied in generating high-quality images[49] and have shown great potential in material design, enabling the exploration of vast design spaces and the creation of novel compounds.[50–53] GANs are a type of deep learning model comprising two neural networks—a generator and a discriminator[54]—that are trained simultaneously in a competitive process called adversarial training (Fig. 2a). The generator aims to create synthetic data (*e.g.*, images, molecular structures) that resemble real-world data, while the discriminator works as a "judge," attempting to distinguish between the real data and the generator's synthetic outputs. This setup forms a zero-sum game: the generator tries to "fool" the discriminator, while the discriminator becomes increasingly skilled at detecting fakes. In mathematical terms, the objective of the GAN is to optimize the following loss function through adversarial training:[54]

$$\min_G \max_D V(D,G) = \mathbb{E}_{x\sim p_{\text{data}}(x)}[\log D(x)] + \mathbb{E}_{z\sim p_z(z)}[\log(1-D(G(z)))]$$

where $x$ represents the real data, $z$ represents the latent vector, $p_{\text{data}}(x)$ models the distribution of the real data, and $p_z(z)$ models the distribution of the latent vector. $\mathbb{E}_{x\sim p_{\text{data}}(x)}[\log D(x)]$ is the loss function that encourages the discriminator $D$ to assign high probabilities to real data samples and $\mathbb{E}_{z\sim p_z(z)}[\log(1-D(G(z)))]$ is the loss function that encourages the discriminator to assign low probabilities to fake data generated by the generator $G$. The discriminator aims to maximize the loss function to correctly classify real and fake data, while the generator aims to minimize the loss function by producing synthetic data that the discriminator misclassifies as real. This adversarial dynamic encourages the generator to create outputs that become increasingly indistinguishable from the real data.

In the context of porous material design, GANs are known for their ability to produce highly realistic samples.[50,56] The generator proposes new frameworks meeting specific criteria, such as optimal pore size,[57] chemical stability,[51] or surface area, while the discriminator ensures that these proposed designs resemble real frameworks. This adversarial setup allows GANs to explore expansive chemical spaces and generate novel porous frameworks that might be overlooked by human intuition. For example, Kim *et al.*[55] developed a zeolite GAN, named ZeoGAN, to generate pure silica zeolite structures (Fig. 2b). The input features for training include material grids representing fixed silicon and oxygen atom distributions, and energy grids representing the methane–host interaction potential derived from classical force fields. The workflow of ZeoGAN involves feeding structured grids into the generator, which attempts to create realistic zeolites while the critic evaluates their plausibility. The model iteratively refines its outputs using adversarial training. In this work, the Earth mover's distance (EMD)[58] which represents the minimum cost required to transform one probability distribution into another, is used to quantify the difference between the distribution of generated data and that of the training data. The goal of optimizing EMD is to make the generated data distribution increasingly similar to the training data distribution, ensuring that the generated samples are realistic and physically meaningful. Using this approach, trained on 31 173 methane-accessible zeolites, ZeoGAN generated 1 million potential structures. After screening for proper bond connectivity and maintaining the correct Si : O ratio, eight unique zeolites were identified that were not present in the

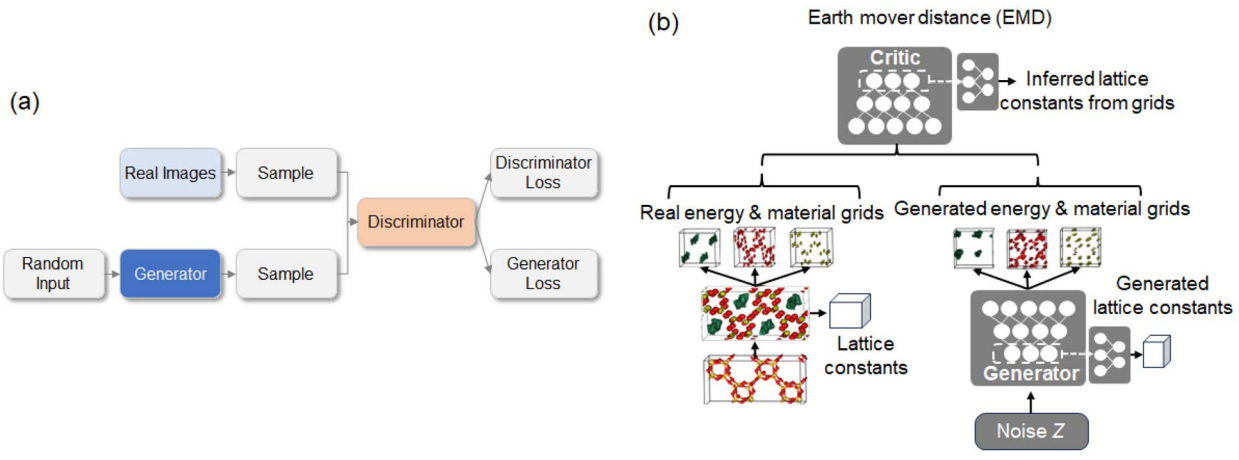


Fig. 2 (a) Basic architecture of a GAN, featuring two neural networks: the generator and discriminator, which work adversarially to generate realistic data. (b) Overview of the ZeoGAN model. Energy (green) refers to the potential energy for methane adsorbate molecules, and material grids indicate silicon (yellow) and oxygen (red) atoms. Adapted with permission from ref. 55. Copyright 2020 American Association for the Advancement of Science (AAAS).

training dataset, suggesting that ZeoGAN generated structures beyond the scope of its training data. ZeoGAN was further refined to generate structures with specific user-desired properties, by biasing its learning process to generate materials within a specific heat of adsorption range (18–22 kJ $mol^{-1}$), resulting in 121 feasible zeolites with the desired adsorption properties.

GANs offer significant flexibility in porous design because of their ability to learn and model complex data distributions. Unlike traditional methods like descriptor-based regression models that assume relatively simple structure property relationships, GANs can adapt to a wide variety of data patterns. For instance, Mao *et al.*[53] leveraged GANs to design 2D porous materials with optimized isotropic elastic properties by generating configurations based on crystallographic symmetries and porosity constraints. They constructed datasets representing different symmetry groups, each containing around one million configurations with varying pixel matrices, Young's modulus, and isotropy. By training GANs on these various datasets, they produced 400 configurations that achieved over 94% of the theoretical maximum Young's modulus across different porosities, demonstrating the ability of GANs in generating near-optimal designs without extensive trial-and-error.[59]

While GANs have been successfully used for designing materials with relatively simple compositions, such as zeolites (especially all-silica zeolites),[53,55,60] their application to more complex materials like MOFs and COFs remains challenging. The primary difficulty stems from the significant structural diversity of these materials, as traditional GAN architectures struggle to capture the vast range of topologies, bonding patterns, and coordination environments present in MOFs and COFs.[61] Unlike zeolites, these materials incorporate a wide variety of atom types, metal–ligand interactions and the complexity of organic molecules, which GANs find difficult to encode in a latent space and accurately reconstruct during generation. Another fundamental challenge lies in mode collapse, a well-known limitation of GANs, where the model tends to generate only a limited subset of structures rather than fully exploring the diverse chemical space. Given the complexity of MOFs and COFs, this issue is exacerbated as the model struggles to balance long-range periodicity with local coordination constraints, often leading to unrealistic or repetitive frameworks.

To mitigate these challenges, some studies have used advanced versions of GANs, such as deep convolutional GANs (DCGANs), to better manage these complexities. For example, Long *et al.*[51] developed a constrained crystal DCGAN (CCDCGAN), integrating deep convolutional layers, to learn hierarchical features from the input data.[62] By leveraging deep convolutional layers, the model progressively extracts hierarchical features from input data. Early layers focus on simple geometric details, such as edges or corners, while deeper layers learn more complex representations, such as the spatial arrangements and symmetries that define crystal lattices. This layered approach enables the model to capture both local bonding environments and global structural characteristics. The CCDCGAN further incorporates constraints directly into the generative process, ensuring that the generated structures meet thermodynamic stability and symmetry requirements. By embedding these constraints, the model not only adheres to physical and chemical principles but also explores a broader latent space to identify novel configurations. This combination of hierarchical feature learning and constraint integration allows CCDCGAN to overcome the limitations of traditional GANs in capturing the vast structural diversity and complex connectivity of porous materials.

We note that traditional GANs also face challenges with training instability, where the generator and discriminator fail to converge properly,[63] or with mode collapse, where the generator fails to capture the full diversity of the target distribution and repeatedly produces only a limited subset of samples.[64] These issues also hinder discovering new materials that may differ significantly from the training data, such as MOFs and COFs with similar building blocks yet different topologies. To mitigate these challenges, some studies[52,55,65] have adopted Wasserstein GANs (wGANs),[66] which replace the traditional GAN loss function with the EMD introduced earlier. This leads to more stable training and helps the model converge more effectively.

### 2.2 Variational autoencoders (VAEs)

Variational Autoencoders (VAEs) are another type of generative model increasingly used for material discovery. They encode high-dimensional data, such as material structures, into a lower-dimensional latent space that captures the essential features of the data, which is then decoded back into the original data space.[67,72] Additionally, the decoding step allows for the reconstruction of material structures, enabling the generation of new, chemically and structurally valid materials based on the learned latent space representation.[68,69] This can be particularly useful for designing new materials with targeted properties, as it facilitates the exploration of large design spaces while maintaining computational efficiency. The training of a VAE involves two main components: the encoder, which compresses the material data into the latent space, and the decoder, which reconstructs the material data from this latent space. Instead of learning a single deterministic encoding, the encoder maps the input data $x$ to a probabilistic distribution in the latent space, specifically a Gaussian distribution $q(z|x)$ characterized by a mean $\mu(x)$ and variance $\sigma^2(x)$. A latent vector $z$ is sampled from this distribution and passed through the decoder to reconstruct $x$.

The training objective of VAEs is to maximize the Evidence Lower Bound (ELBO) $\mathcal{L}$:

$$\mathcal{L} = \mathbb{E}_{q(z|x)}[\log p(x|z)] - \mathrm{KL}(q(z|x)\|p(z))$$

The equation includes two parts – the reconstruction loss: $\mathbb{E}_{q(z|x)}[\log p(x|z)]$ and the Kullback–Leibler (KL) divergence: $\mathrm{KL}(q(z|x)\|p(z))$. The reconstruction loss ensures that the decoder learns to reconstruct input data $x$ that closely matches the original input data. The KL divergence regularizes the latent space $z$ to follow a smooth, structured distribution. Maximizing the ELBO results in a minimization of the KL divergence.

Minimizing the KL divergence ensures that the learned latent space is close to the desired prior distribution, which is typically a standard Gaussian. This encoding-decoding process learns a probabilistic mapping from the input data to a latent space, enabling the generation of plausible new material structures by sampling from this latent distribution[70,71] (Fig. 3a).

One of the major advantages of VAEs is their ability to create a smooth and continuous latent space, which makes it easier to explore new material structures and discover materials with specific properties. This latent space represents the complex, high-dimensional data of material structures in simpler, lower-dimensional form. The continuous nature of this latent space is particularly beneficial for exploring and interpolating between different material designs. Additionally, optimization in the continuous latent space is more tractable than optimizing discrete structures, as it allows for the use of gradient-based methods.

In contrast, discrete optimization is often challenging due to the combinatorial nature and non-differentiability of the structure space. A notable example of this is the supramolecular variational encoder (SmVAE) developed by Yao *et al.*[72] which aimed to design new MOFs with enhanced properties for $CO_2$/$N_2$ and $CO_2$/$CH_4$ separation. The structural training data came from the CoRE MOF 2019-ASR database,[73] which contains experimentally synthesized MOFs. The dataset was augmented to approximately two million MOF structures by applying random functionalization to known molecular fragments. The features extracted for input into the model included the MOF edges, vertices (both inorganic and organic), and topologies defining the reticular framework connectivity. Grand canonical Monte Carlo (GCMC) simulations were performed on 45 000

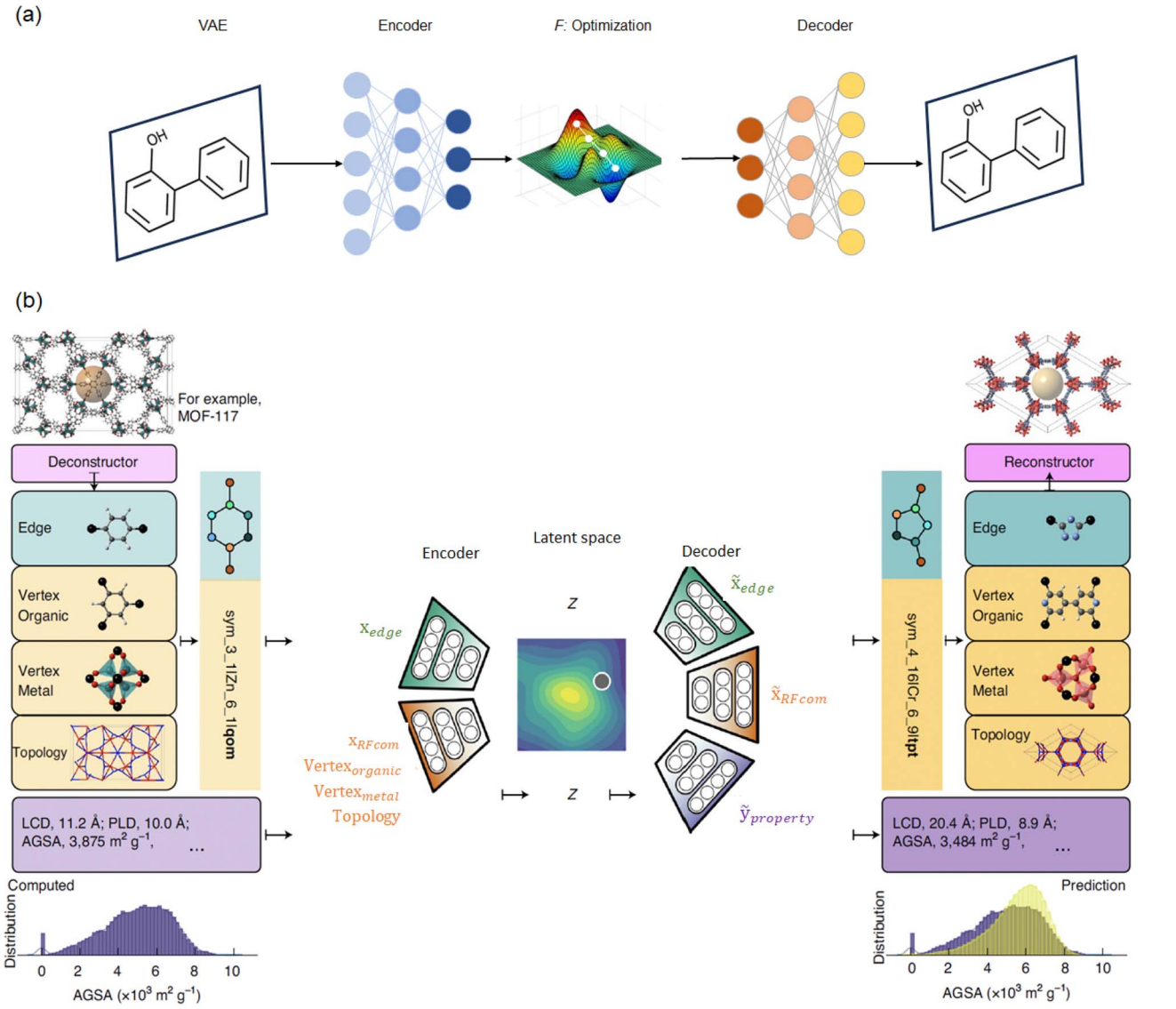


Fig. 3 (a) Basic architecture of a VAE with an encoder-decoder structure for molecular or material design. Adapted from ref. 71. Licensed under CC BY 4.0. (b) Automated porous framework discovery platform using the supramolecular variational autoencoder (SmVAE). Reprinted from ref. 72, with permission from Springer Nature Copyright 2021.

randomly selected MOFs to obtain the gas adsorption properties. Four textural properties (pore-limiting diameter (PLD), largest cavity diameter (LCD), density, and accessible gravimetric surface area (AGSA)) were computed geometrically for these 45 000 structures. The workflow of the SmVAE consists of an encoder that maps discrete framework representations (RFcodes) into a continuous latent vector space and a decoder that reconstructs MOFs from this space. RFcode is an extension of MOFid,[74] which is a unique identifier string that encodes the metal node, organic linker, and topology information of a MOF. Similarly, RFcode[72] represents the structure as a tuple of edges (represented by SMILES), vertices, and topology of the decomposed MOF. The model was trained in a semi-supervised manner using both structures with known properties (45 000 MOFs) and those without property data (the remaining dataset). A Gaussian Process (GP) model was then trained on the latent space to guide optimization towards structures with improved properties. The optimization was achieved by navigating the latent space and generating new MOFs predicted to have superior $CO_2$ separation capabilities. Using this approach, the SmVAE successfully identified candidates with high $CO_2$ capacity and selectivity, with the top-performing MOF achieving a $CO_2$ capacity of 7.55 mol $kg^{-1}$ and a selectivity of 16.0 for $CO_2$/$CH_2$ separation, making it strongly competitive against the best performing materials in the literature for this separation.

In a related study, Zhou *et al.*[69] developed a VAE called Cage-VAE, specifically designed for generating porous organic cages (POCs). Cage-VAE encodes the structural features of existing POCs into a continuous latent space, effectively capturing their geometric and stability characteristics. By sampling different points in the latent space of the model, the authors found that Cage-VAE was highly effective at creating new POCs, particularly in biasing the generation process toward a specific desired property, such as shape persistence, which refers to the ability of a cage to retain its three-dimensional geometry without collapsing. Cage-VAE achieved a high success rate for producing valid, novel, and unique POC structures, with validity, novelty, and uniqueness scores all exceeding 0.900. Here, validity refers to the proportion of chemically valid molecules, as determined by whether the generated SMILES strings can be successfully parsed into molecular graphs. Novelty measures the fraction of valid molecules that do not appear in the training dataset. Uniqueness represents the proportion of valid molecules that are non-duplicated within the generated batch. Additionally, the study incorporated advanced techniques like Bayesian optimization and spherical linear interpolation to explore the latent space more efficiently, demonstrating how VAE, when integrated with other ML methods, can enhance the targeted design of functional materials by guiding generative processes toward specific chemical and structural goals.

Another advantage of VAEs is their stability during training. Unlike GANs that need much fine-tuning, VAEs tend to converge consistently because of their well-defined loss function. This loss function balances how well the model reconstructs the original data with a regularization term that shapes the structure of the latent space. As a result, VAEs are less likely to experience issues like mode collapse, which is a common problem with GANs where the model fails to capture the full diversity of the training data. Furthermore, the latent space created by VAEs allows researchers to generate new structures with combined or intermediate properties.

In recent years, variants of VAEs have been increasingly applied to assist porous materials design. For instance, Sun *et al.*[47] developed a VAE-like encoder-decoder architecture within a meta-learning framework to extract structural fingerprints of nanoporous materials and predict their hydrogen adsorption behavior. Their study leveraged high-throughput MC simulations to generate adsorption data for a diverse set of materials, including MOFs, hyper-cross-linked polymers (HCPs), and zeolites, across a broad range of temperatures and pressures. By encoding the adsorption loading surface into a latent fingerprint representation, their model enabled accurate prediction of hydrogen uptake while circumventing the limitations of traditional adsorption isotherm fitting approaches. Instead of training separate models for different materials, the authors developed a single meta-learning model that generalizes across material classes and effectively predicts their hydrogen adsorption performance, demonstrating improved accuracy and transferability compared to conventional methods.[54]

A common problem with VAEs is an insufficient disentangling effect. This issue arises when the VAE learns a latent space where multiple factors are entangled or overlapping in a single latent dimension, making it difficult to control or interpret specific features of the data. This happens because the VAE's decoding process is probabilistic, which can blend different features together and smooth out important details.[75] In the context of materials design, this means that the VAE may not be able to differentiate between subtle variations in properties like chemical composition, pore structure, or topology required for practical applications.[76,77] As a result, additional refinement steps, such as using further computational or experimental validations[71,75,78] may be required to ensure that the generated materials meet the desired performance and exhibit clearly defined and controllable structural and chemical features necessary for real-world synthesis and application.

### 2.3 Diffusion models

Diffusion models (DMs), initially developed for high-quality image generation, are now being used in porous material design because they can learn from existing structures and generate new ones that are both diverse and chemically reasonable.[79] These models are grounded in a probabilistic framework and operate through a two-step process:[89] a forward process and a reverse process (Fig. 4a). In the forward process, noise is incrementally added to the original data over a series of discrete time steps. At each step $t$, the data become noisier, progressively approaching a standard Gaussian distribution. This process can be mathematically expressed as:

$$q(x_t|x_{t-1}) = \mathcal{N}(x_t; \sqrt{\alpha_t}\, x_{t-1}\,,\ \beta_t I)$$

where the variable $x$ represents a data sample in the diffusion process, such as structural or property-related features of

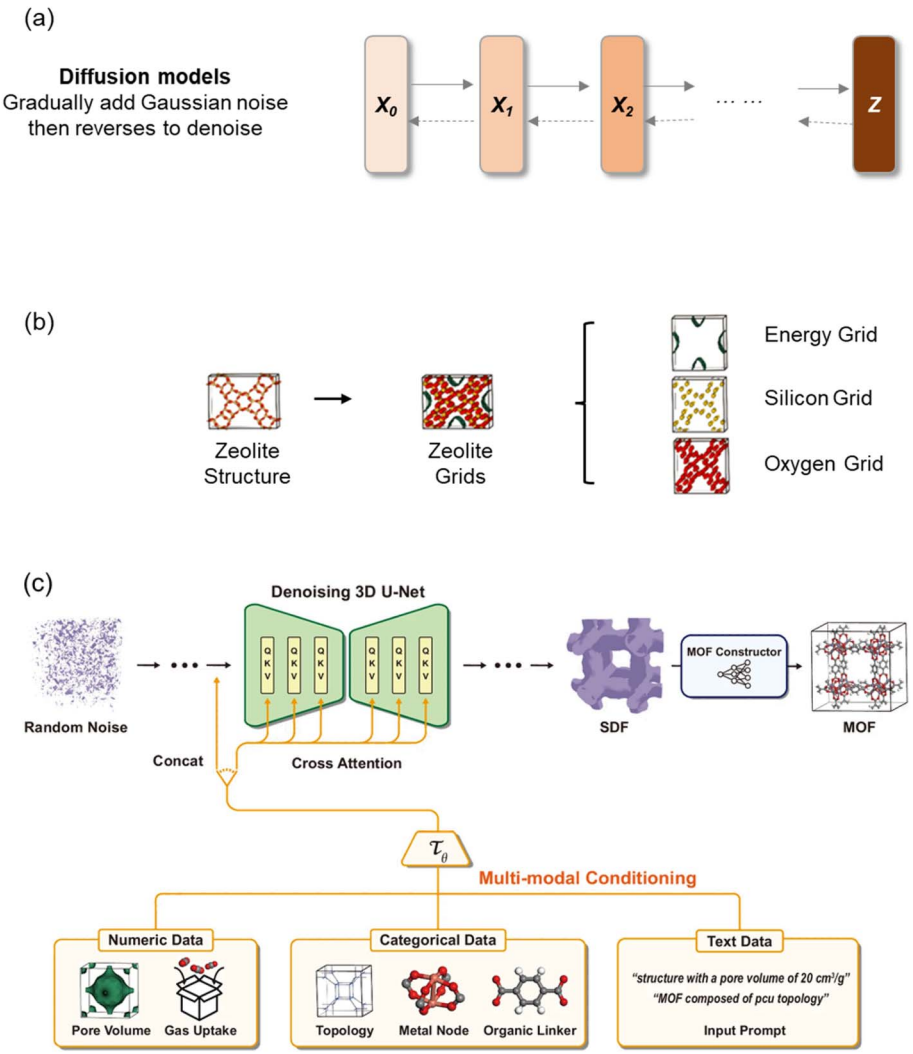


Fig. 4 (a) Overview of the diffusion model, which begins with random noise and iteratively denoises the input through learned probabilistic transitions to generate outputs resembling the original data distribution. (b) Graphical representation of the diffusion process for zeolite generation using ZeoDiff. Adapted from ref. 80. Licensed under CC BY 3.0. (c) Model architecture of MOFFUSION. Within MOFFUSION, a denoising 3D U-Net is used for the diffusion process. Adapted from ref. 81. Licensed under CC BY-NC.

a material. $q(x_t|x_{t-1})$ represents a conditional probability distribution (a Gaussian distribution $\mathcal{N}$) that defines how $x_{t-1}$, a version of $x$ at timestep $t-1$, transitions to $x_t$, a slightly noisier version, in the forward process. $\mathcal{N}(x_t;\mu,\Sigma)$ represents a multivariate Gaussian distribution where $\mu$ is the mean of the distribution and $\Sigma$ is the covariance matrix. In this case, the mean $\mu = \sqrt{\alpha_t}x_{t-1}$ carries forward the signal from the previous step, where $\sqrt{\alpha_t}$ is a scaling factor controlling the contribution of the original data, to ensure that the new state $x_t$ is primarily influenced by $x_{t-1}$. The parameter $\alpha_t$ is defined as $\alpha_t = 1 - \beta_t$, where $\beta_t$ is the variance of the Gaussian noise added at timestep $t$. The choice of hyperparameter $\beta_t$ determines the noise schedule. $I$ represents the identity matrix and the covariance matrix $\Sigma = \beta_t I$ introduces isotropic Gaussian noise at each time step, progressively corrupting the data.

This gradual corruption encodes the data into a form that is easy to model statistically but retains traces of the original structure. Next, the reverse process learns to reverse the noise addition by iteratively denoising the data to recover the original distribution. Using a trained neural network, the model predicts the noise added at each step and refines the data accordingly. The reverse process can be approximated as:

$$p_\theta(x_{t-1}|x_t) = \mathcal{N}\left(x_{t-1}; \mu_\theta(x_t,t), \sum_\theta(x_t,t)\right)$$

where $p_\theta(x_{t-1}|x_t)$ is the learned reverse process distribution with parameters $\theta$. $\mu_\theta(x_t,t)$ is the predicted mean of the Gaussian distribution. It represents the most likely denoised value of $x_{t-1}$ given $x_t$ and the current timestamp $t$. $\sum_\theta(x_t,t)$ is the variance of the Gaussian distribution, which can be either fixed or learned. By starting from Gaussian noise, the trained diffusion model gradually generates realistic data through this reverse denoising

process, making it particularly suitable for generative material discovery.[79]

In generative discovery, DMs have been shown to create high-performing, complex material structures, including MOFs. For example, Park *et al.*[82] utilized a diffusion model named DiffLinker to generate chemically diverse MOF linkers for enhanced $CO_2$ capture. The model was trained on the hMOF dataset,[83] which contains 137 652 hypothetical MOFs with geometric features and adsorption data for various gases. The training data included high-performing MOF linkers, which were extracted and decomposed into molecular fragments serving as input features. DiffLinker employed a generative diffusion process, where Gaussian noise was iteratively added to the molecular fragments and then removed through a denoising network, enabling the generation of chemically diverse and unique linkers. These linkers were subsequently assembled with pre-selected metal nodes (Cu paddlewheel, Zn paddlewheel, $Zn_4O$ nodes) into MOFs with a primitive cubic (pcu) topology. To evaluate these AI-generated MOFs, the study employed a comprehensive screening workflow that included MD and GCMC simulations. This process ensured that the MOFs not only met structural validity and stability requirements but also demonstrated high $CO_2$ adsorption capacities. Among the generated candidates, six MOFs exhibited $CO_2$ adsorption capacities exceeding 2 mmol $g^{-1}$ at 0.1 bar pressure and room temperature, outperforming 96.9% of the MOFs in the reference dataset.

Researchers have also worked to enhance the robustness of DMs by combining them with other generative algorithms, such as VAEs. For example, the Crystal Diffusion Variational Autoencoder (CDVAE) was introduced by Xie *et al.*[84] in 2021 to generate realistic 3D periodic structures of stable crystalline materials. They integrated a VAE with a diffusion model, specifically a noise conditional score network (NCSN), by encoding material structures into a latent space and using the NCSN in the decoder to refine noisy structures (a process that predicts adjustments needed to move towards a stable state) through Langevin dynamics. This integration embeds physical inductive biases, such as energy minimization and bonding preferences, ensuring that the generation process respects stability constraints and invariances, thus improving model robustness. Since then, it has been adapted for various applications. For example, Lyngby *et al.*[85] adapted CDVAE to generate 2D materials, training it on 2615 known stable materials. Their model predicted 11 630 new 2D materials, many of which were more complex than the training examples. Among these, over 8500 materials were found to be chemically stable, with formation energies within 0.3 eV per atom of the convex hull (reference energy), and over 2000 were potentially synthesizable, within 50 meV per atom of the convex hull. In another study, Pakornchote *et al.*[86] employed a different approach called the denoising diffusion probabilistic model (DDPM) in the diffusion model component of the CDVAE. They found that this modified model generated structures that were closer to their true ground states, as predicted by DFT, with an improvement of around 68.1 meV per atom compared to the original CDVAE.

One reason that DMs are effective is that they can introduce diversity in the generated samples, which is crucial for discovering materials that might be overlooked by human intuition. Park *et al.*[80] developed a diffusion model named ZeoDiff to generate all-silica zeolites. ZeoDiff significantly outperformed a previously developed GAN model, ZeoGAN,[55] in terms of structural validity, achieving a 2000-fold increase in the ratio of valid to total generated structures. Specifically, after post-processing, only 0.0008% of the structures generated by ZeoGAN were valid, whereas ZeoDiff achieved a validity rate of 1.83%, highlighting its enhanced capability in producing physically realistic and synthesizable materials. ZeoDiff introduces diversity in the generated samples through its stochastic diffusion-denoising process. Its workflow begins with a representation of zeolite structures as three-dimensional grids composed of energy, silicon, and oxygen channels (Fig. 4b), akin to RGB channels in image processing. These grids are progressively noised and then denoised by the model to generate new, realistic zeolite frameworks. To ensure the validity of generated structures, a post-processing procedure corrects atomic connectivity and Si/O ratios, further refining the outputs. Using this approach, ZeoDiff successfully generated a variety of complex zeolite structures that were previously unknown. Among the 183 generated structures, 84 were entirely new and featured unique geometric properties (Fig. 4b).

In another study, Alverson *et al.*[52] compared the performance of Wasserstein GANs, Vanilla GANs, and DMs in generating crystal structures that are both synthesizable and chemically stable, as determined by predicted formation energy using a pre-trained ML model and stability analysis through iterative DFT relaxation calculations. They found that the diffusion model greatly outperformed the GAN models, creating symmetrical and realistic-looking structures that were validated through energy relaxation calculations. Importantly, the DMs did not suffer from mode collapse, a common problem with GAN models where diversity in generated samples is lost. Instead, the DMs produced a wide range of lattice parameters, lattice angles, and space groups. The ability of DMs to effectively process and accurately reconstruct complex data distributions ensures that the generated frameworks not only meet a variety of design requirements but also maintain structural stability.

One challenge for DMs is their high computational cost. Despite offering high fidelity and rich structure generation, training a DM can require several days on multiple high-performance GPUs, with reported carbon emissions reaching ~9 kg of $CO_2$ equivalent for training alone, and up to hundreds of kilograms for large-scale data generation depending on resolution and sample size.[87] Although efficient sampling methods[88–90] such as the DDPM[88] employed by Pakornchote *et al.*[86] can help reduce some of this cost by speeding up the inference process, the overall computational demands are still significant. For example, when comparing regular DMs, DDPMs, and GANs in image synthesis on the ImageNet 256 × 256 dataset, regular DMs and DDPMs have significantly higher computational demands compared to GANs. Regular DMs require the longest training time—7 million steps—and have the largest model size, with 675 million parameters.[91] In

contrast, GANs offer the fastest inference time at 0.07 seconds (ref. 92) and the smallest model size, with 166.3 million parameters.[93] Although DDPMs are 3× faster than regular DMs, they still require substantial computational resources compared to GANs.[93]

This challenge has driven researchers to develop innovative approaches that balance computational efficiency and generative performance in DMs. A notable example is the work by Park *et al.*,[81] who developed MOFFUSION, a denoising diffusion probabilistic model for MOF structure generation designed to efficiently explore the vast chemical space of MOFs while ensuring structural validity and tunable properties (Fig. 4c). A key innovation of MOFFUSION is its use of the signed distance function (SDF) representation for MOFs, a mathematical framework that encodes geometric shapes by measuring the shortest distance from any point in space to the nearest surface. SDF provides a highly effective way to describe the intricate pore structures of MOFs, but its high dimensionality and large data volume ($32^3$ grid points) pose significant computational challenges, making it infeasible for conventional DMs to process efficiently. To address this issue, the authors incorporated a vector quantized-VAE (VQ-VAE), a discrete latent representation variant of VAE, for feature compression and latent space mapping. By reducing the input data dimensionality from $32^3$ to $8^3$ before feeding it into the diffusion model and subsequently scaling the generated data back up to $32^3$, this compression-decompression process significantly reduces the computational load. As a result, MOFFUSION enables the efficient processing of high-dimensional feature space containing diverse modalities of data including 3D structural data, numeric, categorical, and text data, making large-scale MOF generation computationally affordable.

DMs also require large amounts of high-quality training data to cover the diversity of materials, typically on the order of tens of thousands of examples.[84,94] As introduced by Xie *et al.*,[84] the Perov-5 dataset consists of 18 928 perovskite materials with 56 elements and 5 atoms per unit cell. The carbon-24 dataset[95] contains 10 153 carbon-based materials with 6–24 atoms per unit cell, while the MP-20 dataset[96] from the materials project includes 45 231 materials with up to 89 elements and 1–20 atoms per unit cell. These datasets highlight the scale and diversity needed for training DMs. Datasets for generative discovery of nanoporous materials are often quite limited,[97] especially when targeting novel or difficult-to-compute properties. One solution to this challenge is to use data augmentation techniques to expand the training dataset[98] or to apply transfer learning, leveraging existing data from related materials.[99–101]

### 2.4 Genetic algorithms (GA)

GAs are optimization techniques inspired by natural selection and genetic principles. They are particularly well-suited for generative materials discovery, where the goal is to explore vast design spaces while minimizing the need for sampling and to identify material structures that optimize specific properties or performance criteria. As depicted in Fig. 5a, the process begins with an initial population of randomly generated material configurations, where each configuration, or "individual," represents a potential solution. These individuals are evaluated using a fitness function, $f(x_i)$, which quantifies their performance based on desired material properties such as gas adsorption capacity or thermal stability. Individuals with higher fitness scores are probabilistically selected to contribute to the next generation, ensuring that the best solutions are carried forward.

In the context of porous material design, *e.g.*, MOFs, to generate new individuals, genetic operators like crossover and mutation are applied. Crossover, or recombination, combines the structural building blocks ("genes") of two parent configurations to create offspring. For instance, a typical crossover involves exchanging structural units between two selected MOFs, creating new combinations of inorganic nodes, organic linkers, and functional groups. Mutation introduces random changes to the offspring to create diversity and explore new regions of the design space.[102,103] It occurs with a predefined

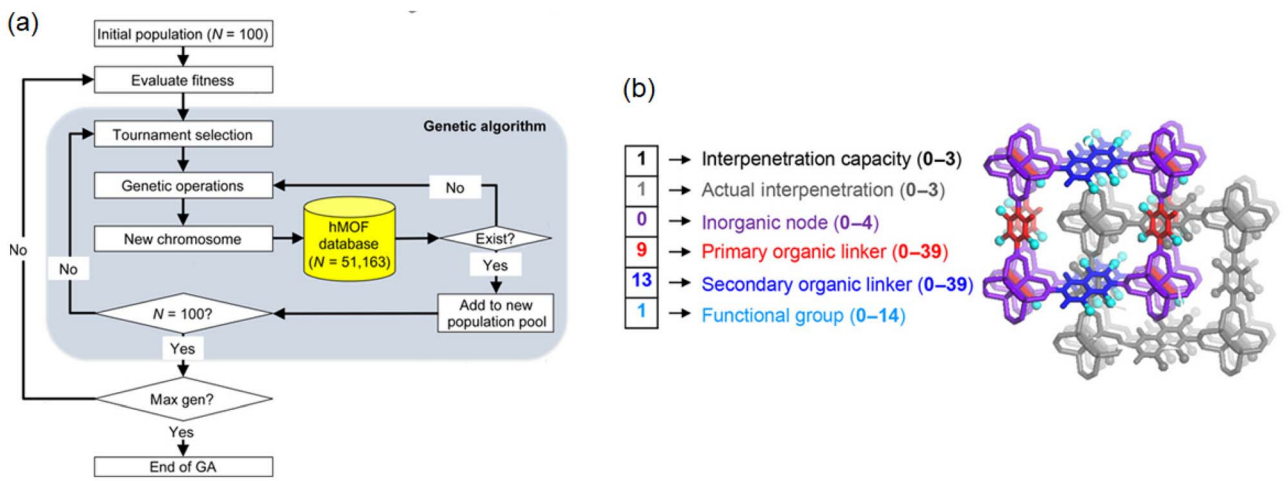


Fig. 5 (a) Workflow of GA and (b) An example chromosome and the corresponding hMOF structure. Colors help illustrate the correspondence between the genes and the hMOF structural features. Adapted from ref. 107. Licensed under CC BY-NC.

probability (*e.g.*, 5%) for each gene, where a randomly chosen gene (such as the type of metal node, organic linker, or functional group) is altered to a different valid option from the dataset. This introduces structural variations that help the algorithm explore novel MOF configurations and avoid premature convergence to suboptimal solutions. The iterative process of crossover and mutation continues for a fixed number of generations or until a material achieving a desired fitness is found. The inherent parallelism of GAs allows them to evaluate multiple solutions simultaneously,[104,105] significantly speeding up the search process, especially when using computationally expensive molecular simulations and DFT calculations to evaluate the fitness of the generated candidates.[106] GAs are particularly advantageous when the design space is vast and not easily navigable by traditional methods. In contrast to DMs, GAs rely on simulation-based fitness scoring and do not involve neural network training, which is the major contributor to the carbon footprint of DMs. However, since each GA evaluation involves simulations that may take hours, whether GAs have a lower carbon footprint than DMs ultimately depends on the specific application and computational setup.

The effectiveness of GAs in discovering superior porous frameworks has been demonstrated in various studies. For example, Chung *et al.*[107] used a GA to identify high-performance MOFs for precombustion $CO_2$ capture. As depicted in Fig. 5a, the search space consisted of 51 163 unique structures from the hMOF database,[108] where each MOF was represented by a chromosome of six integers (Fig. 5b), encoding key structural units such as inorganic nodes, organic linkers, and functional groups. The GA workflow began with an initial population of 100 MOFs, selected to ensure diversity. The algorithm then evolved these MOFs over multiple generations through tournament selection, crossover, and mutation. Crossover was applied with a 65% probability, where a single-point crossover mechanism was used to exchange structural units (*e.g.*, inorganic nodes, organic linkers, and functional groups) between two selected parent MOFs. A random crossover point was chosen along the chromosome, and the genes beyond this point were swapped between the two parent MOFs. This process helped preserve beneficial traits while introducing new combinations. Following this, mutation was introduced with a 5% probability, where one or more structural units were randomly modified. This step enabled the algorithm to explore novel configurations and avoid premature convergence to local optima. In each generation, high-performing MOFs were identified based on $CO_2$ working capacity and $CO_2/H_2$ selectivity, evaluated using GCMC simulations. These high-performing MOFs were then recombined and mutated to create new candidates, and the process was repeated for 10 generations. Using this approach, Chung *et al.* identified and experimentally validated NOTT-101/OEt, a MOF with a $CO_2$ working capacity of 3.8 mol $kg^{-1}$ and a $CO_2/H_2$ selectivity of 60, outperforming previously reported MOFs under the same conditions. Additionally, their GA model reduced computational effort by over 99% compared to a brute-force screening of the entire database, demonstrating the efficiency of AI-driven material discovery.

In another instance, Lee *et al.*[109] employed genetic algorithms to explore over 100 trillion potential MOFs for methane gas storage. By utilizing GCMC simulations and Artificial Neural Networks (ANN) to assess the working capacity of these MOFs, their algorithm successfully identified 964 MOFs with methane working capacities exceeding 200 $cm^3/cm^3$, with 96 of them surpassing the existing world record of 208 $cm^3$ (gas at STP)/$cm^3$ (MOF). Lim *et al.*[110] used a similar approach, combining genetic algorithms with GCMC and ANN, to identify two MOFs that outperformed the current benchmark for xenon/krypton separation. Moreover, their research enhanced the genetic algorithm by considering additional properties such as the cost and selectivity of the frameworks, demonstrating its capability not only to identify optimal materials but also to ensure the practical applicability of MOFs.

Collins *et al.*[111] developed a GA-based approach, named MOFF-GA, to optimize functional groups within MOFs for enhanced $CO_2$ capture. Focusing on experimentally characterized MOFs, the algorithm employs tailored crossover and mutation schemes to efficiently explore the vast search space of possible functional group combinations. This approach was applied to 141 parent MOFs, resulting in 1035 functionalized derivatives with $CO_2$ uptake capacities exceeding 3 mmol $g^{-1}$ at 0.15 atm and 298 K evaluated using GCMC simulations, outperforming the original MOFs by an average of 3.7 times. Remarkably, the MOFF-GA was effective even when working with a small search space of fewer than 1000 structures.

GAs can be applied to a wide range of material design problems, which makes them versatile tools that can be combined with other ML algorithms for better results. For example, Jennings *et al.*[103] combined an on-the-fly trained Gaussian Process (GP) regression model with a GA. The GP serves as a computationally inexpensive surrogate to predict the energy of candidate materials, significantly reducing the need for time-consuming energy calculations using DFT. This hybrid approach, termed ML-accelerated GA (MLaGA), incorporates two levels of evaluation: the ML-predicted energy for quick screening and DFT calculation for final verification. By allowing the GP model to rapidly eliminate less promising candidates, the MLaGA achieved a 50-fold reduction in the number of required energy evaluations compared to a traditional GA.

It should be noted that in several of the examples described above, the GA is not really generative; instead, the GA was used as an optimization tool on an existing set of structures. However, by combining MOF features in new combinations, it is possible to generate new structures that have not previously been considered. One drawback of GAs is their slow convergence in complex and high-dimensional search spaces.[103] Also, since GAs are heuristic, they do not guarantee finding the global optimum. Instead, they rely on stochastic processes that may converge to local minima in the search space.[112,113] This heuristic nature requires careful tuning of parameters, such as mutation rate, crossover rate, and population size, to find the right balance between exploring new solutions and refining existing ones.[114,115] Poorly chosen parameters may lead to premature convergence, a loss of diversity, or an inefficient search process.[116] Additionally, evaluating the fitness of each

individual in a population can be computationally expensive, especially when dealing with large populations or many generations. To address this, many recent applications of GAs in materials design integrate surrogate models such as neural networks to predict the performance of generated materials.[117–120] This combination reduces the need for costly computational simulations to evaluate material performance, thus lowering overall resource requirements and speeding up the optimization process.

### 2.5 Reinforcement learning (RL)

RL is a machine learning approach that enables an agent to learn optimal strategies for decision-making through interactions with an environment. In the context of material design, RL can be applied to optimize material properties by sequentially adjusting design parameters based on feedback from simulated or experimental evaluations. As illustrated in Fig. 6a, the workflow involves three key components: the agent, the environment, and the reward signal.[121] The agent represents the model tasked with proposing material designs. The environment evaluates these designs, either through simulations or experiments, and provides feedback to the agent in the form of a reward signal. The reward quantifies how well a material meets the desired target properties, such as gas adsorption capacity, thermal stability, or mechanical strength. The workflow begins with the agent proposing an initial material design, which is evaluated by the environment. Guided by a policy, the agent then modifies the material's design parameters to map the current design state to the next action. After each action, the agent receives a reward, which measures the success of the

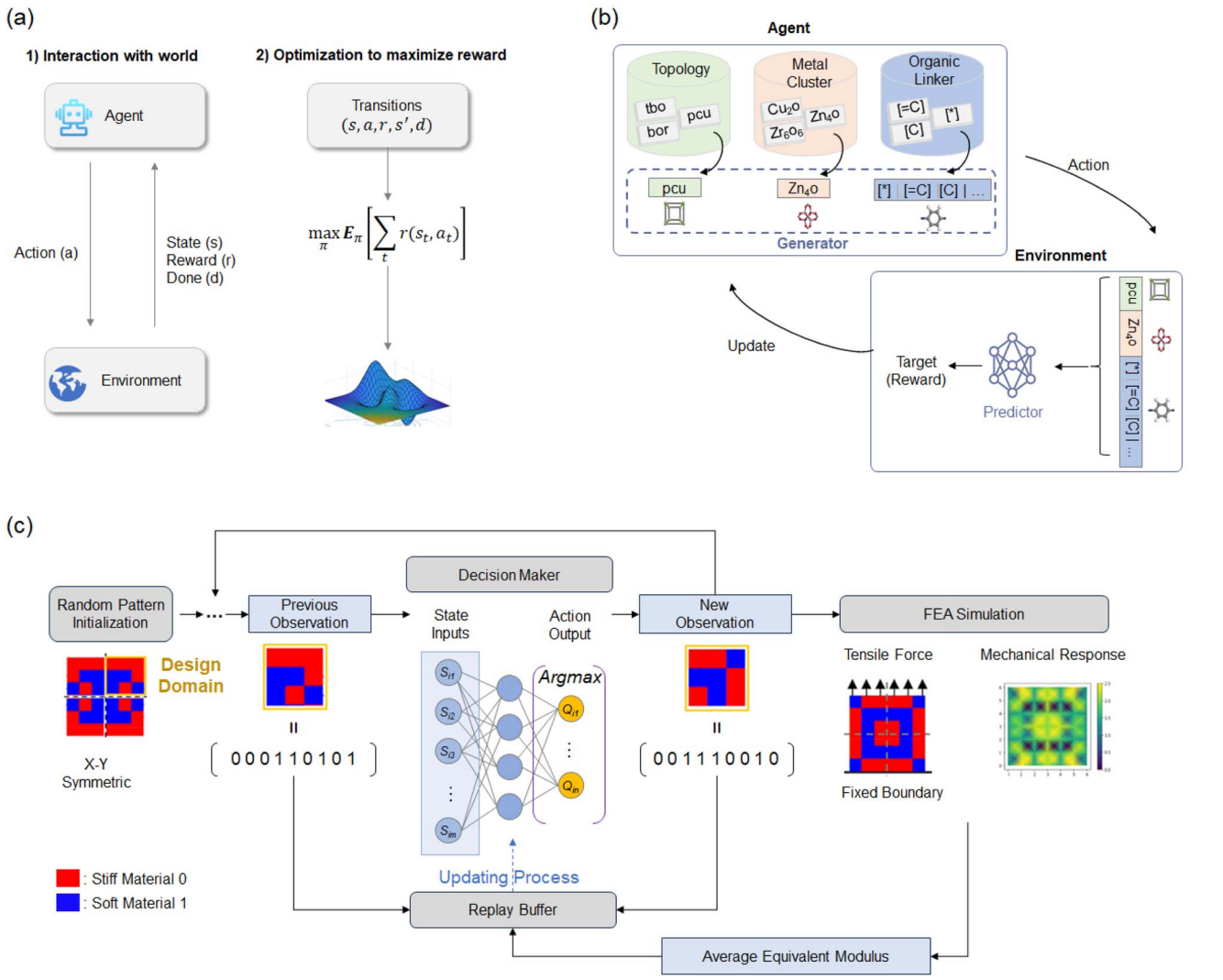


Fig. 6 (a) In RL, an agent learns to make decisions by interacting with an environment, receiving rewards or penalties, and adjusting its strategy through trial and error to improve outcomes. (b) Schematic of the RL framework for generative design of MOFs for direct air capture of $CO_2$. The agent (generator) generates a MOF structure, which the environment (predictor) evaluates to return a reward. The agent uses this feedback to iteratively generate improved MOF structures with desirable properties. Adapted from ref. 122. Licensed under CC BY 3.0. (c) Schematic of the collaborative deep RL system pipeline for optimal digital material discovery, using a 3 × 3 design space of 2D soft and stiff material components. Adapted with permission from ref. 123. Copyright 2021 American Chemical Society.

modified design in achieving the target properties. Over time, the agent uses this feedback to refine its policy, improving its ability to predict which actions are likely to yield better designs.

Mathematically, the agent's goal is to find the optimal policy $\pi^*$ that maximizes the expected cumulative reward:

$$\pi^* = \underset{\pi}{\operatorname{argmax}} \mathbb{E}_\pi \left[ \sum_{t=0}^{\infty} \gamma^t r_{t+1} | s_0 = s \right]$$

The policy $\pi$, often denoted as $\pi(a|s)$, specifies the agent's behavior, defining the probability of taking action $a$ in state $s$. The expected value of the cumulative reward, donated as $\mathbb{E}_\pi$, is calculated over all possible trajectories (*i.e.*, sequences of states and actions) that are generated by following the policy $\pi$. The cumulative reward $\sum_{t=0}^{\infty} \gamma^t r_{t+1}$ is the discounted sum of rewards over time, where the immediate reward $r_{t+1}$ is received by the agent after taking action $a_t$ in state $s_t$ and transitioning to state $s_{t+1}$. The discount factor $\gamma \in [0,1]$, balances short-term and long-term rewards, with the agent only considering immediate rewards if $\gamma = 0$ and giving equal weight to immediate and future rewards if $\gamma = 1$.

RL treats the discovery process as a series of interdependent decisions, where each step builds upon the previous one to optimize the overall outcomes. This makes RL well-suited for handling complex, multi-step synthesis or optimization tasks. A key challenge in RL for material design is balancing exploration and exploitation. Exploration seeks novel material configurations, while exploitation refines known high-performing structures. Too much exploration increases computational costs and inefficiency, while excessive exploitation risks missing superior materials. Striking this balance is crucial for optimizing both efficiency and discovery.

Park *et al.*[122] used a deep RL model to design MOFs for direct air capture of $CO_2$. Their RL model consists of two key components: a generator (agent) that proposes MOF structures and a predictor (environment) that evaluates these structures based on their estimated $CO_2$ heat of adsorption and $CO_2/H_2O$ selectivity. The training data was derived from computationally generated MOFs, constructed using PORMAKE,[109] a tool developed by the authors to assemble MOF structures from predefined metal nodes, organic linkers, and topologies. The RL workflow begins with a pre-training phase, where the generator learns how to construct chemically valid MOFs by analyzing a large dataset of MOFs. The predictor is trained separately from GCMC-computed target properties. Once pre-trained, the RL process starts, with the generator sequentially selecting a topology, metal cluster, and organic linker to propose new MOF structures. These structures are then evaluated by the predictor, which estimates their adsorption properties and provides a reward signal to refine the generator's design strategy. To balance the trade-off between exploitation and exploration, the RL model employs a dual-generator system: one biased toward existing high-performance structures and another encouraging novel MOF exploration. The RL process iterates over multiple rounds, each time refining the generator's ability to propose MOFs that meet the dual objectives of strong $CO_2$ adsorption and high $CO_2/H_2O$ selectivity—a significant challenge due to the strong water affinity of many materials. Their study demonstrated that with each round of training, the generated MOFs increasingly met the desired property criteria. The RL-optimized MOFs exhibited some of the highest reported values for $CO_2$ heat of adsorption (~62 kJ mol) and $CO_2/H_2O$ selectivity, indicating a strong affinity for $CO_2$ under atmospheric conditions (400 ppm, 1 bar, 298.15 K) for direct air capture (DAC). Further chemical analysis of the generated MOFs revealed distinctive features in top-performing structures, such as Mn and Eu-based metal clusters in MOFs with high $CO_2$ adsorption, and Cu and Zn-based clusters in MOFs with high $CO_2/H_2O$ selectivity.

Zheng *et al.*[124] applied a policy-gradient RL framework to iteratively distribute hydroxyl and epoxide groups on the basal plan of graphene to maximize material toughness. This approach successfully addressed the combinatorial complexity of the problem, achieving optimized designs within a vast solution space of up to $10^{16}$ possibilities. Additionally, RL can incorporate different objectives during its learning process, allowing it to optimize multiple properties simultaneously.[125] For example, Sui *et al.*[123] used a deep RL framework to optimize two mechanical properties of complex materials, specifically targeting both stiffness and toughness (Fig. 6c). The authors demonstrated how RL can balance conflicting design objectives and explore vast design spaces efficiently. These studies, although not directly focused on porous materials, demonstrate the efficiency and innovation of RL in multi-objective-driven design.

The process of learning through trial and error, which is central to RL, typically requires a large number of samples or simulations to find an optimal solution.[126–129] This issue is further compounded in material design applications, where the state space (*i.e.*, the possible configurations of materials) is extremely large[123,130] and the relationship between actions (design decisions) and rewards (material properties) is highly non-linear.[131] For instance, the deep RL framework developed by Park *et al.*[122] required extensive computational resources due to the sheer scale of data and iterative training. The generator was trained on 1 540 889 MOFs, validated on 385 223, and tested on 10 000, running for 50 epochs with a batch size of 128. The predictor, trained separately over 100 epochs, relied on ~33 000 MOFs for $CO_2$ heat of adsorption and ~24 000 for $CO_2/H_2O$ selectivity, requiring costly GCMC simulations for data generation. Based on our group's recent benchmarks,[132] such simulations take on average 3–4 hours per MOF using the CPU-based RASPA2 code. Even with our recently developed gRASPA code,[132] which achieves a 20-fold speedup on a single A100 GPU node, generating these datasets still requires ~2000 GPU-hours for $CO_2$ heats of adsorption and ~1500 GPU-hours for $CO_2/H_2O$ selectivity. The RL phase further increased the burden, with each policy gradient training epoch selecting 8000 MOFs and running over 20 epochs. The repeated evaluations, training cycles, and dependence on high-fidelity simulation data made this RL approach computationally expensive.

### 2.6 Large language models (LLMs)

The advance in generative AI best known to the general public is LLMs like GPT[133] and BERT,[134] which have gained significant attention across various fields due to their ability to process and generate human-like text. These models are pre-trained on extensive bodies of text, often containing billions of words, enabling them to learn complex patterns in language, such as grammar, semantics, and context. Central to their function is the concept of a token, which refers to a unit of text (*e.g.*, words, prefixes, or punctuation) that the model uses to understand and generate language. Longer text is broken down into these smaller tokenized units for processing. LLMs operate by using a transformer architecture, which excels at capturing contextual relationships in sequential data. A key component of the transformer is the attention mechanism, which allows the model to focus on relevant parts of the input when generating output. For example, in text generation, the attention mechanism helps the model decide which words in a sentence are most relevant for predicting the next word. This mechanism enables the model to weigh the relevance of each token dynamically, improving its ability to generate coherent and contextually accurate outputs. While the primary applications of LLMs are in natural language processing, their versatility has expanded significantly, and they are rapidly finding applications in materials research.

Recently, LLMs have been applied to understand and predict material properties, generate new material compositions, and suggest synthesis pathways based on literature and databases. Their versatility, combined with their integration with other generative models, makes them a promising tool for advancing material design. Adapting LLMs for material design involves fine-tuning them on specialized datasets containing information about suitable material features like their chemical compositions and desired properties. One key aspect of fine-tuning LLMs is prompt engineering, where the researcher interacts with the LLM through carefully designed prompts to elicit specific and meaningful responses. By crafting prompts that guide the model's reasoning and knowledge retrieval, researchers can optimize LLM outputs for specific tasks, such as synthesis planning and material property prediction. Once fine-tuned, LLMs can carry out several important tasks within the material design process (Fig. 7a).[135] For instance, LLMs can search for known materials and provide detailed descriptions of their structures and properties.[136] In this role, LLMs serve as highly sophisticated encyclopedias, offering researchers comprehensive and easily accessible information on existing materials.[136–138]

A key challenge in human-AI collaborative materials design lies in enabling AI to effectively learn and utilize existing human knowledge. LLMs have shown significant potential in organizing and interpreting data extracted from the literature. Zheng *et al.*[142] employed prompt engineering to guide GPT-3.5-turbo in automating the extraction of MOF synthesis conditions from scientific publications, addressing the common issue of information hallucination in LLMs. They developed a Chem-Prompt Engineering strategy, which integrates principles such as minimizing hallucination through carefully designed queries, providing explicit and structured instructions, and ensuring standardized output formats for reliable data extraction. To achieve this, they constructed a multi-step workflow that enables ChatGPT to parse, filter, and summarize synthesis data with high accuracy. Their approach combined direct summarization of preselected experimental sections, automated classification of synthesis-related paragraphs, and embedding-based filtering to enhance processing efficiency. Applying this system, they extracted 26 257 synthesis parameters for approximately 800 MOFs with an accuracy of 90–99%. The extracted dataset was further used to train a machine learning model that achieved over 87% accuracy in predicting MOF crystallization. Further, they developed a data-driven MOF chatbot capable of answering chemistry-related queries based on literature-derived synthesis conditions and applied it to linker design for water harvesting applications.[143] These studies demonstrate how LLMs can be effectively harnessed for automated knowledge extraction and predictive modeling in chemistry, requiring no coding expertise. This makes them particularly accessible to researchers who may lack coding training.

ChatGPT has also been applied to assist in the design and synthesis of porous materials. For instance, Zheng *et al.*[139] proposed a framework integrating GPT-4 into chemical experimentation to enhance the collaborative dynamic between humans and AI in the synthesis and characterization of MOFs. The system leverages GPT-4's natural language capabilities to streamline complex processes and make design guidance accessible to humans. This collaborative platform is designed to operate in iterative cycles where researchers execute tasks based on GPT-4's suggestions and provide feedback, enabling the model to refine its understanding and recommendations over time. The framework comprises three interconnected phases (Fig. 7b). The first phase, Reticular ChemScope, establishes a detailed research blueprint by breaking the project into manageable activities. The second phase, Reticular ChemNavigator, serves as the central hub, assessing progress and suggesting three possible actions for the researcher to undertake. These suggestions are developed using human feedback, ensuring they align with experimental results. Lastly, the Reticular ChemExecutor offers step-by-step procedural guidance tailored to the selected task, enabling precise execution. The iterative process enables GPT-4 to adapt and learn from both successes and failures, effectively acting as a virtual mentor.

Jablonka *et al.*[144] demonstrated that GPT-3, originally trained on diverse text data, can be fine-tuned for material property prediction. Notable examples involved predicting Henry coefficients, heat capacities, and water stability of MOFs, using datasets as small as hundreds of samples. GPT-3 achieved these predictions with errors lower than conventional ML models in low-data scenarios, which is remarkable.

Another advantage of LLMs in material design is their versatility. LLMs can be fine-tuned for a variety of tasks, ranging from generating textual descriptions of known material structures to predicting the properties of new materials.[145] For

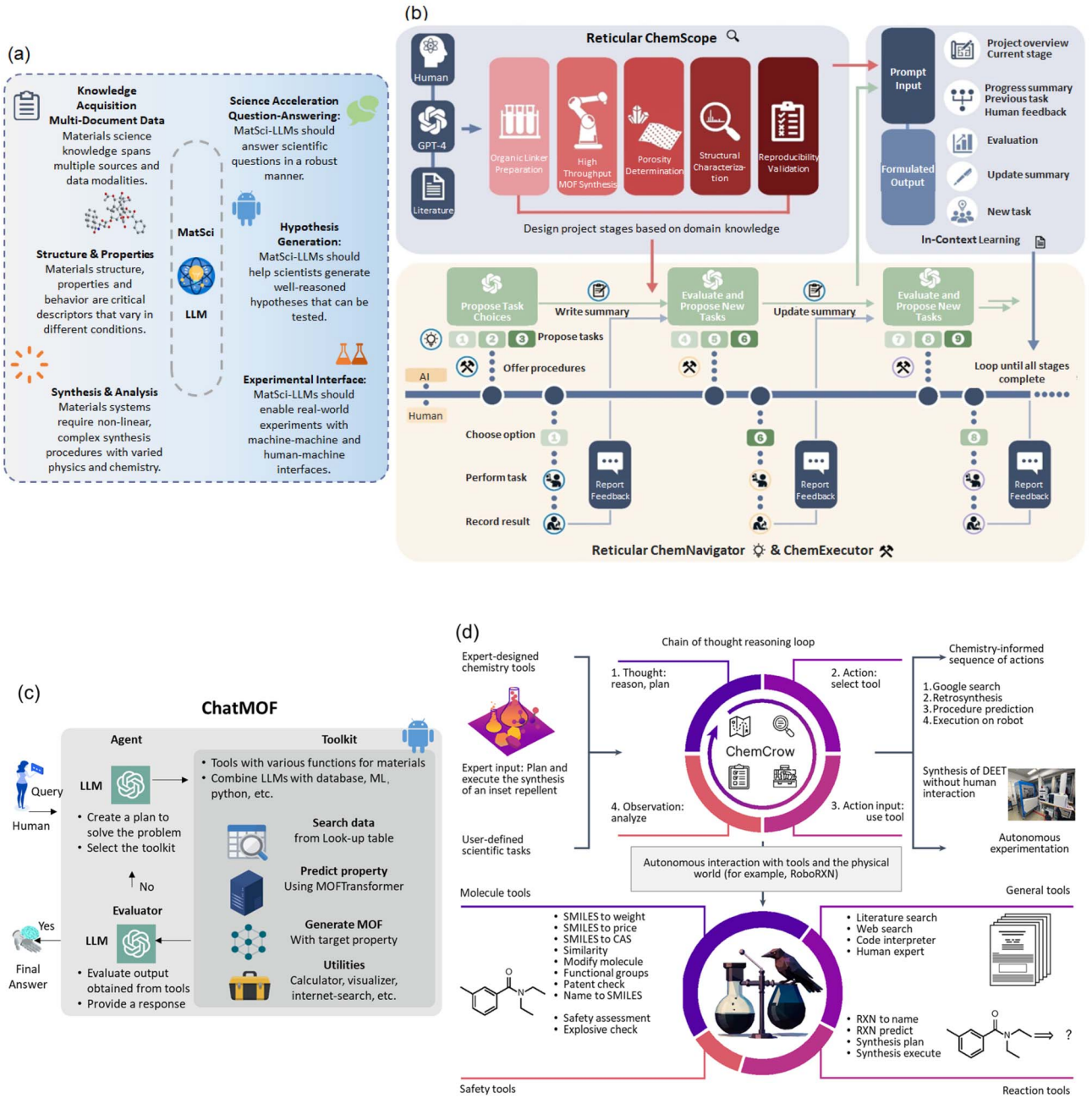


Fig. 7 (a) Overview of materials science (Mat. Sci.) LLM requirements for knowledge acquisition and science acceleration. Adapted from ref. 135. Licensed under CC BY 4.0. (b) Schematic of the GPT-4 Reticular Chemist, which includes three states: "ReticularChemScope," "ReticularChemNavigator," and the "ReticularChemExecutor." Each state uses GPT-4 with distinct prompts, operating entirely through natural language, without coding. Adapted with permission from ref. 139. Copyright 2023 Wiley-VCH. (c) Schematic of ChatMOF featuring three core components: agent, toolkit, and evaluator. The agent formulates a plan based on a user query, selects an appropriate toolkit, and the evaluator provides the final response. Adapted from ref. 140. Licensed under CC BY 4.0. (d) Overview of the task-solving process in ChemCrow, which employs an automated, iterative chain-of-thought process to select tools, define inputs, and determine solution pathways. Toolsets in ChemCrow include modules for molecules, safety, reactions, and general-purpose tasks. Adapted from ref. 141. Licensed under CC BY 4.0.

example, Kang *et al.*[140] developed ChatMOF, a LLM specifically designed for predicting and generating MOFs. They employed ChatMOF as a central coordinator, facilitating appropriate responses to user requests through three main components – an agent, a toolkit, and an evaluator (Fig. 7c). The agent breaks down queries, selects the best approach, and selects an appropriate tool from the toolkit. The evaluator then determines if the results are sufficient or if further refinement is needed. The toolkit consists of four categories: Searcher, retrieving information from existing MOF data; Predictor, using the MOFTransformer[146] model to predict desired material properties; Generator, applying a genetic algorithm to create

new MOFs; and Utilities, handling general tasks like internet queries and calculations. ChatMOF achieves high accuracy rates by leveraging specialized tools for specific tasks: 96.9% for search tasks, 95.7% for prediction tasks, and 87.5% for structure generation. This model represents a significant step toward greater AI autonomy in nanoporous design.

As a more general tool, Bran *et al.*[141] introduced ChemCrow, a chemistry-focused LLM agent designed to tackle tasks in organic synthesis, drug discovery, and materials design. By integrating 18 expert-developed tools with GPT-4, ChemCrow enhances the LLM's chemistry capabilities (Fig. 7d). ChemCrow successfully planned and executed the synthesis of various compounds, including an insect repellent and organocatalysts, and aided in discovering a novel chromophore. Expert chemists found that ChemCrow outperformed GPT-4 in chemical accuracy, logical reasoning, and response completeness, especially when handling complex problems.

Inspired by these advancements, experimental chemists can begin integrating pre-trained LLM assistants into their lab workflows for tasks such as literature text mining and synthesis planning. For example, Zheng *et al.*[142] used ChatGPT to extract MOF synthesis conditions including temperature, solvent, concentration, and time parameters from published papers without requiring coding expertise, achieving high accuracy through carefully designed prompts. In another study, Zheng *et al.*[139] integrated GPT-4 into the experimental design process to propose actionable synthesis steps and provide step-by-step procedural guidance for MOF preparation. More advanced use cases may involve combining LLMs with lab management or automation tools to suggest experimental designs, plan sequential workflows, or automate documentation, where the LLM acts as an accessible interface translating textual instructions into structured experimental plans, as demonstrated by the ChemCrow framework.[141]

Despite their powerful capabilities, LLMs pose challenges related to interpretability. The decision-making process within these models is often seen as a "black box," making it difficult for researchers to understand why a particular material structure was suggested by the model.[135] This lack of transparency can be a hurdle in scientific research, where understanding the rationale behind a prediction is often as important as the prediction itself. Furthermore, training and deploying LLMs from scratch is extremely expensive, making it prohibitively costly for most research groups. A common approach is to leverage pre-trained models such as GPT-4.0. However, there are two key points to keep in mind. First, these models are typically trained on publicly available data rather than the full body of scientific literature, which often resides behind publisher paywalls. To adapt them for specific materials design tasks, researchers need to input relevant datasets and conduct meticulous prompt engineering. Second, some of these models are not free and operate on a token-based pricing system, meaning that for research topics requiring extensive materials data or involving multiple complex prompts, the associated costs can become substantial. It is also important to note that when using LLMs for literature-based data mining, one must consider that most published studies predominantly report positive results while omitting negative or less favorable results. This imbalance introduces a "survivorship bias," potentially skewing the model's understanding of structure–property relationships.[147] As a result, the model may overestimate the effectiveness of certain design strategies while overlooking potentially valuable insights hidden in unreported or unpublished data. Addressing this issue requires careful curation of training datasets, including efforts to incorporate negative results from supplementary materials in published articles, preprints, or experimental databases to improve model robustness and reliability.

# 3. Discussion

## 3.1 Data requirements of generative AI algorithms

The effectiveness of generative AI algorithms in material design is highly dependent on their data requirements. The previously introduced models, excluding GAs and LLMs, generally require large volumes of high-quality data to achieve optimal performance. This reliance on extensive training data poses a challenge for porous material design, where data availability is often limited, especially for applications beyond adsorption, like catalysis. In contrast, GAs are more flexible and can work well with a smaller sample size. However, their performance can be compromised if the initial population lacks depth or variety. For instance, if the dataset includes only a narrow range of features or insufficiently diverse samples, it may fail to represent the broader design space effectively. LLMs benefit from large textual datasets but still need fine-tuning using prompt engineering based on specialized material data to achieve good results.[148]

Some approaches have been adopted to mitigate data limitations. Data augmentation, for example, involves generating new material samples by applying functionalization to existing samples[72] or by permuting and combining structural building blocks and topologies to create a vast number of new structures. For instance, the number of possible MOF structures can reach up to 247 trillion.[109] This enhances data diversity and improves the generative capability of the models. Similarly, transfer learning leverages pre-trained models trained on large, general-purpose datasets, and adjusting them for specific tasks can potentially reduce the need for extensive data. Accelerating the computation of material properties is another promising direction. This can be achieved by developing faster and more accurate force field-based methods (including machine-learned interatomic potentials) or leveraging machine learning models (surrogate models) for direct and rapid property prediction.[149,150]

## 3.2 User-friendliness and scalability

The user-friendliness of generative AI methods varies depending on how they are implemented and the level of technical expertise required. Diffusion models, VAEs, and GANs are accessible to many users, as they typically involve working with pre-written scripts or platforms which require only basic programming skills. Among these, GANs may appear more

approachable, as many pre-trained models are available, and generating outputs can be as simple as modifying parameters. GAs are intuitive to use due to their heuristic nature and relatively simple setup, making them accessible to users with limited machine learning experience. RL, on the other hand, typically presents a steeper learning curve, as designing reward functions and configuring interactive environments can be complex. While pre-existing frameworks can simplify RL implementation, effective use often demands a deeper understanding of training dynamics and policy optimization. In comparison, LLMs are becoming more user-friendly with advancements in tools and interfaces, such as Hugging Face's Transformers library,[151] though effective fine-tuning and deployment of LLMs often still demands familiarity with model architecture, data preprocessing, and prompt engineering.

Scalability is another critical factor in applying these algorithms effectively. Diffusion models, while able to generate chemically viable samples, can require significant computational resources when handling large datasets, with the training process taking multiple GPU days.[93] GANs are also resource-intensive, particularly during training, although they become more efficient for generating samples once trained. For instance, Dan *et al.* introduced MatGAN, which was trained on more than 380 000 inorganic materials. Once trained, MatGAN reached a novelty of 92.5% and a validity of 84.5% when generating more than 2 million samples, demonstrating the model's efficiency in producing viable materials following extensive training.[65] GAs are inherently scalable due to their parallel nature, allowing the evaluation of multiple candidate solutions simultaneously. However, their performance may decrease when working on very large populations or many generations, as the computational cost can become prohibitive. RL can optimize multiple objectives through iterative learning, but the complexity of environments often necessitates considerable amounts of agent interactions with the environment and advanced hardware.[152] VAEs are somewhat more scalable compared to GANs, as they can generate new samples even with limited data, though they still benefit from larger datasets for improved performance. LLMs, while highly scalable and able to process large amounts of text data, demand substantial computational resources for training and deployment. As these models grow, the need for resources also increases, which can limit their use for many research groups.

A comparative summary of the strengths and limitations of these six generative AI approaches is provided in Table 2 to guide their selection for different material design tasks.

### 3.3 Guidance for future material design

Designing nanoporous materials using generative AI requires a systematic approach that begins with ensuring the quality and representativeness of the training data. High-quality datasets that capture the structural diversity and property relationships

Table 2 Comparison of the strengths and limitations of the generative AI methods utilized for nanoporous materials design

| Generative AI method | Strengths | Limitations/challenges |
|---|---|---|
| Generative adversarial networks (GANs) | • Generates realistic, high-quality structures<br>• Effective at modeling complex data distributions<br>• Conditional GANs can target specific properties | • Training instability and potential mode collapse<br>• Difficulty capturing structural diversity in complex materials like MOFs and COFs<br>• Requires large datasets and careful hyperparameter tuning |
| Variational autoencoders (VAEs) | • Smooth and continuous latent space for interpolation and optimization<br>• Stable and efficient training | • May fail to generate valid or realistic structures<br>• Limited disentanglement in latent representations |
| Diffusion models (DMs) | • Effective at learning complex distributions without mode collapse<br>• Generates diverse and complex structures like MOFs | • Computationally expensive due to iterative denoising<br>• Requires large high-quality training datasets |
| Genetic algorithms (GAs) | • No requirement for gradient information<br>• Effective at exploring vast and discrete design spaces<br>• Simple concept and relatively easy to implement | • Convergence can be slow, especially in high-dimensional spaces<br>• May converge to locally optimal material structures rather than the global optimum<br>• Computationally expensive when combined with simulation-based fitness evaluations |
| Reinforcement learning (RL) | • Enables sequential decision-making for goal-directed design<br>• Can optimize multiple objectives and incorporate feedback<br>• Flexible for integration with experimental workflows | • Typically requires a large number of samples and evaluations<br>• Designing effective reward functions can be challenging |
| Large language models (LLMs) | • Versatile in tasks such as literature mining, property prediction, and structure generation<br>• User-friendly *via* natural language prompts<br>• Can integrate with other AI models as AI agent or assistant | • Limited interpretability ("black box" outputs)<br>• Training from scratch is resource-intensive<br>• Prompt engineering and fine-tuning for specialized tasks can be challenging |

of porous materials are crucial. Researchers can rely on domain-specific databases, such as the CoRE MOF database[73] for structures and MOFX-DB[83] for adsorption data, or develop custom datasets tailored to their objectives. To address limitations in data availability, techniques introduced above like data augmentation, which introduce noise or transformations, and transfer learning, can help diversify datasets and improve model robustness.

Building on this foundation, the choice of a suitable generative AI algorithm is critical and should align with the specific design task. For instance, DMs are effective for generating high-resolution structures with complex pore architectures, such as MOFs designed for $CO_2$ capture. GAs are well-suited for early-stage exploration of vast design spaces. RL is particularly advantageous for sequential design tasks, as it iteratively refines designs based on feedback. LLMs can streamline literature review, propose initial material structures, and guide synthesis planning based on textual inputs.

As described in the corresponding sections above, different generative models exhibit varying strengths in generating materials with defined target properties (this is sometimes referred to as conditional design or inverse design; in this review, we have simply referred to it as design or material design). A short summary is provided in Table 2. VAEs are well suited for conditional generation due to their continuous latent space, enabling property optimization through latent space navigation.[72] GANs can incorporate property conditions through conditional GAN architectures, although training stability remains a challenge.[55] DMs can implement conditioning to guide generation toward desired properties but often require large datasets and significant computational resources.[80,82] Reinforcement learning inherently supports conditional design by optimizing reward functions defined by target properties, while genetic algorithms impose conditions through fitness functions, acting more as optimization rather than true generative conditioning. Large language models can provide conditional outputs *via* prompt engineering,[142] but their application in directly generating material structures conditioned on quantitative properties is still emerging. Improving conditional generation capabilities across these models will accelerate the effective design of materials with tailored functionalities.

To further enhance the material design process, hybrid and ensemble approaches can be adopted. For example, the MOFFUSION model,[81] introduced in Section 2.3, combines the generative power of DMs with the dimensionality reduction and reconstruction capabilities of VQ-VAE, making it computationally feasible for DMs to process high-dimensional data. Likewise, LLMs have recently been explored as powerful tools for the early stages of material design, where they can generate initial material concepts by drawing on patterns from large scientific literature and databases.[153] Studies have demonstrated that these models can suggest candidate compositions and synthesis routes,[154] as well as assist in property prediction.[155] Building on this emerging capability, such initial outputs may be further refined using downstream algorithms like genetic algorithms or diffusion models. This combination can leverage the unique strengths of each algorithm to enable innovative solutions. Additionally, LLMs can be trained as AI assistants capable of making decisions, automating the selection of suitable models, and mining datasets tailored to specific applications.[140] These hybrid strategies allow researchers to address complex design challenges more effectively.

Another important limitation of current generative AI models for MOFs and COFs is their restricted ability to generate new topologies. Most existing approaches use topologies from the training dataset, focusing primarily on varying building blocks or functional groups. While this strategy enables the generation of chemically valid and potentially synthesizable structures, it limits the discovery of frameworks with novel topologies, which may become a bottleneck in advancing reticular material design. Future improvements could focus on developing models that integrate topology generation as part of the design process. However, given that mathematicians have identified thousands of topologies, a simpler strategy might incorporate these topologies, which are known mathematically but are new to MOFs.

In addition, a critical task for generative AI methods is careful selection of appropriate descriptors to distinguish one material from another. Defining relevant evaluation metrics for specific applications to ensure accurate and meaningful results is also critical. For example, in adsorption separations, there is often a tradeoff between selectivity, working capacity, and other properties that should be considered. Finally, establishing an iterative feedback loop between AI predictions and experimental or computational validations is essential for refining models and ensuring reliability. Outputs from generative models can be validated using computational methods such as DFT, MD, or GCMC simulations. In addition, integrating experimental workflows allows researchers to verify the performance of AI-generated materials, enabling continuous improvement of the models over time based on real-world data. This iterative refinement process bridges the gap between computational predictions and practical implementation. Currently, experimental validation rates for AI-generated materials remain low, due to synthesis challenges and stability issues. However, there are successful cases, such as the synthesis of MOF NOTT-101/OEt reported by Chung *et al.*,[107] that demonstrate the promising future of AI-enabled materials discovery and its potential to accelerate the design-to-synthesis process. Improving the translation of generative AI outputs into experimentally accessible synthesis procedures and validated nanoporous materials remains a critical task, and it presents an exciting opportunity to integrate AI design with automated synthesis and high-throughput experimental workflows in the future.

## 4. Conclusions and perspective

In this review, we provided an overview of six promising generative AI approaches for designing new porous materials: GANs, VAEs, DMs, GAs, RL, and LLMs. We highlighted the unique advantages and challenges of each. DMs and GANs are excellent for generating chemically viable samples and diverse

outputs, making them suitable for complex design tasks. GAs, with their heuristic nature, are well-suited for exploring broad design spaces and optimizing specific material properties, even with limited initial data. VAEs are effective for exploring and interpolating between different material designs. RL is particularly useful for multi-step processes or dynamic design objectives by balancing the trade-offs between different properties and optimizing synthesis pathways. LLMs offer versatility in generating new materials based on textual input and are becoming very user-friendly. The success of these generative AI approaches depends heavily on the quality of training data, the expertise applied to fine-tuning and implementation, as well as the specific nature of the design task.

Generative AI is shaping new trends in material design, revolutionizing the way we design and discover new materials like zeolites and MOFs. Looking forward, several promising research directions could significantly advance the field of generative AI in material design. One important focus is to improve the interpretability of generative AI models, particularly for LLMs and deep learning based methods. Developing frameworks to explain the reasoning behind generated suggestions will enhance user experience and increase trust in automated design processes. Another exciting direction is integrating generative AI models with experimental workflows in real time, enabling rapid feedback between computational predictions and laboratory results to accelerate material discovery. As these methods become more powerful and user-friendly, they are poised to become a transformative tool to accelerate the discovery and optimization of the next generation of nanoporous materials.

## Data availability

This is a review article and does not report new data.

## Conflicts of interest

R. Q. S. has a financial interest in Numat, a company that is commercializing metal–organic frameworks. The other authors have no conflicts of interest to declare.

## Acknowledgements

This work was supported by the U.S. Department of Energy, Office of Basic Energy Sciences, Division of Chemical Sciences, Geosciences and Biosciences under Award No. DE-SC0023454, as part of the Computational and Theoretical Chemistry Program. H.C. acknowledges Southern Methodist University for start-up funds.

## References

1 H. L. Nguyen, Reticular design and crystal structure determination of covalent organic frameworks, *Chem. Sci.*, 2021, **12**(25), 8632–8647.
2 O. M. Yaghi, M. O'Keeffe, N. W. Ockwig, H. K. Chae, M. Eddaoudi and J. Kim, Reticular synthesis and the design of new materials, *Nature*, 2003, **423**(6941), 705–714.
3 B. Sanchez-Lengeling and A. Aspuru-Guzik, Inverse molecular design using machine learning: Generative models for matter engineering, *Science*, 2018, **361**(6400), 360–365.
4 Y. J. Colón and R. Q. Snurr, High-throughput computational screening of metal–organic frameworks, *Chem. Soc. Rev.*, 2014, **43**(16), 5735–5749.
5 E. Ren, P. Guilbaud and F. X. Coudert, High-throughput computational screening of nanoporous materials in targeted applications, *Digital Discovery*, 2022, **1**(4), 355–374.
6 Y. Y. Song and Y. Lu, Decision tree methods: applications for classification and prediction, *Shanghai Arch. Psychiatry*, 2015, **27**(2), 130–135.
7 L. Breiman, Random forests, *Mach. Learn.*, 2001, **45**(1), 5–32.
8 T. Chen, and C. Guestrin, XGBoost: A scalable tree boosting system, in *KDD'16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 785–794.
9 R. T. Yang, *Adsorbents: Fundamentals and applications*, Wiley, 2003.
10 P. A. Jacobs, E. M. Flanigen, J. C. Jansen, and H. van Bekkum. *Introduction to zeolite science and practice*Elsevier Science, 2nd edn, vol. 137, 2001.
11 R. Devi, V. Kumar, S. Kumar, M. Bulla, A. Jatrana, R. Rani, A. K. Mishra and P. Singh, Recent advancement in biomass-derived activated carbon for waste water treatment, energy storage, and gas purification: a review, *J. Mater. Sci.*, 2023, **58**(30), 12119–12142.
12 S. Yuan, L. Zou, J. S. Qin, J. Li, L. Huang, L. Feng, X. Wang, M. Bosch, A. Alsalme, T. Cagin and H. C. Zhou, Construction of hierarchically porous metal–organic frameworks through linker labilization, *Nat. Commun.*, 2017, **8**(1), 15356.
13 Y. Xie, W. Wang, Z. Zhang, J. Li, B. Gui, J. Sun, D. Yuan and C. Wang, Fine-tuning the pore environment of ultramicroporous three-dimensional covalent organic frameworks for efficient one-step ethylene purification, *Nat. Commun.*, 2024, **15**(1), 3008.
14 A. Nagai, Z. Guo, X. Feng, S. Jin, X. Chen, X. Ding and D. Jiang, Pore surface engineering in covalent organic frameworks, *Nat. Commun.*, 2011, **2**(1), 536.
15 X. L. Lv, M. Tong, H. Huang, B. Wang, L. Gan, Q. Yang, C. Zhong and J. R. Li, A high surface area Zr(IV)-based metal–organic framework showing stepwise gas adsorption and selective dye uptake, *J. Solid State Chem.*, 2015, **223**, 104–108.
16 T. Düren, F. Millange, G. Férey, K. S. Walton and R. Q. Snurr, Calculating geometric surface areas as a characterization tool for metal–organic frameworks, *J. Phys. Chem. C*, 2007, **111**(42), 15350–15356.
17 H. Wang, X. Dong, J. Lin, S. J. Teat, S. Jensen, J. Cure, E. V. Alexandrov, Q. Xia, K. Tan, Q. Wang, D. H. Olson, D. M. Proserpio, Y. J. Chabal, T. Thonhauser, J. Sun,

Y. Han and J. Li, Topologically guided tuning of Zr-MOF pore structures for highly selective separation of C6 alkane isomers, *Nat. Commun.*, 2018, **9**(1), 1745.

18 Y. Xie, J. Li, C. Lin, B. Gui, C. Ji, D. Yuan, J. Sun and C. Wang, Tuning the topology of three-dimensional covalent organic frameworks *via* steric control: From pts to unprecedented ljh, *J. Am. Chem. Soc.*, 2021, **143**(19), 7279–7284.

19 X. Kang, X. Han, C. Yuan, C. Cheng, Y. Liu and Y. Cui, Reticular synthesis of tbo topology covalent organic frameworks, *J. Am. Chem. Soc.*, 2020, **142**(38), 16346–16356.

20 M. P. Suh, H. J. Park, T. K. Prasad and D. W. Lim, Hydrogen storage in metal–organic frameworks, *Chem. Rev.*, 2012, **112**(2), 782–835.

21 Y. F. Zhang, Z. H. Zhang, L. Ritter, H. Fang, Q. Wang, B. Space, Y. B. Zhang, D. X. Xue and J. Bai, New reticular chemistry of the rod secondary building unit: Synthesis, structure, and natural gas storage of a series of three-way rod amide-functionalized metal–organic frameworks, *J. Am. Chem. Soc.*, 2021, **143**(31), 12202–12211.

22 H. Li, A. Dilipkumar, S. Abubakar and D. Zhao, Covalent organic frameworks for $CO_2$ capture: from laboratory curiosity to industry implementation, *Chem. Soc. Rev.*, 2023, **52**(18), 6294–6329.

23 R. V. Listyarini, J. Gamper and T. S. Hofer, Storage and diffusion of carbon dioxide in the metal organic framework MOF-5-A semi-empirical molecular dynamics study, *J. Phys. Chem. B*, 2023, **127**(43), 9378–9389.

24 J. R. Li, J. Sculley and H. C. Zhou, Metal–organic frameworks for separations, *Chem. Rev.*, 2012, **112**(2), 869–932.

25 S. K. Firooz and D. W. Armstrong, Metal-organic frameworks in separations: A review, *Anal. Chim. Acta*, 2022, **1234**, 340208.

26 X. J. Xie, H. Zeng, W. Lu and D. Li, Metal–organic frameworks for hydrocarbon separation: Design, progress, and challenges, *J. Mater. Chem. A*, 2023, **11**(38), 20459–20469.

27 D. Yang and B. C. Gates, Catalysis by metal organic frameworks: Perspective and suggestions for future research, *ACS Catal.*, 2019, **9**(3), 1779–1798.

28 A. Bavykina, N. Kolobov, I. S. Khan, J. A. Bau, A. Ramirez and J. Gascon, Metal–organic frameworks in heterogeneous catalysis: Recent progress, new trends, and future perspectives, *Chem. Rev.*, 2020, **120**(16), 8468–8535.

29 J. Lin, J. Ouyang, T. Liu, F. Li, H. H. Y. Sung, I. Williams and Y. Quan, Metal-organic framework boosts heterogeneous electron donor–acceptor catalysis, *Nat. Commun.*, 2023, **14**(1), 7757.

30 C. S. Diercks, Y. Liu, K. E. Cordova and O. M. Yaghi, The role of reticular chemistry in the design of CO2 reduction catalysts, *Nat. Mater.*, 2018, **17**(4), 301–307.

31 H. D. Lawson, S. P. Walton and C. Chan, Metal–organic frameworks for drug delivery: A design perspective, *ACS Appl. Mater. Interfaces*, 2021, **13**(6), 7004–7020.

32 X. Gao, X. Hai, H. Baigude, W. Guan and Z. Liu, Fabrication of functional hollow microspheres constructed from MOF shells: Promising drug delivery systems with high loading capacity and targeted transport, *Sci. Rep.*, 2016, **6**(1), 37705.

33 M. Wu and Y. Yang, Metal–organic framework (MOF)-based drug/cargo delivery and cancer therapy, *Adv. Mater.*, 2017, **29**(23), 1606134.

34 Y. Sun, L. Zheng, Y. Yang, X. Qian, T. Fu, X. Li, Z. Yang, H. Yan, C. Cui and W. Tan, Metal–organic framework nanocarriers for drug delivery in biomedical applications, *Nano-Micro Lett.*, 2020, **12**(1), 103.

35 G. Zhang, X. Li, Q. Liao, Y. Liu, K. Xi, W. Huang and X. Jia, Water-dispersible PEG-curcumin/amine-functionalized covalent organic framework nanocomposites as smart carriers for *in vivo* drug delivery, *Nat. Commun.*, 2018, **9**(1), 2785.

36 M. Pardakhti, E. Moharreri, D. Wanik, S. L. Suib and R. Srivastava, Machine Learning Using Combined Structural and Chemical Descriptors for Prediction of Methane Adsorption Performance of Metal Organic Frameworks (MOFs), *ACS Comb. Sci.*, 2017, **19**(10), 640–645.

37 J. Cui, F. Wu, W. Zhang, L. Yang, J. Hu, Y. Fang, P. Ye, Q. Zhang, X. Suo, Y. Mo, X. Cui, H. Chen and H. Xing, Direct prediction of gas adsorption *via* spatial atom interaction learning, *Nat. Commun.*, 2023, **14**(1), 7043.

38 S. M. Moosavi, A. Nandy, K. M. Jablonka, D. Ongari, J. P. Janet, P. G. Boyd, Y. Lee, B. Smit and H. J. Kulik, Understanding the diversity of the metal-organic framework ecosystem, *Nat. Commun.*, 2020, **11**(1), 4068.

39 Q. Zhu, Y. Gu and J. Ma, Digital descriptors in predicting catalysis reaction efficiency and selectivity, *J. Phys. Chem. Lett.*, 2025, **16**(9), 2357–2368.

40 X. Bai, Y. Li, Y. Xie, Q. Chen, X. Zhang and J. R. Li, High-throughput screening of CO2 cycloaddition MOF catalyst with an explainable machine learning model, *Green Energy Environ.*, 2025, **10**(1), 132–138.

41 M. Ducamp and F. X. Coudert, Prediction of Thermal Properties of Zeolites through Machine Learning, *J. Phys. Chem. C*, 2022, **126**(3), 1651–1660.

42 J. Lin, H. Zhang, M. Asadi, K. Zhao, L. Yang, Y. Fan, J. Zhu, Q. Liu, L. Sun, W. J. Xie, C. Duan, F. Mo and J. H. Dou, Machine learning-driven discovery and structure–activity relationship analysis of conductive metal–organic frameworks, *Chem. Mater.*, 2024, **36**(11), 5436–5445.

43 A. S. Rosen, S. M. Iyer, D. Ray, Z. Yao, A. Aspuru-Guzik, L. Gagliardi, J. M. Notestein and R. Q. Snurr, Machine learning the quantum-chemical properties of metal–organic frameworks for accelerated materials discovery, *Matter*, 2021, **4**(5), 1578–1597.

44 G. Borboudakis, T. Stergiannakos, M. Frysali, E. Klontzas, I. Tsamardinos and G. E. Froudakis, Chemically intuited, large-scale screening of MOFs by machine learning techniques, *npj Comput. Mater.*, 2017, **3**(1), 40.

45 P. Yang, H. Zhang, X. Lai, K. Wang, Q. Yang and D. Yu, Accelerating the selection of covalent organic frameworks with automated machine learning, *ACS Omega*, 2021, **6**(27), 17149–17161.

46 Z. Shi, W. Yang, X. Deng, C. Cai, Y. Yan, H. Liang, Z. Liu and Z. Qiao, Machine-learning-assisted high-throughput

computational screening of high performance metal–organic frameworks, *Mol. Syst. Des. Eng.*, 2020, **5**(4), 725–742.

47 Y. Sun, R. F. DeJaco, Z. Li, D. Tang, S. Glante, D. S. Sholl, C. M. Colina, R. Q. Snurr, M. Thommes, M. Hartmann and J. I. Siepmann, Fingerprinting diverse nanoporous materials for optimal hydrogen storage conditions using meta-learning, *Sci. Adv.*, 2021, **7**(30), eabg3983.

48 D. M. Anstine and O. Isayev, Generative models as an emerging paradigm in the chemical sciences, *J. Am. Chem. Soc.*, 2023, **145**(16), 8736–8750.

49 B. Zhang, S. Gu, B. Zhang, J. Bao, D. Chen, F. Wen, Y. Wang, and B. Guo: StyleSwin: Transformer-based GAN for High-resolution Image Generation, *arXiv*, 2021, preprint, arXiv:211210762, DOI: **10.48550/arXiv.211210762**.

50 D. Menon and R. Ranganathan, A Generative Approach to Materials Discovery, Design, and Optimization, *ACS Omega*, 2022, **7**(30), 25958–25973.

51 T. Long, N. M. Fortunato, I. Opahle, Y. Zhang, I. Samathrakis, C. Shen, O. Gutfleisch and H. Zhang, Constrained crystals deep convolutional generative adversarial network for the inverse design of crystal structures, *npj Comput. Mater.*, 2021, **7**(1), 66.

52 M. Alverson, S. G. Baird, R. Murdock, H. E. Sin-Hang, J. Johnson and T. D. Sparks, Generative adversarial networks and diffusion models in material discovery, *Digital Discovery*, 2024, **3**(1), 62–80.

53 Y. Mao, Q. He and X. Zhao, Designing complex architectured materials with generative adversarial networks, *Sci. Adv.*, 2020, **6**(17), eaaz4169.

54 I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville and Y. Bengio, Generative Adversarial Networks, *arXiv*, 2014, preprint, arXiv:1406.2661, DOI: **10.48550/arXiv.1406.2661**.

55 B. Kim, S. Lee and J. Kim, Inverse design of porous materials using artificial neural networks, *Sci. Adv.*, 2020, **6**(1), eaax9324.

56 A. K. Shargh and N. Abdolrahim, An interpretable deep learning approach for designing nanoporous silicon nitride membranes with tunable mechanical properties, *npj Comput. Mater.*, 2023, **9**(1), 82.

57 A. Gayon-Lombardo, L. Mosser, N. P. Brandon and S. J. Cooper, Pores for thought: generative adversarial networks for stochastic reconstruction of 3D multi-phase electrode microstructures with periodic boundaries, *npj Comput. Mater.*, 2020, **6**(1), 82.

58 Y. Rubner, C. Tomasi and L. J. Guibas, The earth mover's distance as a metric for image retrieval, *Int. J. Comput. Vis.*, 2000, **40**(2), 99–121.

59 S. Kim, J. Noh, G. H. Gu, A. Aspuru-Guzik and Y. Jung, Generative Adversarial Networks for Crystal Structure Prediction, *ACS Cent. Sci.*, 2020, **6**(8), 1412–1420.

60 P. C. H. Nguyen, N. N. Vlassis, B. Bahmani, W. Sun, H. S. Udaykumar and S. S. Baek, Synthesizing controlled microstructures of porous media using generative adversarial networks and reinforcement learning, *Sci. Rep.*, 2022, **12**(1), 9034.

61 J. Park, H. Kim, Y. Kang, Y. Lim and J. Kim, From data to discovery: Recent trends of machine learning in metal–organic frameworks, *JACS Au*, 2024, **4**(10), 3727–3743.

62 A. Radford, L. Metz, and S. Chintala, Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks, *arXiv*, 2015, preprint, arXiv:1511.06434, DOI: **10.48550/arXiv.1511.06434**.

63 M. Arjovsky, and L. Bottou, Towards Principled Methods for Training Generative Adversarial Networks, *arXiv*, 2017, preprint, arXiv:1701.04862, DOI: **10.48550/arXiv.1701.04862**.

64 Y. Kossale, M. Airaj, and A. Darouichi, Mode Collapse in Generative Adversarial Networks: An Overview, in *2022 8th International Conference on Optimization and Applications (ICOA)*, IEEE, 2022, pp. 1–6.

65 Y. Dan, Y. Zhao, X. Li, S. Li, M. Hu and J. Hu, Generative adversarial networks (GAN) based efficient sampling of chemical composition space for inverse design of inorganic materials, *npj Comput. Mater.*, 2020, **6**(1), 84.

66 M. Arjovsky, S. Chintala, and L. Bottou, Wasserstein GAN, *arXiv*, 2017, preprint, arXiv:1701.07875, DOI: **10.48550/arXiv.1701.07875**.

67 D. P. Kingma, and M. Welling, Auto-Encoding Variational Bayes, *arXiv*, 2013, preprint, arXiv:1312.6114, DOI: **10.48550/arXiv.1312.6114**.

68 N. Kim, D. Lee and Y. Hong, Data-efficient deep generative model with discrete latent representation for high-fidelity digital materials, *ACS Mater. Lett.*, 2023, **5**(3), 730–737.

69 J. Zhou, A. Mroz and K. E. Jelfs, Deep generative design of porous organic cages *via* a variational autoencoder, *Digital Discovery*, 2023, **2**(6), 1925–1936.

70 D. P. Kingma, and M. Welling, An Introduction to Variational Autoencoders, *arXiv*, 2019, preprint, arXiv:1906.02691, DOI: **10.48550/arXiv.1906.02691**.

71 R. Pollice, G. dos Passos Gomes, M. Aldeghi, R. J. Hickman, M. Krenn, C. Lavigne, M. Lindner-D'Addario, A. Nigam, C. T. Ser, Z. Yao and A. Aspuru-Guzik, Data-Driven Strategies for Accelerated Materials Design, *Acc. Chem. Res.*, 2021, **54**(4), 849–860.

72 Z. Yao, B. Sánchez-Lengeling, N. S. Bobbitt, B. J. Bucior, S. G. H. Kumar, S. P. Collins, T. Burns, T. K. Woo, O. K. Farha, R. Q. Snurr and A. Aspuru-Guzik, Inverse design of nanoporous crystalline reticular materials with deep generative models, *Nat. Mach. Intell.*, 2021, **3**(1), 76–86.

73 Y. G. Chung, J. Camp, M. Haranczyk, B. J. Sikora, W. Bury, V. Krungleviciute, T. Yildirim, O. K. Farha, D. S. Sholl and R. Q. Snurr, Computation-ready, experimental metal–organic Frameworks: A tool to enable high-throughput screening of nanoporous crystals, *Chem. Mater.*, 2014, **26**(21), 6185–6192.

74 B. J. Bucior, A. S. Rosen, M. Haranczyk, Z. Yao, M. E. Ziebel, O. K. Farha, J. T. Hupp, J. I. Siepmann, A. Aspuru-Guzik and R. Q. Snurr, Identification schemes for metal–organic frameworks to enable rapid search and cheminformatics analysis, *Cryst. Growth Des.*, 2019, **19**(11), 6682–6697.

75 Y. Liu, Z. Yang, Z. Yu, Z. Liu, D. Liu, H. Lin, M. Li, S. Ma, M. Avdeev and S. Shi, Generative artificial intelligence and its applications in materials science: Current situation and future perspectives, *J. Materiomics*, 2023, **9**(4), 798–816.
76 Y. Wang, Z. Fan, P. Qian, T. Ala-Nissila and M. A. Caro, Structure and Pore Size Distribution in Nanoporous Carbon, *Chem. Mater.*, 2022, **34**(2), 617–628.
77 Y. Zhao, E. M. D. Siriwardane, Z. Wu, N. Fu, M. Al-Fahdi, M. Hu and J. Hu, Physics guided deep learning for generative design of crystal materials with symmetry constraints, *npj Comput. Mater.*, 2023, **9**(1), 38.
78 R. Gómez-Bombarelli, J. N. Wei, D. Duvenaud, J. M. Hernández-Lobato, B. Sánchez-Lengeling, D. Sheberla, J. Aguilera-Iparraguirre, T. D. Hirzel, R. P. Adams and A. Aspuru-Guzik, Automatic Chemical Design Using a Data-Driven Continuous Representation of Molecules, *ACS Cent. Sci.*, 2018, **4**(2), 268–276.
79 L. Yang, Z. Zhang, Y. Song, S. Hong, R. Xu, Y. Zhao, W. Zhang, B. Cui and M. H. Yang, Diffusion Models: A Comprehensive Survey of Methods and Applications, *ACM Comput. Surv.*, 2023, **56**(4), 1–39.
80 J. Park, A. P. S. Gill, S. M. Moosavi and J. Kim, Inverse design of porous materials: a diffusion model approach, *J. Mater. Chem. A*, 2024, **12**(11), 6507–6514.
81 J. Park, Y. Lee and J. Kim, Multi-modal conditional diffusion model using signed distance functions for metal-organic frameworks generation, *Nat. Commun.*, 2025, **16**(1), 34.
82 H. Park, X. Yan, R. Zhu, E. A. Huerta, S. Chaudhuri, D. Cooper, I. Foster and E. Tajkhorshid, A generative artificial intelligence framework based on a molecular diffusion model for the design of metal-organic frameworks for carbon capture, *Commun. Chem.*, 2024, **7**(1), 21.
83 N. S. Bobbitt, K. Shi, B. J. Bucior, H. Chen, N. Tracy-Amoroso, Z. Li, Y. Sun, J. H. Merlin, J. I. Siepmann, D. W. Siderius and R. Q. Snurr, MOFX-DB: An online database of computational adsorption data for nanoporous materials, *J. Chem. Eng. Data*, 2023, **68**(2), 483–498.
84 T. Xie, X. Fu, O. E. Ganea, R. Barzilay, and T. Jaakkola, Crystal Diffusion Variational Autoencoder for Periodic Material Generation, *arXiv*, 2021, preprint, arXiv:211006197, DOI: **10.48550/arXiv.211006197**.
85 P. Lyngby and K. S. Thygesen, Data-driven discovery of 2D materials by deep generative models, *npj Comput. Mater.*, 2022, **8**(1), 232.
86 T. Pakornchote, N. Choomphon-anomakhun, S. Arrerut, C. Atthapak, S. Khamkaeo, T. Chotibut and T. Bovornratanaraks, Diffusion probabilistic models enhance variational autoencoder for crystal structure generative modeling, *Sci. Rep.*, 2024, **14**(1), 1275.
87 M. Seyfarth, S. U. H. Dar, and S. Engelhardt, Latent pollution model: The hidden carbon footprint in 3D image synthesis, *arXiv*, 2024, preprint, arXiv:240714892, DOI: **10.48550/arXiv.240714892**.
88 T. Salimans, and J. Ho, Progressive Distillation for Fast Sampling of Diffusion Models, *arXiv*, 2022, preprint, arXiv:220200512, DOI: **10.48550/arXiv.220200512**.
89 D. Watson, J. Ho, M. Norouzi, and W. Chan, Learning to Efficiently Sample from Diffusion Probabilistic Models, *arXiv*, 2021, preprint, arXiv:210603802, DOI: **10.48550/arXiv.210603802**.
90 J. Song, C. Meng, and S. Ermon, Denoising Diffusion Implicit Models, *arXiv*, 2020, preprint, arXiv:201002502, DOI: **10.48550/arXiv.201002502**.
91 W. Peebles, and S. Xie, Scalable diffusion models with transformers, *Proceedings of the IEEE/CVF international conference on computer vision*, 2023, pp. 4195–4205.
92 A. Sauer, K. Schwarz, and A. Geiger, StyleGAN-XL: Scaling StyleGAN to large diverse datasets, *ACM SIGGRAPH 2022 Conference Proceedings*, 2022, pp. 1–10.
93 A. Ulhaq, and N. Akhtar, Efficient diffusion models for vision: A survey, *arXiv*, 2022, preprint, arXiv:221009292, DOI: **10.48550/arXiv.221009292**.
94 Z. Dai, and D. K. Gifford, Training Data Attribution for Diffusion Models, *arXiv*, 2023, preprint, arXiv:230602174, DOI: **10.48550/arXiv.230602174**.
95 C. J. Pickard, *AIRSS data for carbon at 10GPa and the C+N+H+O system at 1GPa. Materials Cloud Archive 2020.0026/v1*, 2020.
96 A. Jain, S. P. Ong, G. Hautier, W. Chen, W. D. Richards, S. Dacek, S. Cholia, D. Gunter, D. Skinner, G. Ceder and K. A. Persson, Commentary: The Materials Project: A materials genome approach to accelerating materials innovation, *APL Mater.*, 2013, **1**(1), 011002.
97 K. M. Jablonka, A. S. Rosen, A. S. Krishnapriyan and B. Smit, An Ecosystem for Digital Reticular Chemistry, *ACS Cent. Sci.*, 2023, **9**(4), 563–581.
98 Y. Kim, Y. Kim, C. Yang, K. Park, G. X. Gu and S. Ryu, Deep learning framework for material design space exploration using active transfer learning and data augmentation, *npj Comput. Mater.*, 2021, **7**(1), 140.
99 H. Yamada, C. Liu, S. Wu, Y. Koyama, S. Ju, J. Shiomi, J. Morikawa and R. Yoshida, Predicting Materials Properties with Little Data Using Shotgun Transfer Learning, *ACS Cent. Sci.*, 2019, **5**(10), 1717–1730.
100 G. M. Cooper and Y. J. Colón, Metal–organic framework clustering through the lens of transfer learning, *Mol. Syst. Des. Eng.*, 2023, **8**(8), 1049–1059.
101 Y. Lim and J. Kim, Application of transfer learning to predict diffusion properties in metal–organic frameworks, *Mol. Syst. Des. Eng.*, 2022, **7**(9), 1056–1064.
102 S. Y. Chen, F. Zheng, S. Q. Wu and Z. Z. Zhu, An improved genetic algorithm for crystal structure prediction, *Curr. Appl. Phys.*, 2017, **17**(4), 454–460.
103 P. C. Jennings, S. Lysgaard, J. S. Hummelshøj, T. Vegge and T. Bligaard, Genetic algorithms for computational materials discovery accelerated by machine learning, *npj Comput. Mater.*, 2019, **5**(1), 46.
104 A. Nigam, R. Pollice and A. Aspuru-Guzik, Parallel tempered genetic algorithm guided by deep neural networks for

inverse molecular design, *Digital Discovery*, 2022, **1**(4), 390–404.
105 M. J. Vainio and M. S. Johnson, Generating Conformer Ensembles Using a Multiobjective Genetic Algorithm, *J. Chem. Inf. Model.*, 2007, **47**(6), 2462–2474.
106 S. Q. Wu, M. Ji, C. Z. Wang, M. C. Nguyen, X. Zhao, K. Umemoto, R. M. Wentzcovitch and K. M. Ho, An adaptive genetic algorithm for crystal structure prediction, *J. Phys.: Condens. Matter*, 2014, **26**(3), 035402.
107 Y. G. Chung, D. A. Gómez-Gualdrón, P. Li, K. T. Leperi, P. Deria, H. Zhang, N. A. Vermeulen, J. F. Stoddart, F. You, J. T. Hupp, O. K. Farha and R. Q. Snurr, In silico discovery of metal-organic frameworks for precombustion $CO_2$ capture using a genetic algorithm, *Sci. Adv.*, 2016, **2**(10), e1600909.
108 C. E. Wilmer, M. Leaf, C. Y. Lee, O. K. Farha, B. G. Hauser, J. T. Hupp and R. Q. Snurr, Large-scale screening of hypothetical metal–organic frameworks, *Nat. Chem.*, 2012, **4**(2), 83–89.
109 S. Lee, B. Kim, H. Cho, H. Lee, S. Y. Lee, E. S. Cho and J. Kim, Computational Screening of Trillions of Metal–Organic Frameworks for High-Performance Methane Storage, *ACS Appl. Mater. Interfaces*, 2021, **13**(20), 23647–23654.
110 Y. Lim, J. Park, S. Lee and J. Kim, Finely tuned inverse design of metal–organic frameworks with user-desired Xe/Kr selectivity, *J. Mater. Chem. A*, 2021, **9**(37), 21175–21183.
111 S. P. Collins, T. D. Daff, S. S. Piotrkowski and T. K. Woo, Materials design by evolutionary optimization of functional groups in metal-organic frameworks, *Sci. Adv.*, 2016, **2**(11), e1600954.
112 G. Takasao, T. Wada, H. Chikuma, P. Chammingkwan, M. Terano and T. Taniike, Preventing Premature Convergence in Evolutionary Structure Determination of Complex Molecular Systems: Demonstration in Few-Nanometer-Sized $TiCl_4$ -Capped $MgCl_2$ Nanoplates, *J. Phys. Chem. A*, 2022, **126**(31), 5215–5221.
113 H. M. Pandey, A. Chaudhary and D. Mehrotra, A comparative review of approaches to prevent premature convergence in GA, *Appl. Soft Comput.*, 2014, **24**, 1047–1077.
114 M. A. Albadr, S. Tiun, M. Ayob and F. AL-Dhief, Genetic Algorithm Based on Natural Selection Theory for Optimization Problems, *Symmetry*, 2020, **12**(11), 1758.
115 S. Katoch, S. S. Chauhan and V. Kumar, A review on genetic algorithm: past, present, and future, *Multimed. Tool. Appl.*, 2021, **80**(5), 8091–8126.
116 T. D. Pham and R. Q. Snurr, Implementation of genetic algorithms to optimize metal–organic frameworks for $CO_2$ capture, *Langmuir*, 2025, **41**(7), 4585–4593.
117 Y. Kwon, S. Kang, Y. S. Choi and I. Kim, Evolutionary design of molecules based on deep learning and a genetic algorithm, *Sci. Rep.*, 2021, **11**(1), 17304.
118 C. Qian, R. K. Tan and W. Ye, Design of architectured composite materials with an efficient, adaptive artificial neural network-based generative design method, *Acta Mater.*, 2022, **225**, 117548.
119 K. J. DeMille, R. Hall, J. R. Leigh, I. Guven and A. D. Spear, Materials design using genetic algorithms informed by convolutional neural networks: Application to carbon nanotube bundles, *Composites, Part B*, 2024, **286**, 111751.
120 T. K. Patra, V. Meenakshisundaram, J. H. Hung and D. S. Simmons, Neural-Network-Biased Genetic Algorithms for Materials Design: Evolutionary Algorithms That Learn, *ACS Comb. Sci.*, 2017, **19**(2), 96–107.
121 A. K. Shakya, G. Pillai and S. Chakrabarty, Reinforcement learning algorithms: A brief survey, *Expert Syst. Appl.*, 2023, **231**, 120495.
122 H. Park, S. Majumdar, X. Zhang, J. Kim and B. Smit, Inverse design of metal–organic frameworks for direct air capture of $CO_2$ *via* deep reinforcement learning, *Digital Discovery*, 2024, **3**(4), 728–741.
123 F. Sui, R. Guo, Z. Zhang, G. X. Gu and L. Lin, Deep Reinforcement Learning for Digital Materials Design, *ACS Mater. Lett.*, 2021, **3**(10), 1433–1439.
124 B. Zheng, Z. Zheng and G. X. Gu, Designing mechanically tough graphene oxide materials using deep reinforcement learning, *npj Comput. Mater.*, 2022, **8**(1), 225.
125 A. M. Gopakumar, P. V. Balachandran, D. Xue, J. E. Gubernatis and T. Lookman, Multi-objective Optimization for Materials Discovery *via* Adaptive Design, *Sci. Rep.*, 2018, **8**(1), 3738.
126 N. K. Brown, A. P. Garland, G. M. Fadel and G. Li, Deep reinforcement learning for the rapid on-demand design of mechanical metamaterials with targeted nonlinear deformation responses, *Eng. Appl. Artif. Intell.*, 2023, **126**, 106998.
127 S. Li, C. Hu, S. Ke, C. Yang, J. Chen, Y. Xiong, H. Liu and L. Hong, LS-MolGen: Ligand-and-Structure Dual-Driven Deep Reinforcement Learning for Target-Specific Molecular Generation Improves Binding Affinity and Novelty, *J. Chem. Inf. Model.*, 2023, **63**(13), 4207–4215.
128 N. K. Brown, A. Deshpande, A. Garland, S. A. Pradeep, G. Fadel, S. Pilla and G. Li, Deep reinforcement learning for the design of mechanical metamaterials with tunable deformation and hysteretic characteristics, *Mater. Des.*, 2023, **235**, 112428.
129 P. Rajak, A. Krishnamoorthy, A. Mishra, R. Kalia, A. Nakano and P. Vashishta, Autonomous reinforcement learning agent for chemical vapor deposition synthesis of quantum materials, *npj Comput. Mater.*, 2021, **7**(1), 108.
130 P. Rajak, B. Wang, K. Nomura, Y. Luo, A. Nakano, R. Kalia and P. Vashishta, Autonomous reinforcement learning agent for stretchable kirigami design of 2D materials, *npj Comput. Mater.*, 2021, **7**(1), 102.
131 S. Levine, Z. P. Popovíc, and V. Koltun, Nonlinear Inverse Reinforcement Learning with Gaussian Processes, in *NIPS'11: Proceedings of the 25th International Conference on Neural Information Processing Systems*, 2011, pp. 19–27.
132 Z. Li, K. Shi, D. Dubbeldam, M. Dewing, C. Knight, Á. Vázquez-Mayagoitia and R. Q. Snurr, Efficient implementation of Monte Carlo algorithms on graphical processing units for simulation of adsorption in porous

materials, *J. Chem. Theory Comput.*, 2024, **20**(23), 10649–10666.

133 T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever and D. Amodei, Language Models are Few-Shot Learners, *Adv. Neural Inf. Process. Syst.*, 2020, **33**, 1877–1901.

134 J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, in *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, 2019, pp. 4171–4186.

135 S. Miret, and N. M. A. Krishnan, Are LLMs Ready for Real-World Materials Discovery?, *arXiv*, 2024, preprint, arXiv:2402.05200, DOI: **10.48550/arXiv.2402.05200**.

136 Q. Ai, F. Meng, J. Shi, B. Pelkie and C. W. Coley, Extracting structured data from organic synthesis procedures using a fine-tuned large language model, *Digital Discovery*, 2024, **3**(9), 1822–1831.

137 G. Lei, R. Docherty and S. J. Cooper, Materials science in the era of large language models: a perspective, *Digital Discovery*, 2024, **3**(7), 1257–1272.

138 J. Choi and B. Lee, Accelerating materials language processing with large language models, *Commun. Mater.*, 2024, **5**(1), 13.

139 Z. Zheng, Z. Rong, N. Rampal, C. Borgs, J. T. Chayes and O. M. Yaghi, A GPT-4 Reticular Chemist for Guiding MOF Discovery, *Angew. Chem., Int. Ed.*, 2023, **62**(46), e202311983.

140 Y. Kang and J. Kim, ChatMOF: an artificial intelligence system for predicting and generating metal-organic frameworks using large language models, *Nat. Commun.*, 2024, **15**(1), 4705.

141 M. Bran A, S. Cox, O. Schilter, C. Baldassari, A. D. White and P. Schwaller, Augmenting large language models with chemistry tools, *Nat. Mach. Intell.*, 2024, **6**(5), 525–535.

142 Z. Zheng, O. Zhang, C. Borgs, J. T. Chayes and O. M. Yaghi, ChatGPT Chemistry Assistant for Text Mining and the Prediction of MOF Synthesis, *J. Am. Chem. Soc.*, 2023, **145**(32), 18048–18062.

143 Z. Zheng, A. H. Alawadhi, S. Chheda, S. E. Neumann, N. Rampal, S. Liu, H. L. Nguyen, Y. h. Lin, Z. Rong, J. I. Siepmann, L. Gagliardi, A. Anandkumar, C. Borgs, J. T. Chayes and O. M. Yaghi, Shaping the water-harvesting behavior of metal–organic frameworks aided by fine-tuned GPT models, *J. Am. Chem. Soc.*, 2023, **145**(51), 28284–28295.

144 K. M. Jablonka, P. Schwaller, A. Ortega-Guerrero and B. Smit, Leveraging large language models for predictive chemistry, *Nat. Mach. Intell.*, 2024, **6**(2), 161–169.

145 Z. Zheng, Z. He, O. Khattab, N. Rampal, M. A. Zaharia, C. Borgs, J. T. Chayes and O. M. Yaghi, Image and data mining in reticular chemistry powered by GPT-4V, *Digital Discovery*, 2024, **3**(3), 491–501.

146 Y. Kang, H. Park, B. Smit and J. Kim, A multi-modal pre-training transformer for universal transfer learning in metal–organic frameworks, *Nat. Mach. Intell.*, 2023, **5**(3), 309–318.

147 T. Taniike and K. Takahashi, The value of negative results in data-driven catalysis research, *Nat. Catal.*, 2023, **6**(2), 108–111.

148 Y. Liu, J. Cao, C. Liu, K. Ding, and L. Jin, Datasets for large language models: A comprehensive survey, *arXiv*, 2024, preprint, arXiv:2402.18041, DOI: **10.48550/arXiv.2402.18041**.

149 F. L. Oliveira, C. Cleeton, B. F. R. Neumann, B. Luan, A. H. Farmahini, L. Sarkisov and M. Steiner, CRAFTED: An exploratory database of simulated adsorption isotherms of metal-organic frameworks, *Sci. Data*, 2023, **10**(1), 230.

150 H. Daglar, H. C. Gulbalkan, G. O. Aksu and S. Keskin, Computational simulations of metal–organic frameworks to enhance adsorption applications, *Adv. Mater.*, 2024, 2405532.

151 T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. L. Scao, S. Gugger, M. Drame, Q. Lhoest, and A. M. Rush, HuggingFace's transformers: State-of-the-art natural language processing, *arXiv*, 2019, preprint, arXiv:1910.03771, DOI: **10.48550/arXiv.1910.03771**.

152 D. Xu, Q. Zhang, X. Huo, Y. Wang and M. Yang, Advances in data-assisted high-throughput computations for material design, *Mater. Genome Eng. Adv.*, 2023, **1**(1), e11.

153 M. Schilling-Wilhelmi, M. Ríos-García, S. Shabih, M. V. Gil, S. Miret, C. T. Koch, J. A. Márquez, and K. M. Jablonka, From text to insight: Large language models for materials science data extraction, *arXiv*, 2024, preprint, arXiv:2407.16867, DOI: **10.48550/arXiv.2407.16867**.

154 R. Okabe, Z. West, A. Chotrattanapituk, M. Cheng, D. C. Carrizales, W. Xie, R. J. Cava, and M. Li, Large language model-guided prediction toward quantum materials synthesis, *arXiv*, 2024, preprint, arXiv:2410.20976, DOI: **10.48550/arXiv.2410.20976**.

155 K. M. Jablonka, Q. Ai, A. Al-Feghali, S. Badhwar, J. D. Bocarsly, A. M. Bran, S. Bringuier, L. C. Brinson, K. Choudhary, D. Circi, S. Cox, W. A. de Jong, M. L. Evans, N. Gastellu, J. Genzling, M. V. Gil, A. K. Gupta, Z. Hong, A. Imran, S. Kruschwitz, A. Labarre, J. Lála, T. Liu, S. Ma, S. Majumdar, G. W. Merz, N. Moitessier, E. Moubarak, B. Mouriño, B. Pelkie, M. Pieler, M. C. Ramos, B. Ranković, S. G. Rodriques, J. N. Sanders, P. Schwaller, M. Schwarting, J. Shi, B. Smit, B. E. Smith, J. Van Herck, C. Völker, L. Ward, S. Warren, B. Weiser, S. Zhang, X. Zhang, G. A. Zia, A. Scourtas, K. J. Schmidt, I. Foster, A. D. White and B. Blaiszik, 14 examples of how LLMs can transform materials science and chemistry: a reflection on a large language model hackathon, *Digital Discovery*, 2023, **2**(5), 1233–1250.