Materials Horizons

ROYAL SOCIETY OF CHEMISTRY

REVIEW

View Article Online
View Journal | View Issue

Cite this: *Mater. Horiz.*, 2026, **13**, 1694

Machine learning to design metal–organic frameworks: progress and challenges from a data efficiency perspective

Diego A. Gómez-Gualdrón, *[ab] Tatiane Gercina de Vilas, [a] Katherine Ardila, [a] Fernando Fajardo-Rojas [ab] and Alexander J. Pak *[abc]

This review critically examines work at the intersection of machine learning (ML) and metal–organic frameworks (MOFs). The modular nature of MOFs enables immense design flexibility and applicability to a wide range of applications. However, the combinatorially large design space also stresses the resource-intensive nature of traditional high-throughput screening approaches. Due to the increasing availability of data in the form of experimental and hypothetical MOF structures and their properties, ML methods have emerged as a promising solution to accelerate MOF discovery, yet successful application of these methods will require strategies that maximize data and resource efficiency. This work surveys approaches to reduce data and resource burdens for MOF property prediction and design through feature engineering, model architecture choices, transfer learning, active learning, and generative models. We also discuss challenges related to data quality and scalability, as well as future opportunities for ML-empowered methods that, up to this point, have primarily focused on MOF adsorption properties. By focusing on efficiency at every stage (from data generation to model inference), we identify future pathways for making ML-aided MOF design more robust and accessible to both theorists and experimentalists alike.

Received 1st August 2025,
Accepted 4th December 2025

DOI: 10.1039/d5mh01467k

rsc.li/materials-horizons

Wider impact

Metal–organic frameworks (MOFs) are materials with the potential to revolutionize numerous areas of research and technology. MOFs are modular materials combining inorganic and organic building blocks. The premise in MOF research is that there are specific building block combinations that can yield breakthrough-enabling properties. The challenge is thus to identify these combinations out of a vast "design space" spanning trillions of possibilities. Since the early days of high throughput computational screening, artificial intelligence and machine learning (AI/ML) have helped explore this vast MOF design space. However, with the recent explosion of all things AI/ML, there is a lot of excitement about the prospect of AI/ML touching nearly all aspects of MOF design and development, but there are also important questions about where or how AI/ML can make the biggest impact. Aiming to help provide such perspective, this review discusses how AI/ML involvement in MOF research has evolved, but with data efficiency as the guiding underlying theme. Data efficiency is an aspect of ML research in MOFs that has not received much attention and only been implicitly discussed in the past, but that now is coming to the forefront due to the increasingly complex AI/ML models/methods at one's disposal, more ambitious tasks for AI/ML, and the desire to explore new aspects/properties of MOFs.

1. Introduction

Metal–organic frameworks (MOFs) are some of the most fascinating materials under development in the 21st century.[1] MOFs were originally studied for their potential use in hydrogen[2] and methane storage[3] but are now broadly studied for their potential in diverse applications, such as in electronics,[4] catalysis,[5] sensors,[6] medicine,[7] and molecular separations,[8] among others. Conceptually, MOFs can be thought of as porous, ordered, modular materials where each MOF arises from a specific combination of organic linkers and metal-based nodes interconnected into a network that follows a specific pattern or "topology[9]". With hundreds of nodes, thousands of topologies, and billions of linkers to choose from, MOFs have almost unlimited tunability potential. However, the latter comes at the expense of an overwhelmingly large "design space" that comprises (at least) trillions of MOFs.[10]

[a] *Department of Chemical and Biological Engineering, Colorado School of Mines, 1601 Illinois St, Golden, CO 80401, USA. E-mail: dgomezgualdron@mines.edu, apak@mines.edu*
[b] *Materials Science Program, Colorado School of Mines, 1601 Illinois St, Golden, CO 80401, USA*
[c] *Quantitative Biosciences and Engineering Program, Colorado School of Mines, 1601 Illinois Street, Golden, Colorado 80401, USA*

For more than a decade, computation has sought to help experimentalists navigate the design space of MOFs by predicting relevant properties for as many "prototypes" as possible, so that lab efforts and resources are only directed towards the most promising ones.[11] This paradigm is now pervasive in materials science and is known as high throughput computational screening (HTCS). The early vision for HTCS was to exploit the "boom" in computational power to simply automate the prediction of MOF properties using "standard" prediction methods underpinned by classical, quantum, and statistical mechanics (*e.g.*, molecular simulation). However, with so many prototypes, candidate applications, and operating conditions for each application to consider, it became clear that inherently faster prediction methods were needed. Not surprisingly, efforts to predict MOF properties *via* machine learning (ML) started to emerge soon after the first prominent efforts in MOF HTCS came to light.[12–14]

Along with other developments in artificial intelligence (AI), the success of AI/ML tools such as ChatGPT is arguably reshaping society, increasing awareness about AI/ML among the broader public, and creating a sense that maybe "anything" is possible with AI/ML. This "hope" surrounding AI/ML has also extended to the field of computational development of materials in general, and MOFs in particular. However, it is important to recognize the "special" circumstances around the development of ChatGPT. For instance, GPT-3 and GPT-4, *i.e.*, the large language models (LLMs) under ChatGPT's "hood," are believed to have been trained on (at least) 300B tokens (*i.e.*, text-based "data points") using a large cluster of GPUs and costing over millions of US dollars. This is a scale of data and resources that academic research labs do not routinely have access to. For instance, the most ambitious property prediction efforts in MOFs have usually hit a "wall" at around one million structures, even for relatively inexpensive properties to predict, such as methane adsorption or void fraction.[15] In other words, while the development of AI/ML is "hungry" for data and resources, academic research labs in the MOF field (and across materials science in general) must adapt to circumstances of data and resource "scarcity."

With the above in mind, let us note that this review does not aim for an exhaustive listing of the numerous ML efforts that have been reported to date in the MOF field. Rather, this critical review aims to highlight ML efforts in a way that showcases the lead up to current strategies to maximize data and resource utilization efficiency for MOF development. Broadly speaking, these strategies tend to impact one or more of three phases of the ML-based discovery pipeline: (i) the data processing phase, which pertains to the acquisition and preparation of data to be fed to the ML model, (ii) the model training phase, which pertains to the selection of model architecture and the training approach, and (iii) the materials discovery phase, which pertains to the utilization of the ML model to explore the MOF design space. Accordingly, Fig. 1 provides an overview of how the topics discussed in different sections in this review relate to these phases.

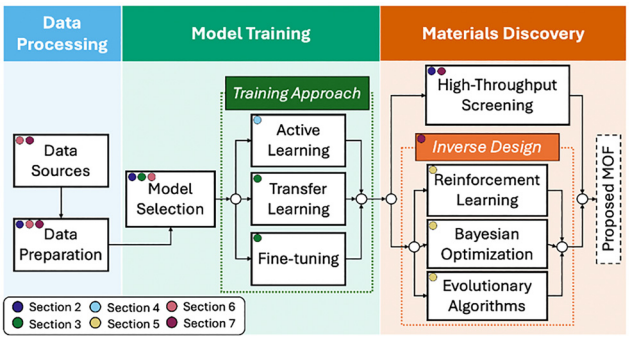

Fig. 1 Overview of some possible paths in a ML-based pipeline seeking the discovery of a promising MOF. The colored dots indicate which sections in this review (see color code bottom-left) discuss different aspects of the pipeline.

2. Data efficiency through feature or model architecture engineering

Two of the most important decisions affecting the learning efficiency of an ML model that predicts MOF properties are choosing features (or model inputs) and the model architecture (or model functional form). To rationalize this fact, imagine a MOF property given by a scalar y (the intended model output) that happens to depend on MOF feature x_1 in the form $f(x)$:

$$y = a_1 {x_1}^2 = f(x) \quad (1)$$

where a_1 is the linear combination of feature x_1. In the usual scenario where one does not actually know $f(x)$, one would aspire to approximate $f(x)$ using a vector of n proposed input features $x \in \mathcal{R}^n$ transformed by a set of basis functions, *e.g.*:

$$f(x) \approx \sum_{i=1}^{n} \sum_{m=0}^{\infty} b_{i,m} x_i^m \quad (2)$$

representing an infinite polynomial basis. If one had infinite basis functions and infinite training data, this approximation of $f(x)$ would be exact. But outside of this unrealistic scenario, one must leverage known aspects of $f(x)$ to limit the space of solutions that need to be investigated, hoping to reduce the amount of training data required to sufficiently approximate $f(x)$.

In the case where $f(x)$ is given by eqn (1), knowledge that $f(x)$ only depends on x_1 – and what x_1 looks like – is an example of feature engineering. On the other hand, knowledge that the functional form of $f(x)$ depends on x^2 is analogous to model architecture engineering. Either case is an example of inductive bias that narrows the range of possible ML model solutions, requiring domain knowledge to impose useful assumptions.

Over the last two decades, researchers have investigated how to design ML models for efficient screening of MOFs mostly based on property predictions related to gas storage (*e.g.*, CH_4, CO_2, H_2) and separations (*e.g.*, CO_2/H_2, CO_2/N_2); interested readers can refer to in-depth reviews[16,17] on this topic. As our focus is on the data efficiency gained through feature and model architecture engineering, we will limit our discussion to the context of CO_2 adsorption predictions for MOFs, which is

one of the prediction tasks that has remained active since the early MOF days until now. Since features and model architectures are inherently coupled, we distinguish ML models that focus on global statistics from those that focus on local (usually microscopic) statistics.

2.1. Model learning efficiency using global statistics

We define global statistics as those that characterize attributes of the MOF in an inherently low-dimensional space, *i.e.*, without focusing on individual atomic or linker/node properties. Borrowing from experimental MOF studies, common sets of descriptors include those that focus on geometric aspects (Fig. 2a) of the MOF (*e.g.*, largest pore diameter, specific surface area, void fraction, topology, *etc.*) and processing conditions (*e.g.*, temperature, pressure, *etc.*). There are also physicochemical aspects of the MOF (*e.g.*, maximum/minimum charges, populations of specific elements, *etc.*) that have been introduced. Froudakis and coworkers[18,19] have shown that augmenting five to six simple geometric features with either 20 chemical descriptors describing the presence of atom types or four energetic descriptors describing the probabilities that generic probe particles would be adsorbed results in improved ML model performance with R^2 = 0.93 compared to R^2 = 0.84 (hMOFs at 298 K and 2.5 bar) and R^2 = 0.87 compared to R^2 = 0.69 (CoREMOFs at 300 K and 2 bar), respectively; in both cases, random forests (RFs) were used as the ML model and the augmented features required an order of magnitude reduction in data to achieve comparable performance to geometric-only features. Energetic descriptors,[19–22] especially those that focus on electrostatics, have been shown to be particularly effective at CO_2 adsorption and selectivity predictions, which can be explained by the polarization of CO_2 and its attractive interaction with metal sites. It is not enough to simply add more descriptors to the model – those that have a mechanistic connection to the output property of interest will impose the most useful inductive bias.

Across many studies,[20,21,23] one common theme has been that the relationship between adsorption properties and global statistics is expectedly nonlinear, as evident by improved predictions using support vector regressors, decision trees (and related methods), and artificial neural networks (ANNs) compared to linear regression. However, while the fidelity of CO_2 adsorption predictions tends to be high at the upper end of tested pressures ($0.86 < R^2 < 0.96$ at pressures greater than 2 bar),[18,19] predictions at lower pressures have had room for improvement ($0.69 < R^2 < 0.84$ at pressures below 0.1 bar).[19,22] One direction to improve ML performance is to focus on descriptors that characterize higher-resolution information about MOFs, which we describe next. Nonetheless, we emphasize that one benefit of global statistics features is for interpretability, which ultimately informs design principles (*e.g.*, defining structure–property relationships) for MOFs.[19,22]

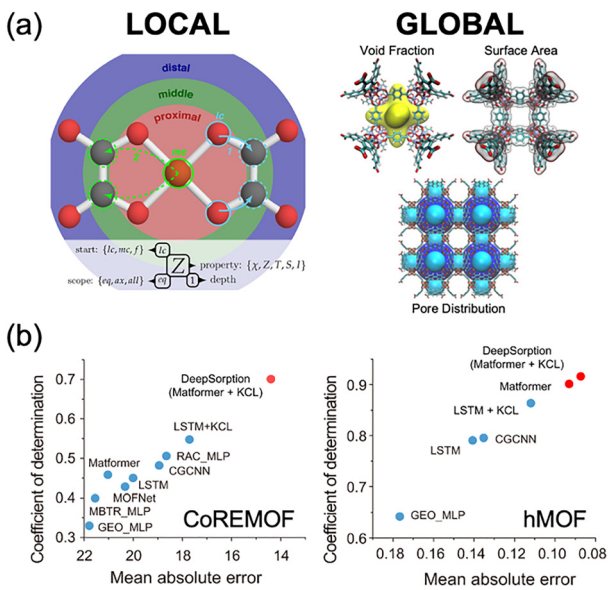

Fig. 2 (a) Examples of local and global features. The depicted local feature is the revised autocorrelation (RAC for ${}^{\text{start}}_{\text{scope}}Z_{\text{depth}}$) function that quantifies the discrete correlation of Z (electronegativity, nuclear charge, topology, covalent atomic radius, and identity for χ, Z, T, S, and I, respectively) between atoms separated up to depth I. The start index (ligand-centered, metal-centered, and full for lc, mc, and f, respectively) refers to the reference of the RAC summation and the scope index (axial, equatorial, and all for ax, eq, and all, respectively) refers to which neighboring ligand atoms are included in the summation. Reprinted with permission from ref. 24. Copyright 2017, American Chemical Society. The depicted global features show commonly used geometric descriptors. Adapted with permission from ref. 25. Copyright 2020, American Chemical Society. (b) Comparison of machine learning model performance using different architectures and features for CO_2 adsorption using the (left) CoREMOF and (right) hMOF datasets. In general, increasingly complex models (from MLPs trained on geometric features (GEO_MLP) to transformers trained on chemical and positional encodings (Matformer)) that learn "long-range" spatial correlations between local features tend to improve model performance. Adapted from ref. 26.

2.2. Model learning efficiency using local statistics

We refer to descriptors that represent higher fidelity aspects of MOFs, often at molecular or atomic scale, as local statistics. These features can also be thought of as higher-dimensional representations of global statistics. One of the earliest examples of local statistics is the atomic property weighted radial distribution function (AP-RDF) as demonstrated by Woo and coworkers,[13,27,28] which represents the pairwise correlation of atomic properties of interest (*e.g.*, electronegativity or polarizability) over a wide range of discretized distances (*e.g.*, from 0.2 to 3.0 nm). These features significantly increase the input feature space by at least an order of magnitude compared to global statistics and were shown to outperform CO_2 working capacity predictions (using multilayer perceptrons (MLPs)) compared to using global geometric features alone with R^2 = 0.94 and R^2 = 0.71, respectively.[27] The revised autocorrelation functions (RACs) introduced by Kulik and coworkers[29] can be thought of as a discretized version of AP-RDFs, instead focusing on atomic property correlations between close atoms (those within a specified depth of a connectivity graph where edges represent bonds) (Fig. 2a). These authors showed that RAC features combined with geometric features improved the

prediction of CO_2 uptake in MOFs (both under low and high pressures) compared to using geometric features alone.

Along these lines, several researchers have aimed to leverage the molecular structure of MOFs in other ways to improve model performance, leading to the adoption of higher-dimensional features and more varied model architectures. The three-dimensional structure (and atomic properties) of MOFs can be voxelized into a three-dimensional (3D) discrete space, with local nonlinear correlations learned through a 3D convolutional neural network (3D-CNN). Froudakis and coworkers[30] demonstrated this approach using a voxelized potential energy surface describing the Lennard-Jones interaction between a probe atom and the framework, also showing that the 3D-CNN required two orders of magnitude less data compared to a RF model with geometric features to achieve comparable performance. Relatedly, Lin and coworkers[31] showed that 3D-CNNs trained using voxelized features containing Lennard-Jones parameters and partial charges are useful for CO_2 adsorption screening.

Alternate model architectures have been proposed that still aim to leverage the structure of MOFs with reduced memory requirements. One approach[32] is to featurize the MOF as an unstructured point cloud described by Cartesian coordinates and any atomic properties of interest (*e.g.*, atomic number, electronegativity, van der Waals radius, *etc.*). Predictions are trained through the permutation-invariant PointNet architecture,[33] which extracts point-wise features through MLPs before applying global pooling, and this approach has been shown to improve CO_2 uptake predictions at low pressure compared to conventional geometric features. Others have opted to directly enforce local structural correlations by representing MOFs as graphs with atoms as nodes and bonds as edges. Reported graph neural networks, such as the crystal graph convolutional neural network (CGCNN)[26] and the atomistic line graph neural network (ALIGNN),[34] use atomic properties (*e.g.*, electronegativity, valence electrons, covalent radius, *etc.*) as node features and bond distances as edge features, then learn how to predict properties *via* message passing along the graph topology. However, as shown by Cui *et al.*,[26] GNN model learning is biased toward local structural characteristics and CO_2 adsorption predictions can be enhanced through learning from global structural awareness (Fig. 2b), *e.g.*, using the attention mechanism popularized by the transformer model[35] (discussed further later). In summary, while increasing the input space dimensionality through local statistics is a promising strategy, it is important to also identify the proper model architectures that bias learning towards the types of feature relationships (*e.g.*, spatial correlations) one believes is most relevant for the prediction task of interest.

2.3. Preventing overfitting during model training

In the previous two subsections, we discuss the benefit of introducing more complex features or model architectures to aid training, which is productive when the additional complexity aligns with an inductive bias related to the task of interest. However, it should also be noted that increasing the number of features or increasing the capacity of the model can increase the risk of overfitting, which is when the model learns to "memorize" patterns in your training data and fails to generalize to new "unseen" data. When using high-dimensional features, the model must learn over a large and potentially sparse input space, making it more prone to fitting irregular or coincidental patterns rather than fundamental underlaying trends. Likewise, models with large capacity can eventually become powerful enough to fit to nearly any training data point, including noise. The broader ML community has adopted several mitigation strategies, including the use of rigorously separated training, validation, and test splits (for early stopping assessment of generalizability),[36] hyperparameter tuning[37] *via* cross-validation to control model complexity, regularization,[38] and feature down-selection or dimensionality reduction[39] to reduce the noise or redundancy in the input space. Despite the routine use of these methods in ML applications, their use, to our knowledge, is rarely the focus of dedicated studies in the MOF literature. In particular, the common training/validation/test splits paradigm requires having a sufficiently large and diverse enough dataset that all three are statistically representative of the scientific task of interest. Therefore, a systematic investigation into how model capacity and feature design interact with data availability and model generalizability remains an open opportunity.

3. Data burden reduction exploiting previously trained models

An alternative to human-based engineering of MOF features is to let the computer engineer MOF features itself, which then can be used as input to train a ML model for a target prediction task. Computer-engineered features usually emerge as a byproduct of training deep learning models. As these models usually involve ANNs (also called MLPs), the desired computer-engineered features are taken to be the output from one of the (wisely chosen) model internal layers. In principle, these features encode (in the form of a vector, matrix, or tensor) the pieces of information (*e.g.*, MOF traits) that were most critical to make a prediction. Although these features are not easily interpretable by humans, their numerical form allows them to be easily reused "as is" by humans as input to other ML models.

3.1. Transfer learning from specialized models

The act of borrowing computer-engineered features emerging from the training of one model (to perform a source task) to use as input for the training of another model (to perform a target task) is the most common example of transfer learning (TL). To successfully perform this kind of transfer learning (*i.e.*, so that limited data is enough to train the model for the target task), one must first identify a source task that facilitates the emergence of computer-engineered features that are highly significant to the target task. It stands to reason that this scenario is more likely to occur when the source task

and target task share some degree of similarity, as they are more likely to depend on similar MOF traits and/or governing equations. Additionally, as computer-engineered features emerge from deep-learning, the ideal source task is one for which training data can be easily and inexpensively generated.

In 2017, the first exploration of transfer learning for MOFs was reported by Ma *et al.*[40] These authors studied to what extent transfer learning was possible with the prediction of H_2 adsorption loadings at high pressure/temperature as the source task, and prediction of H_2 adsorption loadings at high pressure/low temperature, CH_4 adsorption loadings, and Xe/Kr selectivity as the target tasks. Five simple MOF textural traits were used as model inputs, and target task datasets were about ten times smaller than the source task dataset. All models shared the same MLP architecture, which consisted of two hidden layers. Transfer learning was formally done by keeping the parameters up to the first hidden layer of the target task model the same as in the source task model and optimizing the parameters of the second hidden layer and output layer (Fig. 3a). Indicative of the importance that the source and target prediction tasks are governed by similar MOF traits, the computer-engineered features emerging from the source task proved useful for the H_2 and CH_4 adsorption prediction target tasks, which averaged R^2 values of 0.991 and 0.980, respectively, but not so for Xe/Kr selectivity prediction, for which R^2 values averaged around −0.092.

In 2023, Cooper and Colón[41] further examined the efficacy of transfer learning between H_2 and CH_4 adsorption prediction tasks from the perspective of the similarity (based on either textural properties or topologies) between the MOFs in the source and target task datasets. Not surprisingly, transfer learning worked better (*i.e.*, higher accuracy, smaller dataset size requirements) when the MOF datasets used for the source and target tasks were more similar, *e.g.*, as measured by distance in principal component space. But more interestingly, these authors found CH_4 adsorption (and some MOF datasets) to work better as the source task (and as the source MOF dataset) compared to that of H_2 adsorption. Thus, their work underlines the importance of choosing source tasks and MOF datasets that are informative for the target tasks, although guidelines to accomplish this goal are not well-established.

Although in the previous examples, the ''transfer of knowledge'' was done sequentially and explicitly, this transfer can also occur simultaneously and implicitly through multitask learning (MTL). In MTL, which is usually done with neural networks, a single model is trained on various tasks. The first n layers of the model are shared by all the tasks, resulting in internally generated ''shared'' features that feed into subsequent independent layers, which take each prediction task to completion. In one recent example, Zhang *et al.*[44] showed MTL to result in a more accurate CGCNN to predict various MOF stability metrics (*e.g.*, water and thermal stability, among others) compared to any CGCNN (or any other model) trained on a single stability metric.

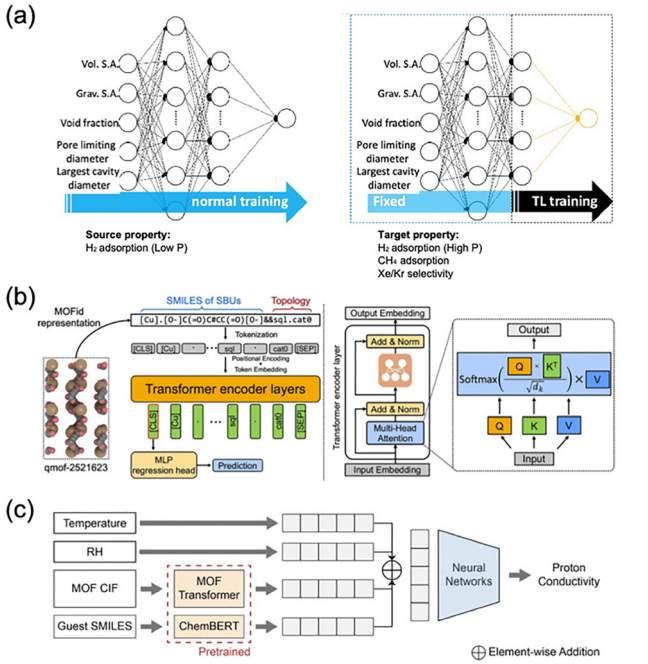

Fig. 3 (a) Schematic representation of transfer learning. First (left), an artificial neural network is trained on a source task with a source dataset. Then (right), the parameters of the hidden layers are frozen except for the final hidden layer, which is trained using a target task and target dataset. Adapted with permission from ref. 40. Copyright 2020, American Chemical Society. (b) The pipeline of the self-supervised MOFormer model for representation learning. The tokenized MOFid representation is embedded and augmented with a positional encoding before entering the transformer encoder layers (see right for a schematic of these layers). The learned embedding of the first token is to be used in downstream prediction tasks. Adapted from ref. 42. (c) Pretrained models (here, MOF transformer and ChemBERT) are used as inputs to downstream prediction tasks. For the prediction of proton conductivity using neural networks, input pretrained representations are augmented with embeddings for temperature and relative humidity and only the neural network and embeddings are trained (the pretrained models are frozen). Adapted with permission from ref. 43. Copyright 2024, American Chemical Society.

3.2. Transfer learning and fine-tuning from foundation models

To mitigate the sensitivity of TL to source datasets and tasks, the MOF field has seen a rise in the development of pre-trained foundation models, which are large models trained using huge and diverse datasets to internally learn (usually in a self-supervised fashion) general representations useful for broad tasks. Foundation models can be used as a common starting point to train new models for a variety of prediction tasks using a small task-specific dataset or limited training steps (or both), making model training more data-efficient. As in many other fields, the pursuit of foundation MOF models has been propelled by the advent of transformer models (the underlying model behind GPT-4), which started in 2017 with the work of Vaswani *et al.*[35] The first transformer model for MOFs was the 2022 MOFnet,[45] which was followed in 2023 by MOFormer[42] and MOF transformer,[46] and in 2024 by Uni-MOF.[47] Transformers are a neural network architecture that includes (trainable)

matrix operations that embody the concept of attention. In MOF transformers, the attention mechanism enhances or attenuates features (*e.g.*, atom or bond encodings) critical to the prediction task in a way that is influenced by other features (*e.g.*, other atom or bond encodings) unrestricted by ''proximity''. In the context of MOFs, this spatially unrestricted nature could be useful to simultaneously learn from aspects such as local chemistry and long-range structure (*e.g.*, pore size distributions). Typically, a MLP is trained to aggregate this attention-weighted ''transformation'' into the final prediction.

MOF atoms and bonds have been originally presented to transformers based on graph-like representations of the whole MOF (MOF transformer, Uni-MOF) or a representative MOF unit (MOFnet). Atom identities and bond topologies are also present in string representations, such as the SMILES of MOF secondary building units used in MOFid (MOFormer). Additionally, complementary global features meant to summarize pore structure have been added to the transformer either directly (*e.g.*, MOFnet with void fraction, surface area, largest pore diameter, and other textural properties) or indirectly (*e.g.*, MOF transformer with flattened representations of adsorption energy grids created *via* molecular mechanics calculations within the MOF unit cell).

Transformers are well-suited to create foundation models because they easily allow the creation of ''data-abundant'' self-supervised learning tasks that allow each MOF atom and/or bond feature, through the trainable attention operations, to focus on understanding the ''context'' in which they exist within the MOF. For example, the attention mechanism in the transformer could be trained to predict the identity and/or properties of a masked (*i.e.*, hidden) atom given the identity and/or properties of other atoms in the MOF (as in MOFormer, see Fig. 3b). Nevertheless, supervised learning tasks can also be added to further influence what aspects of their environment atoms and bonds pay more attention. For instance, looking for the influence of MOF global structural aspects, MOF transformer used predictions of topology and void fraction as part of the transformer training, where the prediction of multiple properties by the model indicates the exploitation of the MLT approach discussed at the end of Section 3.1.

All the above transformers have shown promise as a starting point for new tasks. In 2024, Han *et al.*[43] kept MOF transformer as-is in a new model (*i.e.*, transfer learning), enabling predictions of proton conductivity about 8% more accurate than training standard ML models from scratch (Fig. 3c). This work additionally suggests a transfer learning strategy with a lot of potential but not much explored up to date. Namely, the transfer of knowledge from models trained on simulation data to those trained on experimental data. The premise here is that models trained on experimental data are much more appealing, but that generating a data point from experiments is generally more costly and time-consuming than generating one from simulation.

Transformer parameters can all undergo optimization (initialized with the original parameters) for a new task in what is referred to as fine-tuning. In 2024, the Uni-MOF transformer was used as part of a ML model to predict adsorption in multiple molecules. The authors showed that fine-tuning the Uni-MOF part (as opposed to training the whole ML model from scratch), led to about 18% increase in accuracy. Nonetheless, fine-tuning of the current MOF transformers can still be outperformed by training of standard ML models using wisely chosen input features, as recently shown by Mao *et al.*[48] for predicting free energy in a set of polymorphic sulfur-based MOFs. This suggests there is still room for developing MOF transformers that generalize better upon fine-tuning. Additionally, some of the current MOF transformers require significant work/expertise/preprocessing to generate their inputs, which hinders their widespread use as a foundation model.

4. Efficient construction of training datasets

In ML, a data efficiency test consists of subsampling the training dataset at different fractions, training the best possible model with each subsample, and then evaluating the ML model prediction error against training dataset size. Usually, after an early rapid error drop with an increase in training dataset size, error plateaus after a critical dataset size. As prominently discussed by Moosavi *et al.*[29] in 2020 in the context of diversity in MOF databases, this critical size occurs when additional data points do not provide the model with ''new'' information. The above trends suggest that the training dataset size can be minimized without significant model accuracy loss, as long as each sample point in the dataset is as informative as possible. This minimization becomes more important the more computationally expensive it is to obtain training data points. Finding the most informative points and progressively adding them to the training dataset is the essence of active learning (AL).

4.1. Active learning based on Gaussian processes

AL starts with the training of an initial ML model (or models) to predict the MOF property of interest, using a purposely small initial training dataset. An acquisition function calculated on each potential training point is used to decide which are (likely) the most informative points to add to the extant training dataset. With the expanded training dataset, a new ML model (or models) is (are) trained, the acquisition function is recalculated, the training dataset is expanded again, and so on iteratively until a predetermined stopping criterion is met. For AL in MOFs, the most common acquisition function has simply been the uncertainty of the ML prediction, which simply results in adding to the training set the points for which model is ''less sure.'' There are acquisition functions that focus on diversity (important when there are numerous similar points in the training set) or on maximizing the impact on ML model parameters, but neither kind has been really explored in MOFs.

Thus, in 2022, Mukherjee *et al.*[51] reported the first exploration of AL in MOFs, focusing on predicting CO_2 and CH_4 adsorption, respectively, in Cu-BTC. Then, in 2023, these

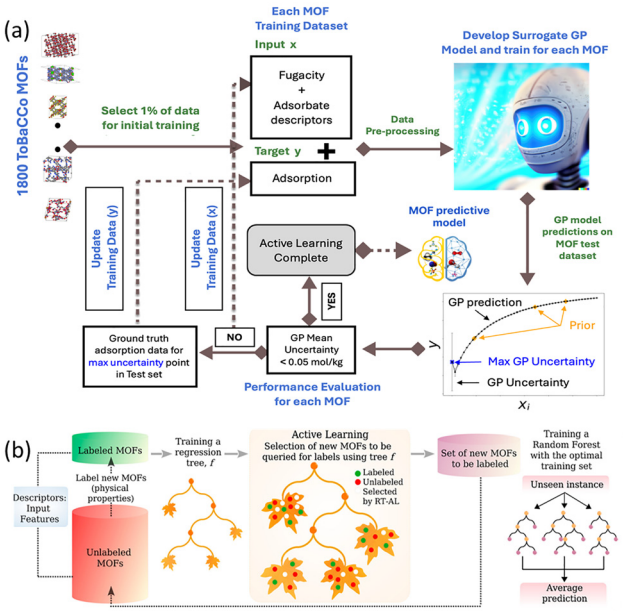

Fig. 4 (a) Schematic of an active learning workflow for alchemical adsorbates. A Gaussian process (GP) regression model is trained on an initial dataset to predict adsorption loading from five input features. The data point from the test set with the largest predicted GPR uncertainty is selected for adsorption calculation then added to the training dataset. The model is retrained and the loop continues until the uncertainty is below 0.05 mol kg^{-1}. Adapted from ref. 49. Copyright Royal Society of Chemistry. (b) Schematic of the regression tree active learning (RT-AL) workflow. During each cycle of training, new samples are selected *via* regression tree leaves with high uncertainty (based on variance) and the ratio of unexplored data points. A separate random forest is trained using the tailored training set for MOF property prediction. Reprinted with permission from ref. 50. Copyright 2024, American Chemical Society.

authors expanded their efforts to the prediction of CO_2/CH_4, Xe/Kr and H_2S/CO_2 mixture adsorption, respectively, in the above MOF.[52] One of the points made in these works was the influence of the initial dataset on final data savings. Interestingly, these authors reported boundary-informed sampling as the best way to choose the initial data points, which is a strategy where heuristics and human expertise can have a significant impact.

The potential data savings AL can achieve are apparent in the 2024 work by Osaro *et al.*,[49] which was directed to the prediction of adsorption isotherms for multiple molecules in MOFs using a single ML model (Fig. 4a). These authors examined the data requirements to train a ML model that uses pressure along with MOF and molecule features to make the relevant adsorption predictions. These authors reduced the training dataset size by a factor of about 2 when using AL to select the most informative (pressure, adsorbate) combinations for each MOF. A further reduction by a factor of about 500 was reported when AL was used to select the most informative (pressure, adsorbate, MOF) combinations, albeit with some loss in prediction accuracy.

In all the above-mentioned works, the AL cycle (*i.e.*, selection of training points) was driven by Gaussian processes (GPs), even if in some cases the final trained ML model was not itself a GP. Because GP predictions are inherently accompanied by a measure of uncertainty, GPs are a natural choice for AL in many fields. However, GP training becomes computationally intractable after a few thousand data points, which probably means that widespread application of AL in MOF research will require the exploration of GP alternatives that scale better with the number of training points.

4.2 Active learning based on Gaussian process alternatives

In 2024, several AL efforts with GP alternatives were reported. An obvious alternative to GPs are Bayesian neural networks (BNNs), which can provide uncertainty because every time inference is made (even for the same input values), the prediction can change. The reason is that, in contrast to regular NNs, in BNNs each node–node connection is described by a probability distribution of weight values instead of by a specific weight value. BNNs scale similar to regular NNs, making them appealing for AL work with large datasets and complex inputs. Still, obtaining the true probability distribution is computer-intensive, so approximations to the distribution are necessary.[53] For instance, Thaler *et al.*[54] used a BNN approximation to perform AL towards the prediction of MOF partial charges using a GNN as the core ML model, with the uncertainty of the prediction measured by having the GNN making predictions multiple times, each time randomly turning off neurons in a procedure known as dropout Monte Carlo. These authors found AL to be twice as efficient in terms of training point selection compared to random selection and showed that only about 13% of the MOFs (for which partial charges from density functional theory (DFT) calculations were available from databases) were needed for the GNN to reach desirable prediction accuracy.

The quantification of uncertainty by repeating predictions with the same input can be extended beyond NNs. Thus, Leverant *et al.*[55] used AL towards the prediction of MD-calculated diffusion coefficients using RFs as the core ML model, and the variance of the predictions from the different trees as the measure of uncertainty. These authors observed the usual improvement in accuracy as training points were added. However, as a reminder that training datasets can be too small even for an AL framework, these authors ran out of training data before desirable accuracies were reached.

An alternative method coined regression tree AL (RT-AL) uses a regressor tree (as the core model) that divides the putative feature space into regions, each one associated with a tree leaf (Fig. 4b). The prediction uncertainty for (potential and extant) training points in a given region corresponds to the variance associated with the corresponding leaf. The acquisition function selects a region based on its associated uncertainty and proportion of unexplored points and then randomly draws points from it. As shown by Jose *et al.*,[50] an advantage of RT-AL is that one can use the regressor tree to select training points, but then train a more powerful ML model (RFs for these authors) for the actual MOF property prediction task. Working on the prediction of band gaps and CO_2 and H_2 adsorption,

these authors found RT-AL to usually outperform GPs and other AL methods to efficiently construct the training set. An interesting byproduct of this work was a clear demonstration that the most efficient features for AL (and ML model training) can depend not only on the property to be predicted but also on the size of the training set. This work serves as an important reminder that AL selects training points by navigating a feature space with an efficacy that (at this point) is still contingent on the chosen input features.

5. Efficient exploration of the MOF design space

ML predictions (*i.e.*, inference), while (almost) instantaneous, still have a non-zero cost that may become relevant if these predictions were to be done trillions of times. On the other hand, a trained ML model is not guaranteed to be accurate across the whole MOF design space (increasingly so the farther one extrapolates beyond the regions covered by the training data). Additionally, training a ML model at all may not be possible if there is not enough accurate training data for it. Therefore, efficient ways to probe the MOF design space are needed, even if not relying on ML evaluation of MOF properties.

5.1. Evolutionary algorithms

The first methods to improve the efficiency of MOF exploration consisted of evolutionary algorithms (EAs). In this family of methods, a small subset of MOFs is initially evaluated, and then progressively "evolved" through rules that mimic biological evolution. These rules are known as genetic operations and are used to create new generations of MOFs, tending to favor traits that appear in the high-performance MOFs from preceding generations (*i.e.*, exploitation), while allowing new (lost) traits to spontaneously appear (or reappear) randomly (*i.e.*, exploration). An essential genetic operation is selection, which mimics evolutionary pressure by biasing MOF selection for subsequent operations based on MOF performance, as embodied by a fitness function f (to be maximized). A common genetic operation is crossover, which mixes the traits of two selected (usually high-fitness) MOFs. To perform these genetic operations, the MOF must be represented by a chromosome (vector), which encodes MOF traits as values of its genes (vector components).

In 2015, Bao *et al.*[59] introduced an EA to MOFs by evolving MOF linkers toward high CH_4 adsorption, using reaction-mimicking genetic operations. In 2016, Collins *et al.*[60] evolved MOF functionalization towards high CO_2 adsorption, while Chung *et al.*[61] evolved MOFs toward high CO_2/H_2 separation. The latter authors experimentally validated the high predicted performance of an EA-identified MOF, which was found by exploring less than 1% of the target search space. In all of the above studies, the fitness function was assessed using grand canonical Monte Carlo (GCMC) simulations, which can be a rate-limiting step that restricts the total number of generations explored. However, easily computed surrogate models that approximate fitness can dramatically improve throughput. To this end, in 2021, Lee *et al.*[10] combined EA and ANN predictions (as a surrogate for fitness) to explore a presumed search space of 247 trillion MOFs towards high CH_4 adsorption. Note that while EAs have hyperparameters, the above studies did not focus on their optimization, but rather on finding incrementally better MOFs than those reported at the time for the application of interest. Thus, there is significant room to improve the efficacy of EAs for MOFs.

Recently, exploring EA efficacy, Pham and Snurr[56] studied hyperparameter effects on the search of MOFs for CO_2/N_2 separation (Fig. 5a). Indicative of the importance of balancing exploitation and exploration in EAs, these authors found the probability of mutation to drastically impact search efficiency. Additionally, supported by a 25-fold reduction in computational cost, these authors proposed the execution of parallel EA runs, each with different initial MOF populations, as a way to improve EA efficiency. Nevertheless, an unsolved issue in EAs for MOF search is the restrictive rules needed to avoid attempting to make nonsensical structures, which hinders pairing EA with on-the-fly MOF construction. A common source of "nonsense" is the incompatibility of EA-proposed building blocks and topology combinations. Thus, a common solution is to restrict EA runs to a particular topology[10,56] or base structure.[62] This creates inefficiency as MOF topology (a critical MOF trait) is not optimized by the EA, and also precludes the discovery of new MOF topologies.

5.2. Bayesian optimization

EAs are intuitive (partly due to the modular structures of MOFs) and trivially adaptable to optimization of any MOF property as the means to evaluate the fitness f has no bearing on the EA. However, EAs are not as sample-efficient as other popular methods in the ML community, such as Bayesian optimization (BO).[63] This fact is important in MOF search, especially if evaluation of f requires quantum mechanical methods, long and involved MD simulations, or even experiments. Thus, recent years have seen the rise of BO in MOF search. BO shares a lot of similarities with AL (Section 4), differing in its goal of finding the x with the best f, as opposed to training a model that predicts f the best for any x (Fig. 5b). However, in both cases an iteratively trained surrogate ML model still predicts f for all x along with a corresponding uncertainty. The prediction and uncertainty still inform the acquisition function to select the next x to properly evaluate f. The new (f, x) pairs are still used to update the ML model.

In 2022, Taw and Neaton[64] presented the first BO example in MOFs, showing that BO would have found the best MOFs for CH_4 adsorption evaluating fewer than 1% of the target search space. However, standard BO (and standard EAs for that matter) may not account for all aspects relevant to MOF development. For instance, various (potentially conflicting) MOF properties may be important for a MOF application. Thus, in 2023 Comlek *et al.*[65] presented a multiobjective BO framework for MOFs, looking to improve the Pareto front that highlights the tradeoff between CO_2 uptake and selectivity. Their

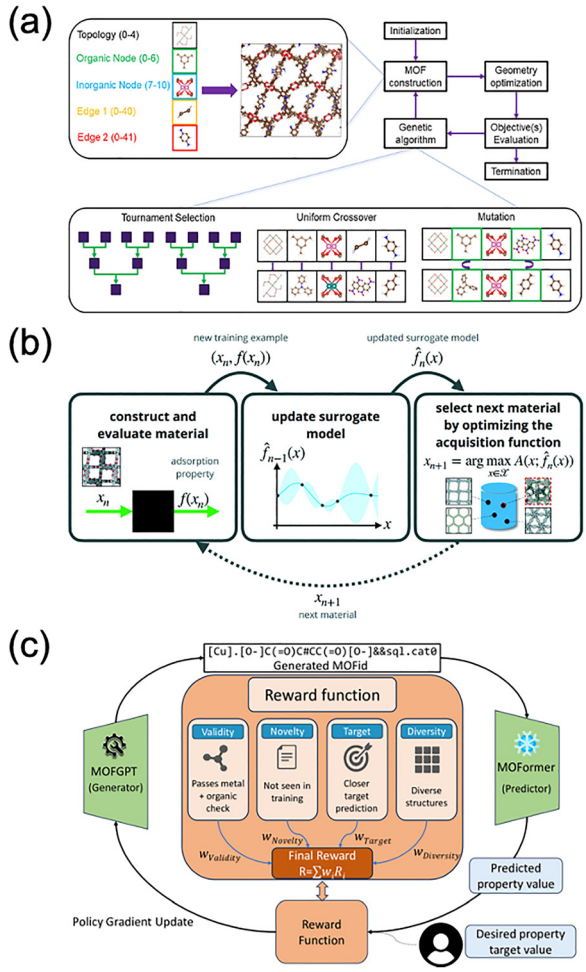

Fig. 5 (a) Schematic of a genetic algorithm (a form of evolutionary algorithm) workflow. A series of candidate MOFs are constructed, each represented as a chromosome with labels for topology, edges, and inorganic/organic nodes. After each generation, new candidates are proposed using evolutionary rules (*i.e.*, mutation, crossover, and tournament selection). The process is repeated until the specified objective (*e.g.*, MOF performance) is achieved. Reprinted with permission from ref. 56. Copyright 2025, American Chemical Society. (b) Overview of Bayesian optimization strategy to maximize adsorption property $f(x)$ of nanoporous materials. After evaluation of $f(x)$ for the current candidate, a surrogate model with uncertainty (*e.g.*, Gaussian process) is updated and the next candidate material is selected *via* an acquisition function, *e.g.*, the candidate that maximizes the upper confidence bound of the surrogate model. Used with permission of Royal Society of Chemistry, from ref. 57. (c) Reinforcement learning framework for property-guided MOF generation using MOFGPT and MOFormer. The reward function assesses the quality of the generated MOF *via* validity, novelty, diversity, and proximity to the target property. The reward is also used to update the policy model, which then selects the next MOF candidate. Reprinted from ref. 58.

key modification was to use the expected maximin improvement (EMMI) acquisition function, which chose MOFs for evaluations seeking to improve one of the two objectives doing worse at decision time. Another consideration is that experimental testing of a presumed best MOF design can fail due to unaccounted for factors, *e.g.*, the structure may not be stable or the prediction may be wrong. Thus, in 2024, Liu *et al.*[66] developed Vendi BO, aiming to find MOFs with similar (presumed) optimal performance but with different structure and chemistry. Their key modification was to use the Vendi score (a measure of diversity)[66] to iteratively eliminate parts of the search space that were too similar to the set of MOFs already evaluated.

As with EAs, there is significant room to improve the efficiency of BO in MOFs through hyperparameter choices[53,67,68] (*e.g.*, which acquisition function is used) or making the predictive ML framework (*i.e.*, surrogate model) more accurate. Due to its robustness, a common acquisition function in MOF search is the upper confidence boundary (UCB), which balances uncertainty with the improvement of the predicted property by adding some of the (positive) uncertainty to the property prediction for a given MOF. But as the impact of acquisition function choice is underexplored, other functions may be more efficient. For instance, Aqib *et al.*[53] showed expected improvement (EI) to outperform UCB. EI is a function that focuses more directly on improving the property as fast as possible, with the caveat of needing a highly reliable surrogate model. On the other hand, functions such as the previously mentioned EMMI and expected hypervolume improvement (EHVI)—which consider the pareto front of MOF properties—may facilitate multi-objective MOF optimization despite their higher computational cost.

As for the accuracy of the surrogate model, it is inherently tied to MOF feature choices but can also be improved by exploiting similar ideas to hierarchical screening, allowing the ML model to see more data. For instance, Gantzlet *et al.*[69] applied multifidelity BO to the search of MOFs for Xe/Kr separation, training the ML framework with many cheaply acquired selectivities based on Henry's constants and fewer expensive selectivities based on adsorption loadings.

5.3. Other alternatives

Aiming to efficiently find MOFs for NH_3 permeable membranes leveraging MOF expertise, Liu *et al.*[70] proposed a search framework that boosts expertise-driven hierarchical screening with an iteratively trained standard ML model. Briefly, the presumed top n set from hierarchical screening is used to initialize an ML model, which is applied to the whole search space to identify m presumed better MOFs than in the current top n set. In each iteration, the m MOFs are fully evaluated and used to improve the ML model and (if possible) update the top n set. Exploring less than 10% of the search space, this approach improved 80% of the top-200 predictions, improving MOF performance metrics by a factor of two for NH_3 adsorption loading and by an order of magnitude for selectivity-weighted NH_3 adsorption.

Leaning more into the ML side, reward-based methods such as Monte Carlo tree search (MCTS) and reinforcement learning (RL) are also emerging as alternatives to search MOFs. These methods seek sequences of MOF modifications that lead to optimal MOFs, with modifications that tend to result in higher "rewards" tending to be favored (some randomness is allowed to balance exploitation with exploration). Zhang *et al.*[71] used MCTS to find hydrophobic MOFs for CO_2 capture. In MCTS, each path through a tree represents a sequence of MOF

modification decisions. To construct the trees (*i.e.*, sequence of decisions), these authors predicted the reward (essentially a MOF performance metric) using a recurrent neural network (RNN) to process the SMILES strings used to define MOF linkers.

Kim *et al.*[72] used RL to search MOFs for CO_2 capture from air. In their RL framework, the MOFs were represented as a sequence of categorical variables (metal node and topology) and linker SMILES. Candidate MOF representations were generated by a transformer model, which along with a policy-gradient algorithm, acted as the decision-making agent. During the process, the agent decided on the strings to add to the MOF representation to maximize the corresponding predicted reward (either proportional to CO_2 heat of adsorption or to CO_2/H_2O selectivity). During RL (Fig. 5c), through policy updates, the agent learns to make "good decisions." The rewards were predicted by corresponding neural networks, each using as input the embedding of the MOF representation learned by the transformer. Promising MOF representations found to be "valid" were turned into actual MOF computational prototypes for which properties were calculated by molecular simulation. RL was clearly shown to propose increasingly better MOFs, with the caveat that the requirement of simulated property data for ~30k MOFs (stated by the authors as necessary to have the predictor ready to initialize RL) may pose challenges for some properties.

6. Data generation and utilization

The quality and diversity in MOFs datasets impact how efficiently a model can learn to predict MOF properties. Briefly, low-quality data points can blur the true relationship between MOF features x and property y, slowing down learning, and/or potentially yielding a model that predicts incorrect property values. Meanwhile, low-diversity datasets can slow down learning as well (see Section 4) while potentially producing models that may seem to work well but only for MOFs in some specific dataset. On the other hand, the efficient advancement of ML in MOFs requires well-curated, standardized, and easily shareable datasets that facilitate benchmarking. This so that efforts by the wider pool of researchers starting to contribute to ML development in MOFs are focused, concerted, complementary and synergistic, as opposed to unfocused, overlapping, and redundant. Therefore, with the importance of datasets in mind, this section provides an overview of the data landscape in MOFs, revealing data-related strengths, weaknesses and opportunities.

6.1. Structural data overview

MOFs in datasets are either hypothetical (*i.e.*, yet-to-be-synthesized) or experimental (*i.e.*, already synthesized). Hypothetical MOFs are important to push the boundaries of MOF development into experimentally unexplored design space, opening the door to dramatically different properties. The first source of hypothetical MOFs was the 2011 hMOF database by Wilmer *et al.*,[73] which contained 137k+ computer-generated structures, but only featuring six out of 2k+ possible topologies. The latter limitation spurred efforts to generate more diverse hypothetical MOF datasets in terms of topologies and inorganic nodes. These efforts usually use codes such as ToBaCCo[74] and ToBasCCo,[75] which map MOF building blocks onto topological templates. In one of the most recent efforts, Rubungo *et al.*[76] used ToBaCCo to create around one million MOFs (MOFMinE dataset) in 1393 topologies. But templates prevent diversification beyond known topologies. AI/ML methods could open the door to new topologies. For instance, MOFFlow by Kim *et al.*[77] uses conditional flow matching, where a learned vector field rotates and translates rigid MOF building blocks and transforms the lattice vectors simultaneously into MOF unit cells without a predetermined topology template. But topology considerations aside, inorganic node diversification in MOF generation is also needed.[78] Notably, Gibaldi *et al.*[79] recently created the HEALED SBU Library, which collects approximately 952 manually selected inorganic nodes, opening new possibilities for more diverse MOF generation, even with existing templates.

As for experimental MOFs, a popular source are the CoRE MOFs, first reported by Chung *et al.* in 2014,[80] and featuring ~40k structures in a recent update.[81] CoRE MOFs are processed versions of MOFs extracted from the more general CSD database[82] (which includes its own subset of ~10k computation-ready MOFs).[83] CoRE MOFs are relatively diverse in topologies and inorganic nodes but are not systematically modified (*e.g.*, in functionalization), which may result in "lots of classes but few examples per class," hindering ML model generalizability. Still, experimental MOFs are appealing because of the (presumed) barrierless transition between computational screening and experimental testing, despite practical concerns such as their general instability (*e.g.*, some authors estimate only 384 of the original CoRE MOFs are stable).[84] Perhaps, "best-of-both-worlds" efforts aggregating hypothetical and experimental MOFs, such as in the ARCMOF database (Fig. 6a), are a wise strategy going forward.

6.2. Property data overview

Property data for hypothetical MOFs must come from computation. Currently, most of it corresponds to adsorption data obtained *via* GCMC simulations. Aggregation efforts to create larger datasets of hypothetical MOFs from different databases[42,46] must consider the "patchwork" of properties available across the aggregated structures. Nevertheless, the most common property data for hypothetical MOFs are adsorption loadings at specific thermodynamic conditions for CH_4,[73] H_2, Xe/Kr,[85] CO_2/CH_4 and CO_2/N_2 mixtures.[86] Whereas, mostly driven by their relevance for GCMC simulations involving CO_2, DFT-calculated MOF partial charges are also common.[87]

Although property data for experimental MOFs can come from computation or experiment, most of it is also computational adsorption data and DFT-calculated partial charges.[88,89] Some efforts breaking with this common trend are the ~20k QMOFs by Rosen *et al.*,[90] which include DFT-calculated band gaps (among other electronic properties), and the NIST/ARPA-E

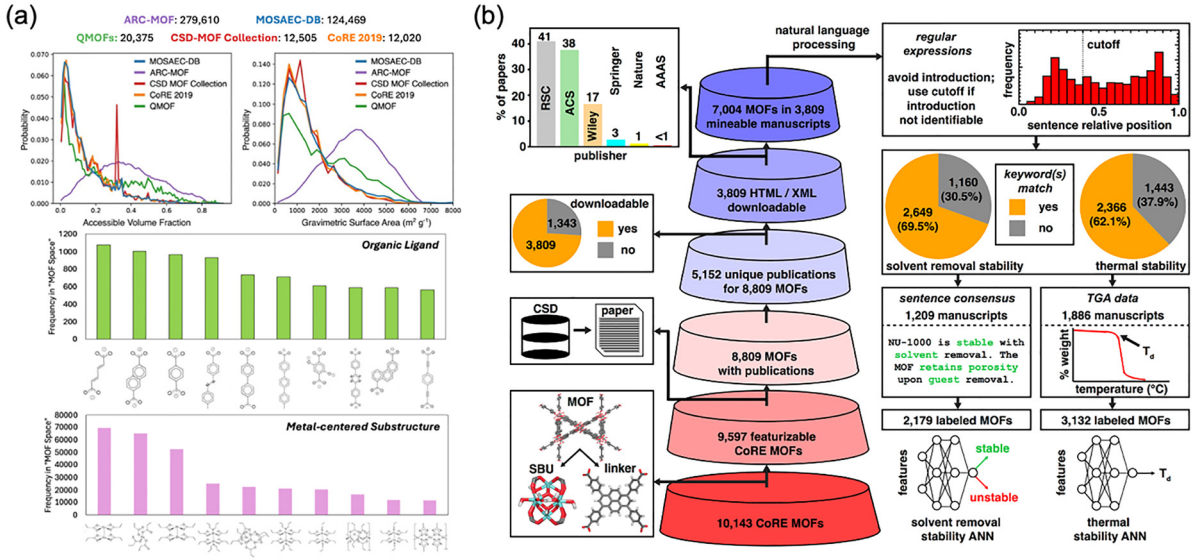

Fig. 6 (a) The diversity of MOFs in databases is varied. The top panel shows the probability density of accessible volume fraction and gravimetric surface area across MOFs in each of the listed databases; the numbers indicate the size of each database. Adapted from ref. 89. The bottom panel shows the diversity in organic ligands (green bars) and in metal-centered substructures (pink bars) present in the ARC-MOF dataset. Adapted with permission from ref. 87. Copyright 2023, American Chemical Society. (b) Mining experimental data on solvent removal and thermal stability of MOFs from the literature, as implemented in the MOFSimplify framework. Sanitized MOF structures from the literature and filtered for featurizability, and their associated manuscripts are retrieved and prepared for natural language processing. Text mining is then used to extract mentions of solvent removal stability and thermogravimetric analysis (TGA) data, including digitization of TGA traces from documents containing relevant keywords. Reprinted from ref. 93.

database for experimental adsorption data.[91] As typical with experimental data, the latter is less systematically varied but has a much wider diversity of "classes" covered than usually done by simulation, such as for more adsorbates, pressures, and temperatures.[92] The NIST/ARPA-E effort is, however, an example of imminent low-cost opportunities to create repositories for other experimentally measured MOF properties by mining reported data from the MOF literature.

6.3. Data from literature extraction

Indeed, with tens of thousands of MOF publications available, the prospect of extracting literature data to create low-cost datasets has been recognized for at least a decade. While earlier efforts required significant manual intervention—*e.g.*, the NIST-ARPA-E database for experimental adsorption isotherms and the WS14 dataset by Burtch *et al.*[94] for water stability of 207 MOFs—natural language processing (NLP) tools promise to automatize these efforts. To be sure, some form of data may still require manual intervention. For instance, the data used by Han *et al.*[43] in 2024 to predict proton conductivity *via* ML in 248 MOFs was typically reported in plots against temperature or relative humidity, so the authors manually extracted the data with the help of plot digitizers. Likewise, data to train a ML model to predict MOF thermal decomposition temperature had to be manually extracted by Nandy *et al.*[95] Still, a lot of important information exists as text, making it easier to mine, since relevant articles can be identified using existing application programming interfaces (APIs) and their text downloaded and parsed from their XML format for scrutiny *via* NLP.

Earlier NLP efforts in MOFs were primarily rule- and pattern-based. For instance, in 2017, Park *et al.*[96] developed a rule-based text mining algorithm that identified surface area and pore volume values by scanning for associated units like "$m^2\ g^{-1}$," and "$cm^3\ g^{-1}$." Despite the simplicity of the rule, the method achieved ~88% accuracy, with most errors stemming from inconsistent formatting or ambiguous naming conventions.

A recurrent, primarily rule-based, NLP tool is ChemDataExtractor,[97] which was used recently to extract MOF synthesis data for the DigiMOF database,[98] water stability information for the WS24 dataset,[99] and synthesis procedures for ZIF-8.[100] However, Glasby *et al.*[98] only managed to extract synthesis data for 9705 MOFs out ~15 000 MOF candidates, whereas Manning and Sarkisov[100] extracted data from only ~20% of the reports, despite their narrow focus on ZIF-8. Relatedly, Terrones *et al.*[99] used the tool to identify candidate sentences in articles for 1092 MOFs out of 5489 articles tied to the CoRE MOF 2019 database, but had to perform manual review to assign water stability classifications to these 1092 MOFs. These cases collectively reflect the Achilles' heel of rule-based NLP methods: the lack of standardized language in synthesis reporting.

Most recently, ML has been brought into NLP of MOF literature, recognizing the large variability in reporting language. The ML model tends to be in the form of RNNs or transformers, whose sequence-awareness and self-attention mechanisms, respectively, allow them to create context-aware representation of words/tokens. For instance, Nandy *et al.*[93] used Stanza,[101] an NLP toolkit based on RNNs, to help analyze nuances in sentences previously processed with ChemDataExtractor as containing information on

stability to solvent removal, which was used to generate training data for an ML model predicting this MOF quality (Fig. 6b). In another instance, Park *et al.*[102] complemented rule-based tools with the training of a named entity recognition (NER) model based on SciBERT,[103] a transformer-based language model pre-trained on scientific text, to mine data for some MOF synthesis aspects from 28 565 publications. However, this effort required extensive manual labeling of hundreds of literature paragraphs.

Literature extraction has heavily focused on MOF synthesis data. The appeal is the quantity of data (after all, every experimental MOF paper should report a synthesis procedure) and the potential use of the data to train ML models to anticipate synthesis outcomes,[104] which is crucial to bridge the gap between computation and experiment. But as language variability is exacerbated in synthesis reporting, LLMs such as GPT-4 are emerging as powerful literature extraction tools. Thanks to their immense pre-training, LLMs are better positioned to recognize synthesis procedures with little or no fine-tuning.

To this end, Zheng *et al.*[105] focused on prompt engineering, finding the "right way" to ask GPT-4, so that the LLM would accurately extract and organize synthesis data. Although the approach achieved high accuracy in extracting specific synthesis parameters (with F1 scores of 90–99%), it was intentionally limited to a fixed set of details (such as solvents, temperatures, and precursor amounts) formatted into tables, which constrained its ability to capture more nuanced or varied synthesis descriptions. Building on this prompt-driven approach, the L2M3 (large language model MOF miner) framework[106] used a series of GPT-based models to extract a broad range of synthesis conditions and material properties from over 40 000 MOF articles. While it primarily relied on updating prompts to adapt to new tasks, L2M3 also incorporates light fine-tuning for specific tasks within its pipeline to improve performance. This combination improves consistency and task-specific accuracy across large-scale, multi-step extraction workflows, addressing limitations in robustness that pure prompting can face.

Despite persistent challenges with inaccurate or inconsistent reporting, NLP extraction has shown promise by producing unified, large-scale MOF datasets that have been actively used to train ML models predicting synthesis outcomes and material properties.

6.4. Considerations on data quality

In describing MOF datasets and properties therein in the preceding subsections, data diversity is implicitly discussed. Thus, now we discuss quality aspects of the data itself. One concern, particularly important for computed data, is MOF structural errors. To be sure, not all structural errors are equally significant for computation, and their impact ultimately depends jointly on the structural error itself and the computed property. For instance, a missing hydrogen on a Zr node would hardly affect CH_4 adsorption, but a missing linker can significantly affect the calculation of partial charges or mechanical stability. Regardless, recognizing the existence of structural errors in MOFs, some authors have started to work on understanding the impact of those errors,[107,108] while others are starting to focus more on the detection and correction of those errors.[109]

Other concerns with respect to computed data stem from the calculation method, whose choice is primarily driven by the goal of facilitating large-scale data generation. For instance, for adsorption data, the predominant use of generic force fields (*e.g.*, UFF for MOF atoms) to describe adsorption interactions raises concerns, especially when the key adsorption interactions involve chemisorption. It has been possible to derive DFT-parameterized force fields to properly model particular MOF-adsorbate combinations,[110–112] but approaches to correctly describe chemisorption interactions during HTCS are needed. For electronic MOF properties, the concern is tied to the use of DFT as the workhorse to generate data, because strictly speaking, DFT is not adequate to model MOF metals. Still, DFT may be acceptable for certain properties such as partial charges and adsorbate binding energies, but more worrisome for properties such as band gap, which DFT is well-known to underestimate (although somewhat systematically across similar materials).[90] The case of MOF electronic properties truly underscores the data scarcity issue in MOFs. Accurate electronic calculations *via* quantum mechanical calculations are so expensive in MOFs that alternatives such as ML models are truly desired. Yet such ML models are not easily trainable because the training data is so expensive to obtain.

Relevant to literature extraction, experimental data is not free from concerns, which primarily arise from the variability in quality of both MOF samples and property measurement methods across labs. For example, the variability in reported Brunauer–Emmett–Teller (BET) surface areas for the same MOF may be reflective of material quality variations.[113] But regardless of the reasons, variability in measured properties is apparent, for instance, when examining differences across experimentally measured isotherms for the same MOF.[114] The obvious question is then: "what is the correct measurement to use for ML?" Empirical correction factors based on perceived MOF quality (as those used by some authors to fairly compare measured and simulated isotherms[115,116]) might be a first step towards unifying experimental data for a given MOF. But given the importance of mining experimental data from the literature to create low-cost datasets, efforts to standardize reported experimental measurements should be beneficial for the ML endeavor in MOFs.

6.5. Exploiting potentially unreliable computational data as synthetic datasets

If computed property data is inaccurate and the ML model uses chemistry-based descriptors (*e.g.*, counts of a given atom) as input, resources will be wasted learning an incorrect chemistry-property relationship. But this issue can be bypassed with chemistry-agnostic models, turning potentially unreliable computational data (from a chemistry perspective) into useful data (providing payoff for their generation cost). Examples of chemistry-agnostic models include models using energy histograms as input, as those introduced by Snurr and coworkers.[117–119] One of these

histograms is built from a MOF-representative grid of adsorption sites and their corresponding energies. While the histogram implicitly corresponds to a certain MOF chemistry, what the model learns is the relationship between the distribution of adsorption energies and adsorption loading. Even if force field refinement later shows that the histogram for a given MOF does not really look as initially thought, the learned energy-loading relationship remains valid and the model useful to predict adsorption for the given MOF based on the new histogram. Another chemistry agnostic model is the 3D-CNN introduced by Lin and coworkers,[31,120] which uses MOF-representative grids of adsorption sites embedded with their force field parameters as input. The latter approach echoes work in other areas that use ML to solve complicated simulation models by using the simulation model parameters as input.[121]

Chemistry-agnostic models are also compatible with synthetic data generation, which can be used to bypass the data generation bottleneck incurred when one must first find the simulation model parameters that accurately describe a specific chemistry. For the adsorption case, instead of running expensive DFT calculations to fit a force field, one may focus on producing large simulation datasets with a variety of simulation parameters. Moreover, the decoupling from specific chemistry also allows one to choose parameters that are most informative to let the ML model learn more efficiently. Anderson *et al.*[25] used this strategy to create ''alchemical'' molecules to train a ML model capable of predicting single adsorption isotherms for a variety of real molecules. Fanourgakis *et al.*[122] extended this idea to the creation of artificial MOFs to train a ML model to predict CH_4 adsorption. The accuracy achieved in the above works can be partly explained by the synthetic data boosting the interpolation capabilities of the ML models. Nevertheless, the generation of synthetic data beyond MOF adsorption properties is yet to be explored.

6.6. Scalability

Based on the evolution in the types of ML models used for MOF research, one may be tempted to expect a smooth pathway towards more accurate property predictions simply by generating more training data and using more complex ML architectures. However, some caution must be exercised to this expectation, as the computational resources needed to handle datasets of a certain size or work with some architectures can impose some seldom thought about constraints. For instance, the training cost of the (now popular) GPs scales according to $O(N^3)$, where N is the number of datapoints. This scaling seems to impose a practical limit of 5k to 10k datapoints. On the other hand, the large unit cells of MOFs and their correspondingly high number of atoms can hinder straightforward translation of architectures successful in other material development areas to MOFs, specifically due to memory footprint.

For instance, for models that use voxelization of a cubic MOF region as input to 3D-CNNs, this footprint increases as $O(L^3)$, where L is the length of the cube. In practice,[31,120] the spatial resolution (*i.e.*, voxel size) limits L up to around 3 nm as atomic information intuitively requires $\sim$1 Å resolution; larger unit cells (*e.g.*, MOFs with lattice dimensions that can go up to $\sim$170 Å)[123] will likely require loss in resolution. For models based on GNNs, this footprint increases as $O(n,d^2)$ where n is the number of nodes and d is the size of the feature vector embedded in each node. As nodes usually correspond to MOF atoms, unit cells can reach up to tens of thousands of atoms while embeddings usually have dimensionality on the order of tens to thousands. A look into reducing training costs of GNNs was given by Korolev and Mitrofanov,[124] who trained coarse-grained GNNs to predict various MOF properties with promising accuracy. These authors coarse-grained the model by basing the MOF graph on the corresponding topological template and using the pre-established mol2vec embeddings of molecules to indicate which MOF building block was occupying a given graph node or edge.

7. Frontiers

We now highlight frontiers in ML-aided MOF development, noting their impact on data-efficiency aspects discussed so far and/or highlighting their own data-efficiency challenges.

7.1. Generative inverse MOF design

In the first wave of computational MOF discovery, MOF databases would be created and then the property of interest would be predicted across all structures to find if one had a desirable property value. In the second wave, MOFs started to be evaluated and/or modified sequentially (from a database or from a ''virtual'' design space), hoping to evolve towards a MOF design that has a desirable property value.[49,68,72] An ambitious emerging paradigm is generative inverse design, in which one would establish a desired property value and a generative model would yield the design with the desired value. As a naïve approach, one could leverage any of the ML models that predict MOF performance using a vector of features as input to optimize these features to maximize performance.[21] One issue is that the specific ''optimal'' combination of features emerging from such an exercise is usually unattainable in an actual MOF. But in a generative model, the vectorial MOF representation is meant to always correspond to realistic structures, ensuring that an optimized vectorial MOF representation also corresponds to a valid MOF design.

An example of a generative model is variational autoencoders (VAE), which are made of an encoder and a decoder. VAEs usually use a neural network as the encoder to learn a continuous representation of, say, MOFs as vectors in a so-called latent space while the decoder (also usually a neural network) is trained to reconstruct, in this case MOFs, from their representation in latent space. A ML property predictor can then be coupled with the encoder to learn the relationship between the latent vectors and the property of interest. With these elements in place, one can simply optimize the latent vector based on the property of interest and use the decoder to recover the corresponding optimal MOF. Yao *et al.*[125] demonstrated the use of VAEs to optimize MOFs to separate CO_2/N_2

and CO_2/CH_4 mixtures. As input to the VAE, these authors used a MOF representation based on categorical variables to constrain topology and modular building blocks and self-referenced embedded strings (SELFIES) to represent connecting building blocks.

Denoising diffusion probabilistic models (DDPMs) are also generative models that learn to predict valid MOFs out of "noise." To train DDPMs, one first iteratively adds (usually Gaussian) noise to valid MOFs until the original representation is reduced to pure noise. Then a neural network (SE(3)-equivariant) learns to reverse the process (*i.e.*, denoise) to recover the valid MOFs. The trained DDPM can thus learn to generate valid MOFs out of randomly sampled noise. Although not used for inverse design, Park *et al.*[126] used a DDPM to generate new linkers but constrained to the isoreticular MOF series, whereas Duan *et al.*[127] expanded this approach to also generate nodes (although constrained to four topologies) and showed the validity of generated structures by synthesizing one of them.

Crucially, Fu *et al.*[128] and Park *et al.*[129] demonstrated the amenability of DDPMs for inverse design by conditioning the learning of the denoising process on a property of interest, which was leveraged to have DDPM generate MOFs with optimal values for said property. Fu *et al.*[128] focused on CO_2 adsorption (a numerical property), while Park *et al.*,[129] by jointly training the diffusion model on conditional and unconditional tasks, showed that conditioning can also be done on categorical properties or text-input without training an external classifier (Fig. 7a). The inverse design of Park *et al.* was focused on "pore surfaces," which were matched (and thus constrained) to pre-existing building blocks and topologies. On the other hand, Fu *et al.*[128] used a coarse-grained representation of the MOF, in principle generating positions for building block centers unconstrained by topological templates, but still indirectly constrained by the validity of the structure once actual building blocks are denoised and mapped onto the building block center positions.

Seeking higher data-efficiency by eliminating the large-scale pre-training phase of VAEs and DDPMs, Cleeton and Sarkisov,[130] in what seems an evolution of the naïve approach of optimizing inputs in a property-predictive ML model, proposed deep dreaming (DD). Thus, these authors first trained

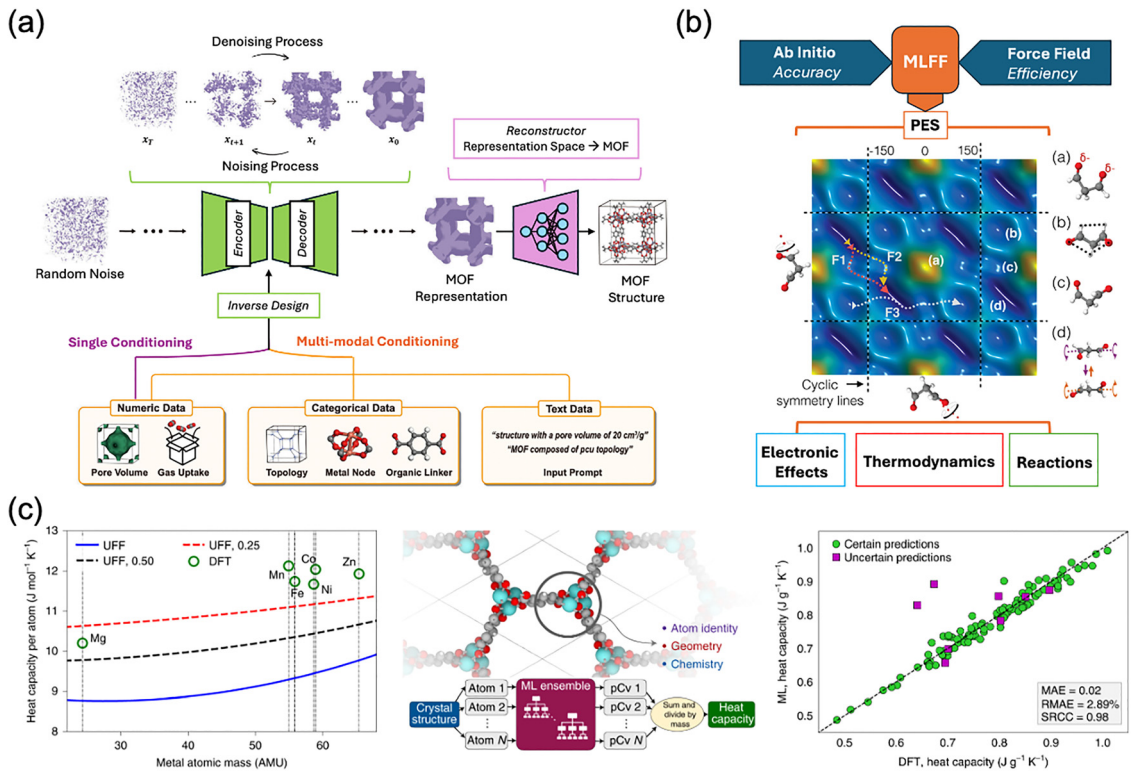

Fig. 7 (a) Inverse design of MOF *via* conditional diffusion model. Schematic of a general diffusion architecture that uses single and/or multi-modal conditioning to guide structure generation. The pair encoder–decoder learns to denoise MOF structures guided by the conditioning criteria while an external model transforms the MOF representation to a material structure. Adapted from ref. 129. (b) Machine learning force fields (MLFFs) bridge the accuracy of *ab initio* methods with the efficiency provided by classical force fields. This combination enables fast and reliable approximations of potential energy surfaces (PESs) to unlock the study of multiple phenomena (*e.g.*, electronic effects, thermodynamics, and reactions). Adapted from ref. 131. (c) The prediction of MOF heat capacities using machine learning models. The left panel compares DFT-computed heat capacities (circles) to those from the classical universal force field (UFF, lines); the dashed lines are results from the metal-linker force constants in UFF scaled by the listed factors. The inconsistent heat capacities computed from the classical model motivate a machine learned model (middle panel) based on the contribution of each atom to the total heat capacity. The correlation between the machine learned model predictions and DFT calculations are shown in the right panel. Adapted from ref. 132. Copyright 2022, Springer Nature.

transformer-inspired ML models to predict target properties. Then DD was applied by freezing the model parameters, reversing the propagation direction, and using gradient ascent to modify the input vector to maximize the target property, with their use of SELFIES-based inputs helping maintain the validity of the proposed input. Nonetheless, this effort was restricted to linker generation.

7.2. Machine learning force fields

Machine learning force fields (MLFFs) can have a huge impact on the kind and quality of data that can be affordably generated at large-scale. Briefly, simulation data for numerous MOF properties depend on statistical sampling of the underlying potential energy surface (PES). As discussed in Section 6.5, some concerns on computed data quality stem from the use of generic classical force fields to approximate the PES. MLFFs[131,133] can be trained on quantum mechanics (QM) data to reproduce (ideally) the entire PES of a MOF system, yielding QM accuracy-level energies and forces at low-cost, to guide MC or MD sampling (Fig. 7b).[134,135] One key factor is the ''flexibility'' of the MLFFs, which allows them to capture features of the PES (*e.g.*, directionality) in ways that the functional forms of classical potentials (even ReaxFF)[136] may not capture without fortuitous error cancellation.

The key to MLFF training is to assign a force and a contribution to total energy to each atom in the system based on their atomic environment. To do so, most MLFFs decompose total energy into atom-centered contributions based on descriptors of each atom's surrounding environment, ensuring additivity and size extensivity, as initially demonstrated by Behler *et al.*[133] in 2007 and later refined by DeepMD. Early models like DeepMD use local descriptors that are invariant to permutation, rotation, and translation (local symmetry functions or descriptor-based encodings) as shown in Zhang *et al.*,[137] while more recent architectures such as NequIP and MACE employ message passing and equivariant neural networks to capture directional interactions and preserve these physical symmetries, as shown by Vandenhaute *et al.*[138] and Elena *et al.*[139] These architectures contribute to data efficiency, having demonstrated strong learning performance from relatively small training sets.

Indeed, while MLFF can address generation challenges for MOF property data, MLFF development can face data challenges itself. For instance, the large unit cells, structural flexibility, and hybrid metal–organic bonding in MOFs introduce challenges for both data generation and model transferability, as noted by Eckhoff *et al.*[140] For instance, flexible MOFs with rotating linkers, as discussed by Dürholt *et al.*[141] and Zhao *et al.*,[142] or guest-induced transitions, as discussed by Bucior *et al.*,[117] demand potentials that respect rotational symmetries and long-range interactions. Even with recent advances, applying MLFFs to MOFs still requires domain-specific strategies, such as training on nodes and linkers separately,[140] using temperature-driven active learning to reduce DFT sampling, as demonstrated by Sharma *et al.*,[143] and hybrid force fields that integrate classical physics with ML components, as presented by Wieser *et al.*[144] These techniques aim to make MLFFs more than just DFT replacements, enabling them to simulate MOF flexibility, guest diffusion, and even decomposition under real-world conditions, as recently discussed by Castel *et al.*[145]

To overcome the scalability bottleneck of training MLFFs directly on full MOF unit cells, fragment-based strategies have emerged that treat chemically meaningful substructures, such as linkers and nodes, as independent learning units.[145] This approach allows the development of transferable potentials with reduced data requirements while maintaining fidelity to periodic properties, as shown by Tayfuroglu *et al.*[146] Recent efforts further integrate active learning with fragment selection to prioritize diverse and data-efficient training sets, as demonstrated by Shi *et al.*[147] Although fragment-based models may underrepresent long-range coupling effects, they offer a practical route to generalizable and scalable MLFFs for large and flexible MOFs.

Hybrid ML/classical approaches also enhance data efficiency by embedding ML corrections, like learned charges or dispersion terms, into existing classical force fields to refine interactions without retraining entire potentials, as demonstrated by Thürlemann *et al.*[148] This has been demonstrated in MOFs, where ML models correct non-bonded terms to achieve better electrostatics and van der Waals behavior, as shown by Korolev *et al.*[149] Additionally, hybrid MLFFs that combine neural short-range potentials with classical electrostatics have been shown to achieve near-DFT accuracy for MOF relaxations and phonons.[144]

Inspired by advances in foundation models for molecules and materials, emerging efforts are exploring pretrained machine learning potentials for MOFs.[138] These models are trained on diverse atomic environments to produce generalizable force fields that can be fine-tuned with minimal new data. For example, MACE MP MOF0, which combines pretrained MACE with targeted MOF fine-tuning, enables rapid adaptation to new MOF fragments and accurate phonon and thermomechanical predictions with very little data.[139] Though no universal MOF MLFF yet exists, the strategy of pretraining on building blocks, such as nodes, linkers, or secondary building units (SBUs), followed by system-specific tuning has been demonstrated for both porous and flexible frameworks.

The impact of ML potentials extends beyond accurate potential energy predictions, as they can serve as core engines for simulations of dynamic MOF behavior. For example, MLFF-driven MD simulations have been used to explore guest diffusion in MOFs: a NequIP-like neural potential accurately modeled H_2 binding and diffusion in open-metal-site frameworks, predicting kinetics and isotherms previously inaccessible *via* DFT, as recently reported by Liu *et al.*[150] In flexible MOFs, where linker motion and node distortion critically influence framework behavior, MLFFs have been shown to reproduce temperature-driven structural and vibrational changes that classical force fields struggle to capture.[138,143] As these models mature and benchmark data improve, MLFF-powered simulations could become indispensable for capturing the full complexity of MOF behavior in real-world scenarios.

7.3. Machine learning for properties beyond physisorption

As noted in Section 6.2, published MOF datasets (and hence ML efforts in MOF) are dominated by simulated adsorption data.[91,151–153] This is a result of MOF development having been primarily driven by adsorption applications. But even in adsorption applications, properties beyond adsorption are relevant to choose a MOF for experimental testing, raising the need for ML to predict properties beyond adsorption. For instance, thermal and mechanical stability are important to know if the MOF can withstand operating conditions.[154–156] Heat capacity is important to know the MOF tendency to suffer performance-degrading as temperature increases upon adsorption.[132,157] Thermal conductivity is important to inform heat dissipation strategies during MOF utilization.[158,159] Moreover, free energy is important to gauge if a MOF computational prototype is even synthesizable.[160]

To develop MOFs for applications beyond adsorption, prediction of properties beyond adsorption is obviously needed. For instance, diffusion properties are relevant for drug delivery, low thermal conductivity for thermoelectrics,[161] high electrical conductivity for electrocatalysis and energy storage,[162] high proton conductivity for fuel cells,[163] and so on. One significant barrier to generating data for properties beyond adsorption is their usually higher simulation cost. This is obvious if quantum mechanical methods are needed (*e.g.*, electronic structure, bond breaking/formation events), but it can also be the case with classical simulations. For instance, computing diffusion coefficients may require MD coupled with enhanced sampling,[164] free energy may require coupling with thermodynamic integration,[160] or thermal conductivity may require large supercells to mitigate finite size effects or extended simulations for convergence (*e.g.*, *via* the Green–Kubo method[159]). For some MOF aspects, the adequate simulation method may not even be clear (*e.g.*, MOF decomposition or formation) or has not been fully developed.[165,166]

Based on the above, a combination of simulation advances, literature extraction, and approaches for data-efficient training is likely needed to develop reliable ML models beyond adsorption. Encouragingly, where enough data has been generated by pushing simulation resources or through literature extraction, ML predictions beyond adsorption have emerged with promising results. From simulation data, models to predict mechanical stability,[156] heat capacity,[132] and diffusion coefficients[167–169] have emerged (Fig. 7c). Similarly, the publication of the QMOF database has spurred a number of ML models trained to predict band gaps.[42,90,170–172] Although in this case the training data is not fully accurate (*i.e.*, based on DFT), these models could offer a starting point for transfer learning or fine-tuning once more accurate but probably scarcer band gap data emerges. Related to this strategy, Rubungo *et al.* showed that fine-tuning a ML model originally trained on low-cost strain energy was key to achieving accurate ML predictions for high-cost free energy.[76] On the other hand, literature extraction has enabled data to train ML models to predict thermal[93] and water stability,[99,173] whereas a combination of literature extraction and fine-tuning enabled ML predictions for proton conductivity.[43,174]

On a final note, while MOF data has been dominated by adsorption, it has been specifically dominated by physisorption, although chemisorption is likely relevant to numerous target applications.[175–177] Thus, efforts to extend data generation to chemisorption are needed. Since the simulation adsorption cost is usually not prohibitive for training data generation, the challenge here is an accurate description of interactions. Although accurate force fields have been parameterized for specific adsorbate-MOF cases,[178–180] HTCS-compatible (*i.e.* transferable) accurate force fields to describe chemisorption interactions are necessary. Force fields aside, an adsorption case for which the simulation data generation is notoriously challenging is water,[181–185] which will require significant simulation advances or reliance on experimental data as an alternative. Nonetheless, water is a case that merits special attention due to its ubiquitous presence in many applications and its direct relevance to applications such as water harvesting.[186–188]

7.4. Possibilities with LLMs

With the advent of the LLM era, it is natural to wonder what can be accomplished with these types of models. With their expected role in data generation *via* literature extraction already noted in Section 6.3, it is worth noting that they can also play a role in efficient MOF property predictions. Recently, Rubungo *et al.*[189] showed a LLM model (LLM-prop) to generally outperform GNN-based models (whose training is more computationally demanding) in the prediction of a variety of crystal properties. These authors posit that this efficiency is due to the ''expressiveness'' of natural language in describing key aspects of crystals that influence their properties. The potential impact of frameworks such as LLM-prop in MOF research was recently shown by prediction of MOF free energies using a string-based MOF representation called MOFSeq as input.[76] Similarly, Wu and Jiang[190] fine-tuned Gemini-1.5 (Google's LLM) to predict MOF hydrophobic character simply using SMILES/SELFIES as input with comparable or superior accuracy compared to other models with highly engineered features.

Given the existence of SMILES/SELFIES, the string-based representation used as input to LLMs is also particularly amenable to MOF linker generation. For instance, by fine-tuning GPT-3, Zheng *et al.*[191] generated new candidate linkers for water-harvesting. Other representations can facilitate other tasks. For instance, by conveying MOF information into textual document form, Zhang *et al.*[192] used the unsupervised Doc2Vec model to create a MOF representation that was used to develop a MOF recommendation system. This recommendation strategy, which was introduced earlier by Sturluson *et al.*,[193] and followed by Zhang *et al.*,[194] was inspired by the Netflix movie recommendation system, and suggests promising (extant) MOFs for applications of interest by analyzing similarities to user-endorsed MOFs.

LLMs have also been shown to work as assistants coordinating and streamlining computational work. For instance, ChatMOF,[195] which integrates GPT-3 and GPT-4 with more specialized ML models (*e.g.*, for property predictions, MOF generation, *etc.*), has been shown capable of recommending

MOF structures for properties of interest. Beyond predictive modeling, and more on the synthesis side, LLMs are increasingly being deployed as "interactive research assistants," capable of orchestrating and accelerating complex experimental workflows. For instance, the GPT-4 Reticular Chemist[196] integrates GPT-4 into a cooperative loop with human researchers, where the model proposes synthesis steps, receives outcome feedback, and adapts its guidance through prompt-based in-context learning. This iterative process allows GPT-4 to refine its recommendations much like an experienced chemist. Similarly, the ChatGPT Chemistry Assistant[105] employed prompt engineering to automate text mining of MOF synthesis conditions across diverse literature formats, eventually leading to a ML model predicting crystallization outcomes with 87% accuracy. Expanding on these capabilities, the ChatGPT Research Group[197] introduces a multi-agent framework comprising seven specialized LLMs responsible for tasks ranging from literature review and synthesis design to robotic control and data interpretation. By combining these agents with BO, the system rapidly identified optimal synthesis conditions, significantly accelerating materials development. These assistant-type applications demonstrate how LLMs can bridge diverse aspects of the scientific process, functioning not just as tools for analysis, but as collaborators in experimental strategy and execution.

8. Conclusions and outlook

AI/ML is reshaping how researchers explore the vast design space of MOFs. From early applications using geometric descriptors and linear regressors to recent advances using transformers, foundation models, and generative architectures, ML tools now touch nearly every stage of the MOF characterization and discovery pipeline. However, the high-dimensional, modular, and (often) sparsely labeled nature of MOF data imposes persistent challenges. As this review has highlighted, progress in ML-aided MOF design has relied not only on increasing model complexity but also on improving data and resource efficiency through feature engineering, architectural choices, transfer learning, active learning, and data curation strategies.

A central theme observed throughout this review is the need to match model sophistication with the quality and diversity of available data. Models that incorporate inductive biases grounded in chemistry and physics often outperform black-box approaches in data-limited settings. For properties that are expensive or difficult to compute, hybrid workflows leveraging either literature mining or active learning combined with ML offer a promising path forward. At the same time, large foundation models and generative models are beginning to offer a pathway for generalized representation learning and *de novo* MOF prototype design. However, both of these approaches will be constrained by the scope of training data and care should be taken to expand the diversity of node/linker chemistries and topologies within these datasets.

Despite these advances, several key challenges remain. Currently, standardized benchmark datasets, similar to those seen in the small molecule development community, do not exist for MOFs, which makes it difficult to compare ML methodologies and critically assess progress over time. In addition, MOF property data and prediction tasks are dominated by gas adsorption whereas the promise of MOFs extends to far more application areas. Therefore, datasets containing transport properties (*e.g.*, diffusion, thermal conductivity), stabilities, and free energies, to name a few, and methods to compute these properties accurately and efficiently are still needed. Nonetheless, for many MOF properties, the quality (accuracy) of the datasets needs to be improved, which creates an opportunity where ML force fields are called to make a significant impact. On the other hand, as interest in MOFs expands to those with increasing complexity (*i.e.*, larger unit cells or flexible topologies), new strategies will be needed to address computational challenges related to data representation and scaling.

Looking forward, the integration of ML models with awareness of synthesis feasibility, simulation-informed priors, or human-in-the-loop design will transform ML pipelines from simply predictive tools into generative, decision-making partners, especially for inverse design. While the true potential of ML-aided MOF design has yet to be realized, the hope is that future ML-mediated workflows will enable the creation of MOFs that defy conventional human intuition, including those with previously unseen topologies, properties, and function. For instance, to our knowledge, only one MOF topology (**nun**) not already present in the RCSR database has been discovered in the past 20 years.[123] Nonetheless, the foundation is now in place for ML to become a critical driver of innovation in MOF materials science.

While all the described methods enable more options to unbiasing explore the MOF design space, a latent challenge emerges strongly. These methods are not aware of the synthesis accessibility of the proposed structures and/or building blocks. Therefore, including efforts that guide the design along synthesizable structures is a key to unlocking the spread of this AI/ML-based inverse design approaches in MOFs.

Conflicts of interest

The authors declare no conflict of interest.

Data availability

No primary research results, software or code have been included and no new data were generated or analyzed as part of this review.

Acknowledgements

The authors acknowledge funding from the National Science Foundation (NSF) *via* grants OAC-2118201 (HDR: Institute for Data-Driven Dynamic Design) and CBET-2450909.

References

1 H. Furukawa, K. E. Cordova, M. O'Keeffe and O. M. Yaghi, *Science*, 2013, **341**, 1230444.
2 M. P. Suh, H. J. Park, T. K. Prasad and D.-W. Lim, *Chem. Rev.*, 2012, **112**, 782–835.
3 Y. He, W. Zhou, G. Qian and B. Chen, *Chem. Soc. Rev.*, 2014, **43**, 5657–5678.
4 V. Stavila, A. A. Talin and M. D. Allendorf, *Chem. Soc. Rev.*, 2014, **43**, 5994–6010.
5 Metal–Organic Frameworks in Heterogeneous Catalysis: Recent Progress, New Trends, and Future Perspectives | Chemical Reviews, https://pubs.acs.org/doi/10.1021/acs.chemrev.9b00685, (accessed July 31, 2025).
6 J. F. Olorunyomi, S. T. Geh, R. A. Caruso and C. M. Doherty, *Mater. Horiz.*, 2021, **8**, 2387–2419.
7 D. S. R. Khafaga, M. T. El-Morsy, H. Faried, A. H. Diab, S. Shehab, A. M. Saleh and G. A. M. Ali, *RSC Adv.*, 2024, **14**, 30201–30229.
8 S. K. Firooz and D. W. Armstrong, *Anal. Chim. Acta*, 2022, **1234**, 340208.
9 M. O'Keeffe, M. A. Peskov, S. J. Ramsden and O. M. Yaghi, *Acc. Chem. Res.*, 2008, **41**, 1782–1789.
10 S. Lee, B. Kim, H. Cho, H. Lee, S. Y. Lee, E. S. Cho and J. Kim, *ACS Appl. Mater. Interfaces*, 2021, **13**, 23647–23654.
11 Y. J. Colón and R. Q. Snurr, *Chem. Soc. Rev.*, 2014, **43**, 5735–5749.
12 M. Fernandez, T. K. Woo, C. E. Wilmer and R. Q. Snurr, *J. Phys. Chem. C*, 2013, **117**, 7681–7689.
13 M. Fernandez, N. R. Trefiak and T. K. Woo, *J. Phys. Chem. C*, 2013, **117**, 14095–14105.
14 C. M. Simon, R. Mercado, S. K. Schnell, B. Smit and M. Haranczyk, *Chem. Mater.*, 2015, **27**, 4459–4475.
15 C. M. Simon, J. Kim, D. A. Gomez-Gualdron, J. S. Camp, Y. G. Chung, R. L. Martin, R. Mercado, M. W. Deem, D. Gunter, M. Haranczyk, D. S. Sholl, R. Q. Snurr and B. Smit, *Energy Environ. Sci.*, 2015, **8**, 1190–1199.
16 I.-T. Sung, Y.-H. Cheng, C.-M. Hsieh and L.-C. Lin, *Ind. Eng. Chem. Res.*, 2025, **64**, 1859–1875.
17 C. Altintas, O. F. Altundal, S. Keskin and R. Yildirim, *J. Chem. Inf. Model.*, 2021, **61**, 2131–2146.
18 G. S. Fanourgakis, K. Gkagkas, E. Tylianakis and G. E. Froudakis, *J. Am. Chem. Soc.*, 2020, **142**, 3814–3822.
19 G. S. Fanourgakis, K. Gkagkas, E. Tylianakis and G. Froudakis, *J. Phys. Chem. C*, 2020, **124**, 7117–7126.
20 I.-T. Sung and L.-C. Lin, *J. Phys. Chem. C*, 2023, **127**, 13886–13899.
21 R. Anderson, J. Rodgers, E. Argueta, A. Biong and D. A. Gómez-Gualdrón, *Chem. Mater.*, 2018, **30**, 6325–6337.
22 I. B. Orhan, T. C. Le, R. Babarao and A. W. Thornton, *Commun. Chem.*, 2023, **6**, 214.
23 T. Bailey, A. Jackson, R.-A. Berbece, K. Wu, N. Hondow and E. Martin, *J. Chem. Inf. Model.*, 2023, **63**, 4545–4551.
24 J. P. Janet and H. J. Kulik, *J. Phys. Chem. A*, 2017, **121**, 8939–8954.
25 R. Anderson, A. Biong and D. A. Gómez-Gualdrón, *J. Chem. Theory Comput.*, 2020, **16**, 1271–1283.
26 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, *Nat. Commun.*, 2023, **14**, 7043.
27 J. Burner, L. Schwiedrzik, M. Krykunov, J. Luo, P. G. Boyd and T. K. Woo, *J. Phys. Chem. C*, 2020, **124**, 27996–28005.
28 H. Dureckova, M. Krykunov, M. Z. Aghaji and T. K. Woo, *J. Phys. Chem. C*, 2019, **123**, 4133–4139.
29 S. M. Moosavi, A. Nandy, K. M. Jablonka, D. Ongari, J. P. Janet, P. G. Boyd, Y. Lee, B. Smit and H. J. Kulik, *Nat. Commun.*, 2020, **11**, 4068.
30 A. P. Sarikas, K. Gkagkas and G. E. Froudakis, *Sci. Rep.*, 2024, **14**, 2242.
31 T.-H. Hung, Z.-X. Xu, D.-Y. Kang and L.-C. Lin, *J. Phys. Chem. C*, 2022, **126**, 2813–2822.
32 A. P. Sarikas, K. Gkagkas and G. E. Froudakis, *Sci. Rep.*, 2024, **14**, 27360.
33 C. R. Qi, H. Su, K. Mo and L. J. Guibas, *arXiv*, 2017, preprint, arXiv:1612.00593, DOI: 10.48550/arXiv.1612.00593.
34 K. Choudhary, T. Yildirim, D. W. Siderius, A. G. Kusne, A. McDannald and D. L. Ortiz-Montalvo, *Comput. Mater. Sci.*, 2022, **210**, 111388.
35 A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser and I. Polosukhin, *arXiv*, 2023, preprint, arXiv:1706.03762, DOI: 10.48550/arXiv.1706.03762.
36 Y. Yao, L. Rosasco and A. Caponnetto, *Constr. Approx.*, 2007, **26**, 289–315.
37 L. Yang and A. Shami, *Neurocomputing*, 2020, **415**, 295–316.
38 Y. Tian and Y. Zhang, *Inf. Fusion*, 2022, **80**, 146–166.
39 J. Cai, J. Luo, S. Wang and S. Yang, *Neurocomputing*, 2018, **300**, 70–79.
40 R. Ma, Y. J. Colón and T. Luo, *ACS Appl. Mater. Interfaces*, 2020, **12**, 34041–34048.
41 G. M. Cooper and Y. J. Colón, *Mol. Syst. Des. Eng.*, 2023, **8**, 1049–1059.
42 Z. Cao, R. Magar, Y. Wang and A. Barati Farimani, *J. Am. Chem. Soc.*, 2023, **145**, 2958–2967.
43 S. Han, B. G. Lee, D.-W. Lim and J. Kim, *Chem. Mater.*, 2024, **36**, 11280–11287.
44 S. Zhang, Z. Wang, H. Gao and T. Zhou, *Ind. Eng. Chem. Res.*, 2025, **64**, 14576–14589.
45 Interpretable Graph Transformer Network for Predicting Adsorption Isotherms of Metal–Organic Frameworks | Journal of Chemical Information and Modeling, https://pubs.acs.org/doi/10.1021/acs.jcim.2c00876, (accessed July 28, 2025).
46 Y. Kang, H. Park, B. Smit and J. Kim, *Nat. Mach. Intell.*, 2023, **5**, 309–318.
47 J. Wang, J. Liu, H. Wang, M. Zhou, G. Ke, L. Zhang, J. Wu, Z. Gao and D. Lu, *Nat. Commun.*, 2024, **15**, 1904.
48 J. Mao, N. Jiang, A. Darù, A. S. Filatov, J. E. Burch, J. Hofmann, S. M. Vornholt, K. W. Chapman, J. S. Anderson and A. L. Ferguson, *J. Am. Chem. Soc.*, 2025, **147**, 17651–17667.

49 E. Osaro, F. Fajardo-Rojas, G. M. Cooper, D. Gómez-Gualdrón and Y. J. Colón, *Chem. Sci.*, 2024, **15**, 17671–17684.
50 A. Jose, E. Devijver, N. Jakse and R. Poloni, *J. Am. Chem. Soc.*, 2024, **146**, 6134–6144.
51 K. Mukherjee, A. W. Dowling and Y. J. Colón, *Mol. Syst. Des. Eng.*, 2022, **7**, 248–259.
52 K. Mukherjee, E. Osaro and Y. J. Colón, *Digital Discovery*, 2023, **2**, 1506–1521.
53 M. Aqib, V. Daoo and J. K. Singh, *Energy Fuels*, 2024, **38**, 9381–9394.
54 S. Thaler, F. Mayr, S. Thomas, A. Gagliardi and J. Zavadlav, *npj Comput. Mater.*, 2024, **10**, 86.
55 C. J. Leverant, J. Cooper, D. F. Sava Gallis and J. A. Harvey, *J. Phys. Chem. C*, 2024, **128**, 16250–16257.
56 T. D. Pham and R. Q. Snurr, *Langmuir*, 2025, **41**, 4585–4593.
57 A. Deshwal, C. M. Simon and J. R. Doppa, *Mol. Syst. Des. Eng.*, 2021, **6**, 1066–1086.
58 S. Badrinarayanan, R. Magar, A. Antony, R. S. Meda and A. B. Farimani, *arXiv*, 2025, preprint, arXiv:2506.00198, DOI: 10.48550/arXiv.2506.00198.
59 Y. Bao, R. L. Martin, C. M. Simon, M. Haranczyk, B. Smit and M. W. Deem, *J. Phys. Chem. C*, 2015, **119**, 186–195.
60 S. P. Collins, T. D. Daff, S. S. Piotrkowski and T. K. Woo, *Sci. Adv.*, 2016, **2**, e1600954.
61 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, *Sci. Adv.*, 2016, **2**, e1600909.
62 S. P. Collins, T. D. Daff, S. S. Piotrkowski and T. K. Woo, *Sci. Adv.*, 2016, **2**, e1600954.
63 J. Snoek, H. Larochelle and R. P. Adams, *arXiv*, 2012, preprint, arXiv:1206.2944, DOI: 10.48550/arXiv.1206.2944.
64 E. Taw and J. B. Neaton, *Adv. Theory Simul.*, 2022, **5**, 2100515.
65 Y. Comlek, T. D. Pham, R. Q. Snurr and W. Chen, *npj Comput. Mater.*, 2023, **9**, 170.
66 T.-W. Liu, Q. Nguyen, A. B. Dieng and D. A. Gómez-Gualdrón, *Chem. Sci.*, 2024, **15**, 18903–18919.
67 V. Daoo and J. K. Singh, *ACS Appl. Mater. Interfaces*, 2024, **16**, 6971–6987.
68 N. Gantzler, A. Deshwal, J. R. Doppa and C. M. Simon, *Digital Discovery*, 2023, **2**, 1937–1956.
69 N. Gantzler, A. Deshwal, J. R. Doppa and C. M. Simon, *Digital Discovery*, 2023, **2**, 1937–1956.
70 T.-W. Liu, F. Fajardo-Rojas, S. Addish, E. Martinez and D. A. Gomez-Gualdron, *ACS Appl. Mater. Interfaces*, 2024, **16**, 68506–68519.
71 X. Zhang, K. Zhang, H. Yoo and Y. Lee, *ACS Sustainable Chem. Eng.*, 2021, **9**, 2872–2879.
72 H. Park, S. Majumdar, X. Zhang, J. Kim and B. Smit, *Digital Discovery*, 2024, **3**, 728–741.
73 C. E. Wilmer, M. Leaf, C. Y. Lee, O. K. Farha, B. G. Hauser, J. T. Hupp and R. Q. Snurr, *Nat. Chem.*, 2012, **4**, 83–89.
74 R. Anderson and D. A. Gómez-Gualdrón, *CrystEngComm*, 2019, **21**, 1653–1665.
75 P. G. Boyd and T. K. Woo, *CrystEngComm*, 2016, **18**, 3777–3792.
76 A. N. Rubungo, F. Fajardo-Rojas, D. Gómez-Gualdrón and A. B. Dieng, *ChemRxiv*, 2025, preprint, DOI: 10.26434/chemrxiv-2025-93xmj.
77 N. Kim, S. Kim, M. Kim, J. Park and S. Ahn, *arXiv*, 2025, preprint, arXiv:2410.17270, DOI: 10.48550/arXiv.2410.17270.
78 S. Majumdar, S. M. Moosavi, K. M. Jablonka, D. Ongari and B. Smit, *ACS Appl. Mater. Interfaces*, 2021, **13**, 61004–61014.
79 M. Gibaldi, O. Kwon, A. White, J. Burner and T. K. Woo, *ACS Appl. Mater. Interfaces*, 2022, **14**, 43372–43386.
80 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, *Chem. Mater.*, 2014, **26**, 6185–6192.
81 G. Zhao, L. M. Brabson, S. Chheda, J. Huang, H. Kim, K. Liu, K. Mochida, T. D. Pham, Prerna, G. G. Terrones, S. Yoon, L. Zoubritzky, F.-X. Coudert, M. Haranczyk, H. J. Kulik, S. M. Moosavi, D. S. Sholl, J. I. Siepmann, R. Q. Snurr and Y. G. Chung, *Matter*, 2025, **8**, 102140.
82 P. Z. Moghadam, A. Li, S. B. Wiggin, A. Tao, A. G. P. Maloney, P. A. Wood, S. C. Ward and D. Fairen-Jimenez, *Chem. Mater.*, 2017, **29**, 2618–2625.
83 A. Li, R. B. Perez, S. Wiggin, S. C. Ward, P. A. Wood and D. Fairen-Jimenez, *Matter*, 2021, **4**, 1105–1106.
84 A. Nandy, S. Yue, C. Oh, C. Duan, G. G. Terrones, Y. G. Chung and H. J. Kulik, *Matter*, 2023, **6**, 1585–1603.
85 B. J. Sikora, C. E. Wilmer, M. L. Greenfield and R. Q. Snurr, *Chem. Sci.*, 2012, **3**, 2217–2223.
86 C. E. Wilmer, O. K. Farha, Y.-S. Bae, J. T. Hupp and R. Q. Snurr, *Energy Environ. Sci.*, 2012, **5**, 9849–9856.
87 J. Burner, J. Luo, A. White, A. Mirmiran, O. Kwon, P. G. Boyd, S. Maley, M. Gibaldi, S. Simrod, V. Ogden and T. K. Woo, *Chem. Mater.*, 2023, **35**, 900–916.
88 D. Nazarian, J. S. Camp and D. S. Sholl, *Chem. Mater.*, 2016, **28**, 785–793.
89 M. Gibaldi, A. Kapeliukha, A. White, J. Luo, R. A. Mayo, J. Burner and T. K. Woo, *Chem. Sci.*, 2025, **16**, 4085–4100.
90 A. S. Rosen, S. M. Iyer, D. Ray, Z. Yao, A. Aspuru-Guzik, L. Gagliardi, J. M. Notestein and R. Q. Snurr, *Matter*, 2021, **4**, 1578–1597.
91 Daniel Siderius (2016), NIST/ARPA-E Database of Novel and Emerging Adsorbent Materials – SRD 205, National Institute of Standards and Technology, https://adsorption.nist.gov/srd205/index.php (Accessed 2025-07-28), 2020.
92 P. Iacomi and P. L. Llewellyn, *Chem. Mater.*, 2020, **32**, 982–991.
93 A. Nandy, G. Terrones, N. Arunachalam, C. Duan, D. W. Kastner and H. J. Kulik, *Sci. Data*, 2022, **9**, 74.
94 N. C. Burtch, H. Jasuja and K. S. Walton, *Chem. Rev.*, 2014, 114(20), 10575–10612.
95 A. Nandy, C. Duan and H. J. Kulik, *J. Am. Chem. Soc.*, 2021, **143**, 17535–17547.
96 S. Park, B. Kim, S. Choi, P. G. Boyd, B. Smit and J. Kim, *J. Chem. Inf. Model.*, 2018, **58**, 244–251.
97 M. C. Swain and J. M. Cole, *J. Chem. Inf. Model.*, 2016, **56**, 1894–1904.

98 L. T. Glasby, K. Gubsch, R. Bence, R. Oktavian, K. Isoko, S. M. Moosavi, J. L. Cordiner, J. C. Cole and P. Z. Moghadam, *Chem. Mater.*, 2023, **35**, 4510–4524.
99 G. G. Terrones, S.-P. Huang, M. P. Rivera, S. Yue, A. Hernandez and H. J. Kulik, *J. Am. Chem. Soc.*, 2024, **146**, 20333–20348.
100 J. R. H. Manning and L. Sarkisov, *Digital Discovery*, 2023, **2**, 1783–1796.
101 P. Qi, Y. Zhang, Y. Zhang, J. Bolton and C. D. Manning, *arXiv*, 2020, preprint, arXiv:2003.07082, DOI: 10.48550/arXiv.2003.07082.
102 H. Park, Y. Kang, W. Choe and J. Kim, *J. Chem. Inf. Model.*, 2022, **62**, 1190–1198.
103 I. Beltagy, K. Lo and A. Cohan, *arXiv*, 2019, preprint, arXiv:1903.10676, DOI: 10.48550/arXiv.1903.10676.
104 Y. Luo, S. Bag, O. Zaremba, A. Cierpka, J. Andreo, S. Wuttke, P. Friederich and M. Tsotsalas, *Angew. Chem., Int. Ed.*, 2022, **61**, e202200242.
105 Z. Zheng, O. Zhang, C. Borgs, J. T. Chayes and O. M. Yaghi, *J. Am. Chem. Soc.*, 2023, **145**, 18048–18062.
106 Y. Kang, W. Lee, T. Bae, S. Han, H. Jang and J. Kim, *J. Am. Chem. Soc.*, 2025, **147**, 3943–3958.
107 S. Velioglu and S. Keskin, *Mater. Adv.*, 2020, **1**, 341–353.
108 T. Chen and T. A. Manz, *RSC Adv.*, 2020, **10**, 26944–26951.
109 A. J. White, M. Gibaldi, J. Burner, R. A. Mayo and T. K. Woo, *J. Am. Chem. Soc.*, 2025, **147**, 17579–17583.
110 L.-C. Lin, K. Lee, L. Gagliardi, J. B. Neaton and B. Smit, *J. Chem. Theory Comput.*, 2014, **10**, 1477–1488.
111 E. Haldoupis, J. Borycz, H. Shi, K. D. Vogiatzis, P. Bai, W. L. Queen, L. Gagliardi and J. I. Siepmann, *J. Phys. Chem. C*, 2015, **119**, 16058–16071.
112 D. Dubbeldam, K. S. Walton, T. J. H. Vlugt and S. Calero, *Adv. Theory Simul.*, 2019, **2**, 1900135.
113 K. S. Walton and R. Q. Snurr, *J. Am. Chem. Soc.*, 2007, **129**, 8552–8556.
114 J. Park, J. D. Howe and D. S. Sholl, *Chem. Mater.*, 2017, **29**, 10487–10495.
115 C. McCready, K. Sladekova, S. Conroy, J. R. B. Gomes, A. J. Fletcher and M. Jorge, *J. Chem. Theory Comput.*, 2024, **20**, 4869–4884.
116 D. Ongari, L. Talirz, K. M. Jablonka, D. W. Siderius and B. Smit, *J. Chem. Eng. Data*, 2022, **67**, 1743–1756.
117 B. J. Bucior, N. S. Bobbitt, T. Islamoglu, S. Goswami, A. Gopalan, T. Yildirim, O. K. Farha, N. Bagheri and R. Q. Snurr, *Mol. Syst. Des. Eng.*, 2019, **4**, 162–174.
118 Z. Li, B. J. Bucior, H. Chen, M. Haranczyk, J. I. Siepmann and R. Q. Snurr, *J. Chem. Phys.*, 2021, **155**, 014701.
119 K. Shi, Z. Li, D. M. Anstine, D. Tang, C. M. Colina, D. S. Sholl, J. I. Siepmann and R. Q. Snurr, *J. Chem. Theory Comput.*, 2023, **19**, 4568–4583.
120 E. H. Cho and L.-C. Lin, *J. Phys. Chem. Lett.*, 2021, **12**, 2279–2285.
121 T. D. Burns, K. N. Pai, S. G. Subraveti, S. P. Collins, M. Krykunov, A. Rajendran and T. K. Woo, *Environ. Sci. Technol.*, 2020, **54**, 4536–4544.
122 G. S. Fanourgakis, K. Gkagkas and G. Froudakis, *J. Chem. Phys.*, 2022, **156**, 054103.
123 P. Li, N. A. Vermeulen, C. D. Malliakas, D. A. Gómez-Gualdrón, A. J. Howarth, B. L. Mehdi, A. Dohnalkova, N. D. Browning, M. O'Keeffe and O. K. Farha, *Science*, 2017, **356**, 624–627.
124 V. Korolev and A. Mitrofanov, *J. Chem. Inf. Model.*, 2024, **64**, 1919–1931.
125 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, *Nat. Mach. Intell.*, 2021, **3**, 76–86.
126 H. Park, X. Yan, R. Zhu, E. A. Huerta, S. Chaudhuri, D. Cooper, I. Foster and E. Tajkhorshid, *Commun. Chem.*, 2024, **7**, 21.
127 C. Duan, A. Nandy, S. Liu, Y. Du, L. He, Y. Qu, H. Jia and J.-H. Dou, *arXiv*, 2025, preprint, arXiv:2505.08531, DOI: 10.48550/arXiv.2505.08531.
128 X. Fu, T. Xie, A. S. Rosen, T. Jaakkola and J. Smith, *arXiv*, 2023, preprint, arXiv:2310.10732, DOI: 10.48550/arXiv.2310.10732.
129 J. Park, Y. Lee and J. Kim, *Nat. Commun.*, 2025, **16**, 34.
130 C. Cleeton and L. Sarkisov, *Nat. Commun.*, 2025, **16**, 4806.
131 O. T. Unke, S. Chmiela, H. E. Sauceda, M. Gastegger, I. Poltavsky, K. T. Schütt, A. Tkatchenko and K.-R. Müller, *Chem. Rev.*, 2021, **121**, 10142–10186.
132 S. M. Moosavi, B. Á. Novotny, D. Ongari, E. Moubarak, M. Asgari, Ö. Kadioglu, C. Charalambous, A. Ortega-Guerrero, A. H. Farmahini, L. Sarkisov, S. Garcia, F. Noé and B. Smit, *Nat. Mater.*, 2022, **21**, 1419–1425.
133 J. Behler and M. Parrinello, *Phys. Rev. Lett.*, 2007, **98**, 146401.
134 A. P. Bartók, M. C. Payne, R. Kondor and G. Csányi, *Phys. Rev. Lett.*, 2010, **104**, 136403.
135 A. V. Shapeev, *Multiscale Model. Simul.*, 2016, **14**, 1153–1173.
136 A. C. T. van Duin, S. Dasgupta, F. Lorant and W. A. Goddard, *J. Phys. Chem. A*, 2001, **105**, 9396–9409.
137 L. Zhang, J. Han, H. Wang, R. Car and W. E, *Phys. Rev. Lett.*, 2018, **120**, 143001.
138 S. Vandenhaute, M. Cools-Ceuppens, S. DeKeyser, T. Verstraelen and V. Van Speybroeck, *npj Comput. Mater.*, 2023, **9**, 1–8.
139 A. M. Elena, P. D. Kamath, T. Jaffrelot Inizan, A. S. Rosen, F. Zanca and K. A. Persson, *npj Comput. Mater.*, 2025, **11**, 125.
140 M. Eckhoff and J. Behler, *J. Chem. Theory Comput.*, 2019, **15**, 3793–3809.
141 J. P. Dürholt, B. F. Jahromi and R. Schmid, *ACS Cent. Sci.*, 2019, **5**, 1440–1448.
142 P. Zhao, S. C. E. Tsang and D. Fairen-Jimenez, *Cell Rep. Phys. Sci.*, 2021, **2**, 100544.
143 A. Sharma and S. Sanvito, *npj Comput. Mater.*, 2024, **10**, 237.
144 S. Wieser and E. Zojer, *npj Comput. Mater.*, 2024, **10**, 18.

145 N. Castel, D. André, C. Edwards, J. D. Evans and F.-X. Coudert, *Digital Discovery*, 2024, **3**, 355–368.
146 O. Tayfuroglu and S. Keskin, *ChemRxiv*, 2025, preprint, DOI: 10.26434/chemrxiv-2025-c85xt.
147 Y. Shi and F. A. Shakib, *ChemRxiv*, 2024, preprint, DOI: 10.26434/chemrxiv-2024-9hmsb.
148 M. Thürlemann and S. Riniker, *Chem. Sci.*, 2023, **14**, 12661–12675.
149 V. V. Korolev, Y. M. Nevolin, T. A. Manz and P. V. Protsenko, *J. Chem. Inf. Model.*, 2021, **61**, 5774–5784.
150 S. Liu, R. Dupuis, D. Fan, S. Benzaria, M. Bonneau, P. Bhatt, M. Eddaoudi and G. Maurin, *Chem. Sci.*, 2024, **15**, 5294–5302.
151 F. L. Oliveira, C. Cleeton, R. Neumann Barros Ferreira, B. Luan, A. H. Farmahini, L. Sarkisov and M. Steiner, *Sci. Data*, 2023, **10**, 230.
152 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, *J. Chem. Eng. Data*, 2023, **68**, 483–498.
153 A. Sriram, S. Choi, X. Yu, L. M. Brabson, A. Das, Z. Ulissi, M. Uyttendaele, A. J. Medford and D. S. Sholl, *ACS Cent. Sci.*, 2024, **10**, 923–941.
154 A. J. Howarth, Y. Liu, P. Li, Z. Li, T. C. Wang, J. T. Hupp and O. K. Farha, *Nat. Rev. Mater.*, 2016, **1**, 15018.
155 H. U. Escobar-Hernandez, L. M. Pérez, P. Hu, F. A. Soto, M. I. Papadaki, H. C. Zhou and Q. Wang, *Ind. Eng. Chem. Res.*, 2022, **61**, 5853–5862.
156 P. Z. Moghadam, S. M. J. Rogge, A. Li, C.-M. Chow, J. Wieme, N. Moharrami, M. Aragones-Anglada, G. Conduit, D. A. Gomez-Gualdron, V. Van Speybroeck and D. Fairen-Jimenez, *Matter*, 2019, **1**, 219–234.
157 B. Mu and K. S. Walton, *J. Phys. Chem. C*, 2011, **115**, 22748–22754.
158 J. Huang, X. Xia, X. Hu, S. Li and K. Liu, *Int. J. Heat Mass Transfer*, 2019, **138**, 11–16.
159 M. Islamov, H. Babaei, R. Anderson, K. B. Sezginel, J. R. Long, A. J. H. McGaughey, D. A. Gomez-Gualdron and C. E. Wilmer, *npj Comput. Mater.*, 2023, **9**, 11.
160 R. Anderson and D. A. Gómez-Gualdrón, *Chem. Mater.*, 2020, **32**, 8106–8119.
161 K. Li, J. Wang and H. Wang, *J. Mater. Chem. A*, 2024, **12**, 14245–14267.
162 L. S. Xie, G. Skorupskii and M. Dincă, *Chem. Rev.*, 2020, **120**, 8536–8580.
163 D.-W. Lim and H. Kitagawa, *Chem. Rev.*, 2020, **120**, 8416–8467.
164 R. Anderson, B. Schweitzer, T. Wu, M. A. Carreon and D. A. Gómez-Gualdrón, *ACS Appl. Mater. Interfaces*, 2018, **10**, 582–592.
165 S. R. G. Balestra and R. Semino, *J. Chem. Phys.*, 2022, **157**, 184502.
166 N. T. T. Nguyen, T. T. T. Nguyen, S. Ge, R. K. Liew, D. T. C. Nguyen and T. V. Tran, *Nanoscale Adv.*, 2024, **6**, 1800–1821.
167 E. Ren and F.-X. Coudert, *J. Phys. Chem. C*, 2024, **128**, 6917–6926.
168 Y. Yang, Z. Yu and D. S. Sholl, *Chem. Mater.*, 2023, **35**, 10156–10168.
169 H. Daglar and S. Keskin, *ACS Appl. Mater. Interfaces*, 2022, **14**, 32134–32148.
170 Q. Wei, J. Qiu, R. Wang, Z. Sun, Y. Xie, Y. Chen and Y. Wan, *J. Phys. Chem. C*, 2025, **129**, 7126–7133.
171 Z. Zhang, C. Zhang, Y. Zhang, S. Deng, Y.-F. Yang, A. Su and Y.-B. She, *RSC Adv.*, 2023, **13**, 16952–16962.
172 A. S. Rosen, V. Fung, P. Huck, C. T. O'Donnell, M. K. Horton, D. G. Truhlar, K. A. Persson, J. M. Notestein and R. Q. Snurr, *npj Comput. Mater.*, 2022, **8**, 112.
173 R. Batra, C. Chen, T. G. Evans, K. S. Walton and R. Ramprasad, *Nat. Mach. Intell.*, 2020, **2**, 704–710.
174 I. V. Dudakov, S. A. Savelev, I. M. Nevolin, A. A. Mitrofanov, V. V. Korolev and Y. G. Gorbunova, *Phys. Chem. Chem. Phys.*, 2025, **27**, 6850–6857.
175 N. T. T. Ha, H. T. Thao and N. N. Ha, *J. Mol. Graphics Modell.*, 2022, **112**, 108124.
176 K. Tan, S. Zuluaga, Q. Gong, P. Canepa, H. Wang, J. Li, Y. J. Chabal and T. Thonhauser, *Chem. Mater.*, 2014, **26**, 6886–6895.
177 C.-H. Ho, M. L. Valentine, Z. Chen, H. Xie, O. Farha, W. Xiong and F. Paesani, *Commun. Chem.*, 2023, **6**, 70.
178 J. K. Bristow, D. Tiana and A. Walsh, *J. Chem. Theory Comput.*, 2014, **10**, 4644–4652.
179 T. M. Becker, L.-C. Lin, D. Dubbeldam and T. J. H. Vlugt, *J. Phys. Chem. C*, 2018, **122**, 24488–24498.
180 L. Vanduyfhuys, T. Verstraelen, M. Vandichel, M. Waroquier and V. Van Speybroeck, *J. Chem. Theory Comput.*, 2012, **8**, 3217–3231.
181 S. Chen, Z. Zhang, W. Chen, B. E. G. Lucier, M. Chen, W. Zhang, H. Zhu, I. Hung, A. Zheng, Z. Gan, D. Lei and Y. Huang, *Nat. Commun.*, 2024, **15**, 10776.
182 X. Liu, X. Wang and F. Kapteijn, *Chem. Rev.*, 2020, **120**, 8303–8377.
183 B. Mazur, L. Firlej and B. Kuchta, *ACS Appl. Mater. Interfaces*, 2024, **16**, 25559–25567.
184 D. W. Siderius, H. W. Hatch and V. K. Shen, *J. Phys. Chem. B*, 2024, **128**, 4830–4845.
185 H. Zhang and R. Q. Snurr, *J. Phys. Chem. C*, 2017, **121**, 24000–24010.
186 M. Aghajani Hashjin, S. Zarshad, H. B. Motejadded Emrooz and S. Sadeghzadeh, *Sci. Rep.*, 2023, **13**, 16983.
187 W. Xu and O. M. Yaghi, *ACS Cent. Sci.*, 2020, **6**, 1348–1354.
188 H. Kim, S. Yang, S. R. Rao, S. Narayanan, E. A. Kapustin, H. Furukawa, A. S. Umans, O. M. Yaghi and E. N. Wang, *Science*, 2017, **356**, 430–434.
189 A. Niyongabo Rubungo, C. Arnold, B. P. Rand and A. B. Dieng, *npj Comput. Mater.*, 2025, **11**, 186.
190 X. Wu and J. Jiang, *J. Mater. Chem. A*, 2025, **13**, 19307–19315.
191 Z. Zheng, A. H. Alawadhi, S. Chheda, S. E. Neumann, N. Rampal, S. Liu, H. L. Nguyen, Y. Lin, Z. Rong,

J. I. Siepmann, L. Gagliardi, A. Anandkumar, C. Borgs, J. T. Chayes and O. M. Yaghi, *J. Am. Chem. Soc.*, 2023, **145**, 28284–28295.

192 X. Zhang, K. M. Jablonka and B. Smit, *Digital Discovery*, 2024, **3**, 1410–1420.

193 A. Sturluson, A. Raza, G. D. McConachie, D. W. Siderius, X. Z. Fern and C. M. Simon, *Chem. Mater.*, 2021, **33**, 7203–7216.

194 X. Zhang, S. Sethi, Z. Wang, T. Zhou, Z. Qi and K. Sundmacher, *Chem. Eng. Sci.*, 2022, **259**, 117801.

195 Y. Kang and J. Kim, *Nat. Commun.*, 2024, **15**, 4705.

196 Z. Zheng, Z. Rong, N. Rampal, C. Borgs, J. T. Chayes and O. M. Yaghi, *Angew. Chem., Int. Ed.*, 2023, **62**, e202311983.

197 Z. Zheng, O. Zhang, H. L. Nguyen, N. Rampal, A. H. Alawadhi, Z. Rong, T. Head-Gordon, C. Borgs, J. T. Chayes and O. M. Yaghi, *ACS Cent. Sci.*, 2023, **9**, 2161–2170.