LANGMUIR

pubs.acs.org/Langmuir Perspective

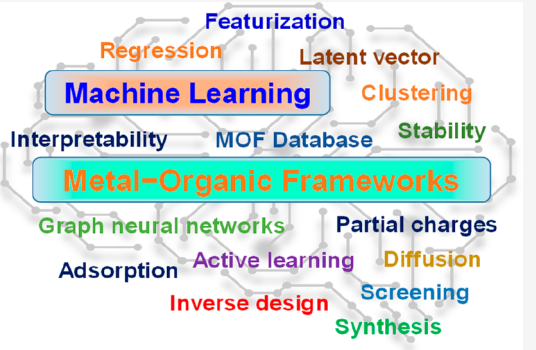
# Leveraging Machine Learning for Metal−Organic Frameworks: A Perspective

Hongjian Tang, Lunbo Duan,* and Jianwen Jiang*

Cite This: *Langmuir* 2023, 39, 15849−15863 Read Online

ACCESS | Metrics & More | Article Recommendations

**ABSTRACT:** Metal−organic frameworks (MOFs) have attracted tremendous interest because of their tunable structures, functionalities, and physiochemical properties. The nearly infinite combinations of metal nodes and organic linkers have led to the synthesis of over 100,000 experimental MOFs and the construction of millions of hypothetical counterparts. It is intractable to identify the best candidates in the immense chemical space of MOFs for applications via conventional trial-to-error experiments or brute-force simulations. Over the past several years, machine learning (ML) has substantially transformed the way of MOF discovery, design, and synthesis. Driven by the abundant data from experiments or simulations, ML can not only efficiently and accurately predict MOF properties but also quantitatively derive structure−property relationships for rational design and screening. In this Perspective, we summarize recent achievements in leveraging ML for MOFs from the aspects of data acquisition, featurization, model training, and applications. Then, current challenges and new opportunities are discussed for the future exploration of ML to accelerate the development of new MOFs in this vibrant field.




## INTRODUCTION

With readily tunable crystalline structures, surface areas, pore sizes, topologies, and chemical functionalities, metal−organic frameworks (MOFs) have received considerable attention for a wide variety of potential applications (e.g., separation, storage, and catalysis).[1] MOFs can be produced through the assembly of metal nodes and organic linkers into extensible topologies, which has resulted in over 100,000 synthesized MOFs.[2] The nearly infinite MOFs cause experimental tests to be technically formidable and economically inviable. With the increase in computational power, high-throughput computational screening (HTCS) has demonstrated great success in establishing structure−property relationships and identifying top MOFs from a database for different applications, particularly gas storage and separation.[3−5] However, brute-force HTCS is not the ultimate solution because exhaustive computations of an entire database are time-consuming and have low efficiency; moreover, the existing database of experimentally synthesized MOFs or in-silico constructed MOFs represents merely a small chemical space and can be easily biased by human intuition.

Over the past several years, machine learning (ML) has brought about a paradigm revolution in scientific development.[6] It transforms traditional ways of thinking, testing, and decision making from physically based to data-driven. Substantial progress has been evidenced with the aid of ML, involving image recognition, natural language processing, genome engineering, and materials discovery. Similarly, we have also witnessed the burgeoning utilization of ML in the field of MOFs.[7−9] For instance, ML has been implemented to guide the synthesis of new MOFs,[10] reveal chemical diversity and similarity across different MOF databases,[11] predict MOF stability,[12] transfer knowledge learned from adsorption to diffusion,[13] and inversely design new MOFs from an unseen chemical space to surpass current MOFs for $CO_2$ capture.[14] In this Perspective, the representative ML studies for MOFs are summarized. Despite the remarkable achievements, limitations and challenges exist in the current ML studies. Thus, we discuss how to move forward and highlight new opportunities that ML can offer to the MOF community.

The general workflow of ML for MOFs is illustrated in Figure 1, involving four critical steps: (a) data acquisition, (b) featurization, (c) model training, and (d) model applications. First, as the basis of ML, a database with MOF structures and properties should be acquired. Second, the structures and sometimes the properties are featurized into machine-readable features via either hand-crafted or data-driven methods. Third, ML models are trained through (i) quantifying the relationships between MOF structures and associated properties, known as

Received: July 13, 2023
Revised: October 16, 2023
Accepted: October 17, 2023
Published: November 3, 2023

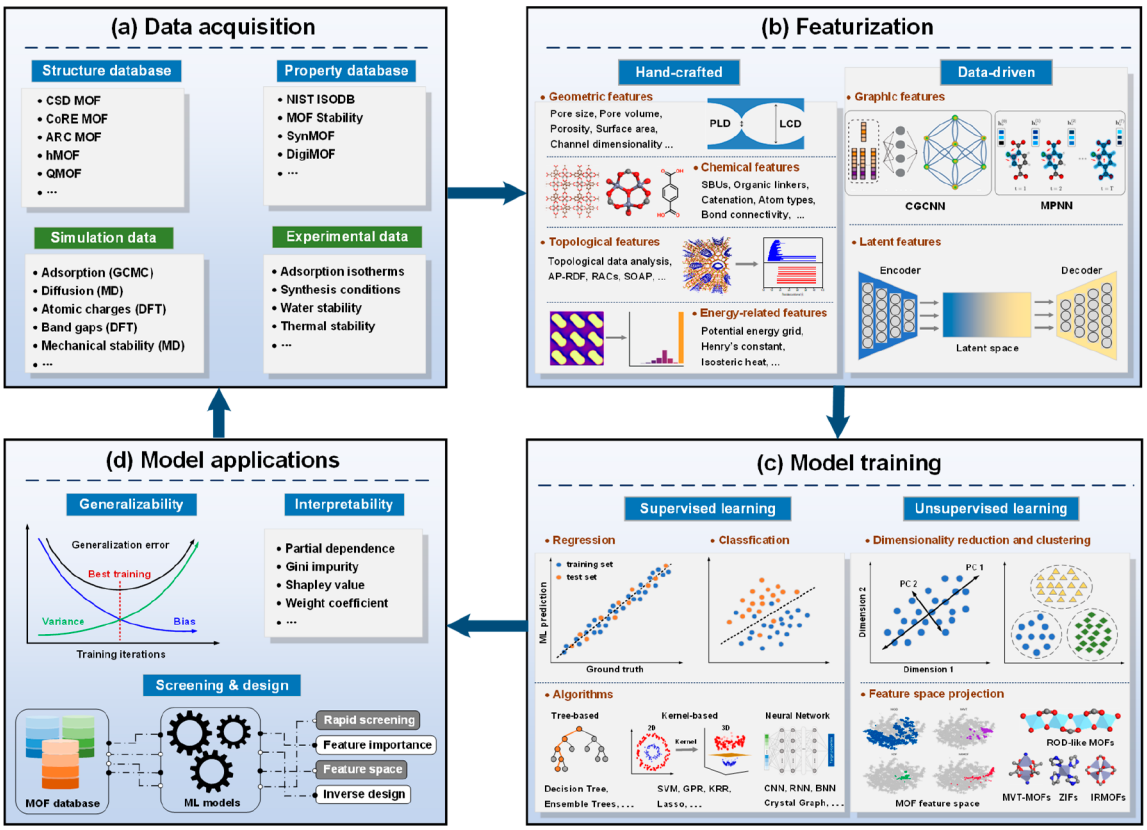



**Figure 1.** General workflow of ML for MOFs. (a) Data acquisition. (b) Featurization (reproduced with permission from ref 15, copyright 2017 Springer Nature; from ref 16, copyright 2019 Royal Society of Chemistry; from ref 47, copyright 2018 American Physical Society; from ref 49, copyright 2020 American Chemical Society). (c) Model training (reproduced with permission from ref 7, copyright 2020 American Chemical Society; from ref 17, copyright 2021 American Chemical Society; from ref 18, copyright 2022 John Wiley and Sons). (d) Model applications.

supervised learning, and (ii) exploring the structural or chemical diversity of MOFs by projecting high-dimensional feature space into a reduced 2D space, known as unsupervised learning. Fourth, the generalizability and interpretability of ML models are applied to the screening and design of MOFs with desired properties. The derived results can be augmented to the existing database to enrich data diversity. Each of the four steps is discussed in detail below.

## MACHINE LEARNING FOR MOFs

**Data Acquisition.** From experimental or computational methods, a massive amount of data has been generated for MOFs including structures and associated properties, as shown in Table 1. The computation-ready experimental (CoRE) MOF database[19] is the most widely used. Its earliest version (CoRE MOF 2014) contains ∼4,700 MOFs that were extracted and curated from the Cambridge Structural Database (CSD),[2] which collects experimentally synthesized MOFs and other materials. In 2019, the CoRE MOF database was updated and expanded to ∼14,000 curated structures.[20] A subset with ∼10,000 MOFs (CSD MOF Collection)[21] was collected from over 100,000 MOFs currently deposited in the CSD MOF repository.[2] Meanwhile, we also witnessed the proliferation of hypothetical MOF databases. The first one is hMOF with ∼137,000 structures in-silico constructed by assembling metal nodes and organic linkers via a bottom-up approach called "Tinkertoy".[22] Despite the enormous number of structures, the hMOF database lacks chemical diversity as only six topologies exist with "pcu" topology as the dominant one. By contrast, the CoRE MOF database contains over 350 unique topologies.[20] Alternatively, Boyd et al. proposed a top-down approach to construct ∼325,000 BW-DB MOFs by exploring different topologies.[23] Based on a similar top-down strategy, hypothetical MOF databases such as ToBaCCo[24] and Diverse ToBaCCo[25] were subsequently developed with an aim to boost the structural and chemical diversity of MOFs. In addition to crystalline structures, these databases also contain geometric information and/or gas adsorption data estimated from simulations and have been widely used in HTCS and ML studies, mainly focused on gas adsorption and separation. Recently, other important properties have also been included in the databases. For example, atomic charges, band gaps, electronic energies, and density of states from density-functional theory (DFT) calculations were available for ∼15,000 MOFs in the quantum MOF (QMOF) database.[26] In the ARC-MOF database, atomic charges and descriptors such as RACs and AP-RDFs were assigned, making it user-friendly for ML.[27] Recently, elastic moduli were computed for nearly 10,000 ultrastable structures in a new hypothetical MOF database.[28]

While the above MOF databases contain simulated or DFT-calculated properties, experimentally measured adsorption, stability, and synthesizability have been collected in different databases. More than 32,000 adsorption isotherms in ∼4,000 synthesized porous materials including MOFs were curated from literature sources and deposited into the NIST isotherm database (ISODB),[29] thus enabling the screening of adsorbents for gas separation. By text mining 3,809 published papers, the thermal stability and solvent-removal stability were compiled for


https://doi.org/10.1021/acs.langmuir.3c01964
Langmuir 2023, 39, 15849−15863

**Table 1. MOF Databases**

| Database | Number of structures | Source | Associated properties | Reference |
|---|---|---|---|---|
| | | With computational properties | | |
| CoRE MOF 2014 | ~4,700 | Experimental | Porosity | 19 |
| CoRE MOF 2019 | ~14,000 | Experimental | Geometric information | 20 |
| CSD MOF subset | ~10,000 | Experimental | Geometric information | 21 |
| hMOF | ~137,000 | Hypothetical | Geometric information, adsorption data of $CO_2$, $N_2$, $CH_4$, $H_2$, Xe, and Kr | 22 |
| BW-DB | ~32,5000 | Hypothetical | Geometric information, adsorption data of $CO_2$ and $N_2$ | 23 |
| ToBaCCo | ~13,000 | Hypothetical | – | 24 |
| Diverse ToBaCCo | ~20,000 | Hypothetical | Porosity, RACs,[a] adsorption data of $CO_2$, $N_2$, and $H_2$ | 25 |
| QMOF | ~15,000 | Experimental | Atomic charges, band gaps, charge densities, density of states | 26 |
| ARC-MOF | ~280,000 | Experimental and hypothetical | Geometric information, atomic charges, RACs, AP-RDFs[b] | 27 |
| Nandy et al. | 50,000 | Hypothetical | Elastic moduli, $CH_4$-deliverable capacities | 28 |
| | | With experimental properties | | |
| NIST ISODB | ~4,000[c] | Experimental | Adsorption isotherms | 29 |
| Nandy et al. | 3132 for thermal stability<br>2179 for solvent-removal stability | Experimental | Stability information | 30 |
| SynMOF | ~900 | Experimental | Synthesis information | 31 |
| DigiMOF | ~15,000 | Experimental | Synthesis information | 32 |

[a]RACs: revised autocorrelation functions. [b]AP-RDFs: atomic property-weighted radial distribution functions. [c]Collected adsorption isotherms in porous adsorbents including zeolites, MOFs, COFs, porous polymers, and carbons.

3,132 and 2,179 CoRE MOFs, respectively.[30] It is formidable to evaluate stability using a computational approach; therefore, the compilation of stability information would facilitate the use of ML to evaluate the MOF stability and benefit the MOF community. MOF synthesizability is another important but largely experience-based topic. Aiming to predict MOF synthesizability, a SynMOF database was constructed by encompassing synthesis information (i.e., temperature, solvent, additive and reaction time) for 983 MOFs.[31] A similar but larger database called DigiMOF was also compiled, which contains 15,501 MOFs and 52,680 pieces of text-mined synthesis information.[32] Such an advance in MOF databases toward synthesizability, though in their infancy, offers an intriguing data-driven approach to comprehending and controlling the sophisticated synthesis of MOFs.

**Featurization.** The structures of MOFs require numerical digitalization by machine-readable features (i.e., descriptors or fingerprints). Featurization should be unique to structures, invariant to any transformation (e.g., translation, rotation, and permutation), and computationally inexpensive. Features are supposed to capture the distinctive structural information on MOFs, also physically and meaningfully correlated with target properties. They are crucial in determining the generalizability and interpretability of the ML models. As summarized in Figure 2, there are three categories of features commonly used for MOFs: (a) hand-crafted descriptors, (b) graph representations, and (c) latent vectors.

*Hand-Crafted Descriptors.* "Hand-crafted" denotes that features are established manually in an experience-oriented way. Consequently, the selection of features requires adequate domain expertise, which is highly tied to a target task. For example, to predict gas adsorption in MOFs, pore geometry is important and would be most reasonable to featurize MOFs with pore size, volume, surface area, and other geometric descriptors. For electronic properties, however, we need to encode the MOFs with atomic-level information. As displayed in Figure 1b, the most common hand-crafted descriptors for MOFs are geometric, chemical, topological, and energy-related.

The geometric descriptors include pore size, volume, surface area, void fraction, channel dimensionality, etc. Several open-source programs such as Zeo++,[34] RASPA,[35] and Poreblazer[36] can be readily used to generate these descriptors. They play a dominant role in gas uptake, particularly at high pressure, whereas they fall short of capturing the holistic pore information in MOFs, neither pore shape nor pore chemistry. In this regard, more comprehensive geometric descriptors are necessary. Lee et al.[15] developed a topological data analysis (TDA) technique to encode complex pore features in a MOF with a persist barcode, which incorporates high-dimensional pore geometry information, spanning from local pore shape to global pore connectivity. The overall similarity of pore shapes among different MOFs can be quantified by comparing the barcodes, thus the TDA has been demonstrated to be a robust descriptor in the discovery of top-performing porous materials including zeolites,[37] porous molecular crystals,[38] and MOFs[39] for gas adsorption. It is worthwhile to note, however, that generating TDA is nontrivial and representing a MOF by a barcode is less intuitive than gross descriptors such as pore size and surface area, thus causing ML models to be less interpretable. Similar drawbacks exist in energy-based descriptors such as the potential energy surface,[16] Boltzmann factor,[40] and Henry's constant,[41] which have been specifically used for gas adsorption in MOFs. These descriptors are computationally expensive to generate and hence are not widely adopted.

In addition to pore geometry, framework chemistry (i.e., metal, linker, and functionality; Figure 2a) is also crucial. The recent ML studies by Jiang and co-workers demonstrated that the incorporation of chemical descriptors including atom types and densities in MOFs would improve the ML prediction accuracy for $C_3H_8/C_3H_6$ separation and water adsorption.[17,42] A commonly used descriptor of MOF chemistry is the revised autocorrelation functions (RACs).[43] The RACs possess two key advantages: (1) decomposing a MOF into subgraphs of respective metal clusters, organic linkers, and functional groups, hence capturing subtle chemical details and (2) considering MOF hierarchy that couples atomic-level properties (i.e., atom


https://doi.org/10.1021/acs.langmuir.3c01964
Langmuir 2023, 39, 15849−15863

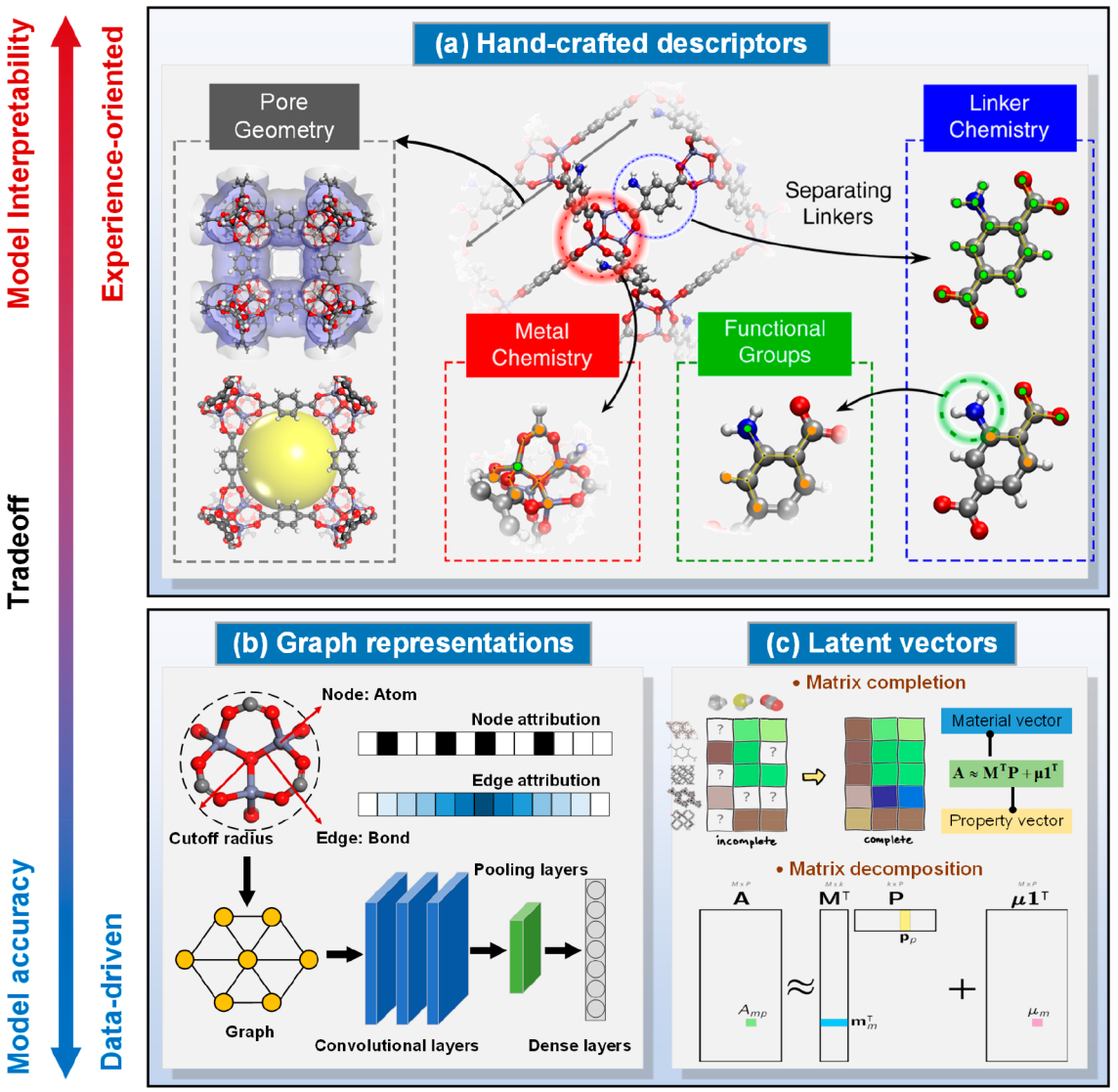



**Figure 2.** Featurization. (a) Hand-crafted descriptors (reproduced with permission from ref 11, copyright 2020 Springer Nature). (b) Graph representations and (c) latent vectors (reproduced with permission from ref 33, copyright 2021 American Chemical Society).

identity, connectivity, Pauling electronegativity, covalent radii, nuclear charge, and polarizability) closely with subgraphs, thus expressing chemical descriptors across the atomic and molecular scales coherently. Consequently, the RACs have been evidenced with high prediction accuracy in ML studies for MOF chemical properties (e.g., chemical diversity and similarity among different MOF databases,[11] thermal stability[30] and synthesizability,[31] colors,[44] and oxidation states of metal centers in MOFs).[45] Similarly, the atomic property-weighted radial distribution functions (AP-RDFs) also capture chemical descriptors hierarchically.[46] By weighting the RDFs with atomic properties (e.g., electronegativity, polarizability, and van der Waals volume), the AP-RDFs concurrently encode the MOF geometry at a crystal level and chemistry at an atomic level. The AP-RDF descriptors have been shown to greatly improve ML prediction for $CH_4$ adsorption in MOFs especially at a low pressure.[46] Batra et al. proposed another hierarchical descriptors for MOFs in which the hierarchy of organic linkers from atomic information to linker size was featurized, in addition to conventional descriptors such as metal properties and molar ratios between metals and organic linkers; the resultant ML models were able to predict the water stability of MOFs with desired accuracy.[12] The success of these hierarchical descriptors suggests that we can include more hierarchies in terms of pore geometry and framework chemistry to develop more interpretable and expressive descriptors.

*Graph Representations.* In contrast to hand-crafted method, data-driven featurization is an automatic way to extract material features, thus bypassing the constraint of human intuition. Molecular graphs in conjunction with deep neural networks constitute one of the most popular data-driven encoding architectures. The crystal structures of MOFs can be approximately represented by crystal graphs, which comprise atomic properties as well as bonding environments. The crystal graph convolutional neural network (CGCNN)[47] is a state-of-the-art data-driven featurization method specifically for periodic structures such as MOFs. As illustrated in Figure 2b, the atoms and bonds in an MOF are represented by nodes and edges. For each atom, a user-defined cutoff radius is imposed to control the maximal bound of the neighboring environment. Representation is then conducted on each atom with a convolutional neural network resolving its local environment iteratively. By


https://doi.org/10.1021/acs.langmuir.3c01964
Langmuir 2023, 39, 15849−15863

optimizing CGCNN hyperparameters through training data, the representation is learned automatically, together with mapping from structures to properties. Different from hand-crafted featurization, CGCNN works in an end-to-end manner, thereby enabling automatic and application-independent representations of MOFs. Being used to predict the band gaps of MOFs, CGCNN-derived models were found to significantly outperform kernel ridge regression models with conventional features such as the sine Coulomb matrix, the orbital field matrix, and the smooth overlap of atomic positions (SOAP).[26] Similarly, the graph neural network (GNN)-based representation was also developed for chemical features, involving SchNet, the message passing neural network (MPNN), or the MatErials graph network (MEGNet).[48] A successful use of MPNN was to assign the atomic charges of MOFs.[49] MPNN efficiently learns the atomic features of MOFs and their bonding environments via a message-passing mechanism between bonded atoms. By employing MPNN, atomic charges of MOFs can be calculated in seconds with DFT-level accuracy. Despite the popularity of GNN representation among crystalline materials, their applications for MOFs are typically limited to quantum-chemical properties. This is because the GNN representation inherently encodes more chemical features but places less emphasis on geometric information.[50] For an application such as gas adsorption, CGCNN prediction was shown to be improved by including pore geometry and other three-dimensional topology information.[51]

*Latent Vectors.* Word embedding is another data-driven featurization technique. It learns to represent each word as a numerical vector from text, wherein the latent attributes of each word are preserved along with the underlying relationships among different words. The potential of word embedding has been highlighted for text mining in recent publications. Based on a vocabulary data set of ∼500,000 words text mined from ∼3.3 million papers for materials, Tshitoyan et al. generated a set of word embeddings through the word2vec algorithm, which could smartly tie specific material formula words (e.g., $Li_2CuSb$, $CuBiS_2$, and $CdIn_2Te_4$) to "thermoelectric".[52] Similar work was conducted by Krishnapriyan et al. to encode the stoichiometric formula of MOFs into word embeddings to capture chemical information, which was found to accurately predict the Henry constants of $CO_2$ in CoRE MOFs.[39] This reveals the potential use of the word embedding in encoding MOFs. Alternatively, Yao et al. designed the RFcode to decompose the building blocks and topologies of MOFs, which were then converted to semantically constrained graph-based canonical sequences; each part of the RFcode was encoded into a latent space and decoded back to a MOF structure bilaterally in conjunction with the supramolecular variational autoencoder (Figure 1b), thereby enabling the inverse design of novel MOFs for $CO_2$ capture.[14] In another recent work, Lee et al. proposed a MOF-NET architecture in which the embeddings of building blocks and topologies were automatically updated during the ML training process; the MOF-NET in tandem with the genetic algorithm was utilized to discover hypothetical MOFs with record-breaking $CH_4$ storage capacity.[53] To mitigate the limitation of MOFid representation in encoding MOF geometric features, Cao et al. developed a transformer model, MOFormer; by introducing a self-supervised pretraining manner with CGCNN on 1 million hypothetical MOFs, the MOFormer was found to surpass CGCNN in the prediciton of MOF band gaps with few training data points (e.g., <1000).[54]

As discussed above, one can readily encode a MOF with a well-defined structure or a stoichiometric formula via either a hand-crafted or data-driven featurization method. Nevertheless, the structure may not be available. For instance, the NIST ISODB comprises ∼4000 synthesized adsorbents whose features are not uniformly described because of the lack of structural information. In this context, Zhang et al. employed the neural collaborative filtering method to learn adsorbents directly from their gas adsorption properties.[55] Based on a matrix filled with the pairwise adsorption uptake at a specific temperature and pressure, the neural collaborative filtering mapped inherent adsorbent−adsorbate relationships into a latent space, from which the latent vectors of adsorbents and adsorbates were learned, respectively, thus bypassing the dependence on adsorbent structures. The resultant neural recommender system could be further used to supplement missing data in the adsorbent−adsorbate matrix, enabling the screening of adsorbents from the NIST ISODB for specific gas adsorption tasks. Similarly, Sturluson et al. completed the missing adsorption data in a matrix for 572 covalent organic frameworks (COFs) and 16 gases. As illustrated in Figure 2c, the latent vectors of COFs (**M**) and gas-adsorption properties (**P**) were directly learned through decomposing the COF-property matrix (**A**); after projecting the latent vectors of COFs onto a 2D space, the proximity between COFs and similar adsorption properties was observed, thus revealing the similarity across different COFs.[33]

**Model Training.** Many ML studies for MOFs aim to establish structure−property relationships via supervised learning, generally known as regression or classification (Figure 1c). The former predicts properties (e.g., gas adsorption) based on a set of features, while the latter learns to differentiate MOFs (e.g., stable or unstable). Interpretability and generalizability are two fundamental requirements for supervised learning. On one hand, if the structure−property relationships are well interpreted, then they can be used together with domain knowledge to guide the design of new MOFs. On the other hand, generalizability guarantees accuracy when applying ML models to predict unknown MOFs. Before model training, it is important to perform feature engineering to select the most relevant features and visualize the feature space, as discussed below.

*Feature and Dimensionality Reduction.* Using high-dimensional and dependent features in ML may increase the model complexity, training time, and probability of overfitting. Therefore, the rational selection of features is crucial because it also affects model interpretability and generalizability as well as out-of-sample prediction performance. In predicting the thermal stability of MOFs, Escobar-Hernandez et al. employed Pearson correlation coefficient matrix to assess the multicollinearity of features and reduced 155 features to 12 independent ones.[56] Feature reduction via Pearson matrix largely relies on human inspection. As an alternative, automatic feature selection methods such as recursive feature addition (RFA),[11] recursive feature elimination (RFE),[12] univariate filtering (UVF),[43] and wrapper feature selection (WFS)[57] are more preferred to bypass human bias. UVF is suitable to handle high-dimensional features as each feature is evaluated individually by discarding poorly behaved ones through a univariate statistical test but it inherently neglects the underlying interactions across different features, and the resultant feature set may poorly correlate with a nonlinear ML model. As shown in the screening of MOFs for sour gas sweetening, WFS may


https://doi.org/10.1021/acs.langmuir.3c01964
Langmuir 2023, 39, 15849−15863

require an exhaustive search of all possible combinations of features, making it feasible only for low-dimensional features.[12] In contrast, RFA and RFE are more amendable to high-dimensional features by a stepwise search. By implementing RFA of 165 features, Moosavi et al. found that the number of optimal features varied largely across different tasks (e.g., 41 for the maximum positive charge, 28 for the Henry coefficient of $CO_2$, and 7 for the $CH_4$ deliverable capacity), which highlights the necessity of feature reduction.[11] For water stability of MOFs, Batra et al. demonstrated that RFE could largely reduce feature dimensionality from 149 to ∼30 and meanwhile increase model accuracy.[12]

In addition to feature reduction for supervised learning, dimensionality reduction is often used for unsupervised learning, where a high-dimensional feature space is projected onto a low-dimensional feature space with the best preserved original neighboring information (Figure 1c). By doing so, materials with structural and chemical similarity are closely clustered in the low-dimensional feature space and can be easily visualized. As a commonly used technique for dimensionality reduction, Principal component analysis (PCA) is designated to project original data into a set of orthogonal vectors or principal components, along which data exhibit the highest variance.[58] Sarkisov et al. applied PCA to analyze the correlations among geometric features of CSD MOFs; the ratio of LCD/PLD, pore volume, and crystal density were revealed to be independent and strongly correlated with the first three principal components from PCA.[36] By projecting the latent vectors of COFs into a 2D space based on PCA, Sturluson et al. observed similar adsorption properties in clustered COFs, hence suggesting the physical robustness of such data-driven latent vectors.[33] Despite being computationally efficient, PCA is subject to a linear assumption across different features and hence inherently falls short in capturing nonlinear relationships. As an improvement over PCA, the nonlinear manifold learning technique assumes that the original high-dimensional feature space can be approximately embedded by low-dimensional manifolds. Typically, t-distributed stochastic neighbor embedding (t-SNE) aims to construct a reduced space with neighborhood probabilities that are most similar to those in the original feature space.[59] Moosavi et al. demonstrated the promise of t-SNE in quantifying and visualizing the similarity across different MOF databases (one experimental and five hypothetical), and the hypothetical MOF databases were clearly observed with insufficient diversity in metal chemistry.[11] The similarity and diversity revealed by t-SNE can be used to interpret the ML models. As highlighted in the study for $C_3H_8/C_3H_6$ separation by Tang et al., the models trained upon CoRE MOFs showed good transferability to CSD MOFs but not to hypothetical MOFs, as attributed to chemical bias in the feature space between experimental and hypothetical MOF databases.[17] Also by implementing t-SNE, Tang and Jiang visualized the chemical space of COFs and addressed a widely concerned question, that is, whether an unknown database deserves in-depth exploration for possible candidates with targeted performance. Moreover, the Genomic COF database with massive unexplored 3D COFs was found to be potentially promising for ultrahigh $CH_4$ storage.[18] As another nonlinear manifold learning technique, uniform manifold approximation and projection (UMAP) exhibits better learning efficiency than t-SNE as well as preserving the global data structure.[60] Rosen et al. employed UMAP to visualize the distributions of the QMOF database for DFT-computed band gaps, wherein a subtle structure−property trend spanning from low to high band gaps was clearly observed.[26]

*Algorithms.* After feature engineering, the next step is to train ML models. One needs to make an optimal choice among a wide range of ML algorithms and optimize hyperparameters for the selected algorithm. As summarized in Figure 1c, the commonly used algorithms for MOF-related studies include tree-based, kernel-based, and an artificial neural network.

Tree-based algorithms, typically decision trees (DTs), employ an "if−then" decision rule and a flowchart-like architecture to make predictions, which are hence facile to interpret. Wu et al. implemented DT classifiers to identify necessary features to render defective UiO-66 with optimal $C_2H_6/C_2H_4$ separation performance, where surface area and pore volume were found to be indicative features.[61] However, DT with excessive depth is inherently subject to an overfitting issue; that is, it works only on training data. Such a drawback can be alleviated by constructing tree ensembles such as the random forest (RF) and gradient-boosted decision tree (GBDT). The former uses bagging as a core idea that the final prediction is made by averaging all trees, thus reducing the prediction variance, and the latter adopts a boosting strategy, wherein trees are operated in sequence but parallel to minimizing prediction bias. The superiority of both RF and GBDT over DT has been demonstrated in ML studies for MOFs.[62,63] Despite these mathematical improvements, it is worthwhile to note that tree-based algorithms generally do not work well for extrapolation. They are incapable of out-of-sample prediction, with features not included in training. As evidenced by Fanourgakis et al., an RF model trained on ∼8000 MOFs with $CH_4$ adsorption capacity of <100 $v_{STP}/v$ completely failed to generalize to ∼2000 MOFs with $CH_4$ adsorption capacity of >100 $v_{STP}/v$; after additionally enriching the training set by artificial MOFs with broader structural diversity, the resultant RF model was improved in prediction.[64]

Kernel-based algorithms, also known as kernel tricks, map original data into a high-dimensional space wherein nonlinearity across data is treated linearly and more efficiently (Figure 1c). The similarity across data is measured by a kernel matrix, based on which the relationships learned from labeled data can be extrapolated to unlabeled ones. The support vector machine (SVM) represents one of the most popular kernel-based algorithms for MOFs. In stark contrast to tree-based algorithms, SVM is capable of circumventing the extrapolation issue and performing multitask prediction.[64] However, kernel-based algorithms rely on the inversion of the kernel matrix to make a prediction, and the computational complexity increases acutely with the size of the training samples, thus largely restricting learning efficiency when handling large-volume data.[18]

Artificial neural networks (ANNs) comprise layers of neurons and work in a way similar to the real neurons in a human brain. In comparison with the aforementioned two types of ML algorithms, ANN exhibits a significant advance in mathematics, that is, concurrently bypassing the extrapolation issue in tree-based algorithms and the expensive matrix inversion in kernel-based algorithms. In addition, deep ANN algorithms such as MPNN, CGCNN, and word2vec typically work in an end-to-end manner by integrating material features with structure−property relationships. Consequently, deep ANN algorithms are preferred to treat large-volume and unstructured data. The promise of ANN for MOFs has been shown in the transfer learning of MOFs[13] and the inverse design of MOFs.[14] Due to the black-box nature of the algorithms, however, the models from ANN lack good interpretability.


https://doi.org/10.1021/acs.langmuir.3c01964
Langmuir 2023, 39, 15849−15863

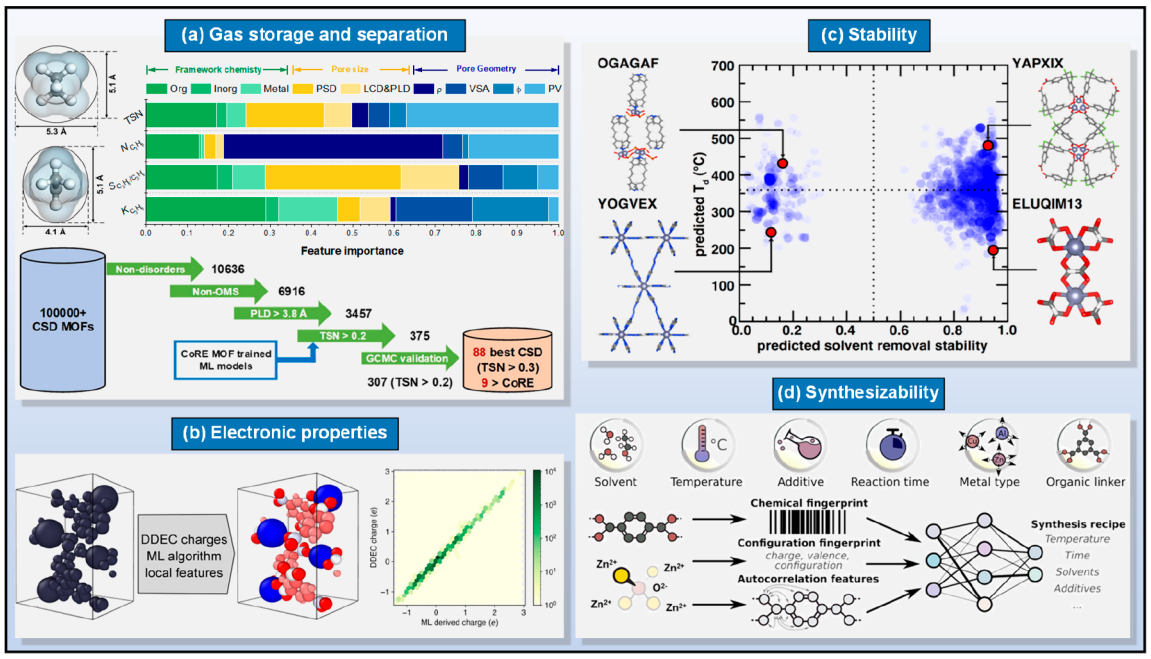



**Figure 3.** Applications of MOFs. (a) Gas storage and separation (reproduced with permission from ref 17, copyright 2021 American Chemical Society). (b) Electronic properties (reproduced with permission from ref 67, copyright 2020 American Chemical Society). (c) Stability (reproduced with permission from ref 30, copyright 2021 American Chemical Society). (d) Synthesizability (reproduced with permission from ref 31, copyright 2022 John Wiley and Sons).

No existing algorithms are generic and applicable to all of the different tasks. For a given task, the predictive performance of different algorithms should be compared in order to select the optimal one. For instance, various ML models including DT, RF, GBDT, SVM, and ANN were benchmarked by Anderson et al. for $CO_2/N_2$ separation in ~400 MOFs, where ANN and GBDT were found to outperform others.[63] Similarly, the RF classifier was observed to surpass DT, SVM, and ANN for $CO_2$ capture in MOFs.[65] Nevertheless, one should be cautious about data leakage and data imbalance, which may cause the optimal model with alleged superiority to fail in the out-of-sample prediction. Data leakage refers to the data split being human-biased rather than randomized, through which extra knowledge from a test set is leaked into a training process. Typically, one may manipulate the training/test split until a desired prediction accuracy is achieved in the test set. Smit and co-workers reported another example of nonhuman-induced data leakage caused by the presence of many duplicates in existing MOF databases (i.e., CoRE MOF 2019, CSD MOF subset, and QMOF), where duplicated MOFs with different CSD refcodes but nearly identical structures appeared in both training and test sets, thus the model performance was overestimated due to the dependence between training and test sets.[66] To alleviate data leakage, it is recommended to conduct cross validation (e.g., 5-fold) and average the predictions of different random splits (e.g., 20 repeats), as demonstrated by Tang et al. for $C_3H_8/C_3H_6$ separation in MOFs.[17] Tang et al. also discussed the data imbalance in $C_3H_8/C_3H_6$ selectivity, which caused heteroscedasticity in prediction.[17] Such a heteroscedasticity issue can be mitigated by increasing the structural diversity of MOFs in the training set. In contrast to augmenting more diverse data to a training set, few-shot learning strategies (e.g., transfer learning and active learning) are robust to tackle data imbalance.

**Model Applications.** In the past several years, ML has been increasingly applied to many different applications of MOFs, as summarized in Figure 3, including (a) gas storage and separation, (b) electronic properties, (c) stability, and (d) synthesizability.

*Gas Storage and Separation.* To date, overwhelming ML studies for MOFs have focused on gas storage and separation. Review articles are available on this topic including $H_2$ or $CH_4$ storage, $CO_2$ capture, and noble gas separation.[7−9,68] Here, we highlight the very recent studies that were largely not discussed in these review articles. Most reported ML studies emphasize the prediction accuracy of ML models within a single MOF database. Nevertheless, cross-database prediction is more interesting and important to evaluate the generalizability of ML models. As illustrated in Figure 3a for $C_3H_8/C_3H_6$ separation, Tang et al. found that the RF models trained upon CoRE MOFs exhibited good generalizability to CSD MOFs due to the large similarity between the two databases; however, less satisfactory prediction performance was observed when extending the RF models to hypothetical MOFs, which lack chemical similarity to CoRE MOFs.[17] Starting from experimentally measured water adsorption isotherms in 285 MOFs, Zhang et al. developed ML models for atmospheric water harvesting; the transferability of the ML models was validated by out-of-sample predictions in newly reported MOFs, and finally the ML models were applied to screen ~8,000 CoRE MOFs and identify top-performing candidates.[42] For gas diffusion in MOFs, Krokidas et al. benchmarked a ML model to predict the diffusivities of 12 different gases in 72 variants of zeolitic-imidazolate frameworks (ZIFs).[69] Daglar and Keskin constructed ML models to predict gas diffusivities and permeabilities in MOF membranes and MOF/polymer mixed-matrix membranes.[70] Moreover, the ML has been utilized to characterize MOFs. It is known that the Brunauer−Emmett−Teller (BET) method tends to overestimate the surface areas of MOFs because the monolayer loading would be erroneously estimated from the adsorption isotherm. In this context, Datar et al.


https://doi.org/10.1021/acs.langmuir.3c01964
Langmuir 2023, 39, 15849−15863

developed a ML model that was able to accurately predict the surface areas, thus overcoming the limitation of the traditional BET method; furthermore, the ML trained with Ar adsorption exhibited good transferability to $N_2$ adsorption.[71] In another study, ML classification models were developed by Pétuya et al. to evaluate guest accessibility in MOFs, and the models could accurately categorize MOFs into nonporous, small pores, medium pores, and large pores without a priori knowledge of MOF structures.[72] Hence, these ML classification models would benefit experimentalists in prioritizing the choice of building units when synthesizing MOFs with desired pore accessibility.

*Electronic Properties.* The electronic properties of MOFs including atomic charges, band gaps, colors, and oxidation states have been examined through ML. In particular, atomic charges are required in molecular simulations to evaluate electrostatic interactions between guest molecules and frameworks. Traditionally, DFT calculations are conducted to evaluate electrostatic potentials from which atomic charges are derived. This method is time-consuming and not transferable among different MOFs. By taking the density-derived electrostatic and chemical (DDEC) charges as a benchmark, Raza et al. proposed a MPNN architecture to estimate the atomic charges of MOFs, found it more accurate than the charge equilibration method, and finally assigned the MPNN charges to 9,122 CoRE MOFs.[49] As illustrated in Figure 3b, Korolev et al. developed a ML model for the atomic charges of MOFs by integrating hand-crafted features (23 elemental properties and 6 structural properties), and the model was more interpretable than the MPNN charge method and was applicable to all 10,140 structures in the CoRE MOF database.[67] Instead of featurizing MOFs by crystal graphs or numerous global parameters, Kancharlapalli et al. presented a simpler and easy-to-train RF model for charge assignment, called partial atomic charges in MOFs (PACMOF); the resultant atomic charges were found to match well with the DDEC counterparts and were significantly faster than the extended charge equilibration (EQeq) method when handling large unit cells.[73] Band gaps represent another type of electronic property and are of great interest when applying MOFs to sensing, catalysis, and energy storage. He et al. reported the earliest ML classification model for the band gaps of MOFs based on inorganic crystals in the Open Quantum Materials Database (OQMD); the model was able to directly transfer learned knowledge to judge whether a given MOF is metallic and was further used to filter out 9 metallic candidates from 2,932 CoRE MOFs.[74] As discussed earlier, the QMOF database also contains band gaps and atomic charges of ∼15,000 MOFs estimated from DFT calculations.[26] Colors are highly indicative of optoelectronic properties, whereas the estimation of colors is challenging via conventional methods. In this regard, a ML model was developed by Jablonka et al. to predict the colors of MOFs (RGB values) with fair accuracy. The model could learn certain chemical intuitions affecting colors; for instance, the addition of an amino group would induce a red shift of UiO-66.[44] The same group leveraged data-driven tools to assign the oxidation states of metal sites in MOFs, which are commonly evaluated with ambiguity by electron-counting rules; supervised by domain expertise in determining oxidation states, the ML model exhibited desirable prediction accuracy and was able to correct ambiguous assignments throughout the CSD MOF database.[45]

*Stability.* A major concern about MOFs is stability, which largely restricts their scale-up and practical applications. The computation of stability is time-consuming, as sophisticated or first-principles-based methods are required, particularly for chemical and thermal stability. ML tools have demonstrated robustness to overcome this challenge. From simulation data, Moghadam et al. developed a ML model for the mechanical stability of MOFs by using geometric features including topology, density, gravimetric surface area, and void fraction; the model was able to predict bulk moduli with accuracy comparable to that of simulation.[75] As mentioned, the chemical and thermal stabilities of MOFs cannot be easily evaluated by simulation. Alternatively, Batra et al. collected 207 MOFs with experimentally measured water stability and constructed ML models by using chemical features; the models were applied to screen water-stable MOF candidates and extract simple stability trends in MOFs.[12] These few ML studies were based on small data sets of MOFs, thus one should be cautious about their prediction capability. As presented in Figure 3c, Nandy et al. employed a natural language processing (NLP) technique to collect a set of diverse MOFs with experimental measures of thermal stability (3132) and solvent-removal stability (2179), respectively, and trained ML models to encode structure−property relationships and predict stability. From the interpretation of important features, strategies to engineer stability were suggested (e.g., linker variations rather than metal substitutions would exert a greater effect on thermal stability).[30] Subsequently, a workflow and a web interface (MOFSimplify) were reported by them to provide the access for stability predictions of new MOFs.[76] Recently, they further constructed a database of ultrastable hypothetical MOFs and computed elastic moduli to confirm good mechanical stability.[28]

*Synthesizability.* As a sophisticated process, the synthesis of MOFs depends on many high-dimensional conditions, as shown in Figure 3d, including metal type, organic linker, solvent, temperature, additive, pH value, reaction time, etc. Hence, predicting MOF synthesizability as well as optimal conditions is a great challenge from either experimental or computational methods. Currently, there is little explicit connection between synthesis conditions and the resultant product. As a preliminary attempt, ML has found potential in predicting MOF synthesizability. Moosavi et al. employed the genetic algorithm (GA) to optimize the synthesis conditions of HKUST-1 and learn chemical intuition to control the crystallinity and purity of HKUST-1, and highlighted the importance of including failed synthesis data in ML to predict MOF synthesizability.[10] From powder X-ray diffraction (PXRD) patterns and scanning electron microscopy (SEM) images, Chen et al. constructed ML models for MOF morphologies and guided MOF growth with desired morphologies during synthesis, hence largely promoting the catalytic activity of resultant MOFs for olefin hydrogenation.[77] A genetic algorithm was used by Domingues et el. to systematically search for the optimal synthesis conditions of Al-PMOF, and excellent crystallinity and yield close to 80% were shown in a short reaction time in just two generations.[78] Kitamura et al. used a data-driven approach to visually map the previously reported synthesis conditions for anionic lanthanide-based MOFs, revealed the existence of unexplored search spaces, and then synthesized a series of new MOFs.[79] The importance of abandoned synthesis data was also emphasized by Hu et al. and shown to benefit the ML predictions of $C_2H_2$, $C_2H_4$, and $CO_2$ adsorption in anion-pillared MOFs.[80] At present, there exist a handful of databases containing MOF synthesis information such as SynMOF[31] and DigMOF.[32] These databases are small, with only successful synthesis information. It is highly desired to construct larger databases including both successful and


https://doi.org/10.1021/acs.langmuir.3c01964
Langmuir 2023, 39, 15849−15863

unsuccessful syntheses to develop more reliable ML models to predict MOF synthesizability.

## MOVING FORWARD

We have witnessed the rapidly increasing utilization of ML for many different applications of MOFs, including gas storage and separation, electronic properties, stability, and synthesizability. Physically meaningful features were proposed, and interpretable ML models were constructed to predict and discover new MOFs. Despite these remarkable achievements, several technical challenges and new opportunities are highlighted for moving forward in this vibrant field.

**Data.** Though the number of possible MOFs is nearly infinite, our understanding of their property boundary is far from complete due to the small chemical space explored in existing ML studies. Emerging data-driven methods have manifested great potential in broadening the current knowledge of MOFs. Nonetheless, we should note that only a small fraction of data produced for MOFs to date are findable, accessible, interoperable, and reusable (FAIR).[81] In other words, an overwhelming majority of MOF-related data are unpublished, unstructured, misplaced, or treated without following the FAIR principles. This urges us to be cautious about data availability and reproducibility.

*Data Availability.* As discussed above, most applications of MOFs have been focused on adsorption-based gas storage and separation. With readily tunable structures and functionalities, MOFs have been envisioned for many other potential applications such as heat pumping, catalysis, and sensing. To implement ML in the future for these applications, it is a prerequisite to construct databases that contain a wide range of relevant properties such as thermal conductivity, catalytic activity, and selectivity. A few open-source platforms such as Materials Project[82] and Materials Cloud[83] greatly facilitate the availability of MOF structures as well as their corresponding properties from simulations. In particular, the QMOF database has been incorporated into Materials Project (https://materialsproject.org/mofs) following the FAIR principles. Certain properties cannot be easily simulated; alternatively, they are measured experimentally. For water stability, ML models have been constructed from experimental data.[12] However, the current database for water stability is small, and more MOFs should be included to develop more reliable and quantitative models. Nowadays, experimental failures are scarcely reported in scientific publications. Nevertheless, ML studies on the MOF synthesizability revealed that including failed experiments would improve model performance. Instead of being abandoned, experimental failures may serve as meaningful and potential outliers to extend our knowledge boundary. Hence, reporting and sharing data of failed experiments are recommended in the MOF community.

*Data Reproducibility.* In the field of MOFs, researchers tend to focus more on new structures and fascinating properties but less on reproducibility. For instance, identical MOFs, if synthesized and measured by separate research groups, may exhibit different $CO_2$ adsorption isotherms.[84] Poor reproducibility is indeed observed in the NIST ISODB, which comprises over 30,000 adsorption isotherms in ∼4,000 adsorbents. This suggests that experimental data collected in the NIST ISODB should be used with great caution. By manually matching MOFs in the NIST ISODB with their crystal structures, Park et al. demonstrated the feasibility of molecular simulation to assess the reproducibility of adsorption isotherms in the NIST ISODB.[84] A similar study was recently conducted by Ongari et al., who automatically linked adsorption isotherms in the NIST ISODB to corresponding structures in the CSD MOF database, thus constructing a data set concurrently encompassing MOF structures and associated isotherms.[85] Another issue impeding reproducibility is that data (e.g., adsorption isotherms in the NIST ISODB) from different sources may vary in unit (e.g., kPa, Pa, bar, and Torr for pressure; mmol/g, cc/g, and cc/cc for capacity), file storage format (e.g., Excel, pdf, and Word), and MOF naming (e.g., Cu-BTC and HKUST-1 for an identical MOF). This necessitates a standard procedure for collecting, storing, and sharing data. For adsorption data, Evans et al. proposed a format called the adsorption information file (AIF), which accepts the transformation from the NIST ISODB JSON format as well as other formats by different adsorption instruments, thus bypassing human intervention and bias when comparing adsorption isotherms.[86] This work has also set a reference to archive and report data in a standardized procedure.

**Featurization.** Featurization is critical to determining the accuracy and interpretability of ML models. The complex hierarchy in the geometry and chemistry of MOFs, spanning from the atomic to crystalline scale, makes featurization challenging. At present, hand-crafted features are common or preferred, as they possess high interpretability and expressivity to reveal quantitative structure−property relationships and design guidelines for new MOFs. Recently, graph-based neural network (e.g., CGCNN) features have received increasing attention because of the end-to-end trait from automatic featurization to model prediction. However, graph-based architecture inherently limits interpretability and requires a large data set to achieve sufficient expressivity. As discussed earlier, CGCNN is less expressive in the MOF geometry, thus hindering its performance for tasks that are sensitive to the MOF geometry. With CGCNN as a starting point to customize novel deep learning architecture, recent attempts such as MOF-CGCNN and k-NAGCN are attractive in the featurization of MOFs for future exploration. Featurization by invertible vectors is another prospective direction. "Invertible" refers to a bidirectional mapping relationship between material structures and latent vectors based on which generative models are trained to enable the inverse design of new materials from a latent space. At present, only a handful of featurization methods have been demonstrated to be invertible such as wording embeddings and molecular graphs, while crystal graph featurization such as CGCNN is technically noninvertible. By applying invertible featurization, Yao et al. inversely designed novel MOFs for $CO_2$ capture.[14] Though in their infancy, invertible featurization together with inverse design is a promising direction in future ML for MOFs and other materials. It is worthwhile to note that featurization may fail in certain MOFs, for example, using Zeo++[34] for pore features, MPNN for atomic charges,[49] lammps_interface[87] for atom types, molSimplify[88] for RACs, and MOFid[89] for nomenclature. In most ML studies, unfeaturizable MOFs are usually discarded, thus inevitably reducing the structural diversity or skipping promising candidates. In this regard, one should attempt an alternative way to featurize and include more MOFs in ML.

**Learning Strategy.** A common practice in ML for MOFs is to train models that can be used to predict the performance of unknown MOFs, which are not within the chemical space of trained MOFs. Such out-of-sample predictions by extrapolating from known to unknown space cannot be achieved well by


https://doi.org/10.1021/acs.langmuir.3c01964
Langmuir 2023, 39, 15849−15863

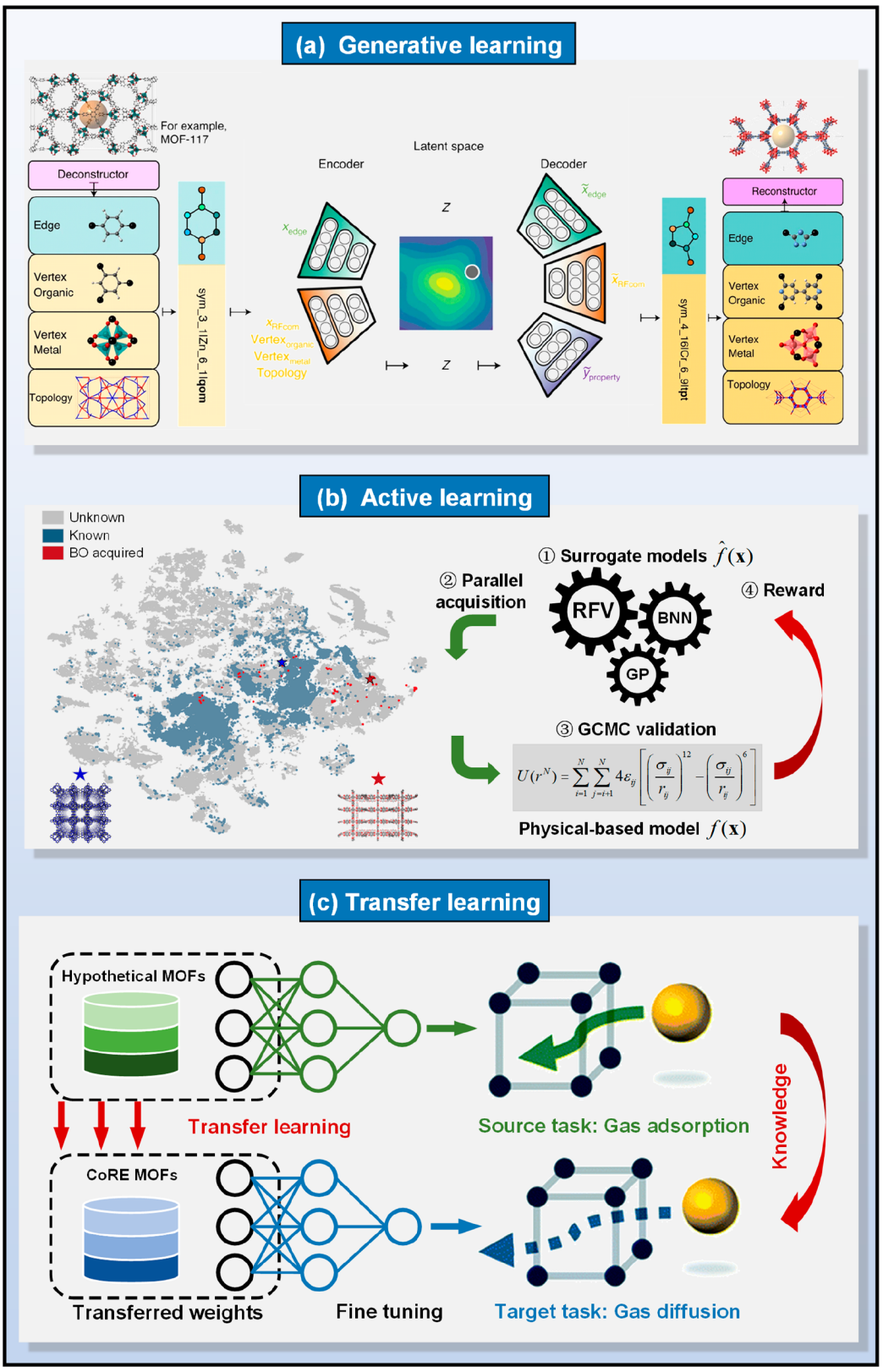



**Figure 4.** Advanced learning strategies. (a) Generative learning for inverse design (reproduced with permission from ref 14, copyright 2021 Springer Nature). (b) Active learning for accelerating discovery (reproduced with permission from ref 18, copyright 2022 John Wiley and Sons). (c) Transfer learning for bridging cross-domain knowledge (reproduced with permission from ref 13, copyright 2022 Royal Society of Chemistry).

conventional learning algorithms (e.g., tree-based). Advanced learning strategies need to be adopted such as (a) generative learning, (b) active learning, (c) transfer learning, and (d) meta learning.



https://doi.org/10.1021/acs.langmuir.3c01964
Langmuir 2023, 39, 15849−15863

Generative learning can largely broaden the limited chemical space of known MOFs (Figure 4a). As evidenced by Yao et al., a generative deep learning model was able to produce a continuous and differentiable latent space based on a discrete and limited MOF data set, from which an optimal structure−property mapping was learned for the inverse design of novel MOFs with $CO_2$ capture performance surpassing that of currently known materials.[14] This learning strategy is expected to be increasingly used in the MOF community by optimizing the generative learning architecture toward the design of new MOFs for various applications. Differently, as illustrated in Figure 4b, active learning (AL) enables an on-the-fly search of a chemical space. It iteratively optimizes a surrogate model based on a known search space and operates an acquisition function to select candidates from an unknown search space. In such a way, exploration and exploitation are balanced in the search space. AL has been adopted in MOF studies, from optimizing MOF synthesis conditions to accelerating computational screening. Xie et al. combined an autonomous synthesis robot and Bayesian optimization (BO) to optimize the crystallinity of ZIF-67 within ∼20 iterations.[91] Instead of sequential acquisition, Tang and Jiang proposed a parallel BO strategy by incorporating the Kriging−Believer-based acquisition and a RF surrogate model to screen COFs for $CH_4$ storage; top-performing COFs with record-breaking high storage capacity were shortlisted by acquiring only 50 out of 445,845 COFs without exhaustively examining undesirable structures.[18] Technically, the parallel BO strategy can be further implemented for multiobjective ML tasks, which is scarcely reported for MOFs. Taking adsorption-based gas separation as an example, it is challenging to find MOFs that can balance the trade-off between adsorption capacity and selectivity. In such a case, multiobjective BO would be the right solution to seek for the Pareto frontier without human bias (i.e., a set of optimal trade-offs among various contradicting properties).[92] Hence, ML in tandem with the multiobjective BO strategy deserves more attention in future ML for MOFs. Apparently, multiobjective learning is applicable only when sufficient multitask data are available. Nevertheless, we may encounter the problem of data sparsity. For instance, overwhelming data of gas adsorption in different MOFs are available, while diffusion data are scarce, as they are more difficult to measure either experimentally or computationally.

Transfer learning (TL) is a useful strategy to tackle data sparsity by bridging cross-domain knowledge. Specifically, TL extends the knowledge learned from a source task to a new target task, which enables TL to make desirable predictions on the new target task even when few-shot data are available for training. Figure 4c exemplifies a typical TL case where knowledge learned from $CH_4$ adsorption in MOFs is transferred to $CH_4$ diffusion.[13] Similarly, by transferring knowledge from $CO_2/CH_4$ separation in MOF-based mixed-matrix membranes, Guan et al. demonstrated that $CO_2/N_2$ separation performance could be also predicted.[93] Very recently, Kim and co-workers introduced a multimodal pretraining transformer (MOFTransformer) by integrating the energy-grid embeddings of MOFs and the CGCNN representation; the MOFTransformer was pretrained with 1 million hypothetical MOFs and transferable to a wide range of MOF properties such as gas adsorption, diffusion, and band gap, which were found to outperform CGCNN and descriptor-based ML models trained from scratch.[94] Improved over TL, meta-learning is another one-shot learning strategy. Instead of performing a single learning task in an exclusive data set, a meta-learning model is trained on many meta-data sets each trained on a subset. By doing so, meta-learning is able to learn the similarity across meta-data sets, measure the relationship among different learning tasks, and achieve a better learning outcome. With the "learning-to-learn" architecture, the meta-learning model is highly generalizable to cross-domain tasks even with little training data. Sun et al. developed meta-learning models, merely trained on $H_2$ adsorption data in zeolites, to finely tune $H_2$ adsorption in hyper-cross-linked polymers and large MOFs with desirable prediction accuracy; their models also learned a latent representation of nanoporous materials, thus allowing generalization to experimental adsorption data even with sparse temperature and pressure points.[90] Currently, few-shot learning strategies such as TL and meta-learning are underdeveloped for MOF applications but deserve more attention as they are promising for handling cross-fidelity data. For instance, adsorption and diffusion data in MOFs from experiments show higher fidelity but are less available, in contrast to low-fidelity data from simulations. We should also note that both TL and meta-learning may fail if training data lack remarkable similarity to a target task.[13,90] In other words, sufficient and diverse data are required for reliable few-shot learning.

**ML-Aided Simulation.** At present, most molecular simulation studies are based on generic force fields, which were developed by fitting to limited experimental data with certain empirical rules and thus may lead to inaccurate predictions. There has been burgeoning interest in deriving force fields (or potentials) from the ML for MOFs. Eckhoff and Göttingen conducted DFT calculations of small molecular fragments and developed a ML-derived high-dimensional neural network potential (HDNNP) for MOF-5.[95] With the quantum-informed ML force field, Zheng et al. predicted that the binding free energy landscape and diffusion coefficient of $CO_2$ in Mg-MOF-74 were close to experimental values.[96] Vandenhaute et al. proposed an incremental learning scheme to construct ML potentials for MOFs, subsequently showing its accuracy and transferability to UiO-66 and MIL-53.[97] An on-the-fly ML force field was used by Huang et al. in a MD simulation to explore the structural disorder of layered COFs, in which an initially eclipsed stacking mode was found to spontaneously distort to form a zigzag configuration.[98] While these ML-derived force fields demonstrated their robustness, they were targeted to specific MOFs. It is highly desired to develop transferable ML-derived force fields for different MOFs. To achieve this, a large volume of DFT calculations is required to cover sufficient chemical diversity of MOFs. Ideally, ML-derived force fields should be integrated into existing simulation packages to allow for easy and practical use. A typical case is the on-the-fly force field generation method, which is now available in the Vienna Ab Initio Simulation Package (VASP), and its effectiveness has been shown in the prediction of crystal melting points.[98,99] We envision increasing attempts in developing new ML-derived force fields to tackle intriguing tasks for MOFs (e.g., dynamic response of flexible frameworks upon external stimuli such as gas adsorption, temperature, and pressure variation).

## CONCLUSIONS

In this Perspective, we summarize the recent representative ML studies in the field of MOFs and particularly discuss data acquisition, featurization, model training, and applications. Though most of the existing studies are on gas storage and separation, there have been increasing efforts to study the electronic properties, stability, and synthesizability of MOFs. It


https://doi.org/10.1021/acs.langmuir.3c01964
Langmuir 2023, 39, 15849−15863

is anticipated that other important properties such as thermal conductivity and catalytic activity will also be explored in the future by leveraging ML. Despite the remarkable progress and significant advance in this rapidly evolving field, implementing ML for MOFs still faces technical challenges in data availability and reproducibility, featurization, and learning strategy. These challenges provide new and exciting opportunities to further leverage ML in revealing new mechanistic insights and discovering new structures.

## AUTHOR INFORMATION


### Corresponding Authors

**Lunbo Duan** − *Key Laboratory of Energy Thermal Conversion and Control of Ministry of Education, School of Energy & Environment, Southeast University, Nanjing 210096, China;* orcid.org/0000-0002-8210-0851; Email: duanlunbo@seu.edu.cn

**Jianwen Jiang** − *Department of Chemical and Biomolecular Engineering, National University of Singapore, 117576, Singapore;* orcid.org/0000-0003-1310-9024; Email: chejj@nus.edu.sg

### Author

**Hongjian Tang** − *Key Laboratory of Energy Thermal Conversion and Control of Ministry of Education, School of Energy & Environment, Southeast University, Nanjing 210096, China; Department of Chemical and Biomolecular Engineering, National University of Singapore, 117576, Singapore*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.langmuir.3c01964


### Notes

The authors declare no competing financial interest.

### Biographies

Hongjian Tang received his undergraduate degree and completed his Ph.D. at Southeast University, China. During his Ph.D., he visited Professor David Sholl's group at the Georgia Institute of Technology sponsored by the CSC scholarship. Thereafter, he joined Professor Jianwen Jiang's group at the National University of Singapore as a research fellow. Currently, he is an associate professor in the School of Energy and Environment, Southeast University. His research interest is in the computational and data-driven design of functional materials for energy- and environment-related applications.

Lunbo Duan received his bachelor's degree in 2005 from the Nanjing Normal University, China and Ph.D. in 2010 from Southeast University, China. From 2014 to 2015, he was a research fellow at Cranfield University. In 2018, he received awards from the Top Young Scholar of National "Ten Thousand Talent Program" and the National Science Fund for Excellent Young Scholars of China. Currently, he is a professor at Southeast University, investigating low-carbon technologies and materials for energy-related applications.

Jianwen Jiang received his bachelor's degree and Ph.D. from the East China University of Science and Technology. Currently, he is a professor in the Department of Chemical and Biomolecular Engineering at the National University of Singapore. His research expertise is computational materials modeling with focus on membranes and nanoporous materials for energy, environmental, and pharmaceutical applications, such as carbon capture and conversion, water desalination, biofuel purification, and solvent recovery.

## ACKNOWLEDGMENTS

The authors gratefully acknowledge the National Natural Science Foundation of China (U22A20435), the scientific and technological innovation project of carbon emission peak and carbon neutrality of Jiangsu Province, China (BK20220001), Fundamental Research Funds for the Central Universities of China (2242023K5001), the A*Star AME IRG grant (A20E5c0092), the A*Star LCER-FI grant (LCERFI01-0015 U2102d2004), the Ministry of Education of Singapore, and the National University of Singapore (R-279-000-578-112, R-279-000-598-114, and R-279-000-574-114) for financial support.

## REFERENCES

(1) Furukawa, H.; Cordova, K. E.; O'Keeffe, M.; Yaghi, O. M. The Chemistry and Applications of Metal-Organic Frameworks. *Science* **2013**, *341*, 1230444.

(2) 2019 CSD Data Updates; https://www.ccdc.cam.ac.uk/discover/blog/2019-csd-data-updates/ (accessed March 16, 2023).
(3) Colón, Y. J.; Snurr, R. Q. High-Throughput Computational Screening of Metal-Organic Frameworks. *Chem. Soc. Rev.* **2014**, *43*, 5735−5749.
(4) Jiang, J. W. Computational Screening of Metal-Organic Frameworks for $CO_2$ Separation. *Curr. Opin. Green Sustain. Chem.* **2019**, *16*, 57−64.
(5) Ren, E.; Guilbaud, P.; Coudert, F. X. High-Throughput Computational Screening of Nanoporous Materials in Targeted Applications. *Digit. Discovery* **2022**, *1*, 355−374.
(6) Jordan, M. I.; Mitchell, T. M. Machine Learning: Trends, Perspectives and Prospects. *Science* **2015**, *349*, 255−260.
(7) Jablonka, K. M.; Ongari, D.; Moosavi, S. M.; Smit, B. Big-Data Science in Porous Materials: Materials Genomics and Machine Learning. *Chem. Rev.* **2020**, *120*, 8066−8129.
(8) Chong, S.; Lee, S.; Kim, B.; Kim, J. Applications of Machine Learning in Metal-Organic Frameworks. *Coord. Chem. Rev.* **2020**, *423*, 213487.
(9) Demir, H.; Daglar, H.; Gulbalkan, H. C.; Aksu, G. O.; Keskin, S. Recent Advances in Computational Modeling of MOFs: From Molecular Simulations to Machine Learning. *Coord. Chem. Rev.* **2023**, *484*, 215112.
(10) Moosavi, S. M.; Chidambaram, A.; Talirz, L.; Haranczyk, M.; Stylianou, K. C.; Smit, B. Capturing Chemical Intuition in Synthesis of Metal-Organic Frameworks. *Nat. Commun.* **2019**, *10*, 539.
(11) Moosavi, S. M.; Nandy, A.; Jablonka, K. M.; Ongari, D.; Janet, J. P.; Boyd, P. G.; Lee, Y.; Smit, B.; Kulik, H. J. Understanding the Diversity of the Metal-Organic Framework Ecosystem. *Nat. Commun.* **2020**, *11*, 4068.
(12) Batra, R.; Chen, C.; Evans, T. G.; Walton, K. S.; Ramprasad, R. Prediction of Water Stability of Metal-Organic Frameworks Using Machine Learning. *Nat. Mach. Intell.* **2020**, *2*, 704−710.
(13) Lim, Y.; Kim, J. Application of Transfer Learning to Predict Diffusion Properties in Metal-Organic Frameworks. *Mol. Syst. Des. Eng.* **2022**, *7*, 1056−1064.
(14) Yao, Z. P.; Sanchez-Lengeling, B.; Bobbitt, N. S.; Bucior, B. J.; Kumar, S. G. H.; Collins, S. P.; Burns, T.; Woo, T. K.; Farha, O. K.; Snurr, R. Q.; Aspuru-Guzik, A. Inverse Design of Nanoporous Crystalline Reticular Materials with Deep Generative Models. *Nat. Mach. Intell.* **2021**, *3*, 76−86.
(15) Lee, Y. J.; Barthel, S. D.; Dlotko, P.; Moosavi, S. M.; Hess, K.; Smit, B. Quantifying Similarity of Pore-Geometry in Nanoporous Materials. *Nat. Commun.* **2017**, *8*, 15396.
(16) Bucior, B. J.; Bobbitt, N. S.; Islamoglu, T.; Goswami, S.; Gopalan, A.; Yildirim, T.; Farha, O. K.; Bagheri, N.; Snurr, R. Q. Energy-Based Descriptors to Rapidly Predict Hydrogen Storage in Metal-Organic Frameworks. *Mol. Syst. Des. Eng.* **2019**, *4*, 162−174.
(17) Tang, H. J.; Xu, Q. S.; Wang, M.; Jiang, J. W. Rapid Screening of Metal-Organic Frameworks for Propane/Propylene Separation by Synergizing Molecular Simulation and Machine Learning. *ACS Appl. Mater. Interfaces* **2021**, *13*, 53454−53467.
(18) Tang, H. J.; Jiang, J. W. Active Learning Boosted Computational Discovery of Covalent-Organic Frameworks for Ultrahigh $CH_4$ Storage. *AIChE J.* **2022**, *68*, No. e17856.
(19) Chung, Y. G.; Camp, J.; Haranczyk, M.; Sikora, B. J.; Bury, W.; Krungleviciute, V.; Yildirim, T.; Farha, O. K.; Sholl, D. S.; Snurr, R. Q. Computation-Ready, Experimental Metal-Organic Frameworks: A Tool to Enable High-Throughput Screening of Nanoporous Crystals. *Chem. Mater.* **2014**, *26*, 6185−6192.
(20) Chung, Y. G.; Haldoupis, E.; Bucior, B. J.; Haranczyk, M.; Lee, S.; Zhang, H. D.; Vogiatzis, K. D.; Milisavljevic, M.; Ling, S. L.; Camp, J. S.; Slater, B.; Siepmann, J. I.; Sholl, D. S.; Snurr, R. Q. Advances, Updates and Analytics for the Computation-Ready, Experimental Metal-Organic Framework Database: CoRE MOF 2019. *J. Chem. Eng. Data* **2019**, *64*, 5985−5998.
(21) Li, A.; Perez, R. B.; Wiggin, S.; Ward, S. C.; Wood, P. A.; Fairen-Jimenez, D. The Launch of a Freely Accessible MOF CIF Collection from the CSD. *Matter* **2021**, *4*, 1105−1106.
(22) Wilmer, C. E.; Leaf, M.; Lee, C. Y.; Farha, O. K.; Hauser, B. G.; Hupp, J. T.; Snurr, R. Q. Large-Scale Screening of Hypothetical Metal-Organic Frameworks. *Nat. Chem.* **2012**, *4*, 83−89.
(23) Boyd, P. G.; Chidambaram, A.; García-Díez, E.; Ireland, C. P.; Daff, T. D.; Bounds, R.; Gładysiak, A.; Schouwink, P.; Moosavi, S. M.; Maroto-Valer, M. M. Data-Driven Design of Metal-Organic Frameworks for Wet Flue Gas $CO_2$ Capture. *Nature* **2019**, *576*, 253−256.
(24) Anderson, R.; Gomez-Gualdron, D. A. Increasing Topological Diversity During Computational "Synthesis" of Porous Crystals: How and Why. *CrystEngComm* **2019**, *21*, 1653−1665.
(25) Majumdar, S.; Moosavi, S. M.; Jablonka, K. M.; Ongari, D.; Smit, B. Diversifying Databases of Metal-Organic Frameworks for High-Throughput Computational Screening. *ACS Appl. Mater. Interfaces* **2021**, *13*, 61004−61014.
(26) Rosen, A. S.; Iyer, S. M.; Ray, D.; Yao, Z. P.; Aspuru-Guzik, A.; Gagliardi, L.; Notestein, J. M.; Snurr, R. Q. Machine Learning the Quantum-Chemical Properties of Metal-Organic Frameworks for Accelerated Materials Discovery. *Matter* **2021**, *4*, 1578−1597.
(27) Burner, J.; Luo, J.; White, A.; Mirmiran, A.; Kwon, O.; Boyd, P. G.; Maley, S.; Gibaldi, M.; Simrod, S.; Ogden, V.; Woo, T. K. ARC-MOF: A Diverse Database of Metal-Organic Frameworks with DFT-Derived Partial Atomic Charges and Descriptors for Machine Learning. *Chem. Mater.* **2023**, *35*, 900−916.
(28) Nandy, A.; Yue, S.; Oh, C.; Duan, C.; Terrones, G. G.; Chung, Y. G.; Kulik, H. J. A Database of Ultrastable MOFs Reassembled from Stable Fragments with Machine Learning Models. *Matter* **2023**, *6*, 1585−1603.
(29) NIST Adsorption Data Resources. https://github.com/nist-isodb (accessed March 16, 2023).
(30) Nandy, A.; Duan, C. R.; Kulik, H. J. Using Machine Learning and Data Mining to Leverage Community Knowledge for the Engineering of Stable Metal-Organic Frameworks. *J. Am. Chem. Soc.* **2021**, *143*, 17535−17547.
(31) Luo, Y.; Bag, S.; Zaremba, O.; Cierpka, A.; Andreo, J.; Wuttke, S.; Friederich, P.; Tsotsalas, M. MOF Synthesis Prediction Enabled by Automatic Data Mining and Machine Learning. *Angew. Chem., Int. Ed.* **2022**, *61*, No. e202200242.
(32) Glasby, L. T.; Gubsch, K.; Bence, R.; Oktavian, R.; Isoko, K.; Moosavi, S. M.; Cordiner, J. L.; Cole, J. C.; Moghadam, P. Z. DigiMOF: A Database of Metal-Organic Framework Synthesis Information Generated via Text Mining. *Chem. Mater.* **2023**, *35*, 4510−4524.
(33) Sturluson, A.; Raza, A.; McConachie, G. D.; Siderius, D. W.; Fern, X. Z.; Simon, C. M. Recommendation System to Predict Missing Adsorption Properties of Nanoporous Materials. *Chem. Mater.* **2021**, *33*, 7203−7216.
(34) Willems, T. F.; Rycroft, C.; Kazi, M.; Meza, J. C.; Haranczyk, M. Algorithms and Tools for High-Throughput Geometry-Based Analysis of Crystalline Porous Materials. *Microporous Mesoporous Mater.* **2012**, *149*, 134−141.
(35) Dubbeldam, D.; Calero, S.; Ellis, D. E.; Snurr, R. Q. Raspa: Molecular Simulation Software for Adsorption and Diffusion in Flexible Nanoporous Materials. *Mol. Simul.* **2016**, *42*, 81−101.
(36) Sarkisov, L.; Bueno-Perez, R.; Sutharson, M.; Fairen-Jimenez, D. Materials Informatics with Poreblazer V4.0 and the CSD MOF Database. *Chem. Mater.* **2020**, *32*, 9849−9867.
(37) Lee, Y. J.; Barthel, S. D.; Dlotko, P.; Moosavi, S. M.; Hess, K.; Smit, B. High-Throughput Screening Approach for Nanoporous Materials Genome Using Topological Data Analysis: Application to Zeolites. *J. Chem. Theory Comput.* **2018**, *14*, 4427−4437.
(38) Zhao, C. X.; Chen, L. J.; Che, Y.; Pang, Z. F.; Wu, X. F.; Lu, Y. X.; Liu, H. L.; Day, G. M.; Cooper, A. I. Digital Navigation of Energy-Structure-Function Maps for Hydrogen-Bonded Porous Molecular Crystals. *Nat. Commun.* **2021**, *12*, 817.
(39) Krishnapriyan, A. S.; Montoya, J.; Haranczyk, M.; Hummelshoj, J.; Morozov, D. Machine Learning with Persistent Homology and Chemical Word Embeddings Improves Prediction Accuracy and Interpretability in Metal-Organic Frameworks. *Sci. Rep.* **2021**, *11*, 8888.
(40) Fanourgakis, G. S.; Gkagkas, K.; Tylianakis, E.; Klontzas, E.; Froudakis, G. A Robust Machine Learning Algorithm for the Prediction


https://doi.org/10.1021/acs.langmuir.3c01964
*Langmuir* 2023, 39, 15849−15863

of Methane Adsorption in Nanoporous Materials. *J. Phys. Chem. A* **2019**, *123*, 6080−6087.
(41) Shi, Z. A.; Yang, W. Y.; Deng, X. M.; Cai, C. Z.; Yan, Y. L.; Liang, H.; Liu, Z. L.; Qiao, Z. W. Machine-Learning-Assisted High-Throughput Computational Screening of High Performance Metal-Organic Frameworks. *Mol. Syst. Des. Eng.* **2020**, *5*, 725−742.
(42) Zhang, Z. M.; Tang, H. J.; Wang, M.; Lyu, B.; Jiang, Z.; Jiang, J. W. Metal-Organic Frameworks for Water Harvesting: Machine Learning-Based Prediction and Rapid Screening. *ACS Sustain. Chem. Eng.* **2023**, *11*, 8148−8160.
(43) Janet, J. P.; Kulik, H. J. Resolving Transition Metal Chemical Space: Feature Selection for Machine Learning and Structure-Property Relationships. *J. Phys. Chem. A* **2017**, *121*, 8939−8954.
(44) Jablonka, K. M.; Moosavi, S. M.; Asgari, M.; Ireland, C.; Patiny, L.; Smit, B. A Data-Driven Perspective on the Colours of Metal-Organic Frameworks. *Chem. Sci.* **2021**, *12*, 3587−3598.
(45) Jablonka, K. M.; Ongari, D.; Moosavi, S. M.; Smit, B. Using Collective Knowledge to Assign Oxidation States of Metal Cations in Metal-Organic Frameworks. *Nat. Chem.* **2021**, *13*, 771−777.
(46) Fernandez, M.; Trefiak, N. R.; Woo, T. K. Atomic Property Weighted Radial Distribution Functions Descriptors of Metal-Organic Frameworks for the Prediction of Gas Uptake Capacity. *J. Phys. Chem. C* **2013**, *117*, 14095−14105.
(47) Xie, T.; Grossman, J. C. Crystal Graph Convolutional Neural Networks for an Accurate and Interpretable Prediction of Material Properties. *Phys. Rev. Lett.* **2018**, *120*, 145301.
(48) Fung, V.; Zhang, J. X.; Juarez, E.; Sumpter, B. G. Benchmarking Graph Neural Networks for Materials Chemistry. *npj Comput. Mater.* **2021**, *7*, 84.
(49) Raza, A.; Sturluson, A.; Simon, C. M.; Fern, X. Message Passing Neural Networks for Partial Charge Assignment to Metal-Organic Frameworks. *J. Phys. Chem. C* **2020**, *124*, 19070−19082.
(50) Wang, R. H.; Zou, Y. R.; Zhang, C. C.; Wang, X.; Yang, M. L.; Xu, D. G. Combining Crystal Graphs and Domain Knowledge in Machine Learning to Predict Metal-Organic Frameworks Performance in Methane Adsorption. *Microporous Mesoporous Mater.* **2022**, *331*, 111666.
(51) Zhang, B. Y.; Zhou, M. S.; Wu, J. Z.; Gao, F. C. Predicting the Materials Properties Using a 3D Graph Neural Network with Invariant Representation. *IEEE Access* **2022**, *10*, 62440−62449.
(52) Tshitoyan, V.; Dagdelen, J.; Weston, L.; Dunn, A.; Rong, Z. Q.; Kononova, O.; Persson, K. A.; Ceder, G.; Jain, A. Unsupervised Word Embeddings Capture Latent Knowledge from Materials Science Literature. *Nature* **2019**, *571*, 95−98.
(53) Lee, S.; Kim, B.; Cho, H.; Lee, H.; Lee, S. Y.; Cho, E. S.; Kim, J. Computational Screening of Trillions of Metal-Organic Frameworks for High-Performance Methane Storage. *ACS Appl. Mater. Interfaces* **2021**, *13*, 23647−23654.
(54) Cao, Z.; Magar, R.; Wang, Y.; Barati Farimani, A. MOFormer: Self-Supervised Transformer Model for Metal-Organic Framework Property Prediction. *J. Am. Chem. Soc.* **2023**, *145*, 2958−2967.
(55) Zhang, X.; Sethi, S.; Wang, Z. H.; Zhou, T.; Qi, Z. W.; Sundmacher, K. A Neural Recommender System for Efficient Adsorbent Screening. *Chem. Eng. Sci.* **2022**, *259*, 117801.
(56) Escobar-Hernandez, H. U.; Perez, L. M.; Hu, P. F.; Soto, F. A.; Papadaki, M. I.; Zhou, H. C.; Wang, Q. S. Thermal Stability of Metal-Organic Frameworks: Concept, Determination, and Model Prediction Using Computational Chemistry and Machine Learning. *Ind. Eng. Chem. Res.* **2022**, *61*, 5853−5862.
(57) Cho, E. H.; Deng, X. P.; Zou, C. L.; Lin, L. C. Machine Learning-Aided Computational Study of Metal-Organic Frameworks for Sour Gas Sweetening. *J. Phys. Chem. C* **2020**, *124*, 27580−27591.
(58) Wold, S.; Esbensen, K.; Geladi, P. Principal Component Analysis. *Chemom. Intell. Lab. Syst.* **1987**, *2*, 37−52.
(59) van der Maaten, L.; Hinton, G. Visualizing Data Using t-SNE. *J. Mach. Learn. Res.* **2008**, *9*, 2579−2605.
(60) Becht, E.; McInnes, L.; Healy, J.; Dutertre, C. A.; Kwok, I. W. H.; Ng, L. G.; Ginhoux, F.; Newell, E. W. Dimensionality Reduction for Visualizing Single-Cell Data Using UMAP. *Nat. Biotechnol.* **2019**, *37*, 38−44.
(61) Wu, Y.; Duan, H. P.; Xi, H. X. Machine Learning-Driven Insights into Defects of Zirconium Metal- Organic Frameworks for Enhanced Ethane-Ethylene Separation. *Chem. Mater.* **2020**, *32*, 2986−2997.
(62) Pardakhti, M.; Moharreri, E.; Wanik, D.; Suib, S. L.; Srivastava, R. Machine Learning Using Combined Structural and Chemical Descriptors for Prediction of Methane Adsorption Performance of Metal-Organic Frameworks. *ACS Comb. Sci.* **2017**, *19*, 640−645.
(63) Anderson, R.; Rodgers, J.; Argueta, E.; Biong, A.; Gomez-Gualdron, D. A. Role of Pore Chemistry and Topology in the $CO_2$ Capture Capabilities of MOFs: From Molecular Simulation to Machine Learning. *Chem. Mater.* **2018**, *30*, 6325−6337.
(64) Fanourgakis, G. S.; Gkagkas, K.; Froudakis, G. Introducing Artificial MOFs for Improved Machine Learning Predictions: Identification of Top-Performing Materials for Methane Storage. *J. Chem. Phys.* **2022**, *156*, 054103.
(65) Fernandez, M.; Barnard, A. S. Geometrical Properties Can Predict $CO_2$ and $N_2$ Adsorption Performance of Metal-Organic Frameworks at Low Pressure. *ACS Comb. Sci.* **2016**, *18*, 243−252.
(66) Jablonka, K. M.; Rosen, A. S.; Krishnapriyan, A. S.; Smit, B. An Ecosystem for Digital Reticular Chemistry. *ACS Central Sci.* **2023**, *9*, 563−581.
(67) Korolev, V. V.; Mitrofanov, A.; Marchenko, E. I.; Eremin, N. N.; Tkachenko, V.; Kalmykov, S. N. Transferable and Extensible Machine Learning-Derived Atomic Charges for Modeling Hybrid Nanoporous Materials. *Chem. Mater.* **2020**, *32*, 7822−7831.
(68) Mukherjee, K.; Colon, Y. J. Machine Learning and Descriptor Selection for the Computational Discovery of Metal-Organic Frameworks. *Mol. Simul.* **2021**, *47*, 857−877.
(69) Krokidas, P.; Karozis, S.; Moncho, S.; Giannakopoulos, G.; Brothers, E. N.; Kainourgiakis, M. E.; Economou, I. G.; Steriotis, T. A. Data Mining for Predicting Gas Diffusivity in Zeolitic-Imidazolate Frameworks. *J. Mater. Chem. A* **2022**, *10*, 13697−13703.
(70) Daglar, H.; Keskin, S. Combining Machine Learning and Molecular Simulations to Unlock Gas Separation Potentials of Mof Membranes and MOF/Polymer MMMs. *ACS Appl. Mater. Interfaces* **2022**, *14*, 32134−32148.
(71) Datar, A.; Chung, Y. G.; Lin, L. C. Beyond the BET Analysis: The Surface Area Prediction of Nanoporous Materials Using a Machine Learning Method. *J. Phys. Chem. Lett.* **2020**, *11*, 5412−5417.
(72) Petuya, R.; Durdy, S.; Antypov, D.; Gaultois, M. W.; Berry, N. G.; Darling, G. R.; Katsoulidis, A. P.; Dyer, M. S.; Rosseinsky, M. J. Machine-Learning Prediction of Metal-Organic Framework Guest Accessibility from Linker and Metal Chemistry. *Angew. Chem., Int. Ed.* **2022**, *61*, No. e202114573.
(73) Kancharlapalli, S.; Gopalan, A.; Haranczyk, M.; Snurr, R. Q. Fast and Accurate Machine Learning Strategy for Calculating Partial Atomic Charges in Metal-Organic Frameworks. *J. Chem. Theory Comput.* **2021**, *17*, 3052−3064.
(74) He, Y. P.; Cubuk, E. D.; Allendorf, M. D.; Reed, E. J. Metallic Metal-Organic Frameworks Predicted by the Combination of Machine Learning Methods and Ab Initio Calculations. *J. Phys. Chem. Lett.* **2018**, *9*, 4562−4569.
(75) Moghadam, P. Z.; Rogge, S. M. J.; Li, A.; Chow, C. M.; Wieme, J.; Moharrami, N.; Aragones-Anglada, M.; Conduit, G.; Gomez-Gualdron, D. A.; Van Speybroeck, V.; Fairen-Jimenez, D. Structure-Mechanical Stability Relations of Metal-Organic Frameworks via Machine Learning. *Matter* **2019**, *1*, 219−234.
(76) Nandy, A.; Terrones, G.; Arunachalam, N.; Duan, C.; Kastner, D. W.; Kulik, H. J. MOFSimplify, Machine Learning Models with Extracted Stability Data of Three Thousand Metal-Organic Frameworks. *Sci. Data* **2022**, *9*, 74.
(77) Chen, P. C.; Tang, Z. Y.; Zeng, Z. M.; Hu, X. F.; Xiao, L. P.; Liu, Y.; Qian, X. D.; Deng, C. Y.; Huang, R. Y.; Zhang, J. Z.; Bi, Y. L.; Lin, R. K.; Zhou, Y.; Liao, H. G.; Zhou, D.; Wang, C.; Lin, W. B. Machine-Learning-Guided Morphology Engineering of Nanoscale Metal-Organic Frameworks. *Matter* **2020**, *2*, 1651−1666.


https://doi.org/10.1021/acs.langmuir.3c01964
*Langmuir* 2023, 39, 15849−15863

(78) Domingues, N. P.; Moosavi, S. M.; Talirz, L.; Jablonka, K. M.; Ireland, C. P.; Ebrahim, F. M.; Smit, B. Using Genetic Algorithms to Systematically Improve the Synthesis Conditions of Al-PMOF. *Commun. Chem.* **2022**, *5*, 170.
(79) Kitamura, Y.; Nakamura, Y.; Sugimoto, K.; Yoshikawa, H.; Tanaka, D. Data-Driven Efficient Synthetic Exploration of Anionic Lanthanide-Based Metal-Organic Frameworks. *Chem. Commun.* **2022**, *58*, 11426.
(80) Hu, J.; Gui, J.; Gao, B.; Yang, L.; Ding, Q.; Li, Y.; Mo, Y.; Chen, H.; Cui, X.; Xing, H. Machine-Learning-Assisted Exploration of Anion-Pillared Metal-Organic Frameworks for Gas Separation. *Matter* **2022**, *5*, 3901−3911.
(81) Scheffler, M.; Aeschlimann, M.; Albrecht, M.; Bereau, T.; Bungartz, H. J.; Felser, C.; Greiner, M.; Gross, A.; Koch, C. T.; Kremer, K.; Nagel, W. E.; Scheidgen, M.; Woll, C.; Draxl, C. Fair Data Enabling New Horizons for Materials Research. *Nature* **2022**, *604*, 635−642.
(82) Jain, A.; Ong, S. P.; Hautier, G.; Chen, W.; Richards, W. D.; Dacek, S.; Cholia, S.; Gunter, D.; Skinner, D.; Ceder, G.; Persson, K. A. Commentary: The Materials Project: A Materials Genome Approach to Accelerating Materials Innovation. *APL Mater.* **2013**, *1*, 011002.
(83) Talirz, L.; Kumbhar, S.; Passaro, E.; Yakutovich, A. V.; Granata, V.; Gargiulo, F.; Borelli, M.; Uhrin, M.; Huber, S. P.; Zoupanos, S.; Adorf, C. S.; Andersen, C. W.; Schutt, O.; Pignedoli, C. A.; Passerone, D.; VandeVondele, J.; Schulthess, T. C.; Smit, B.; Pizzi, G.; Marzari, N. Materials Cloud: A Platform for Open Computational Science. *Sci. Data* **2020**, *7*, 299.
(84) Park, J.; Howe, J. D.; Sholl, D. S. How Reproducible Are Isotherm Measurements in Metal-Organic Frameworks? *Chem. Mater.* **2017**, *29*, 10487−10495.
(85) Ongari, D.; Talirz, L.; Jablonka, K. M.; Siderius, D. W.; Smit, B. Data-Driven Matching of Experimental Crystal Structures and Gas Adsorption Isotherms of Metal-Organic Frameworks. *J. Chem. Eng. Data* **2022**, *67*, 1743−1756.
(86) Evans, J. D.; Bon, V.; Senkovska, I.; Kaskel, S. A Universal Standard Archive File for Adsorption Data. *Langmuir* **2021**, *37*, 4222−4226.
(87) Boyd, P. G.; Moosavi, S. M.; Witman, M.; Smit, B. Force-Field Prediction of Materials Properties in Metal-Organic Frameworks. *J. Phys. Chem. Lett.* **2017**, *8*, 357−363.
(88) Ioannidis, E. I.; Gani, T. Z. H.; Kulik, H. J. molSimplify: A Toolkit for Automating Discovery in Inorganic Chemistry. *J. Comput. Chem.* **2016**, *37*, 2106−2117.
(89) Bucior, B. J.; Rosen, A. S.; Haranczyk, M.; Yao, Z.; Ziebel, M. E.; Farha, O. K.; Hupp, J. T.; Siepmann, J. I.; Aspuru-Guzik, A.; Snurr, R. Q. Identification Schemes for Metal-Organic Frameworks to Enable Rapid Search and Cheminformatics Analysis. *Cryst. Growth Des.* **2019**, *19*, 6682−6697.
(90) Sun, Y.; DeJaco, R. F.; Li, Z.; Tang, D.; Glante, S.; Sholl, D. S.; Colina, C. M.; Snurr, R. Q.; Thommes, M.; Hartmann, M.; Siepmann, J. I. Fingerprinting Diverse Nanoporous Materials for Optimal Hydrogen Storage Conditions Using Meta-Learning. *Sci. Adv.* **2021**, *7*, No. eabg3983.
(91) Xie, Y. C.; Zhang, C.; Deng, H.; Zheng, B. J. D.; Su, J. W.; Shutt, K.; Lin, J. Accelerate Synthesis of Metal-Organic Frameworks by a Robotic Platform and Bayesian Optimization. *ACS Appl. Mater. Interfaces* **2021**, *13*, 53485−53491.
(92) Daulton, S.; Balandat, M.; Bakshy, E. Differentiable Expected Hypervolume Improvement for Parallel Multi-Objective Bayesian Optimization. *Adv. Neural Inf. Process.* **2020**, *33*, 9851−9864.
(93) Guan, J.; Huang, T.; Liu, W.; Feng, F.; Japip, S.; Li, J.; Wang, X.; Zhang, S. Design and Prediction of Metal-Organic Framework-Based Mixed Matrix Membranes for $CO_2$ Capture via Machine Learning. *Cell Rep. Phys. Sci.* **2022**, *3*, 100864.
(94) Kang, Y. H.; Park, H.; Smit, B.; Kim, J. A Multi-Modal Pre-Training Transformer for Universal Transfer Learning in Metal-Organic Frameworks. *Nat. Mach. Intell.* **2023**, *5*, 309−318.
(95) Eckhoff, M.; Behler, J. From Molecular Fragments to the Bulk: Development of a Neural Network Potential for MOF-5. *J. Chem. Theory Comput.* **2019**, *15*, 3793−3809.
(96) Zheng, B.; Oliveira, F. L.; Neumann Barros Ferreira, R.; Steiner, M.; Hamann, H.; Gu, G. X.; Luan, B. Quantum Informed Machine-Learning Potentials for Molecular Dynamics Simulations of $CO_2$'s Chemisorption and Diffusion in Mg-MOF-74. *ACS Nano* **2023**, *17*, 5579−5587.
(97) Vandenhaute, S.; Cools-Ceuppens, M.; DeKeyser, S.; Verstraelen, T.; Van Speybroeck, V. Machine Learning Potentials for Metal-Organic Frameworks Using an Incremental Learning Approach. *npj Comput. Mater.* **2023**, *9*, 19.
(98) Huang, J.; Shin, S. J.; Tolborg, K.; Ganose, A. M.; Krenzer, G.; Walsh, A. Room-Temperature Stacking Disorder in Layered Covalent-Organic Frameworks from Machine-Learning Force Fields. *Mater. Horizons* **2023**, *10*, 2883−2891.
(99) Jinnouchi, R.; Karsai, F.; Kresse, G. On-the-Fly Machine Learning Force Field Generation: Application to Melting Points. *Phys. Rev. B* **2019**, *100*, 014105.


https://doi.org/10.1021/acs.langmuir.3c01964
*Langmuir* 2023, 39, 15849−15863