Journal of CO2 Utilization 89 (2024) 102941

Contents lists available at ScienceDirect

Journal of CO2 Utilization

journal homepage: www.elsevier.com/locate/jcou

Review article

Machine learning of metal-organic framework design for carbon dioxide capture and utilization

Yang Jeong Park [a,b], Sungroh Yoon [b,c,*], Sung Eun Jerng [d,**]

[a] Department of Materials Science and Engineering, Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, 02139, USA
[b] Department of Electrical and Computer Engineering, Seoul National University, Seoul, 08826, Republic of Korea
[c] Interdisciplinary Program in Artificial Intelligence, Seoul National University, Seoul, 08826, Republic of Korea
[d] Department of Environmental Energy Engineering, The University of Suwon, 17, Wauan-gil, Bongdam-eup, Hwaseong-si, Gyeonggi-do, 18323, Republic of Korea

ARTICLE INFO

Keywords:
Metal-organic framework
Carbon capture
Machine learning
High-throughput screening
Generative model

ABSTRACT

Metal-organic frameworks (MOFs) are attractive materials with easily tunable porous structures. Their selective carbon dioxide (CO_2) capture ability can be varied by altering the functionality of the organic ligands. However, rule-based approaches to tuning and developing MOFs with high CO_2 capture and conversion abilities are hindered by the numerous possible combinations of metal ions and organic linkers. Recently, machine learning (ML) has been applied to unravel key descriptors in predicting the performance of MOFs. This review summarizes recent advancements in ML models for MOFs in CO_2 capture and utilization, including high-throughput screening, neural network interatomic potential, and generative models. The development of sophisticated ML models for designing high-performance MOFs will play a critical role in addressing climate change in the future. Finally, the main challenges and limitations of current approaches in designing high-performance MOFs are discussed.

1. Introduction

The atmospheric carbon dioxide (CO_2) level has been identified as the major driver for the era ofglobal boiling. Thus, reducing the CO_2 level has become the most urgent issue to resolve for the future of human race. As a result, the carbon capture, utilization, and storage (CCUS) research field has been remarkably growing [1,2]. Particularly, porous materials such as metal-organic frameworks (MOFs) [3], zeolites [4], and microporous organic polymers (MOPs) [5] have been widely investigated for capturing CO_2 via adsorption and absorption.

Among various porous materials, metal-organic frameworks are composed of metal ions linked by molecular building units to form a reticular structure. MOFs have fascinated materials scientists due to their ability to regulate the pore size and structures on a multi-dimensional scale with large surface area [6,7]. Additionally, through simple and facile synthesis methods, the functionality of MOFs can be tuned by varying the organic linkers [8]. Moreover, MOFs can also be further functionalized via defect engineering [9], linker exchange [10], and mixing metal ions/linkers [11], among other methods. Due to the advantages of MOFs' tunable pore sizes, structure, chemical functionality, and combination with other materials, MOFs have been extensively applied in various energy devices such as lithium-ion batteries [12], gas storage [13,14] and separations [15], catalysis [16], supercapacitors [17], and more. Specifically, MOFs have been applied to selectively capture CO_2 gas from mixed gases [3,18,19]. Furthermore, MOFs have demonstrated excellent performance in electrochemically converting CO_2 into valuable products [20].

However, designing and developing effective MOF is time and labor-intensive due to slow synthesis kinetics, trial-and-error based testing, and variability in reported performance under non-standardized test conditions, which hinder the clarification of scientific performance descriptors [21]. The countless possible MOF structures, achieved by changing metal ions and organic linkers, further complicate this process [22]. Therefore, high-throughput traditional calculations are impractical for screening desired MOF candidates. In this context, machine learning (ML) can play a crucial role in unraveling the relation between MOF structure and performance by screening thousands of compounds in seconds prior to synthesis and test their performances [23,24].

In this review, we briefly discuss applications of artificial intelligence (AI) and ML in MOF discovery for carbon capture and guide readers

* Corresponding author at: Department of Electrical and Computer Engineering, Seoul National University, Seoul, 08826, Republic of Korea.
** Corresponding author.
E-mail addresses: sryoon@snu.ac.kr (S. Yoon), sejerng@suwon.ac.kr (S.E. Jerng).

https://doi.org/10.1016/j.jcou.2024.102941
Received 1 August 2024; Received in revised form 16 September 2024; Accepted 23 September 2024
Available online 21 October 2024
2212-9820/© 2024 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY-NC license (http://creativecommons.org/licenses/by-nc/4.0/).

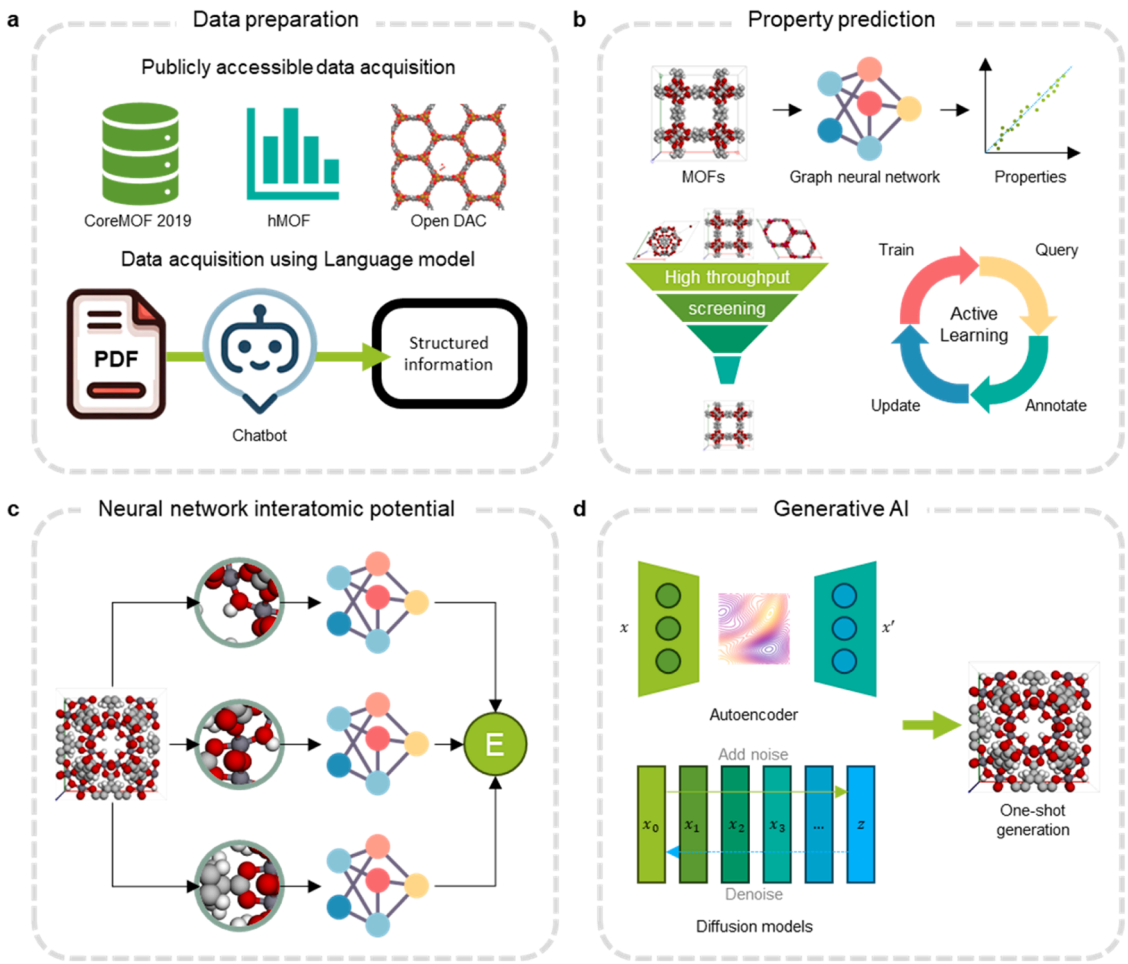

Fig. 1. ML workflows for MOF discovery. (a) Data preparation for training ML models. (b) Property prediction models for high-throughput screening. Active learning for achieving data efficiency also relies on predictive ML models. (c) Accelerating molecular simulations using ML interatomic potential. (d) Expanding chemical space to explore using generative AI.

interested in applying state-of-the-art techniques. We cover the combination of AI and ML techniques with classical quantum chemical simulations or state-of-the-art techniques for capturing structure-performance relationships. In Section 2, we introduce the type of contributions of ML workflows. In Section 3, we present available resources for ML applications. In Section 4, we discuss ML applications for predicting carbon capture properties of MOFs. Methods for accelerating molecular simulations to obtain interpretable and reliable results rather than relying on the black box of AI, are covered in Section 5. Section 6 reviews studies that expand the MOF space that can be explored with generative models. Finally, we outline the directions we consider most important for further progress by addressing the current opportunities and challenges in this field.

2. Machine learning workflows

ML is a subset of AI that involves training algorithms to identify patterns in data and make predictions without explicit programming. It leverages statistical techniques to enable machines to improve their performance on a task through experience. Key types of ML include supervised learning (learning from labeled data), unsupervised learning (finding hidden patterns in unlabeled data), and reinforcement learning (learning through rewards and punishments).

Basically, supervised learning aims to find a function that best describes the relationship between data and labels. Therefore, the most basic and important task is to collect high-quality data to train the model. These data can be divided into computational data based on quantum chemical simulations and experimental data. Traditionally, human scientists analyzed literatures manually and extracted experimental data, but today, chatbots can be used to automatically extract structured data from literatures. The most common task of ML applications for MOF discovery is predicting the properties of a given MOF by developing quantitative structure-property relationships. Trained property prediction models can be applied as filters in large-scale screening tasks. It can also be used in an active learning strategy to reduce data labeling costs by quantifying uncertainties in ML predictions. Next, an ML model can be trained to predict potential energy surface using the results of quantum chemical calculations such as density functional theory (DFT) as labels. A model trained in this way can perform a variety of simulations at a much faster rate than DFT. Finally, generative AI models expand the explorable candidate space to discover new MOFs to the entire chemical space instead of relying on existing databases. These major research streams are shown in Fig. 1.

3. Data preparation

3.1. Available databases

The quantity and quality of data are key factors in the discovery of materials. Data can be collected from available databases (Table 1) or published papers. The database includes various types of data, which originates from experiments, and simulations.

Table 1
Publicly accessible databases of metal-organic frameworks.

Name	Datasize	Description
hMOF [25]	137,953 MOFs	Hypothetical MOFs were generated from a building block library and surface area, pore-size distribution, and methane-storage capacity were calculated.
CSD MOF subset [28]	69,666 MOFs	A collection including 69,666 MOFs from a subset of Cambridge Crystallographic Data Centre (CCDC). Geometric and physical properties were calculated.
Boyd Materials Cloud [29]	324,426 MOFs	A library of hypothetical MOFs screened for CO_2/N_2 selectivity and CO_2 working capacity, with accurate electrostatic potential representation.
CoRE MOF [27]	14,142 MOFs	An expanded database of computation-ready, experimental MOFs derived from experimental data suitable for molecular simulations
QMOF [30]	15,713 MOFs	experimentally characterized MOFs after structure relaxation via DFT, including but not limited to optimized geometries, energies, band gaps, charge densities, density of states, partial charges, spin densities, and bond orders.
Open DAC [31]	38 M DFT calculations from 170 K DFT relaxations	More than 38 M DFT calculations on more than 8412 MOF materials containing adsorbed CO_2 and/or H_2O for direct air capture.
ARC-MOF [32]	279,610 MOFs	A database of 279,610 MOFs which have been either experimentally characterized or computationally generated, spanning all publicly available MOF databases.
DigiMOF [33]	15,510 MOFs	Generated by adapting the chemistry-aware natural language processing tool, ChemDataExtractor, extracting data from unique MOF journal articles.

Due to the continued development of quantum chemical simulation technology and the demand for machine learning research, efforts to secure large-scale data have continued. hMOF [25] is one of early efforts. 137,953 hypothetical MOFs were generated from a library of 102 building blocks and properties such as the surface area, pore-size distribution, and methane-storage capacity were calculated for each one using Grand Canonical Monte Carlo (GCMC). Chung et al. [26] developed a database of 4764 MOF structures based on experimental data, designed to be immediately applicable for molecular simulations, known as the computation-ready, experimental (CoRE) MOF database. CoRE MOF [27] was expanded to include 14,142 MOFs through follow-up studies. Moghadam et al. [28] reported a collection including 69,666 MOFs from subset of Cambridge Crystallographic Data Centre (CCDC). After residual solvent removal via CSD Python API, Geometric and physical properties were calculated such as surface area, pore volume, pore limiting diameter (PLD), the largest cavity diameter (LCD), void fraction, and density. Boyd et al. [29] generated a library of 324, 426 hypothetical MOFs and perform screening each MOF for its CO_2/N_2 selectivity and its CO_2 working capacity. Using a method to assemble these materials with topological blueprints, only materials that could be accurately represented with the MEPO-QEq charge generation method are selected. By ensuring that the electrostatic potential is accurately represented in these materials, screening for CO_2 adsorption properties would result in very few false positives/negatives.

Rosen et al. [30] introduce the Quantum MOF (QMOF) database, a publicly available database of computed quantum-chemical properties, such as those based on the electronic, optical, magnetic, and/or catalytic properties of MOFs, for 14,482 experimentally synthesized MOFs. Sriram et al. [31] introduced a dataset named Open DAC 2023 (ODAC23) consisting of more than 38 M DFT calculations on more than 8412 MOF materials containing adsorbed CO_2 and/or H_2O for direct air capture (DAC). Burner et al. [32] introduces the ab initio REPEAT charge MOF (ARC–MOF) database of 279,610 MOFs which have been either experimentally characterized or computationally generated, spanning all publicly available MOF databases. ARC–MOF contains both experimentally characterized and hMOF structures taken from several different sources. ARC–MOF sufficiently spans the overall chemical space, and that it is sufficiently balanced with respect to geometric properties, as well as ligand chemistry. However, ARC–MOF suffers from being highly unbalanced with respect to metal chemistry, a well-known flaw of current hMOF databases. Glasby et al. [33] generated the DigiMOF database by adapting the chemistry-aware natural language processing tool, ChemDataExtractor (CDE). Using the CDE web scraping package alongside the Cambridge Structural Database (CSD) MOF subset, they automatically downloaded 43,281 unique MOF journal articles, extracted 15,501 unique MOF materials, and text-mined over 52,680 associated properties including the synthesis method, solvent, organic linker, metal precursor, and topology.

Overall, scientists now have access to several publicly available databases containing more than hundreds of thousands of MOFs. Understanding the history of research efforts to build a useful MOF database is essential for developing new ML models. An appropriate database must be selected to identify MOFs that can achieve effective carbon uptake. It also remains challenges to understand the biases remaining in public databases and to build new ones to eliminate them.

3.2. Performance metrics

In supervised ML, evaluating the performance of regression models is crucial to understand how well a model predicts continuous outcomes. Common performance metrics include Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and the coefficient of determination (R^2).

3.2.1. Root Mean Square Error (RMSE)

RMSE measures the square root of the average squared differences between predicted values ($\hat{y}_i$) and actual values (y_i). It is sensitive to large errors, making it useful when larger errors are particularly undesirable. The formula for RMSE is:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}$$

where n is the number of observations.

3.2.2. Mean Absolute Error (MAE)

MAE calculates the average of the absolute differences between predicted and actual values. Unlike RMSE, it does not square the errors, so it provides a linear score that is easy to interpret. The formula for MAE is:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|\hat{y}_i - y_i|$$

MAE is less sensitive to outliers compared to RMSE, making it a better choice when all errors are equally important.

3.2.3. Coefficient of determination (R^2)

R^2 indicates the proportion of the variance in the dependent variable that is predictable from the independent variables. It provides a measure of how well the model's predictions approximate the actual data points. The formula for R^2 is:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{\sum_{i=1}^{n}(\bar{y} - y_i)^2}$$

where $\bar{y}$ is the mean of the actual values. An R^2 value closer to 1 indicates

a better fit of the model to the data, while an R^2 value of 0 means the model does not explain any of the variability in the target variable.

Although RMSE and MAE are frequently used to evaluate model predictions, these metrics alone may not adequately represent the model's generalization capability. In such cases, including the R^2 metric can provide a more comprehensive evaluation of the model's performance.

4. Prediction of CO_2 capture and conversion property of MOFs

4.1. Traditional ML

There are some advantages of traditional ML algorithms over DL. First, traditional MLs can perform effectively with relatively small amount of data while DL often requires large datasets. And traditional MLs are generally easier to interpret. For example, decision trees and linear regression provide clear insights into how the model makes decisions while DL models considered black box. Traditional ML can easily integrate domain knowledge through feature engineering. Moreover, traditional ML models are often simpler to implement and manage. These advantages make traditional ML models more favorable in scenarios with experimental data, which is expensive and limited, when model interpretability is required, or in environments with constrained computational resources. In this review, we will focus on use cases rather than introducing each traditional ML algorithm.

Due to the above advantages, many studies using traditional ML tend to follow the following sequence: 1) Acquire experimental data, 2) develop features based on domain knowledge, 3) develop and compare multiple models instead of a single ML model, and 4) identify variables that have a significant impact on their predictions. This process provides low-barrier accessibility to experiment-oriented researchers with limited computing resources while identifying important features and providing insight into future experiment design.

Bailey et al. [34] implemented a linear model, support vector machine(SVM), decision trees(DT) and gradient boosted decision trees (GBDTs) to predict the gas uptake capability of the MOFs. They collected 589 data from the literatures in which the CO_2, H_2 and CH_4 uptake datapoints were 268, 205, and 115 each. Among the tested models, the GBDTs demonstrated a superior average R^2 value of 0.86. Also, they identified temperature, gas type and pressure as the most important descriptors in the 51 descriptors.

Gheytanzadeh et al. [35] compared Gaussian process regression (GPR) models with various kernel functions such as Matern, Exponential, Squared exponential, and Rational quadratic kernel functions. By training with 506 experimental data of MOFs from literatures, the GPR model based on exponential kernel function presented the most accurate CO_2 uptake predictions. In addition, they evaluated that the pressure and the surface area of MOF are the most critical factors in the sensitivity analysis.

Tsamardinos et al. [36] utilized hMOF database of Snurr group together with GCMC simulations to calculate the carbon dioxide and methane adsorption capacity. They implemented Random Forest with a Just Add Data (JAD) tool to improve the carbon dioxide and methane adsorption prediction. This tool showed accuracy in prediction with a low number of datasets such as 50 data sets.

Li et al. [37] predicted CO_2 adsorption capacity of MOFs using random forest, gradient boosting decision tree, light gradient boosting machine, and eXtreme gradient boosting machine. By training the models with 348 data points from literature, the Random Forest showed best performance with R^2 value of 0.97. Also, the pressure and temperature had the most influential effects in determining the CO_2 adsorption capability. (73 %)

Ozsoysal et al. [38] developed a random forest regression model to predict the photocatalytic CO_2 reduction performance of MOFs. Their model was trained to predict the CO_2 reduced liquid and gas products such as CH_3OH, CO, H_2 and HCOOH. By applying 605 datasets extracted from the published papers, the R^2 values for training, validating, and testing were 0.96, 0.94 and 0.60. Their model suggested that the reactor volume and the volumetric percentage of the catalysts were crucial features in enhancing total gas production rate.

Abdi et al. [39] developed Categorical Boosting (CatBoost), Light Gradient Boosting Machine (LightGBM), Extreme Gradient Boosting (XGBoost), and Random Forest models to predict CO_2 adsorption capability of MOFs. They applied 1191 datasets in various temperatures and pressure extracted from literatures. In addition, the influence of pressure, temperature, specific surface area, and pore volume on the CO_2 adsorption capacity of MOF was evaluated. From their results, the pressure and the specific surface area are the most crucial features in the performance of MOFs. Also, the temperature was negatively related to the CO_2 adsorption capacity. The XGBoost model demonstrated the most accurate prediction with a low RMSE value of 0.5682 as well as a high R^2 value of 0.9955.

Ma et al. [40] utilized random forest to unravel the influential features and predict the CO_2 adsorption capacity of porous carbons. They collected data of various porous carbons such as MOFs, porous organic polymers, biomass and organic salts to train the model and verified by analyzing performance of UC800, ZNC650(ZIF-8), and NPC600. The R^2 of their random forest model was over 0.97. Their discovery showed that the nitrogen groups in the porous carbon affect the CO_2 adsorption capacity significantly at 0–0.15 bar while the micropores are crucial at 0.15–1 bar.

However, there is a famous conventional wisdom in data-driven research: "garbage in garbage out". While choosing the right ML model is crucial, the effectiveness of any ML algorithm fundamentally depends on the quality of the input data. Well-prepared and cleaned data can significantly enhance the performance of even simple models, whereas poor-quality data can lead to inaccurate predictions and unreliable results, regardless of the complexity of the model used. Data cleaning involves removing inconsistencies, handling missing values, and ensuring the data is as accurate as possible. This process helps in uncovering hidden patterns and insights that are critical for ML model training.

4.2. Deep learning

Traditional ML is still powerful, but recently deep learning using deep neural networks has been in the spotlight for the following reasons. Model performance improves when large amounts of data are accessible and provides better results than traditional ML [41]. Advances in parallel processing technology using GPUs have made it possible to train powerful models with performance from massive amounts of data [42]. Additionally, features can be automatically extracted from raw data input, bypassing feature engineering that requires expert knowledge [43]. In this section, we survey studies that have applied different types of deep neural networks.

4.2.1. Graph neural network

Graph neural networks (GNNs) have seen increasing applications in the field of chemistry over the past few years. A graph is a data structure consisting of nodes (also called vertices) and edges (links between nodes), making it useful to represent relationships and interactions within a set of entities. Deep neural networks are designed to pass information from these nodes in the form of messages through the edge, and the structure with this mechanism is called a message passing neural network (MPNN) [44]. These structures have proven successful in encoding material information and predicting properties, not only in molecules but also in crystalline materials [45]. Choudhary et al. [46] developed a model to predict the CO_2 adsorption capability of MOF based on Atomistic Line Graph Neural Network (ALIGNN) method. In their study, a large-scale dataset consisting of 137,953 simulation-based hMOFs was utilized to train the model to fully utilize the potential of the DL model. The developed DL model was used to calculate the CO_2

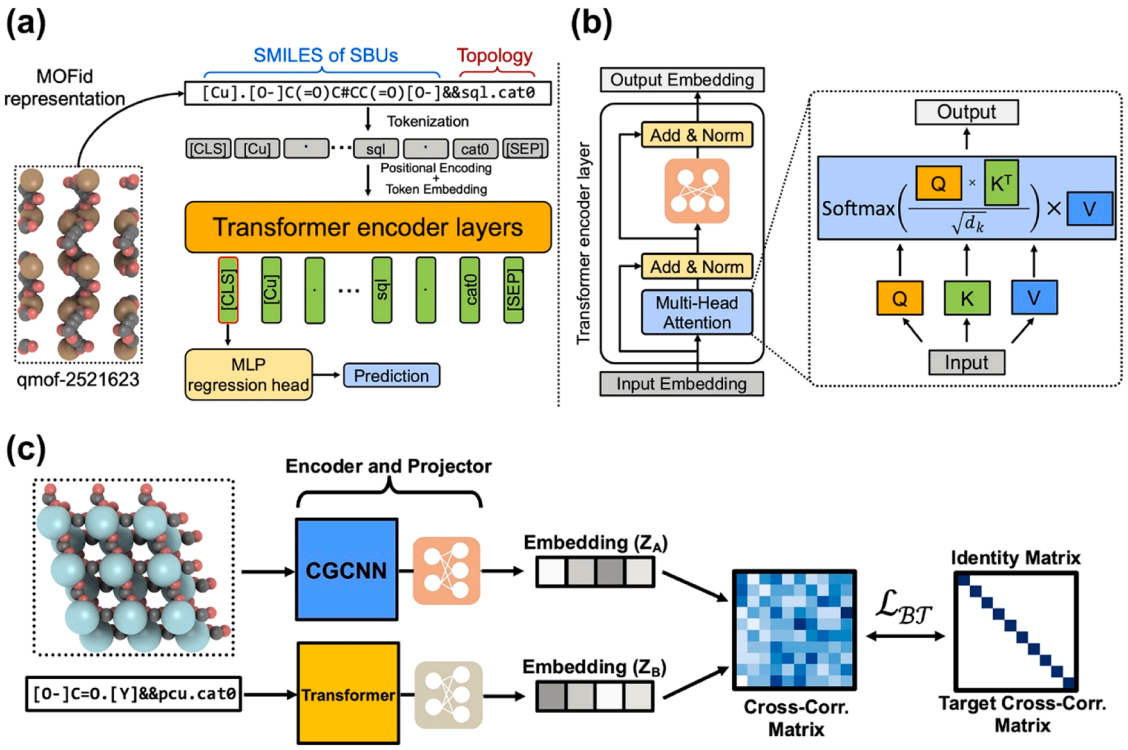

Fig. 2. (a) MOFid representation for MOF structures using SMILES of secondary building units and topology information. The MOF structure is tokenized, embedded with positional encoding, and processed through transformer encoder layers. A multi-layer perceptron (MLP) regression head is used for prediction. (b) Detailed structure of the transformer encoder layer. Input embeddings are processed through multi-head attention mechanisms and normalization layers. The multi-head attention mechanism involves calculating query (Q), key (K), and value (V) matrices to produce the output embeddings through a softmax operation. (c) Cross-modality learning framework combining graph convolutional neural networks (CGCNN) and transformers. MOF structures are encoded into embeddings (Z_A and Z_B) using each encoder. These embeddings are projected into a cross-correlation matrix and optimized to align with an identity matrix, ensuring consistency between the two modalities using Barlow Twins loss.
Reproduced with permission: Copyright 2023, ACS Publications [51].

adsorption capacity for real MOFs reported in the literature, and the results showed a good agreement with the experimental values. Based on this, the model was applied to the Core MOF 2019 dataset, an experimental-based database, to extract MOFs with high CO_2 adsorption capacity. Among the top 10 MOFs predicted by the DL model to exhibit the highest CO_2 uptake in CoRE MOF, 9 MOFs were above 10 mol/kg at 2.5 bar, which is quite high compared to all other well-known experimental MOFs. These findings support the approach of utilizing both simulation-based and experimental-based datasets in future studies, despite the inherent errors in simulations. Moreover, they further applied their model to train pore limiting diameter, surface area, void fraction, electronic bandgaps, and lowest cavity diameter. Also, they performed GCMC validation for the selected candidates.

4.2.2. *Transformer*

Transformer [47] is the most popular neural network structure in machine learning today. In tasks that process sequential data such as natural language processing, it has the advantage of parallel processing compared to existing recurrent neural networks (RNNs) and the ability to explicitly handle long-range dependencies. Transformers allow for parallel processing of input sequences, unlike RNNs which process input sequentially. This leads to significantly faster training times, especially for long sequences Transformers are particularly adept at capturing long-range dependencies in data due to their self-attention mechanism, which allows every token in the input to directly attend to every other token. Transformers have since been proven to have excellent prediction performance and scalability not only for sequential data but also for other structures such as images and graphs. In this section, we will look at several applications of MOF carbon capture, case by case.

Transformer can better handle graph structures by overcoming over-squashing, one of the difficulties experienced by existing GNNs [48]. Transformers are adept at capturing long-range dependencies due to their self-attention mechanism, which considers all pairs of nodes regardless of their distance in the graph. Transformers avoid over-squashing, an information bottleneck common in most GNNs by fully connecting the graph [48]. Chen et al. [49] developed MOFNet which is able to accurately predict the adsorption isotherms of 13MOFs. The MOF structures were encoded using a hierarchical representation. A graph transformer network was implemented to extract chemical features based on atomic-level information. Through a pressure adaptive mechanism, their model could accurately predict adsorption isotherms at various pressure ranges. Moreover, through self-attention mechanism, they canindicate the structure-property relationships. Zhao et al. [50] developed a graph transformer called GC-Trans which utilized the features from the crystal diagram to predict the adsorption performance of MOFs under binary gas components. Particularly, they applied their model to predict the gas adsorption capacity in CO_2/CH_4 mixture gas.

Transformers can be used to process text input as well as the graph structure as an input for MOF property prediction. Cao et al. [51] developed MOFormer which is based on Transformer deep learning method (Fig. 2). MOFormer is able to expedite screening process through text string representation of MOF called MOFid(Identification Schemes for Metal-Organic Frameworks To Enable Rapid Search and Cheminformatics Analysis). Thus, this model does not require gathering time-consuming 3D structural information of MOF data and can perform more accurate predictions than structure-based crystal graph convolutional neural network (CGCNN) [45] with limited data sets. Moreover, they designed self-supervised learning framework that maximizes cross-correlation between the structure-agnostic transformer and the structure-based CGCNN model for the larger combined dataset more

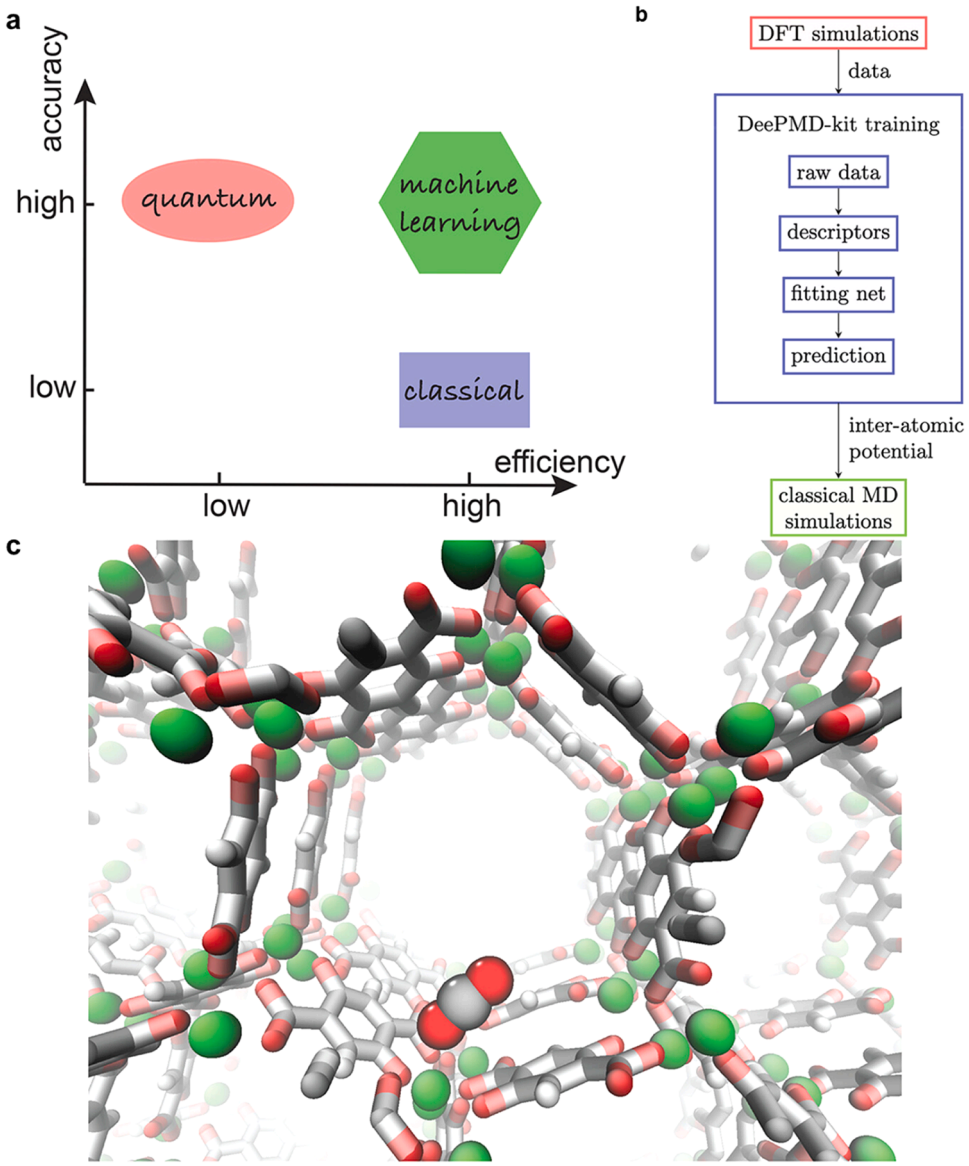

Fig. 3. Comparative analysis of computational workflow for molecular simulations (a) Efficiency and accuracy comparison of computational methods. (b) Workflow for DFT simulation approach based on QMLFF. (c) Visualization of Mg-MOF-74 with a CO_2 molecule chemically adsorbed at an open metal site. Reproduced with permission: Copyright 2023, Publisher ACS Publications. [56].

than 400,000 MOFs. The proposed SSL model between the two modalities have improved prediction performance over models trained on a single modality. The model predicted CO_2 and CH_4 gas adsorption under various pressure conditions included in the hMOF data.

Large-scale self-supervised pre-training can be used to learn useful representations that can be generalized to downstream tasks. Wang et al. [52] developed Uni-MOF which is based on three-dimensional MOF representation learning for various gas adsorption capacity prediction. Uni-MOF is pretrained from two masking tagging task similar with BERT [53]: 1) reconstructing the pristine atomic position and 2) predicting the type of masked atoms. Over 631,000 datasets on MOF and COF were applied to train the model. At the fine-tuning stage, Uni-MOF leverages the pretrained structural representation and encoding of various conditions to predict the adsorption capacity. Kang et al. [54] reported a multi-modal Transformer encoder called MOFTransformer. 1 million hypothetical MOFs were applied to train the model. This multi-modal model utilizes integrated atom-based graph and energy-grid embeddings to capture both local and global features of MOFs, respectively. Through attention scores within the self-attention layers, MOFTransformer can regenerate chemists' intuitions: H_2 uptake and diffusivity rely on global features while bandgap relies on local features. By using transfer learning from the pretrained model, their model could predict diffusion, gas adsorption, electronic properties, and text-mined data as well.

5. Accelerating molecular simulations

Molecular simulations, such as GCMC and molecular dynamics (MD), provide molecular-level details of gas adsorption processes, revealing the interactions between CO_2 molecules and the MOF structures which these details are often difficult to obtain experimentally. However, traditional simulation techniques such as density functional theory and classical molecular dynamics suffer from a cost-accuracy trade-off [55, 56] (Fig. 3**a**). Machine Learning Interatomic Potentials (MLIPs) are computational models that predict the potential energy and forces within atoms using ML. MLIP eliminates this trade-off and allows us to increase time and length scales at lower cost. This allows larger systems to be simulated within a reasonable time. Various ML models, such as

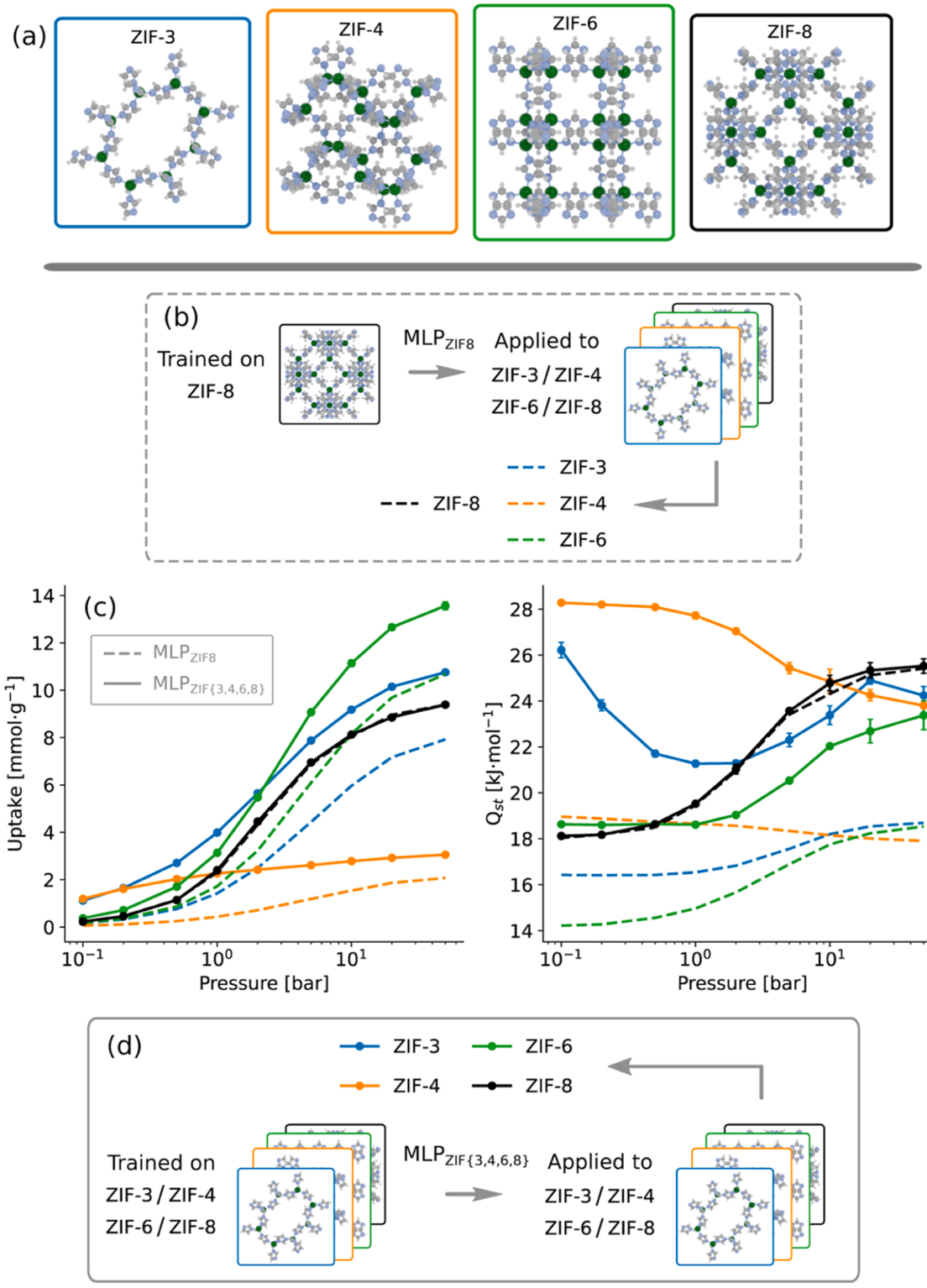

Fig. 4. (a) Diagrams of ZIFs (zeolitic imidazolate framework; a class of MOFs). (b) To perform GCMC simulations in ZIFs above, an MLP was trained on optimizations in ZIF-8 (MLP_{ZIF8}). (c) Uptake and heats of adsorption, calculated by GCMC simulations, is shown as curves as a function of the gas pressure. (d) Optimizations of all four ZIFs are used to train an MLP ($MLP_{ZIF\{3,4,6,8\}}$) and applied accordingly.
Reproduced with permission: Copyright 2023, Publisher ACS Publications [62].

GPR, artificial neural networks, and Random Forest, can be used for MLIP development. Among these, products using neural networks are specifically called neural network interatomic potentials (NNIPs). These models can be categorized into the Behler–Parrinello neural networks [57] style, which applies neural networks to descriptors that describe the local atomic environment of individual atoms to calculate the potential energy surface by summing individual results, and the GNN family, which utilizes the aforementioned graph data structure.

The dynamics and thermal properties of carbon dioxide in such materials are not fully understood in theoretical and experimental

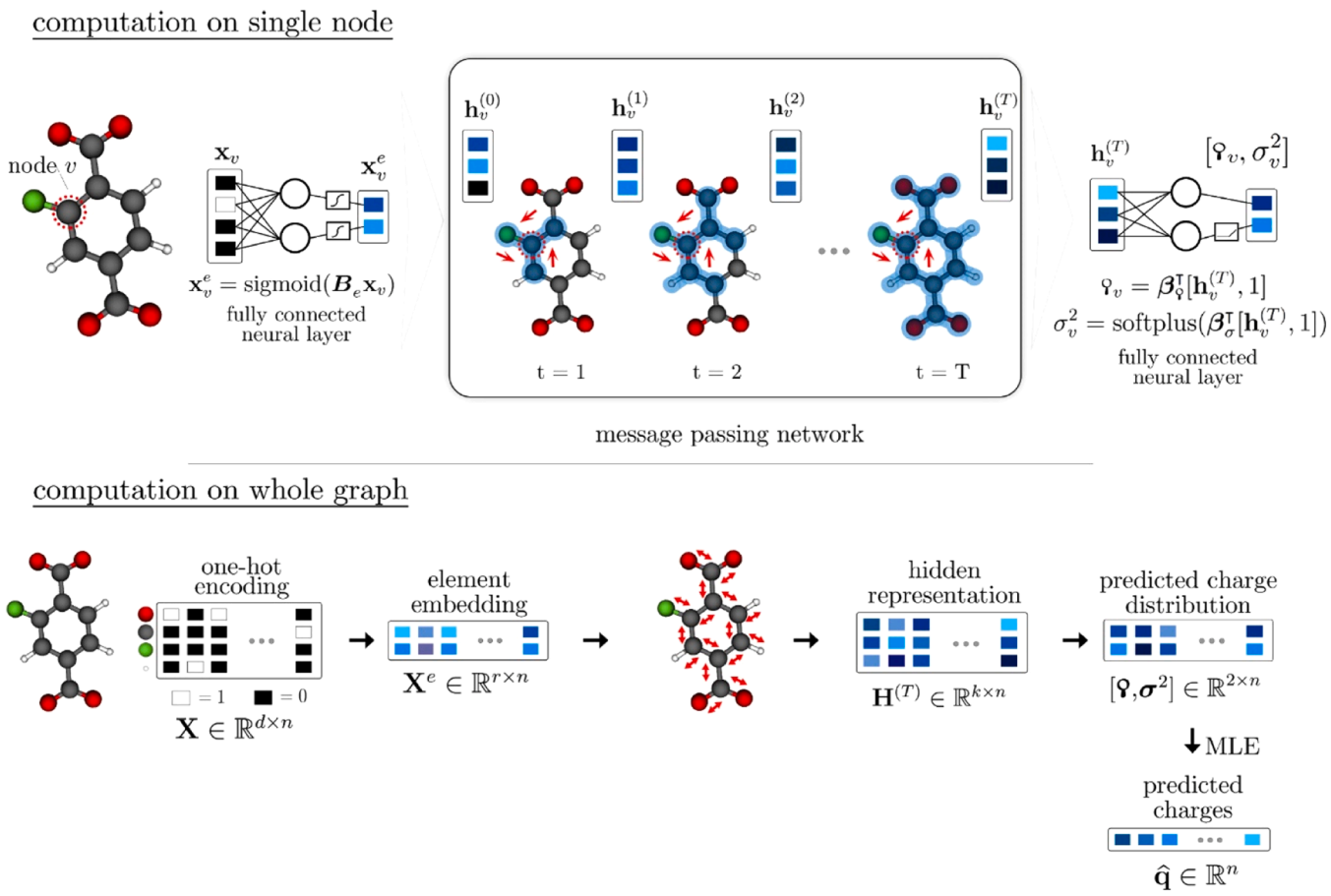

Fig. 5. MPNN architecture for partial charge prediction in the atoms of a MOF. The top panel show the computation on a single node. Each node v in the graph is represented by its initial feature vector $\mathbf{x}_v$, which is passed through a fully connected neural layer to produce a chemical element embedding $\mathbf{x}_v^e$. Node features are iteratively updated over T time steps. The final hidden state $\mathbf{h}_v^{(T)}$ for each node is used to predict the charge distribution parameters through another fully connected layer as Gaussian, returning the mean $\mathfrak{P}_v$ and variance σ_v^2. Bottom panel shows computation on the whole graph. The molecular graph is encoded using one-hot encoding, which is then embedded into a continuous space to produce $\mathbf{X}^e$. These embeddings are processed through the MPNN to obtain a hidden representation H(T). The final output is the predicted charge distribution parameters for the entire crystal structure, and the predicted charges $\hat{\boldsymbol{q}}$ are obtained via maximum likelihood estimation.
Reproduced with permission: Copyright 2020, Publisher ACS Publications. [63].

terms. Because of the complexity of amine-appended MOFs, theoretical simulations are computationally expensive. Furthermore, developing *ab initio*-quality reactive atomistic potentials for organic–inorganic hybrid materials remains a substantial challenge due to the diverse nature of atomic interactions, which range from covalent and ionic bonding to dispersion. NNIP arises to find a breakthrough to treat cost-accuracy trade-off in atomistic simulations. Large *ab initio* computations are still required to construct labels for training, but once trained, substantial time is saved in performing atomistic simulations with ab initio accuracy. The next steps remain are a workflow to construct effective training data for NNIP, a neural network structure to capture quantum chemical patterns, and high-throughput screening using NNIP simulation, and data construction using trained NNIPs.

Eckhoff et al. [58] explored the suitability of NNIP to MOFs, using MOF-5 as a benchmark. They showed that their proposed method derives a high-quality potential from DFT calculations on only small molecular fragments for the periodic MOF-5 crystal. The resulting NNIP named HDNNP, achieved an energy RMSE of 1.6 meV/atom for molecular fragments not included in the training data. HDNNP accurately determined the equilibrium lattice constant of MOF-5 bulk with an error of apporximately 0.1 % compare to DFT and predict the negative thermal expansion behavior. Yang et al. [59] investigated the use of Behler–Parrinello NNs to predict interactions between adsorbates and adsorbents for gas adsorption. This may result in an excessive number of parameters due to chemically diverse environments of host materials since multiple NNs will be required. Instead of directly using the pair distances as the inputs for the DNN model, they employed Born–Mayer distances, Coulombic interaction, and dispersion multipoles to better capture fundamental adsorption behaviors. This approach allows the DNN to be up to 100 times faster than classical non-polarizable potentials. Shaidu et al. [60] develop an NNIP model for amine-appended MOFs and use them to study the vibrational and thermal properties of amine-appended MOFs, including heat capacity and the nature of low thermal expansion coefficients. By combining the NNIP with simulated annealing, a new pathway is proposed to generate initial structures for geometry optimization, enabling an atomistic study of the properties of carbon dioxide insertion in new amine appendages before experimental synthesis. Zheng et al. [56] introduced NNIP named quantum-informed machine-learning force fields (QMLFFs) to simulate atomistic behavior of CO_2 in MOFs (Fig. 3**b**). The proposed method enhances computational efficiency approximately 1000 times compared to the first-principle calculations while preserving the quantum chemical accuracy. As a proof of concept, they performed QMLFF-based MD simulations of CO_2 molecules in Mg-MOF-74 to predict the diffusion coefficient, aligning closely with experimental values (Fig. 3**c**). More importantly, this QMLFF-based approach (free from human intervention) can be automated on high-performance clusters or supercomputers for *in silico* screening of ~1 million MOFs.

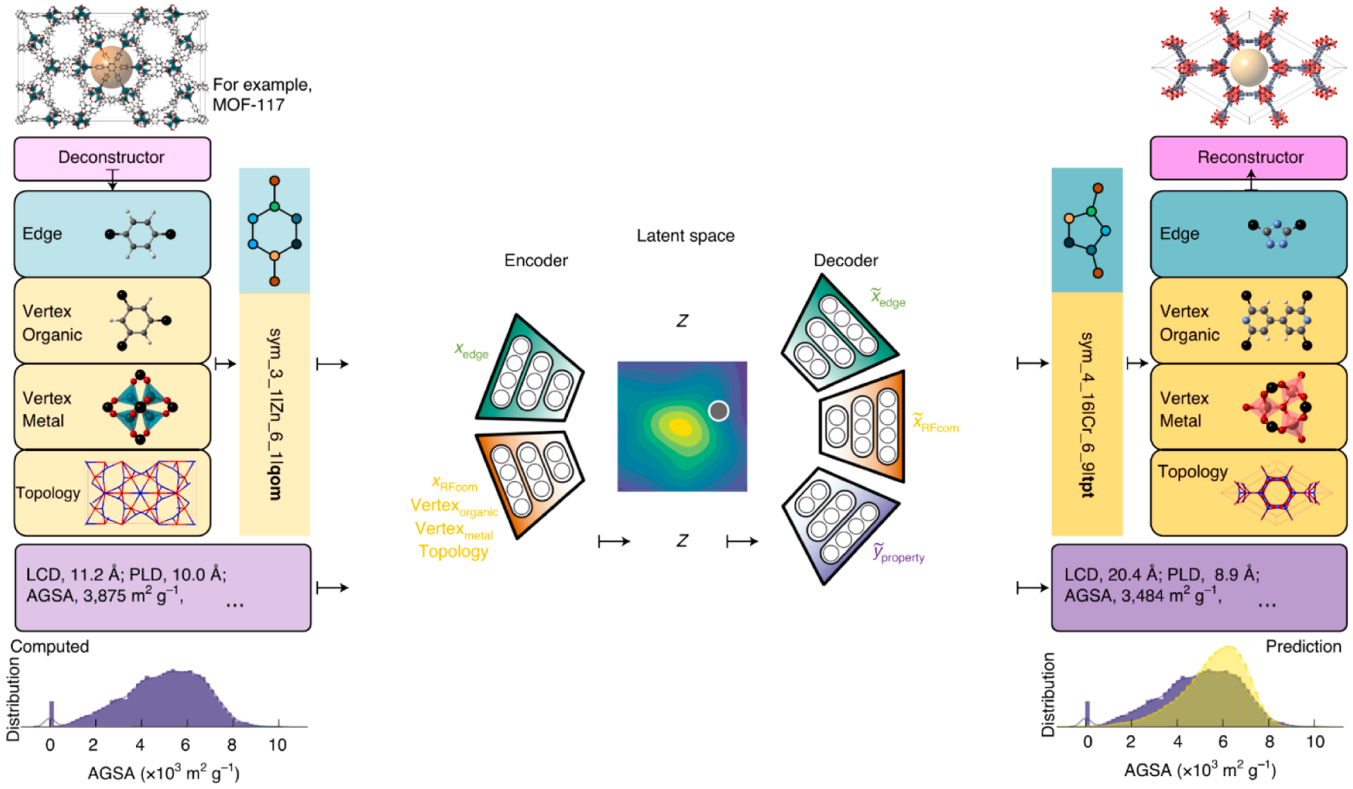

Fig. 6. Diagram of the automated reticular framework discovery platform, named SmVAE, a sophisticated variational autoencoder with multiple modules. Reticular frameworks are encoded into discrete RFcodes, converted into continuous vectors (z), and then decoded back. To structure the latent space based on desired properties, an additional component into the model that utilizes labeled data (y) is incorporated.
Reproduced with permission: Copyright 2021, Publisher Springer Nature. [69].

GNNs that satisfy equivariance conditions under E(3) transformations have been shown to achieve higher performance with greater data efficiency [61]. GCMC simulations have become a well-established tool for computational screenings of the adsorption properties of large sets of MOFs. Goeminne et al. [62] make use of the equivariant NequIP model, a variant of GNN, which has demonstrated excellent data efficiency as shown in Fig. 4, and as such an error on the interaction energies below 0.2 kJ mol^{-1} per adsorbate in ZIF-8 was attained. Its use in GCMC simulations results in highly accurate adsorption isotherms and heats of adsorption.

In molecular simulations of gas adsorption/diffusion in MOFs, the adsorbate–MOF electrostatic interaction is typically modeled by placing partial point charges on the atoms of the MOF. For the virtual screening of large libraries of MOFs, it is critical to develop computationally inexpensive methods to assign atomic partial charges to MOFs. These methods must accurately reproduce the electrostatic potential within pores, particularly for adsorption involving molecules with dipole or quadrupole moments, such as water and CO_2. Raza et al. [63] design and train a MPNN to predict the atomic partial charges on MOFs under a charge neutral constraint (Fig. 5). Kancharlapalli et al. [23] investigated an use of a random forest model to predict the partial atomic charges in MOFs. They utilized small collection of features capturing the elemental properties and the local atomic environment.

By accelerating molecular simulation using ML, large-scale HTS can be realized with relatively inexpensive computational resources. Kancharlapalli et al. [64] carried out a systematic computational HTS of the all-solvent-removed version of the CoRE MOF 2019 database for selective adsorption of CO_2 from a wet flue gas mixture. After initial screening based on the pore diameters, a total of 3703 unique MOFs from the database were considered for screening based on the FF interaction energies of CO_2, N_2, and H_2O molecules with the MOFs. MOFs showing stronger interactions with CO_2 compared to that with H_2O and N_2 were considered for the next level of screening based on the interaction energies calculated from DFT. CO_2-selective MOFs from DFT screening were further screened using two-component (CO_2 and N_2) and finally three-component (CO_2, N_2, and H_2O) GCMC simulations to predict the CO_2 capacity and CO_2/N_2 selectivity. The proposed screening study identified MOFs that show selective CO_2 adsorption under wet flue gas conditions with significant CO_2 uptake capacity and CO_2/N_2 selectivity in the presence of water vapor.

6. Generative model to discover MOF candidates

A significant number of MOFs have been databased, and several screening techniques are available, but much of the MOF space still remains unexplored. Generative models, based on the probability distribution of the data, can extend chemical space to identify MOFs not included in existing databases or directly return MOFs that meet specified conditions, thereby saving computational costs compared to screening the entire database [65,66].

Park et al. [67] implemented a generative artificial intelligence to develop GHP-MOFassemble which can facilitate the design of high CO_2 adsorbing MOFs with synthesizable linkers. They discovered novel linkers with one of the three metal nodes (Zn tetramer, Zn paddlewheel, and Cu paddlewheel). Their model could also validate the uniqueness, structural validity, and synthesizability of the suggested MOFs. In addition, they applied crystal graph neural networks and Grand Canonical Monte Carlo simulations to predict the CO_2 adsorption capacity of the predicted MOFs. Cipcigan et al. [68] applied GFlowNets with to develop 'matgfn' Python package to design MOFs and COFs for CO_2 capture. With matgfn derived data sets, they calculated the adsorption isotherms under single and binary components gas. Moreover, they reported 13 materials with superior performance. Yao et al. [69] demonstrated an automated nanoporous materials discovery platform

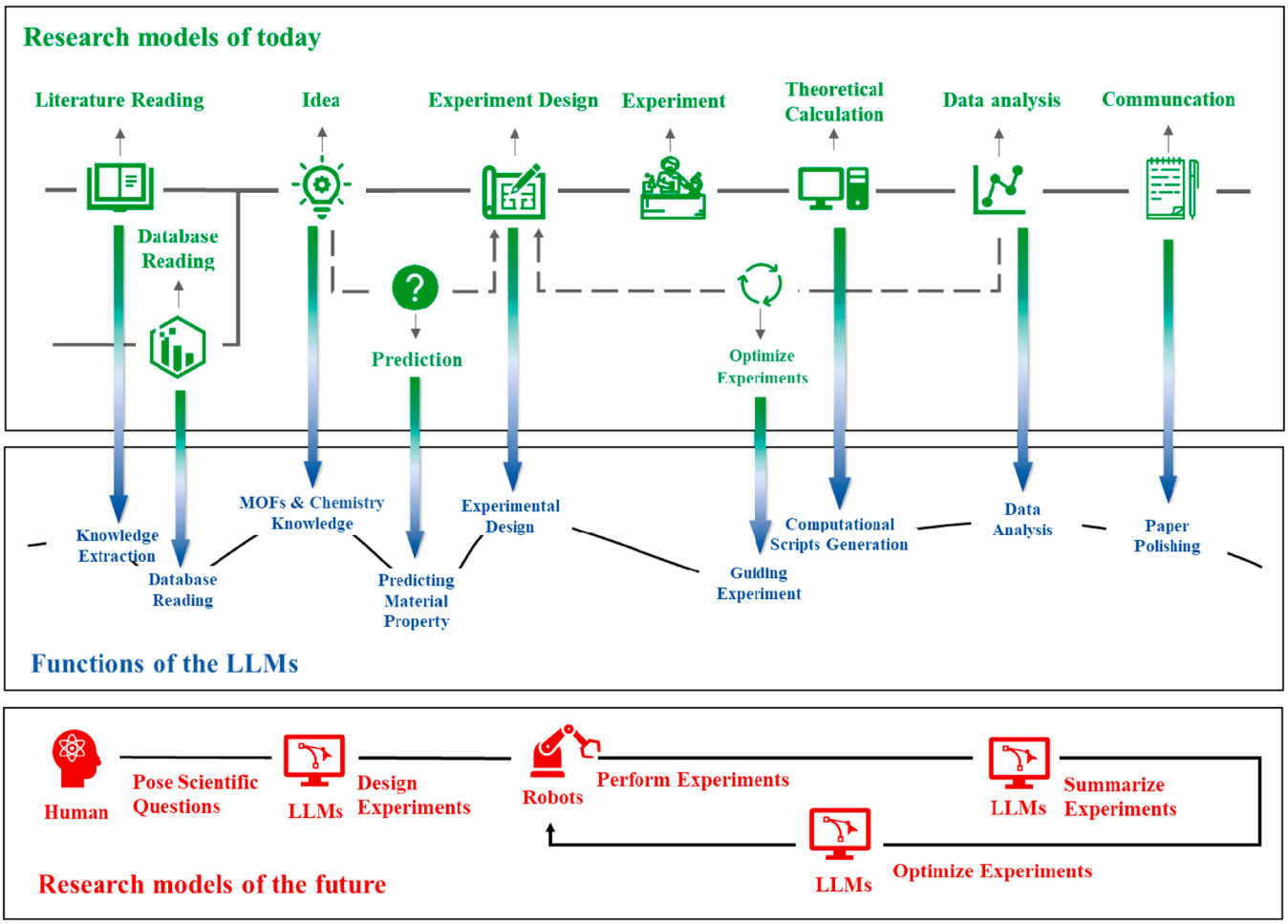

Fig. 7. The current and future roles of large language models (LLMs) in MOFs research encompass their integration into a unified research workflow today, with the potential to evolve into an autonomous research model in the future.
Reproduced with permission: Copyright 2024, Publisher ACS Publications. [71].

via supramolecular variational autoencoder to generate MOFs with selective capture of CO_2 in flue gas (Fig. 6). They used this platform to design novel MOFs with improved capacity and good selectivity for CO_2/N_2 and CO_2/CH_4 separations, which are important clean-energy-relevant applications.

Combining generative models with other machine learning approaches, such as reinforcement learning and genetic algorithms, can enhance the exploration of design spaces and optimize MOF properties more effectively. Park et al. [70] developed an inverse design model of MOFs for direct CO_2 capture using reinforcement learning. They selected validity, scaffold, uniqueness, target and reward as metrics of CO_2 heat of absorption and CO_2/H_2O selectivity. The predicted MOFs with high performance well matched with the literature and the heat of adsorption was the most important feature for identifying high performance MOFs.

The rapid rise in the popularity of transformers and chatbots from the latter half of 2022 can be attributed to advancements in AI technology, increased demand for efficient digital communication tools, and significant improvements in user experience and capabilities. Bai et al. [71] tested six open-source large language models (LLMs) such as GPT 3.5, ChatGLM2–6B, Llama2–7B, Vicuna-7B, Marcoroni-7B, Mistral-7B, Falcon-7B, Llama2–13B, and Vicuna-13B about their applicability on the MOF research (Fig. 7). They specifically divided the criteria as follows: MOFs knowledge, paper polishing, basic chemistry knowledge, knowledge extraction, research assistance, and in-deep chemistry knowledge. From their results, Llama2–7B and ChatGLM2–6B showed the best performance. Integrating the complex structural information of MOFs with the power of pre-trained language models on massive corpora could be a promising research direction. ChatGLM2–6B is also good for many tasks, including experiment design and computational script generation, and its weaknesses in knowledge can be easily improved by feeding it with more literature through fine-tuning. The commendable performance of these models demonstrated their high potential as powerful tools and a central platform to connect all sections of research activity. It is expected that open-source LLMs will play a greater role in assisting chemical and material research and feasibly transform our research paradigm in the near future.

7. Explainable AI

Model interpretability remains a concern, particularly with DL models, which are often considered black boxes. This lack of transparency can hinder the understanding and trust in model predictions. This field, in particular, lies at the intersection of experimental and computational researchers. Therefore, information about the model's decisions can provide experimental researchers with various inspirations and potentially lead to new discoveries.

Random forest provides a built-in measure of feature importance, which can be useful for understanding the data. Bai et al. [72] applied RF feature importance and SHapley Additive exPlanations (SHAP) techniques to identify factors influencing CO_2 adsorption performance. These techniques have the advantage of identifying useful features, but

they present the challenge that experts need to manually engineer the features. The identified importance order of descriptors is ligand characteristics, metal charge, temperature, metal type, co-catalyst type and loading, specific surface area, reaction pressure, pore volume and substrate type. To evaluate the prediction accuracy, they conducted experimental tests on MOF-76(Y), which was identified as one of the most active materials by the prediction.

In the future, proposing more useful descriptors based on expert intuition or applying explainable AI techniques in deep learning could be promising approaches.

8. Challenges and limitations

The application of ML in predicting CO_2 capture and conversion properties of MOFs faces several challenges and limitations. Data quality and availability are significant issues, as experimental data are often expensive and time-consuming to acquire, leading to small datasets that may not capture the full diversity of MOFs. As more and more studies rely on deep neural networks, model interpretability remains an issue. Feature engineering in traditional ML requires substantial domain expertise and can be labor-intensive. Ensuring model generalization to unseen data is another challenge, given the variability in experimental conditions. Additionally, to our knowledge, there is not yet a sufficient consensus on the metrics used to evaluate the performance of models that generate new MOFs, like those in generative AI tasks. While validity is widely used for models that generate molecules or crystals, many deep learning-based generative models achieve near 100 % validity, which fails to reflect actual performance differences. Furthermore, there is not yet enough accumulated research for meaningful comparisons. For instance, Yao et al. [69] developed a VAE-based generative model in their paper but did not compare the metrics for generation performance with other works. Therefore, further research on this topic is necessary. Despite advances in designing MOFs, experimental validation remains a costly and time-consuming bottleneck. Determining whether a proposed MOF can actually be synthesized is important, but this synthesizability has not been deeply explored. Park et al. [73] proposed a thermodynamics-based metric to assess the synthesizability of MOFs in their study. However, being thermodynamically stable does not necessarily imply that a MOF is synthesizable. Integrating such validation processes will be a significant challenge for future research.

9. Conclusion

In this paper, we review recent work applying ML to MOF discovery. Because there are countless possible combinations of MOFs, traditional experimental approaches alone have limitations in quickly discovering the optimal MOF. ML techniques are attracting attention as an innovative method to solve these problems. Through this review, we introduced an accessible MOF database and the latest ML techniques to researchers who want to apply ML to identify new MOFs. The feasibility of predicting the synthesizability of MOFs through ML and designing optimized MOFs for carbon capture applications has been demonstrated. Future research should aim to design and discover more sophisticated and efficient MOFs by combining ML and experimental approaches.

CRediT authorship contribution statement

Sung Eun Jerng: Writing – review & editing, Writing – original draft, Visualization, Supervision, Resources, Investigation, Funding acquisition, Conceptualization. **Yang Jeong Park:** Writing – review & editing, Writing – original draft, Investigation, Funding acquisition, Conceptualization.

Declaration of Competing Interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Yang Jeong Park reports financial support was provided by National Research Foundation of Korea. Sung Eun Jerng reports financial support was provided by Gyeonggi Technology Development Program. Sung Eun Jerng reports financial support was provided by The University of Suwon. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data Availability

No data was used for the research described in the article.

Acknowledgements

This work was partly supported by a grant from the National Research Foundation of Korea (NRF) funded by the Korean government, Ministry of Science and ICT (MSIT) (No. 2021R1A6A3A01086766). This work was partly supported by the BK21 FOUR program of the Education and Research Program for Future ICT Pioneers, Seoul National University in 2024. This work was partly supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) [NO.RS-2021-II211343, Artificial Intelligence Graduate School Program (Seoul National University)]. This work was also partly supported by a grant (code 2023–006) from Gyeonggi Technology Development Program funded by Gyeonggi Province, Republic of Korea and the research grant of The University of Suwon, Republic of Korea in 2023 (2023–0166).

References

[1] Y. Yan, T.N. Borhani, S.G. Subraveti, K.N. Pai, V. Prasad, A. Rajendran, P. Nkulikiyinka, J.O. Asibor, Z. Zhang, D. Shao, L. Wang, W. Zhang, Y. Yan, W. Ampomah, J. You, M. Wang, E.J. Anthony, V. Manovic, P.T. Clough, Harnessing the power of machine learning for carbon capture, utilisation, and storage (CCUS) – a state-of-the-art review, Energy Environ. Sci. 14 (2021) 6122–6157, https://doi.org/10.1039/D1EE02395K.

[2] H. Chen, Y. Zheng, J. Li, L. Li, X. Wang, AI for nanomaterials development in clean energy and carbon capture, utilization and storage (CCUS), ACS Nano 17 (2023) 9763–9792, https://doi.org/10.1021/acsnano.3c01062.

[3] J. Yu, L.-H. Xie, J.-R. Li, Y. Ma, J.M. Seminario, P.B. Balbuena, CO_2 Capture and Separations Using MOFs: computational and experimental studies, Chem. Rev. 117 (2017) 9674–9754, https://doi.org/10.1021/acs.chemrev.6b00626.

[4] A. Velty, A. Corma, Advanced zeolite and ordered mesoporous silica-based catalysts for the conversion of CO_2 to chemicals and fuels, Chem. Soc. Rev. 52 (2023) 1773–1946, https://doi.org/10.1039/D2CS00456A.

[5] A.M. Parvanian, N. Sadeghi, A. Rafiee, C.J. Shearer, M. Jafarian, Application of porous materials for CO_2 reutilization: a review, Energies 15 (2022) 63, https://doi.org/10.3390/en15010063.

[6] R. Freund, O. Zaremba, G. Arnauts, R. Ameloot, G. Skorupskii, M. Dincă, A. Bavykina, J. Gascon, A. Ejsmont, J. Goscianska, M. Kalmutzki, U. Lächelt, E. Ploetz, C.S. Diercks, S. Wuttke, The current status of MOF and COF applications, Angew. Chem. Int. Ed. 60 (2021) 23975–24001, https://doi.org/10.1002/anie.202106259.

[7] Q. Wang, D. Astruc, State of the art and prospects in metal–organic framework (MOF)-based and MOF-derived nanocatalysis, Chem. Rev. 120 (2020) 1438–1511, https://doi.org/10.1021/acs.chemrev.9b00223.

[8] K. Xu, S. Zhang, X. Zhuang, G. Zhang, Y. Tang, H. Pang, Recent progress of MOF-functionalized nanocomposites: from structure to properties, Adv. Colloid Interface Sci. 323 (2024) 103050, https://doi.org/10.1016/j.cis.2023.103050.

[9] O.V. Gutov, M.G. Hevia, E.C. Escudero-Adán, A. Shafir, Metal–organic framework (MOF) defects under control: insights into the missing linker sites and their implication in the reactivity of zirconium-based frameworks, Inorg. Chem. 54 (2015) 8396–8400, https://doi.org/10.1021/acs.inorgchem.5b01053.

[10] A.F. Gross, E. Sherman, S.L. Mahoney, J.J. Vajo, Reversible ligand exchange in a metal–organic framework (MOF): toward MOF-based dynamic combinatorial chemical systems, J. Phys. Chem. A 117 (2013) 3771–3776, https://doi.org/10.1021/jp401039k.

[11] Y. Zhou, R. Abazari, J. Chen, M. Tahir, A. Kumar, R.R. Ikreedeegh, E. Rani, H. Singh, A.M. Kirillov, Bimetallic metal–organic frameworks and MOF-derived composites: recent progress on electro- and photoelectrocatalytic applications, Coord. Chem. Rev. 451 (2022) 214264, https://doi.org/10.1016/j.ccr.2021.214264.

[12] G. Song, Y. Shi, S. Jiang, H. Pang, Recent progress in MOF-derived porous materials as electrodes for high-performance lithium-ion batteries, Adv. Funct. Mater. 33 (2023) 2303121, https://doi.org/10.1002/adfm.202303121.

[13] B. Zhang, Y. Sun, H. Xu, X. He, Hydrogen storage mechanism of metal–organic framework materials based on metal centers and organic ligands, Carbon Neutraliz. 2 (2023) 632–645, https://doi.org/10.1002/cnl2.91.

[14] M.P. Suh, H.J. Park, T.K. Prasad, D.-W. Lim, Hydrogen storage in metal–organic frameworks, Chem. Rev. 112 (2012) 782–835, https://doi.org/10.1021/cr200274s.

[15] H. Demir, G.O. Aksu, H.C. Gulbalkan, S. Keskin, MOF membranes for CO_2 capture: past, present and future, Carbon Capture Sci. Technol. 2 (2022) 100026, https://doi.org/10.1016/j.ccst.2021.100026.

[16] A. Bavykina, N. Kolobov, I.S. Khan, J.A. Bau, A. Ramirez, J. Gascon, Metal–organic frameworks in heterogeneous catalysis: recent progress, new trends, and future perspectives, Chem. Rev. 120 (2020) 8468–8535, https://doi.org/10.1021/acs.chemrev.9b00685.

[17] S.-J. Shin, J.W. Gittins, C.J. Balhatchet, A. Walsh, A.C. Forse, Metal–Organic Framework Supercapacitors: Challenges and Opportunities, Adv. Funct. Mater. n/a (n.d.) 2308497. https://doi.org/10.1002/adfm.202308497.

[18] C. Altintas, O.F. Altundal, S. Keskin, R. Yildirim, Machine learning meets with metal organic frameworks for gas storage and separation, J. Chem. Inf. Model. 61 (2021) 2131–2146, https://doi.org/10.1021/acs.jcim.1c00191.

[19] S. Gupta, L. Li, The potential of machine learning for enhancing CO_2 sequestration, storage, transportation, and utilization-based processes: a brief perspective, JOM 74 (2022) 414–428, https://doi.org/10.1007/s11837-021-05079-x.

[20] D. Narváez-Celada, A.S. Varela, CO_2 electrochemical reduction on metal–organic framework catalysts: current status and future directions, J. Mater. Chem. A 10 (2022) 5899–5917, https://doi.org/10.1039/D1TA10440C.

[21] S. Chong, S. Lee, B. Kim, J. Kim, Applications of machine learning in metal-organic frameworks, Coord. Chem. Rev. 423 (2020) 213487, https://doi.org/10.1016/j.ccr.2020.213487.

[22] H. Demir, S. Keskin, A new era of modeling MOF-based membranes: cooperation of theory and data science, Macromol. Mater. Eng. 309 (2024) 2300225, https://doi.org/10.1002/mame.202300225.

[23] S. Kancharlapalli, A. Gopalan, M. Haranczyk, R.Q. Snurr, Fast and accurate machine learning strategy for calculating partial atomic charges in metal–organic frameworks, J. Chem. Theory Comput. 17 (2021) 3052–3064, https://doi.org/10.1021/acs.jctc.0c01229.

[24] Y. Situ, X. Yuan, X. Bai, S. Li, H. Liang, X. Zhu, B. Wang, Z. Qiao, Large-scale screening and machine learning for metal-organic framework membranes to capture CO_2 from flue gas, Membranes 12 (2022) 700, https://doi.org/10.3390/membranes12070700.

[25] C.E. Wilmer, M. Leaf, C.Y. Lee, O.K. Farha, B.G. Hauser, J.T. Hupp, R.Q. Snurr, Large-scale screening of hypothetical metal–organic frameworks, Nat. Chem. 4 (2012) 83–89, https://doi.org/10.1038/nchem.1192.

[26] Y.G. Chung, J. Camp, M. Haranczyk, B.J. Sikora, W. Bury, V. Krungleviciute, T. Yildirim, O.K. Farha, D.S. Sholl, R.Q. Snurr, Computation-ready, experimental metal–organic frameworks: a tool to enable high-throughput screening of nanoporous crystals, Chem. Mater. 26 (2014) 6185–6192, https://doi.org/10.1021/cm502594j.

[27] Y.G. Chung, E. Haldoupis, B.J. Bucior, M. Haranczyk, S. Lee, H. Zhang, K. D. Vogiatzis, M. Milisavljevic, S. Ling, J.S. Camp, B. Slater, J.I. Siepmann, D. S. Sholl, R.Q. Snurr, Advances, updates, and analytics for the computation-ready, experimental metal–organic framework database: CoRE MOF 2019, J. Chem. Eng. Data 64 (2019) 5985–5998, https://doi.org/10.1021/acs.jced.9b00835.

[28] P.Z. Moghadam, A. Li, S.B. Wiggin, A. Tao, A.G.P. Maloney, P.A. Wood, S.C. Ward, D. Fairen-Jimenez, Development of a cambridge structural database subset: a collection of metal–organic frameworks for past, present, and future, Chem. Mater. 29 (2017) 2618–2625, https://doi.org/10.1021/acs.chemmater.7b00441.

[29] P.G. Boyd, A. Chidambaram, E. García-Díez, C.P. Ireland, T.D. Daff, R. Bounds, A. Gładysiak, P. Schouwink, S.M. Moosavi, M.M. Maroto-Valer, J.A. Reimer, J.A. R. Navarro, T.K. Woo, S. Garcia, K.C. Stylianou, B. Smit, Data-driven design of metal–organic frameworks for wet flue gas CO_2 capture, Nature 576 (2019) 253–256, https://doi.org/10.1038/s41586-019-1798-7.

[30] A.S. Rosen, S.M. Iyer, D. Ray, Z. Yao, A. Aspuru-Guzik, L. Gagliardi, J.M. Notestein, R.Q. Snurr, Machine learning the quantum-chemical properties of metal–organic frameworks for accelerated materials discovery, Matter 4 (2021) 1578–1597, https://doi.org/10.1016/j.matt.2021.02.015.

[31] A. Sriram, S. Choi, X. Yu, L.M. Brabson, A. Das, Z. Ulissi, M. Uyttendaele, A. J. Medford, D.S. Sholl, The open DAC 2023 dataset and challenges for sorbent discovery in direct air capture, arXiv (2023), https://doi.org/10.48550/arXiv.2311.00341.

[32] J. Burner, J. Luo, A. White, A. Mirmiran, O. Kwon, P.G. Boyd, S. Maley, M. Gibaldi, S. Simrod, V. Ogden, others, ARC-MOF: a diverse database of metal-organic frameworks with DFT-derived partial atomic charges and descriptors for machine learning, Chem. Mater. 35 (2023) 900–916.

[33] L.T. Glasby, K. Gubsch, R. Bence, R. Oktavian, K. Isoko, S.M. Moosavi, J. L. Cordiner, J.C. Cole, P.Z. Moghadam, DigiMOF: a database of metal–organic framework synthesis information generated via text mining, Chem. Mater. 35 (2023) 4510–4524.

[34] T. Bailey, A. Jackson, R.-A. Berbece, K. Wu, N. Hondow, E. Martin, Gradient boosted machine learning model to Predict H2, CH4, and CO_2 uptake in metal–organic frameworks using experimental data, J. Chem. Inf. Model. 63 (2023) 4545–4551, https://doi.org/10.1021/acs.jcim.3c00135.

[35] M. Gheytanzadeh, A. Baghban, S. Habibzadeh, A. Esmaeili, O. Abida, A. Mohaddespour, M.T. Munir, Towards estimation of CO_2 adsorption on highly porous MOF-based adsorbents using gaussian process regression approach, Sci. Rep. 11 (2021) 15710, https://doi.org/10.1038/s41598-021-95246-6.

[36] I. Tsamardinos, G.S. Fanourgakis, E. Greasidou, E. Klontzas, K. Gkagkas, G. E. Froudakis, An Automated Machine Learning architecture for the accelerated prediction of Metal-Organic Frameworks performance in energy and environmental applications, Microporous Mesoporous Mater. 300 (2020) 110160, https://doi.org/10.1016/j.micromeso.2020.110160.

[37] X. Li, X. Zhang, J. Zhang, J. Gu, S. Zhang, G. Li, J. Shao, Y. He, H. Yang, S. Zhang, H. Chen, Applied machine learning to analyze and predict CO_2 adsorption behavior of metal-organic frameworks, Carbon Capture Sci. Technol. 9 (2023) 100146, https://doi.org/10.1016/j.ccst.2023.100146.

[38] S. Özsoysal, B. Oral, R. Yıldırım, Analysis of photocatalytic CO_2 reduction over MOFs using machine learning, J. Mater. Chem. A 12 (2024) 5748–5759, https://doi.org/10.1039/D3TA07001H.

[39] J. Abdi, F. Hadavimoghaddam, M. Hadipoor, A. Hemmati-Sarapardeh, Modeling of CO_2 adsorption capacity by porous metal organic frameworks using advanced decision tree-based models, Sci. Rep. 11 (2021) 24468, https://doi.org/10.1038/s41598-021-04168-w.

[40] X. Ma, W. Xu, R. Su, L. Shao, Z. Zeng, L. Li, H. Wang, Insights into CO_2 capture in porous carbons from machine learning, experiments and molecular simulation, Sep. Purif. Technol. 306 (2023) 122521, https://doi.org/10.1016/j.seppur.2022.122521.

[41] Y. LeCun, Y. Bengio, G. Hinton, Deep learning, Nature 521 (2015) 436–444, https://doi.org/10.1038/nature14539.

[42] G.B. Goh, N.O. Hodas, A. Vishnu, Deep learning for computational chemistry, J. Comput. Chem. 38 (2017) 1291–1307, https://doi.org/10.1002/jcc.24764.

[43] C. Janiesch, P. Zschech, K. Heinrich, Machine learning and deep learning, Electron. Mark. 31 (2021) 685–695, https://doi.org/10.1007/s12525-021-00475-2.

[44] J. Gilmer, S.S. Schoenholz, P.F. Riley, O. Vinyals, G.E. Dahl, Neural message passing for quantum chemistry, in: Int. Conf. Mach. Learn., PMLR, 2017: pp. 1263–1272.

[45] T. Xie, J.C. Grossman, Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties, Phys. Rev. Lett. 120 (2018) 145301.

[46] K. Choudhary, T. Yildirim, D.W. Siderius, A.G. Kusne, A. McDannald, D.L. Ortiz-Montalvo, Graph neural network predictions of metal organic framework CO_2 adsorption properties, Comput. Mater. Sci. 210 (2022) 111388, https://doi.org/10.1016/j.commatsci.2022.111388.

[47] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. ukasz Kaiser, I. Polosukhin, Attention is All you Need, in: Adv. Neural Inf. Process. Syst., Curran Associates, Inc., 2017. ⟨https://papers.nips.cc/paper_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html⟩. accessed July 8, 2024.

[48] D. Kreuzer, D. Beaini, W. Hamilton, V. Létourneau, P. Tossou, Rethinking graph transformers with spectral attention, in: Adv. Neural Inf. Process. Syst., Curran Associates Inc, 2021, pp. 21618–21629, in: ⟨https://proceedings.neurips.cc/paper_files/paper/2021/hash/b4fd1d2cb085390fbbadae65e07876a7-Abstract.html⟩ (accessed May 31, 2024).

[49] P. Chen, R. Jiao, J. Liu, Y. Liu, Y. Lu, Interpretable graph transformer network for predicting adsorption isotherms of metal–organic frameworks, J. Chem. Inf. Model. 62 (2022) 5446–5456, https://doi.org/10.1021/acs.jcim.2c00876.

[50] Y. Zhao, Y. Zhao, Q. Gong, Z. Wang, Graph transformer with convolution parallel networks for predicting single and binary component adsorption performance of metal–organic frameworks, ACS Appl. Mater. Interfaces 15 (2023) 49527–49537, https://doi.org/10.1021/acsami.3c10951.

[51] Z. Cao, R. Magar, Y. Wang, A. Barati Farimani, MOFormer: self-supervised transformer model for metal–organic framework property prediction, J. Am. Chem. Soc. 145 (2023) 2958–2967, https://doi.org/10.1021/jacs.2c11420.

[52] J. Wang, J. Liu, H. Wang, M. Zhou, G. Ke, L. Zhang, J. Wu, Z. Gao, D. Lu, A comprehensive transformer-based approach for high-accuracy gas adsorption predictions in metal-organic frameworks, Nat. Commun. 15 (2024) 1904, https://doi.org/10.1038/s41467-024-46276-x.

[53] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, ArXiv Prepr. ArXiv181004805 (2018).

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

[55] V.L. Deringer, M.A. Caro, G. Csányi, Machine learning interatomic potentials as emerging tools for materials science, Adv. Mater. 31 (2019) 1902765, https://doi.org/10.1002/adma.201902765.

[56] B. Zheng, F.L. Oliveira, R. Neumann Barros Ferreira, M. Steiner, H. Hamann, G. X. Gu, B. Luan, Quantum informed machine-learning potentials for molecular dynamics simulations of CO_2's chemisorption and diffusion in Mg-MOF-74, ACS Nano 17 (2023) 5579–5587, https://doi.org/10.1021/acsnano.2c11102.

[57] J. Behler, M. Parrinello, Generalized neural-network representation of high-dimensional potential-energy surfaces, Phys. Rev. Lett. 98 (2007) 146401, https://doi.org/10.1103/PhysRevLett.98.146401.

[58] M. Eckhoff, J. Behler, From molecular fragments to the bulk: development of a neural network potential for MOF-5, J. Chem. Theory Comput. 15 (2019) 3793–3809, https://doi.org/10.1021/acs.jctc.8b01288.

[59] C.-T. Yang, I. Pandey, D. Trinh, C.-C. Chen, J.D. Howe, L.-C. Lin, Deep learning neural network potential for simulating gaseous adsorption in metal–organic frameworks, Mater. Adv. 3 (2022) 5299–5303, https://doi.org/10.1039/D1MA01152A.

[60] Y. Shaidu, A. Smith, E. Taw, J.B. Neaton, Carbon capture phenomena in metal-organic frameworks with neural network potentials, PRX Energy 2 (2023) 023005, https://doi.org/10.1103/PRXEnergy.2.023005.

[61] V.G. Satorras, E. Hoogeboom, M. Welling, E. (n), equivariant graph neural networks, in: Int. Conf. Mach. Learn., PMLR, 2021, pp. 9323–9332.
[62] R. Goeminne, L. Vanduyfhuys, V. Van Speybroeck, T. Verstraelen, DFT-quality adsorption simulations in metal–organic frameworks enabled by machine learning potentials, J. Chem. Theory Comput. 19 (2023) 6313–6325, https://doi.org/10.1021/acs.jctc.3c00495.
[63] A. Raza, A. Sturluson, C.M. Simon, X. Fern, Message passing neural networks for partial charge assignment to metal–organic frameworks, J. Phys. Chem. C. 124 (2020) 19070–19082, https://doi.org/10.1021/acs.jpcc.0c04903.
[64] S. Kancharlapalli, R.Q. Snurr, High-throughput screening of the CoRE-MOF-2019 database for CO_2 capture from wet flue gas: a multi-scale modeling strategy, ACS Appl. Mater. Interfaces 15 (2023) 28084–28092, https://doi.org/10.1021/acsami.3c04079.
[65] N. De Cao, T. Kipf, MolGAN: An implicit generative model for small molecular graphs, ArXiv Prepr. ArXiv180511973 (2018).
[66] B. Sanchez-Lengeling, A. Aspuru-Guzik, Inverse molecular design using machine learning: Generative models for matter engineering, Science 361 (2018) 360–365, https://doi.org/10.1126/science.aat2663.
[67] H. Park, X. Yan, R. Zhu, E.A. Huerta, S. Chaudhuri, D. Cooper, I. Foster, E. Tajkhorshid, A generative artificial intelligence framework based on a molecular diffusion model for the design of metal-organic frameworks for carbon capture, Commun. Chem. 7 (2024) 21.
[68] F. Cipcigan, J. Booth, R.N.B. Ferreira, C.R. dos Santos, M. Steiner, Discovery of novel reticular materials for carbon dioxide capture using GFlowNets, Digit. Discov. 3 (2024) 449–455, https://doi.org/10.1039/D4DD00020J.
[69] 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, A. Aspuru-Guzik, Inverse design of nanoporous crystalline reticular materials with deep generative models, Nat. Mach. Intell. 3 (2021) 76–86, https://doi.org/10.1038/s42256-020-00271-1.
[70] H. Park, S. Majumdar, X. Zhang, J. Kim, B. Smit, Inverse design of metal-organic frameworks for direct air capture of CO_2 via deep reinforcement learning, Digit. Discov. (2024).
[71] X. Bai, Y. Xie, X. Zhang, H. Han, J.-R. Li, Evaluation of open-source large language models for Metal–Organic frameworks research, J. Chem. Inf. Model. (2024).
[72] X. Bai, Y. Li, Y. Xie, Q. Chen, X. Zhang, J.-R. Li, High-throughput screening of CO_2 cycloaddition MOF catalyst with an explainable machine learning model, Green. Energy Environ. (2024), https://doi.org/10.1016/j.gee.2024.01.010.
[73] J. Park, Y. Lim, S. Lee, J. Kim, Computational design of metal–organic frameworks with unprecedented high hydrogen working capacity and high synthesizability, Chem. Mater. 35 (2023) 9–16, https://doi.org/10.1021/acs.chemmater.2c01822.