JACS Au
AN OPEN ACCESS JOURNAL OF THE AMERICAN CHEMICAL SOCIETY

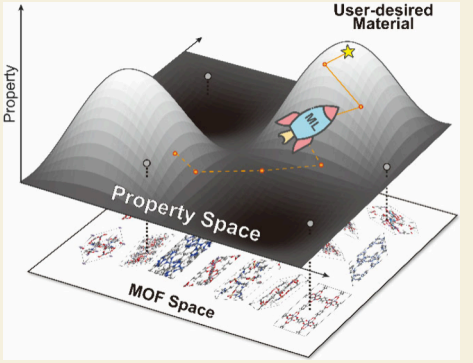
pubs.acs.org/jacsau
Perspective

From Data to Discovery: Recent Trends of Machine Learning in Metal−Organic Frameworks

Junkil Park, Honghui Kim, Yeonghun Kang, Yunsung Lim, and Jihan Kim*

Cite This: JACS Au 2024, 4, 3727−3743

ABSTRACT: Renowned for their high porosity and structural diversity, metal−organic frameworks (MOFs) are a promising class of materials for a wide range of applications. In recent decades, with the development of large-scale databases, the MOF community has witnessed innovations brought by data-driven machine learning methods, which have enabled a deeper understanding of the chemical nature of MOFs and led to the development of novel structures. Notably, machine learning is continuously and rapidly advancing as new methodologies, architectures, and data representations are actively being investigated, and their implementation in materials discovery is vigorously pursued. Under these circumstances, it is important to closely monitor recent research trends and identify the technologies that are being introduced. In this Perspective, we focus on emerging trends of machine learning within the field of MOFs, the challenges they face, and the future directions of their development.



KEYWORDS: *Machine Learning, Metal−Organic Frameworks, Data-Driven, Regression Models, Generative Models, Machine Learning Potentials, Data Mining, Autonomous Lab*

1. INTRODUCTION

Metal−organic frameworks (MOFs), composed of an ordered network of metal ions/clusters and organic linkers, have emerged as an innovative class of crystalline materials in recent decades. Given their expansive interstitial pores and chemical versatility, MOFs have gained attention across a wide range of applications, including gas storage,[1−4] catalysis,[5−7] sensors,[8,9] and drug delivery.[10,11] In particular, the high structural diversity that originates from the countless combination of their building blocks stands out as the most distinctive characteristic of MOFs.[12,13] Numerous MOFs with unique structures and novel functionalities have been identified, with the number of experimentally reported structures exceeding 100,000.[14,15] Furthermore, the relatively simple synthetic motif of MOFs has enabled their generation in silico, offering an unlimited number of hypothetical structures. Such high structural diversity and vast search space keep pushing the boundaries of the applications of MOFs into unexplored territories.[13]

Paradoxically, the high structural diversity endows MOFs with boundless potential, yet at the same time, it presents a challenge in pinpointing the top performing structures for specific applications. While there are clear limitations in experimentally examining a large number of structures, molecular simulations have provided viable alternative opportunities. Various types of simulations, including Density Functional Theory (DFT) calculations, Grand Canonical Monte Carlo (GCMC) simulations, and Molecular Dynamics (MD) simulations, have been actively used in this context. Though simulations contain a certain amount of error, they have enabled much faster evaluation of structures, thereby contributing to the discovery of novel MOFs. However, the employment of molecular simulations within the vast material space of MOFs still faces several limitations. Even though molecular simulations are relatively fast compared to experimental assessment, evaluating every structure in the database (so-called high-throughput screening[16,17]) is inefficient or often not viable.[13] More importantly, interpreting the large volumes of simulated data is not straightforward, and as such, there is a demand for effective methodologies to analyze accumulated data and extract meaningful information from these data.

Data-driven approaches, represented by machine learning, have become an irreplaceable tool in materials design that successfully addresses the aforementioned challenges. By effectively capturing inherent patterns from large amounts of data, machine learning models have successfully unveiled various structure−property relationships, providing a comprehensive understanding of materials chemistry.[18] Chemical

Received: July 10, 2024
Revised: August 28, 2024
Accepted: August 29, 2024
Published: September 12, 2024

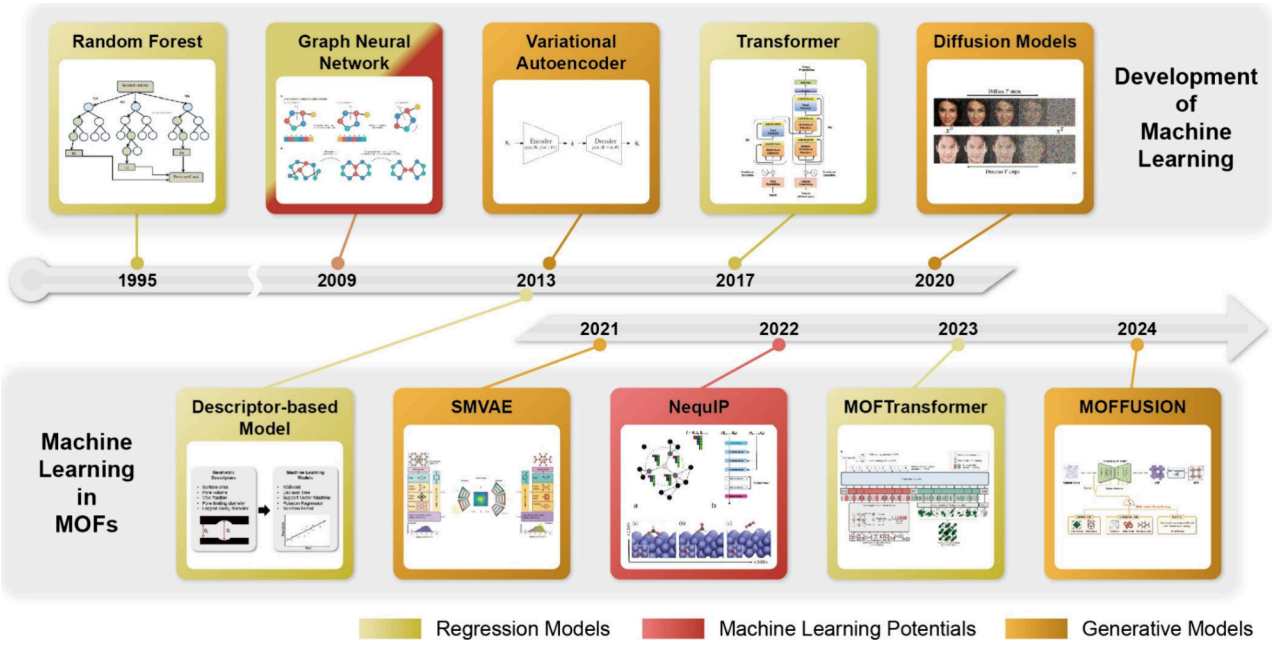


Figure 1. Development of machine learning and its application in MOFs. The emergence of notable machine learning architectures or methodologies and their representative implications in MOFs are presented in chronological order. Elements related to regression models, machine learning potentials, and generative models are colored in yellow, red, and orange, respectively. The graph neural network, a fundamental class of artificial neural networks, is colored in both yellow and red, as it is actively used in developing various machine learning models, including regression models and machine learning potentials. Reprinted with permission from refs 28−36 with the following respective copyrights and licenses. Copyright 2019 Springer Nature (ref 28). Copyright 2024 Springer Nature (ref 29). Copyright 2021 Springer Nature (ref 30). Copyright 2021 Springer Nature (ref 31). licensed under CC BY 4.0 (ref 32). Copyright 2021 Springer Nature (ref 33). Copyright 2022 Springer Nature (ref 34). Copyright 2023 Springer Nature (ref 35). licensed under CC BY-NC (ref 36).

insights provided by these methods have been actively utilized in planning future experiments or in designing new materials with desired characteristics.[19−22] Another recent trend in materials design is the employment of deep generative models, which directly generate chemical structures as outputs without their search space being restricted to training databases.[23,24] Moreover, recently emerging methodologies such as machine learning potentials (MLPs)[25,26] or knowledge transfer between different domains[27] are of great interest, and their incorporation into the various areas of materials science is now encouraging. In this Perspective, we offer a comprehensive outline of recent trends and future directions in machine learning approaches in materials science with a particular focus on MOFs (see Figure 1).

2. STRUCTURE−PROPERTY RELATIONSHIPS AND REGRESSION MODELS

2.1. Importance of Structure−Property Relationship

Elucidating how structural features affect the microscopic/macroscopic properties of materials (so-called structure−property relationships) is one of the primary tasks in the field of materials science.[37,38] Based on the chemical intuitions provided by these relationships, researchers establish design principles, thereby enabling the construction of materials with the desired characteristics for targeted properties. In addition, elucidating the structure−property relationship provides insights into the underlying mechanisms of material behaviors, which can lead to the discovery of new phenomena beyond existing knowledge.

Before the advent of machine learning techniques, elucidating structure−property relationships in materials science relied heavily on experimental characterization and theoretical modeling. Researchers would often hypothesize relationships between the atomic or molecular arrangement within a material and the macroscopic target property and then validate these hypotheses through extensive experimentation and/or theoretical analysis. However, the rapid development and widespread adoption of machine learning methods have provided powerful tools for analyzing complex relationships within vast amounts of data. With the help of various algorithms and architectures, procedures for unveiling structural-property relationships have become much more systematic and efficient over the last few decades. Specifically, regression models,[39] statistical methods used to understand and quantify the relationship between one or more independent variables and a dependent variable, is widely used in this purpose.

Due to the large materials space and high structural complexity, structure−property relationships in MOFs are often complicated and not straightforward. So, elucidating structure−property relationships of MOFs may be more challenging compared with other material domains, which is why utilizing advanced data-driven methodologies becomes even more important. In this section, we focus on the research of MOFs with regression models. We first briefly summarize how the property predictions of MOFs have been conducted so far and pay special attention to recent attempts to utilize the transformer architectures for accurate property predictions.

2.2. Conventional Machine Learning Approaches

With regard to the data-driven approaches to predict material properties, input representation is crucial in determining the overall performance of the model. Early machine learning approaches for MOFs focused on representing structures using

https://doi.org/10.1021/jacsau.4c00618
JACS Au 2024, 4, 3727−3743

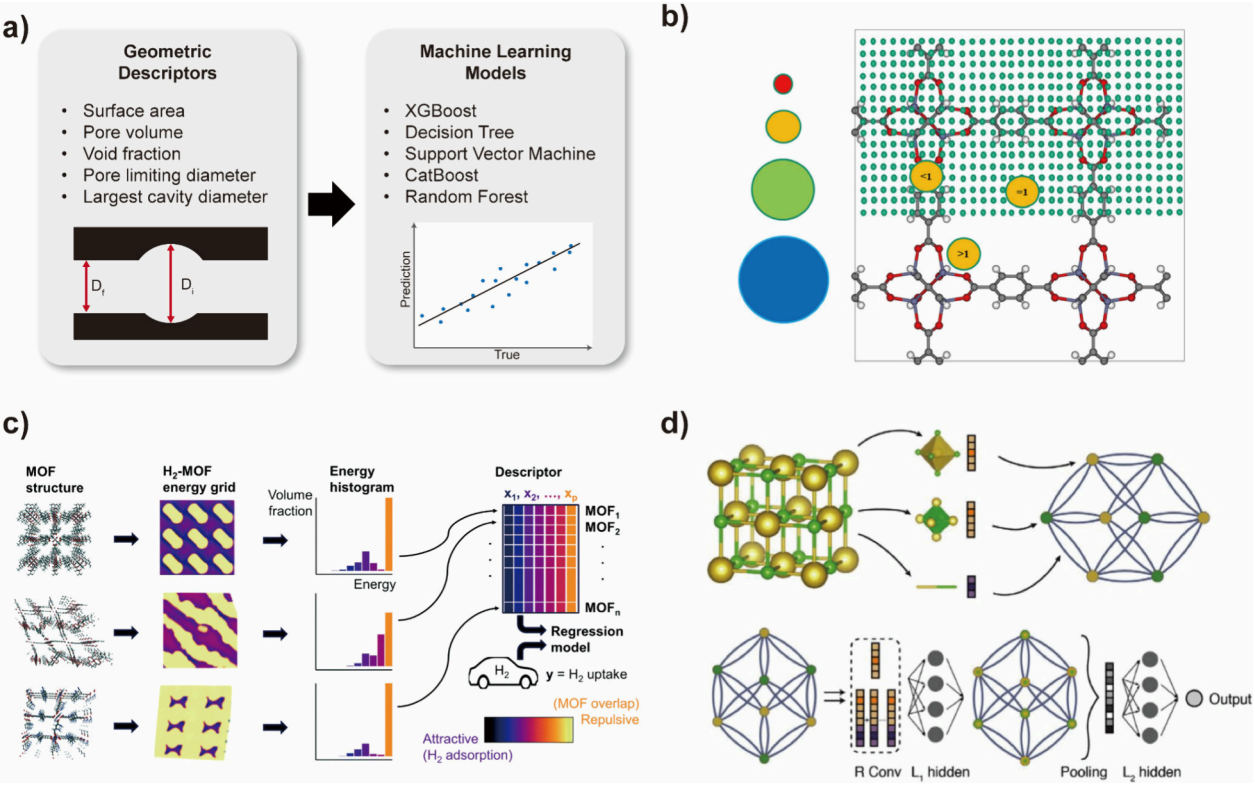


Figure 2. Various machine learning models in MOFs. (a) Scheme of conventional machine learning using geometric descriptors. (b) Example of an energy surface used to calculate the average Boltzmann factor, showing different probe atoms interacting with the scaffold. Reprinted with permission from ref 44. Copyright 2019 American Chemical Society. (c) Model predicting H_2 working capacity using a histogram of energy grids. Used with permission of Royal Society of Chemistry from ref 45. (d) Overview of CGCNN, representing atoms as graph nodes and bonds as graph edges, and using a convolutional neural network to predict properties. Reprinted figure with permission from ref 50. Copyright 2018 American Physical Society.

their geometric descriptors, which is the simplest way of converting MOFs into machine interpretable vectors. These descriptors were typically processed using conventional machine learning models such as the Support Vector Machines (SVM),[40] XGBoost,[41] and Random Forest[42] (Figure 2a). For example, Fernandez et al.[43] introduced a quantitative structure−property relationship (QSPR) for CH_4 uptake in MOFs by utilizing geometric descriptors such as void fraction and volumetric surface area.

Subsequent advances in machine learning models for MOFs involved considering additional relevant chemical descriptors that relied on empirical intuition. Fanourgakis et al.[44] used descriptors derived from the potential energy surface for CH_4 adsorption predictions (Figure 2b). They calculated the average Boltzmann factor by considering different probe atoms interacting with the scaffold and observed a meaningful improvement in the accuracy. Bucior et al.[45] used a histogram of the energy grid to predict the H_2 working capacity of MOFs (Figure 2c). The energy grid here represents the Lennard-Jones and Columbic potential energy between the probe gas and the MOF at each grid points. Using the Least Absolute Shrinkage and Selection Operator (LASSO) model, they achieved a high accuracy (R^2 score of 0.96) and successfully identified 51 promising MOF candidates with a H_2 working capacity greater than 45 g L^{-1}. These attempts highlight the importance of selecting appropriate and innovative descriptors to improve the predictive power of machine learning models in identifying high performing MOFs. Beyond the works mentioned above, there are a number of studies that have utilized intrinsic descriptors for MOF property prediction.[46−48]

Recently, Graph Neural Networks (GNNs[49]) have become one of the most widely used input abstractions in various materials domains, including inorganic compounds and drug-like molecules. By representing atoms as nodes and bonds as edges, graph representations can effectively capture both local atomic environments and long-range structural information. These models have also been expanded to periodic (i.e., crystalline) materials, with the Crystal Graph Convolutional Neural Network (CGCNN)[50] being the most representative example, successively addressing the issue regarding periodic boundary conditions (Figure 2d). As such, Rosen et al.[51] built a QMOF database with more than 20,000 MOF band gap data calculated at the DFT level, and the CGCNN model achieved an R^2 score of 0.876 in predicting band gaps of these structures. In their subsequent work,[52] they achieved even higher performance with the Material Graph Network (MEGNET),[53] another GNN-based prediction model. Recently, Shoghi et al.[54] applied joint multidomain pretraining approach (JMP) based on GemNet-OC[55] and achieved state-of-the-art performance in predicting the bandgap of MOFs. These advancements demonstrate the importance of using GNN models to significantly improve the predictive accuracy of MOF properties by effectively capturing complex interactions within their structures.

2.3. Transformer-Based Regression Models

Lately, property prediction for materials have been advanced with the implementation of transformer-based machine learning models.[35,56−58] Transformers, primarily investigated in natural language processing, are known to efficiently process sequential

https://doi.org/10.1021/jacsau.4c00618
JACS Au 2024, 4, 3727−3743

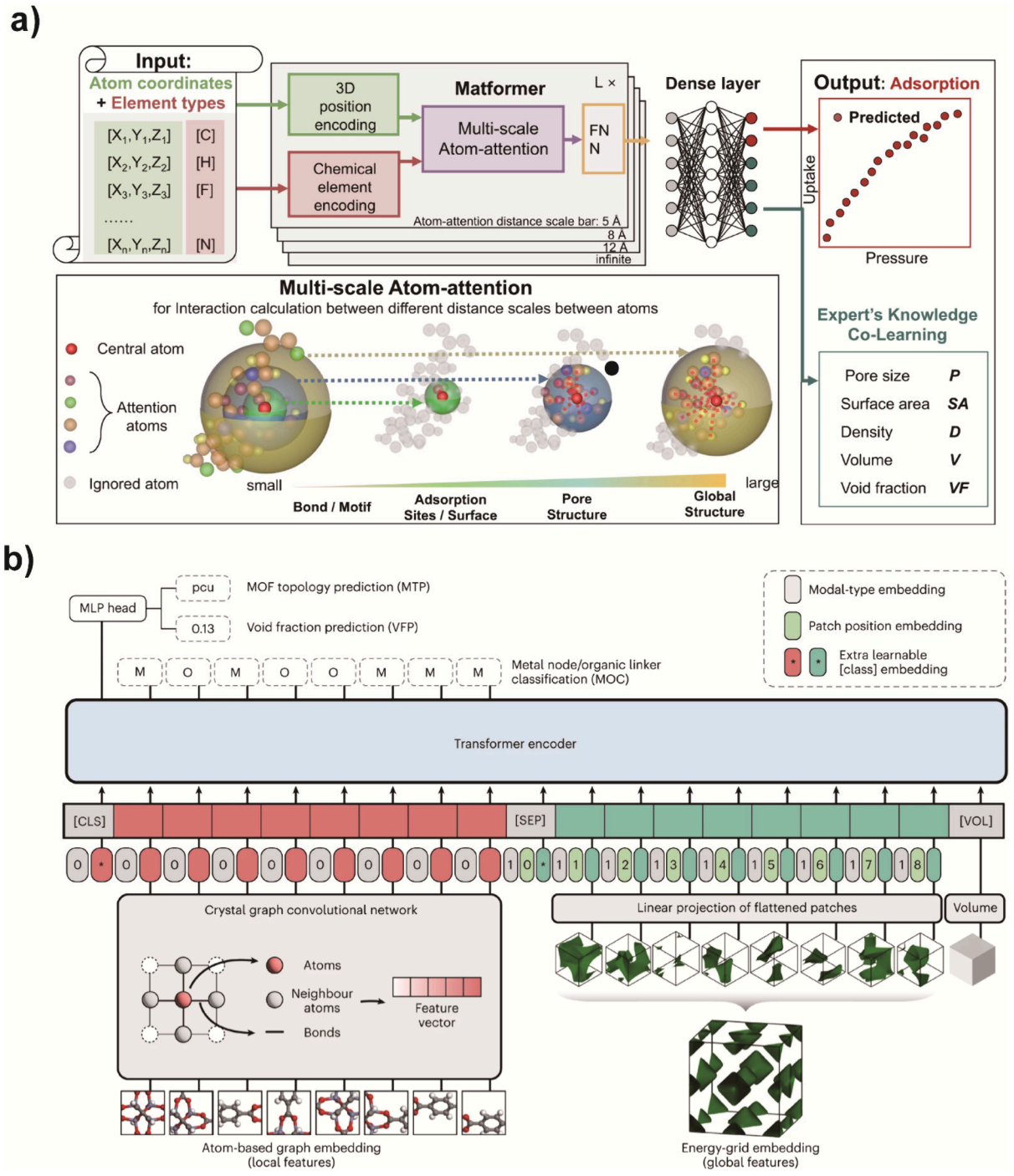


Figure 3. Examples of MOF property prediction models using transformers. (a) Overview of the Deepsorption model, which inputs atom coordinates and element types into a transformer, colearning with expert knowledge on adsorption. Reprinted from ref 57. Copyright 2023 The Authors. (b) Architecture of the MOFTransformer, which considers two different types of input (local features: atom graph, and global features: energy grid) simultaneously in the transformer. The model is pretrained on one million virtual MOFs and fine-tuned to predict desired properties. Reprinted with permission from ref 35. Copyright 2023 Springer Nature.

data using the attention mechanism.[59] Their large number of hyperparameters offers significant potential and versatility, facilitating their wide application in areas such as computer vision and audio processing.

Based on this high potential, there has been a surge of research on MOF property prediction using transformer-based models, capitalizing on their ability to capture complex relationships within data. With regards to MOFs, the MOFormer, developed by Cao et al.[58] used the contextual representation of MOFs as inputs, and pretrained the transformer and CGCNN through self-supervised learning. The pretrained model was fine-tuned for bandgap and gas adsorption, achieving up to 48% higher accuracy compared to models such as stoichiometric-120[60] and those using revised autocorrelations descriptors (RACs[61]), which are discrete correlations between heuristic atomic properties. In DeepSorption, Cui et al.[57] developed a transformer model with atomic coordinates and atom types as inputs (Figure 3a). This model accounts for global structure and local

https://doi.org/10.1021/jacsau.4c00618
JACS Au 2024, 4, 3727−3743

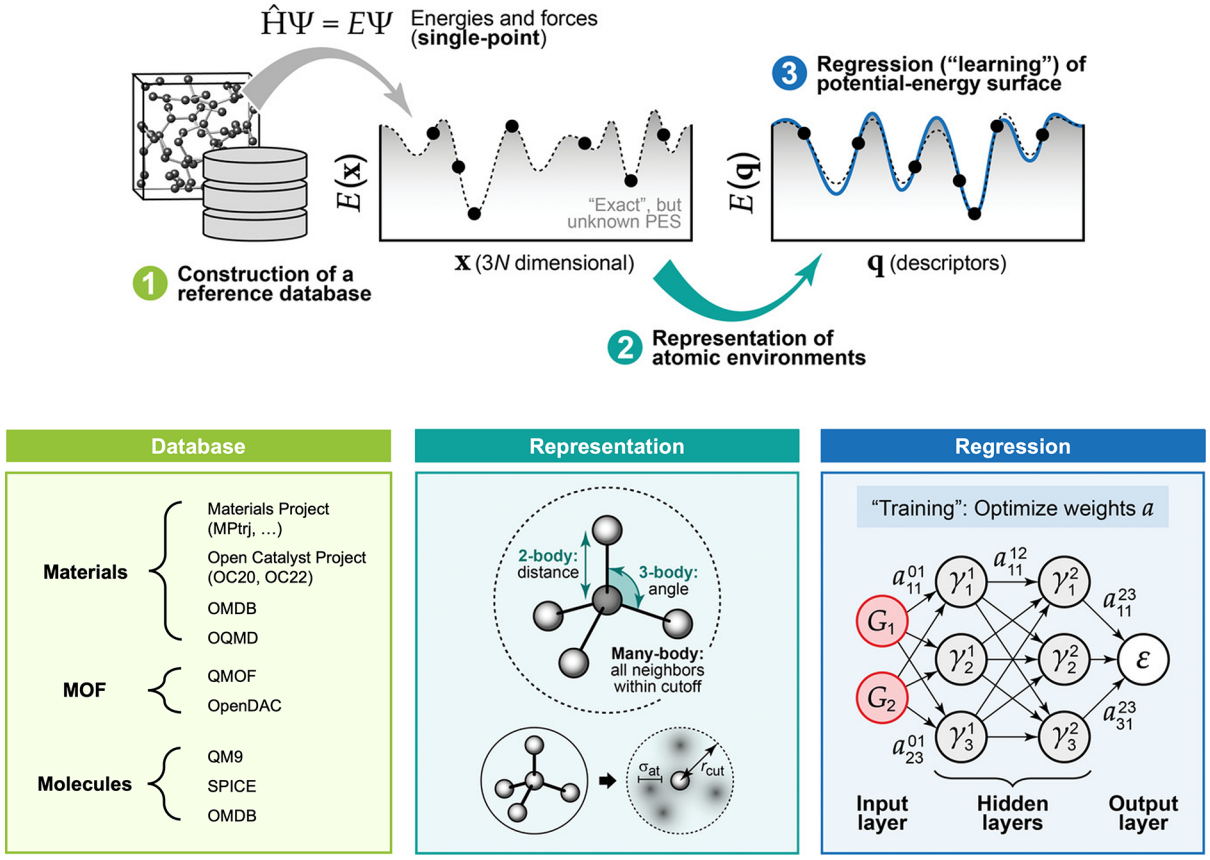


Figure 4. Schematic illustration of constructing an MLP. It consists of three steps: construction of a database through ab initio calculations, representing atomic environments with descriptors, and training a regression model. Used with permission of John Wiley & Sons from ref 70.

spatial atomic interactions using multiscale atomic attention (5, 8, 12, and infinite). In addition, various geometric descriptors were predicted along with the target value during the training process, improving the accuracy and time efficiency. This approach showed a 20−35% reduction in mean absolute error (MAE) compared to those of CGCNN and other descriptor-based models in predicting adsorption curves. Wang et al.[56] developed Uni-MOF by employing self-supervised learning on a transformer model using 631,000 MOFs and COFs. They pretrained the model by predicting the three-dimensional positions of atoms from noisy data and by predicting the masked atoms. This model was then fine-tuned with multi-system properties randomly sampled from various MOF databases, gas types, temperatures, and pressures, achieving gas adsorption predictions with stable R^2 values ranging from 0.83 to 0.98.

Recent research has focused on improving the prediction performance of MOFs using multimodal learning, which considers different types of inputs simultaneously. The MOFTransformer developed by Kang et al.,[35] considered both global features (e.g., geometric and topological descriptors) and local features (e.g., specific bonds and chemistry of the building block) to target various different properties that might being correlated to one regime or another (Figure 3b). Using 1 million hypothetical MOFs, they performed three types of pretraining tasks: topology prediction, void fraction prediction, and metal cluster/organic linker classification. After fine-tuning, their model successfully predicted H_2 uptake, H_2 diffusivity, and bandgap with higher performance than baseline models. Multimodal learning is a new branch that has not been explored in depth for MOFs but has a huge potential for improving the accuracy and robustness of property predictions.

2.4. Potential Weaknesses and Improvements

Despite the significant advancements in property predictions brought about by machine learning, several challenges still remain. Regression models are highly dependent on data, and given that most property data are obtained through molecular simulations, the accuracy of the simulation methods significantly affects the model's performance. Currently, several databases of MOFs, such as CoREMOF[62] and QMOF,[51] exist, but they still contain relatively small amounts of data compared to other domains of materials. Particularly, properties that require DFT calculations, such as band gaps, demand a substantial amount of time to obtain, making it difficult to secure a large data set. Additionally, in the case of MD simulations, including gas diffusivity or bulk modulus calculations, the reproducibility of data is often low, leading to inaccuracies in machine learning predictions. These issues highlight the need for the continued development of more robust databases and improved simulation techniques to increase the reliability and accuracy of machine learning models.

From descriptor-based models to multimodal transformer models, there have been significant advances in property prediction for MOFs. Looking ahead, developing innovative ways to represent MOFs and employing advanced machine learning techniques will be critical for further improving the accuracy and efficiency of these methodologies. In addition, to obtain MOFs with desired properties using regression models, the trained models are often implemented into high-throughput screening (HTS) schemes, replacing traditional molecular

https://doi.org/10.1021/jacsau.4c00618
JACS Au 2024, 4, 3727−3743

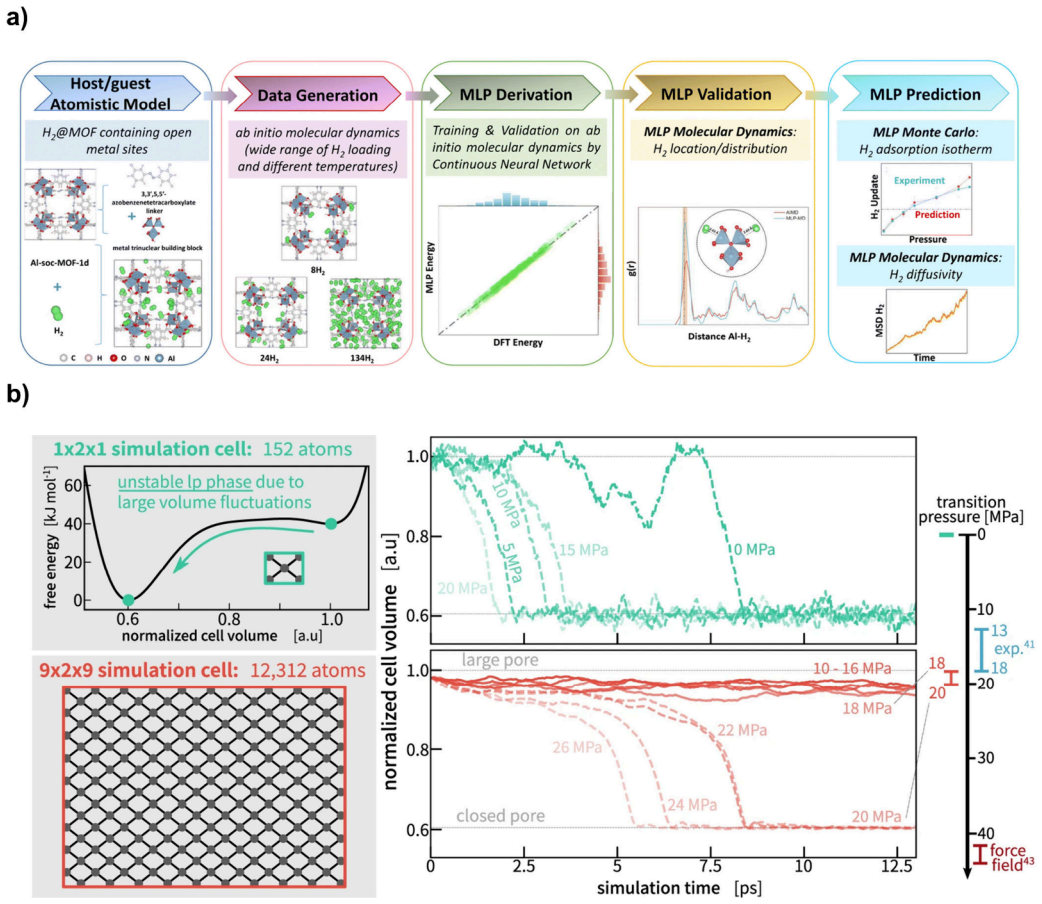


Figure 5. Examples of using MLPs in the context of MOFs. (a) Proposed protocols for conducting molecular simulations using MLP with H_2 molecules and Al-soc-MOF-1d. The trained MLP was used to obtain the H_2 adsorption isotherm and diffusion coefficient. The obtained properties were compared with the experimental values. Reproduced from ref 92 with permission from the Royal Society of Chemistry. (b) The change in volume over time of two different supercell of MIL-53 (Al). The larger simulation cell with more than 10k atoms were simulated using MLP. While there was no large pore (lp) phase for the small supercell, the lp phase was observed for the large supercell, and as such, the transition pressure matched well with the experimental results. Reprinted from ref 96 with permission from Copyright 2023 Springer Nature.

simulations. Frequently, in these endeavors, methods such as genetic algorithms,[13,63,64] active learning,[65] and reinforcement learning[66] are used synchronously to explore the MOF space efficiently. However, there still exists a demand for more accurate and efficient exploration, highlighting the growing need for the development of novel search algorithms.

3. MACHINE LEARNING POTENTIALS

3.1. Brief Introduction to MLPs

Atomistic simulations enable the accumulation of various material property data, but the trade-off between accuracy and computational cost has long been a challenging aspect of exploring large chemical spaces. The widely used DFT calculation is considered to be relatively accurate but slow. For less computational cost, classical (empirical) interatomic potentials are used to simulate molecular dynamics, where Coulomb potential[67] and Lennard-Jones potential[68] are widely used examples. A reactive force filed (ReaxFF)[69] is another notable example that can simulate flexible bond breaking and formation, but nonetheless, these interatomic potentials lack accuracy and are applicable to narrower chemical domains compared to quantum level simulations. In this regard, machine learning potentials (MLPs) overcome the trade-off between accuracy and computational cost by using predictive machine learning models trained on ab initio calculation data, allowing them to learn the potential energy surface at the level of quantum mechanics. These approaches of using MLPs have gained significant attention in various materials domains, and recent works in MLPs are approaching the level of covering a broad range of chemical space with a single transferable model.[25,70,71]

To construct MLP models, one needs to represent atomic environments and train a model to learn the potential energy surface from the selected database (Figure 4). There has been extensive effort to develop descriptors for atomic environments and to train models capable of learning interatomic potentials.[34,72−83] The history and detailed examples are beyond the scope of this article, so one can refer to cited papers to understand the model architecture and concepts. As of the writing this manuscript, MACE[83,84] (specifically MACE-mp-0) demonstrates the highest accuracy among publicly available MLPs in the Matbench Discovery benchmark,[85] which evaluates the performance of models in predicting solid-state thermodynamic stability properties. The MACE architecture combines atomic cluster expansion[73] and message passing neural network,[86,87] where MPTrj data set[81] has been used for training. As neural network architectures have advanced quickly and new ab initio data has been integrated in a compatible manner,[88,89] the field of MLPs in atomistic simulations is growing rapidly.

https://doi.org/10.1021/jacsau.4c00618
JACS Au 2024, 4, 3727−3743

Utilization of MLPs in MOFs is also highly encouraging, as the large and complex chemical environment causes difficulties in their atomistic simulations. In the following sections, examples of MLP implementation for modeling MOFs are demonstrated alongside potential future directions.

3.2. Usage of MLPs in MOFs: Ab Initio Level Energy Calculation

In many MOF systems, generic force field parameters cannot accurately model various adsorption properties, because these force fields were not specifically developed for MOFs, which often contain unique chemical moieties such as open metal sites (OMS). In this regard, MLPs can be an optimal method to calculate adsorption energies within a reasonable time scale and thereby significantly enhance accuracy compared to force field-based calculations.

In particular, Zheng et al.[90] utilized MLP to simulate the potential energy surface along with the insertion of a single CO_2 molecule into Mg-MOF-74, which contains a 5-coordinated Mg−O cluster that induces OMS within the framework. Since quadrupole moments play a role in CO_2, the oxygen part of CO_2 shows strong interaction with the positively charged OMS. Although classical force fields can provide the plausible free energy profile, they still significantly underestimate the interaction between the OMS and the CO_2 molecules. However, the trained MLP in this work accurately accounted for the chemisorption between the OMS and the CO_2 molecule by predicting the intermolecular interaction to be much stronger than that predicted by the classic force field.

Beyond the case of a single molecule, Goeminne et al. trained the MLP with snapshots containing different numbers of guest molecules (CO_2 in this case) from DFT calculations.[91] From this procedure, they efficiently considered the host−guest interactions within the entire framework and captured the subtle changes in host−guest or guest−guest interactions due to the confinement effects of the MOFs. Following the underlying assumption of the widely used conventional GCMC, they collected data with a static framework and rigid adsorbates and extracted interaction energies and forces from single point calculations. They applied this scheme to ZIF-8 and confirmed that MLP can reproduce the experimental adsorption isotherm. However, one limitation encountered in their study was the limited transferability of the MLP. The MLP trained with ZIF-8 cannot be directly used in other polymorphic MOFs that share the same components but exhibit different topologies. From this point of view, the distinct characteristics of MOFs can be a hurdle for the transferability of MLPs in MOFs. Similarly, Liu et al. attempted to run GCMC simulations to accurately reproduce the experimental H_2 adsorption isotherm of Al-soc-MOF-1d, with the aid of MLP (Figure 5a).[92] Given that the MLP was trained using configurations collected from ab initio molecular dynamics (AIMD), one can anticipate that the guest-induced flexibility of the frameworks, which is often overlooked, may be implicitly considered during the adsorption simulation.

Finally, Yu et al. tried to ascertain the stable position of Pt clusters in MOF-808 using MLP.[93] They trained an MLP with a few representative fragments capable of representing the entire system and demonstrated that the trained MLP could accurately predict the optimal position of the Pt clusters and their corresponding energies, closely aligning with the DFT calculation results. Furthermore, given the relatively low computational cost compared to conventional DFT calculations, they were able to readily explore migration pathways, which are typically computationally intensive due to the numerous transition states between the initial and final states.

3.3. Usage of MLPs in MOFs: Ab Initio Level Molecular Dynamics

Given that MLP outputs both interaction energy and atomic forces, it theoretically allows one to simulate MD simulations at the DFT accuracy level. Furthermore, unlike conventional generic force fields widely used for MOFs, MLPs can simulate bond breaking and formation, allowing for the simulation of unpredictable intrinsic structural flexibilities regarding various external stimuli such as temperature, pressure, and gas adsorption. Thus, MLPs may help simulate phenomena observed in experiments that cannot be reproduced using conventional DFT calculations or force field based calculations. Additionally, they may provide new insights into experimentalists by revealing previously undiscovered behavior related to MOFs.

In the early stages of MLP development in MOFs, structures were decomposed into molecular fragments, and the MLP was trained based on these units.[94,95] These trained models were able to accurately predict the equilibrium lattice constant and be applied for more complex applications such as mechanistic and thermal behaviors.[95] However, this method fundamentally has a limitation: the periodicity cannot be explicitly considered at the training level. To address this issue, several follow-up studies used the entire unit cell as a training set for MLPs. Vandenhaute et al.[96] trained the MLP model named NequIP with an incremental learning approach that used the concept of metadynamics to efficiently sample configurations to train the model. In this work, UiO-66 (Zr) and MIL-53 (Al) were selected as representative rigid and flexible MOFs, respectively. With the merits of MLPs, which can simulate large cells with high accuracy, they expanded the simulation cell to $9 \times 2 \times 9$ for MIL-53 (Al) (comprising more than 10 000 atoms) and successfully reproduced the experimental transition pressure data (Figure 5b). In this work, they investigated the framework's flexibility in response to external pressure, but the guest-induced flexibility has not yet been studied using MLP. Therefore, it would be a great suggestion for a future research topic. Similarly, Fan et al.[97] rationally designed two-dimensional MOFs that potentially exhibit phase transitions due to wine-rack motif and observed the mechanistic behaviors using an MLP model. While previous studies have matched simulation results to experimental ones, recent study from the same group utilized MLPs to unravel phenomena not observed in experiments at the molecular level.[98] They investigated the dynamics of CALF-20 and observed negative area compressibility (NAC) and negative thermal expansion (NTE), which are counterintuitive phenomena not observed experimentally with CALF-20.

In addition to the deformation within the crystalline frameworks, Castel et al.[26] used MLP to investigate the amorphous phases of MOFs at the molecular level. They selected ZIF-4 as the benchmark MOF, which has been shown to possess multiple amorphous phases in experiments. The amorphous phases were obtained using melt-quenching methods, which can intuitively be embodied using molecular dynamics. Since bonds should remain intact in conventional generic force fields, the breaking of the crystallinity of MOFs, which involves reversible bond breaking and formation, cannot be simulated. However, as previously mentioned, bond breaking and formation can be simulated using MLP similar to DFT calculations, enabling investigation of the loss of crystallinity.

https://doi.org/10.1021/jacsau.4c00618
JACS Au 2024, 4, 3727−3743

3.4. Potential Applications of MLPs in MOFs

Publicly available quantum calculation data to train MLPs are limited considering the vast chemical space of MOFs that originates from their high tunability and modularity. Hence, employing MLPs in MOFs is slower than in other materials, which makes it worthwhile to explore MLP applications in different domains and adapt them to the MOF.

Most MOF simulations involving quantum calculations are restricted to considering only solid state frameworks due to computational costs. However, this approach can lead to discrepancies compared to experiments, as the fixed framework assumption may not fully capture real-world behaviors. On the other hand, MLPs have been proven to accurately simulate the physicochemical properties of various liquid systems and heterogeneous solid−liquid systems.[99−102] Given that MOFs often interact with liquid systems such as water or other solvents, the simulation of heterogeneous systems including MOFs using MLPs represents an important area for future research. Likewise, MLPs can offer an attractive solution for research on MOFs in heterogeneous catalysis. In other materials, MLPs have been used to predict reaction pathways, free energy diagrams of heterogeneous catalysts, and the chemical environment of surfaces.[103−108] Given the growing interest in using MOFs as potential heterogeneous catalysts, deployment of MLPs can accelerate this research field.

In addition, MLPs increase the accessible scale of simulation systems. Through MLPs, amorphous phase of MOFs or MOFs with defects can be simulated, which have been challenging with conventional methods.[70,109−112] Also, MLPs hold promise for simulating the bottom-up generation process of MOFs, which is anticipated to provide a deeper understanding of MOFs and their synthetic procedure.

Despite potential usages of MLPs in MOFs, the main bottleneck remains the lack of organized quantum calculation data. The QMOF database[51] (containing quantum-level calculated chemical properties of 20,000+ MOFs) and a recently published OpenDAC database[113] (containing 40 M DFT calculations, targeting direct air capture using MOFs), provide enormous help and support to deploy MLPs in MOFs. Nonetheless, a larger amount of high-quality data is required to accelerate the development of MLPs in MOFs. The increased availability of quantum calculation data will drive the development of MLPs targeting MOFs, thereby enhancing the capability of atomistic simulations at the mesoscale, which is crucial for MOFs.

4. GENERATIVE MODELS FOR MOF DESIGN

4.1. Generative Models and Their Applications in Materials Design

In recent years, the emergence of deep generative models has brought about huge success in the field of computer science. Unlike conventional discriminative models, by learning the intrinsic probability distributions of the data, generative models are capable of generating new data points which resemble a given data set. Various generative model architectures, including Variational Autoencoders (VAEs),[114] Generative Adversarial Networks (GANs),[115] and diffusion models,[116] have been developed and have achieved significant advancements in various domains. Notably, computer vision[117,118] and natural language processing[119,120] are areas where the success of generative models is particularly prominent. Their exceptional performance in generating realistic images has led to the development of image generators such as DALL-E[121,122] and Midjourney,[123] while large language models like ChatGPT[124] have become irreplaceable tools in various applications.

The huge success of generative models witnessed in other domains has facilitated their implementation in the field of materials science. Notably, the generation process of these models often initiate with the sampling of a random noise or a random vector within the latent space.[114,116] Therefore, they offer theoretically unlimited search space without being restricted to a given data set, which is a fascinating characteristic in generating novel chemical structures. In addition, generation process of these models can be intentionally biased for the acquirement of desired samples, which is called conditional generation.[125,126] Through such processes, materials with user-desired properties can be selectively generated, achieving the inverse design of materials.

Generative models have been widely investigated for various domains, including drug-like molecules,[127,128] proteins,[129,130] and small crystals.[131] However, as discussed in previous sections, the implementation of new technologies into porous materials is often relatively slow. This is due to their structural complexity, which often comprises a relatively large number of atoms spanning a wider range of atom types.[13,132] In addition, the periodic nature of porous materials is another main bottleneck when dealing with crystalline structures, which slows the introduction of cutting-edge methodologies.

Nonetheless, the industrial importance of porous materials has motivated researchers to continually explore them, and few pioneering studies have reported the implementation of generative models for these materials. Kim et al.[133] have applied GAN architecture for the generation of zeolites, which is a class of porous materials that have received significant industrial attention. A GAN is a representative generative model composed of two components, a generator and a discriminator, that compete with each other to produce realistic data and evaluate its authenticity. They represented the zeolites as three channels of grids containing information on the location of silicon atoms, oxygen atoms, and energy of probe gas molecules, respectively. They successfully generated structures unseen in training sets and validated the capability of generating structures with user-desired properties. A follow-up study by Park et al.[134] replaced the GAN with the diffusion model, while utilizing the same representations as the previous work. The diffusion model is one of the most recent generative model architectures, which iteratively refine noisy data to generate high-quality samples by simulating a diffusion process. Accordingly, Park et al. observed a 2,000-fold improvement in the generation performance, validating that diffusion models exhibit exceptional performance in materials generation. Conditional generation was aimed at void fraction, heat of adsorption, and Henry coefficient, and has successfully generated structures with the desired chemical properties.

4.2. Generative Models for MOFs

Compared with zeolites, MOFs exhibit much greater structural and chemical complexity. There are more than 100 types of atoms known to comprise MOFs,[135] and the average number of atoms in the unit cell is much larger. Consequently, the number of experimentally synthesized MOFs exceeds 100,000,[14] while there are only ∼200 experimentally validated zeolite structures.[136] Therefore, implementing MOFs into generative models is more challenging and may require alternative input

https://doi.org/10.1021/jacsau.4c00618
JACS Au 2024, 4, 3727−3743

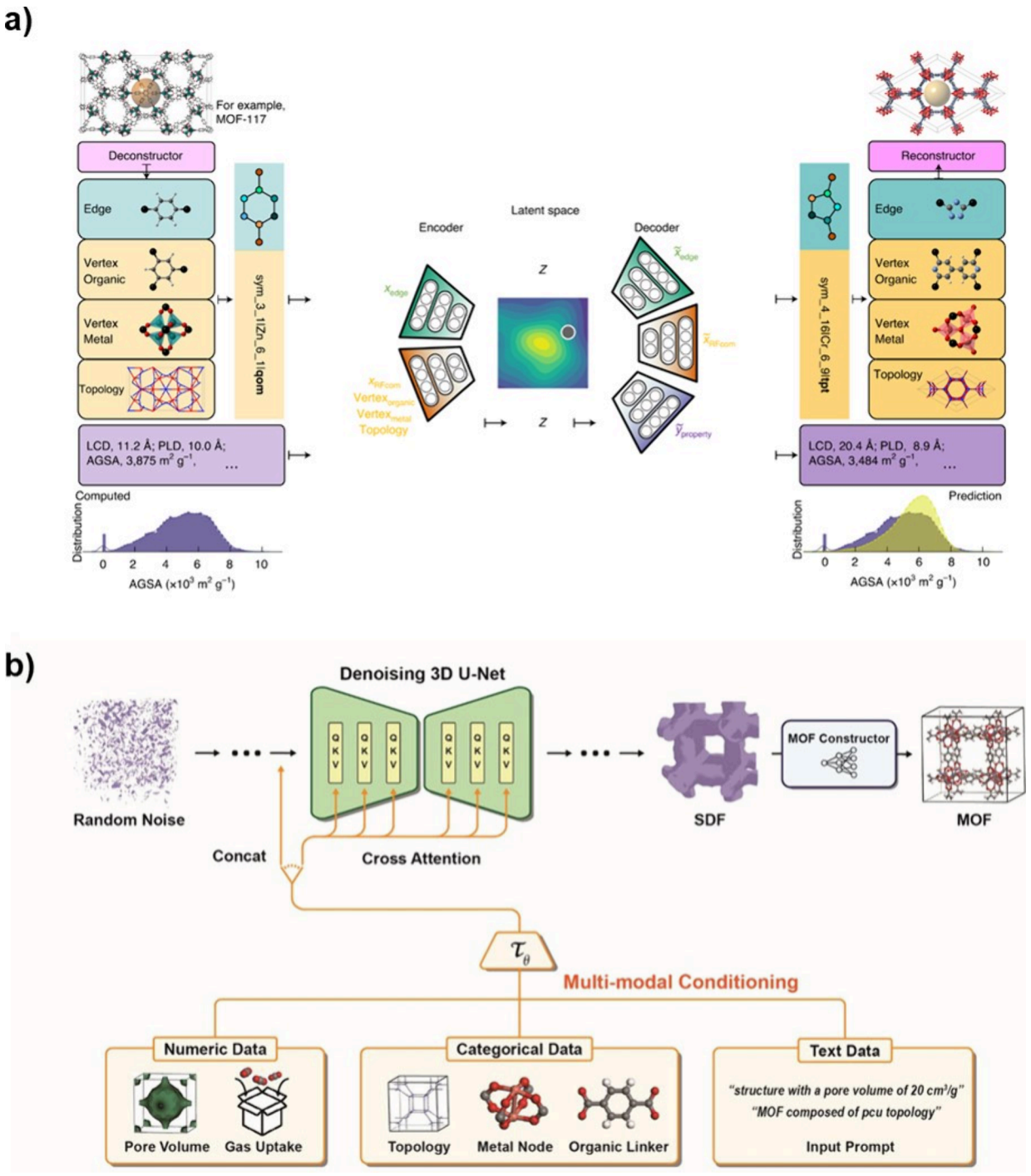


Figure 6. Generative models for MOFs. (a) SMVAE Architecture. A variational autoencoder was trained for the generation of the MOFs. MOFs were represented as a combination of their building components (topology, metal node, organic node, and edge). Reprinted from ref 33 with permission from Copyright 2021 Springer Nature. (b) MOFFUSION Architecture. A diffusion model was trained for MOF generation using a signed distance function (SDF) as an input representation. Reprinted from ref 36 licensed under CC BY-NC.

representations rather than representing them as a whole graph or coordinating each and every atom within the 3D coordinate.

Yao et al.[33] have utilized VAE for the generation of MOFs, named SMVAE, marking the first attempt at implementing generative models for MOFs (Figure 6a). Empowered by the modular nature of MOFs, they represented MOFs as a combination of topology and building blocks (i.e., metal nodes, organic nodes, and edges), which is a simple but effective way of representing complicated MOF structures. SMVAE is composed of two parts: the edge encoder/decoder and reticular framework encoder/decoder. The encoder components map the given MOF structures into latent vectors with reduced dimensions, while the decoders map them back to the original MOF structures. The entire architecture is trained jointly, and during the generation phase, the trained decoders enable the generation of new structures from randomly sampled latent vectors. SMVAE exhibited 61.5% of structure validity, measured by the prior validity of the sampled vectors. (Among randomly

https://doi.org/10.1021/jacsau.4c00618
JACS Au 2024, 4, 3727−3743

sampled latent vectors, 61.5% resulted in valid MOF structures.) By training an additional property prediction model that predicts chemical properties of MOFs from latent vectors, they were able to generate structures with desired properties. Focusing on CO_2 adsorption, they demonstrated that the model is capable of controlling the CO_2 adsorption capacity and selectivity.

Among various generative model architectures, diffusion models have gained huge attention with their remarkable performance in generating realistic samples and versatile conditioning capabilities.[116,118] Application of diffusion models into MOF generation has recently been attempted with several different approaches to deal with the structural complexity of MOFs. Park et al.[137] concentrated in generating MOF linkers (rather than the entire structure) and plugged the generated linkers into pcu topology with fixed types of metal nodes (Cu paddlewheel, Zn paddlewheel, and Zn tetramer). They utilized DiffLinker,[128] which is a predeveloped diffusion model, for the generation of linker molecules based on chemical fragments. They defined the structures whose CO_2 capacity is higher than 2 mmol/g at 0.1 bar as high-performing MOFs, and obtained 540 unique molecular fragments extracted from these high-performing MOFs in pre-existing databases. Using these fragments, they have obtained 12,305 linkers and based on them, obtained 120,000 MOF structures. They screened their MOF space using a separate predictive model and validated six MOFs with a CO_2 capacity higher than 2 mmol/g.

Fu et al.[138] utilized an elegant way of dealing the structural complexity of MOFs by incorporating coarse-grained representation of building blocks. With raising the claim that a template-based approach can restrict the search space and exclude viable materials, they generated MOFs by directly positioning coarse grained building blocks within the 3D coordinate. An additional MOF assembly process was required as a postprocessing step, for the alignment of the orientations of building blocks and determination of the connectivity between them. They trained an additional regression model which predicts CO_2 working capacity from the latent vectors, and utilized this for the selective generation of structure with high CO_2 capacities.

Most recently, Park et al.[36] have developed a diffusion model for MOF generation, with a particular focus on handling diverse modalities of data during the conditional generation (Figure 6b). They pointed out that the target properties considered in inverse design research for porous materials have been restricted so far in terms of flexibility and claimed the necessity of handling diverse data modalities during conditional generation. Notably, they used signed distance functions (SDFs) as an input representation to delicately describe the pore structures of MOFs, and their model achieved a high structural validity of 81.7%. Furthermore, they showcased the capability of conditioning on diverse modalities of data, including numeric, categorical, text data, and even their combinations.

4.3. Strengths and Weaknesses

Materials generation using generative models has several strengths over other material design methodologies. In general, the training process of generative models follows unsupervised learning, where a model is trained on data without explicit labels or annotations. Generative models can learn from unlabeled data by capturing the underlying distribution; therefore, the generative models can be beneficial in cases where labeled data are scarce or expensive to obtain. In addition, generative models can explore vast (often theoretically unlimited) search space, which is an attractive feature when it comes to generating diverse and novel chemcial structures. Lastly, as introduced in previous research, materials with target properties can be designed through conditional generation.

With more diverse and versatile conditioning methodologies being investigated, one interesting research direction is the incorporation of natural language processing as one of the conditions. Recently, large language models have been deployed to construct user-friendly platform for materials design.[139] Notably for MOFs, Kang et al. have developed a module named ChatMOF[140] which enables users to communicate with the module using their natural language. Likewise, handling text data during the conditional generation is a direction where generation models are pursing.[141] This allows users to express their desired features in natural language, significantly lowering the barrier to using these models, especially for those without domain knowledge.

However, there still exist several fundamental challenges when it comes to generative models for materials design. As previously mentioned, generative models work by learning the probability distributions of the given training data, and this fundamentally makes generative models highly dependent on their training data, where a bias in training data set would lead to the same bias in generated samples. This dependency can pose significant challenges for materials generation, where the construction of a perfectly nonbiased data set is frequently difficult to achieve. Therefore, considerable effort should be made in preparing a diverse yet evenly distributed training data set when training generative models.

Another aspect worth sharing is that generative models are less suitable for extrapolation. As these models utilize the learned patterns from the training data, they are capable of generating new samples that are similar to the training data (i.e., interpolation). However, generating data points that lie outside the range of the training data can be challenging. For example, tasks such as generating MOFs with an unprecedentedly high H_2 capacity could be challenging for these models. Nonetheless, research on generation models for extrapolation is actively ongoing,[142] and through this, it is anticipated that more fundamental solutions for such tasks will emerge.

5. FUTURE DIRECTIONS AND EMERGING TRENDS

5.1. Enhancing Speed and Efficiency

While conventional machine learning models have about 10,000 parameters, regression models using the transformer architecture can have up to 100 million parameters.[35,56−58] The increase in the number of parameters improves the model's potential to learn more complex patterns and relationships, thereby increasing its accuracy and flexibility. However, as the number of parameters increases, the training process becomes significantly longer and the amount of data required for training increases exponentially.[143,144] This presents a challenge in terms of computational resources and data acquisition, highlighting the need for efficient training strategies and robust data acquisition methods.

In this context, transfer learning is considered a promising method to reduce both the training time and the amount of data required. Transfer learning is a machine learning technique where a predeveloped model serves as the foundation model that can be applied to other tasks.[145] It typically involves training a model on a large data set (known as the pretrained model) and

https://doi.org/10.1021/jacsau.4c00618
JACS Au 2024, 4, 3727−3743

then initializing the training process for a specific task of interest (referred to as fine-tuning), often with smaller learning rates. This method can be effective when it comes to systems in which the number of data is relatively small. Transfer learning thus holds significant promise for enhancing the efficiency and accuracy of property prediction models for MOFs, which often exhibits complicated data distributions and insufficient data.[35,146,147] In a similar context, knowledge transfer between different domains is also a promising approach, which is discussed in the following section.

Furthermore, lightweight deep learning models can be considered to enhance the efficiency of their use in materials science. Lightweighting of models refers to designing models with a reduced number of parameters while preserving performance.[148] This practice is particularly important for handling large amounts of data, as it can mitigate the slow inference speeds and the resource-intensive pretraining steps. In this context, MLPs are noteworthy, as recent advancements have led to increased accuracy,[149] but also larger model complexity and size, resulting in greater memory usage and computational costs. While MLPs have significant potential, in order to develop a method with high applicability, researchers should also consider the burden of large model sizes in addition to the model accuracy. Various techniques are being explored for lightweighting machine learning models, such as knowledge distillation,[150,151] where smaller models are taught what larger models have learned, and model architectures are modified to reduce their size.[152−155] Although lightweighting has not yet been extensively studied in materials science, focusing research efforts on model efficiency within the trend of increasing model parameters could lead to the development of models with shorter learning times while maintaining high performance.

5.2. Merging Tools/Data Set from Other Materials

Due to complex structures and the need for high-fidelity data, data scarcity is one of the main bottlenecks that data-driven approaches frequently face, especially in porous materials research.[156] Particularly, this problem is prominent in fields where the property of interest is hard to simulate, thus experimental assessment is demanded (e.g., proton conductivity[157,158]), or where simulation data is expensive (e.g., electronic structure calculation[159]). For such cases of having limited data for the specific task, if one has access to a large data set from another field, then transferring that data to your specific task would be a viable approach to remedy the issue of data scarcity. Especially for porous materials, as diverse classes of porous materials (e.g., zeolite, MOFs, COFs, PPNs, etc.) have been investigated for several decades, utilizing the accumulated data across porous materials domains (or even from nonporous materials) could propose a new paradigm in their investigation using data-driven approaches.

One of the representative examples of utilizing cross-domain knowledge is the work by He et al.[160] With the goal of verifying the electrically conductive MOFs, they developed a binary classification model that classifies whether a structure is conductive or not based on its band gap. At the time, there was no database specifically for MOF band gaps. (However, a QMOF database,[51] now provides DFT-calculated band gaps of MOF structures.) To overcome this limitation, they utilized the accumulated data on inorganic compounds. The Open Quantum Materials Database (OQMD)[161] contains band gap information for ∼52,300 inorganic compounds calculated at DFT level. They trained their classification model on this data set and applied it to the MOF database to identify structures likely to be conductive. Their methodology predicted nine conductive MOFs, with subsequent computational validation confirming that six of them were metallic. This serves as an adequate example of using data from another domain to address the data scarcity issue in the field of MOF research.

Transfer learning is a widely used method for the transfer of knowledge between different domains. Cai et al.[162] explored the feasibility of implementing transfer learning in the context of heterogeneous porous materials. Their focus was on addressing the imbalance between the sizes of the computation-ready experimental (CoRE) data sets, where there are relatively few CoRE COF structures (about 1,000 structures) compared to CoRE MOF (with over 10,000 structures).[14] In order to achieve high accuracy in predicting Xe/Kr selectivity with CoRE COFs, the researchers employed transfer learning. Specifically, they pretrained the model using the CoRE MOF database and fine-tuned it with 300 COF structures. They confirmed that transfer learning consistently achieved higher accuracy compared to direct learning, which solely utilized the COF database.

Park et al.[163] presented another compelling application of transfer learning in porous materials. They trained a transformer architecture across various domains of porous materials, encompassing MOFs, COFs, zeolites, and PPNs. Empowered by a synergistic effect among different material classes, they aimed to address data imbalances via transfer learning. Their approach involved proposing a cross-material few-shot learning scheme. By transferring knowledge obtained from MOFs, they observed an enhancement in the accuracy of predicting the band gap of COFs in few-shot learning scenarios. Additionally, the authors further explored the feasibility of a zero-shot learning scheme where a model trained solely on material A could be applied for the prediction of unseen material B.

As another representative example of knowledge transfer between porous materials, Sun et al.[27] utilized a meta-learning scheme to seek for the optimal H_2 storage conditions. Meta-learning and transfer learning are both approaches that aim to improve the efficiency of training by using prior knowledge. However, the goal of meta-learning is slightly different, as it aims to train a model that can generalize well across a variety of tasks. With utilizing high-throughput data for zeolites, MOFs, and hyper-cross-linked polymers, they developed a meta-learning model capable of jointly predicting the adsorption capacity for various materials across a wide range of pressure and temperature. It was verified that a model trained using the meta-learning method exhibits higher accuracy compared to separately trained models for each material.

Several precedent works that utilized knowledge transfer across material domains to address the issue of data scarcity are covered in this section. As more data accumulate rapidly and new databases are constructed, the utilization of data without being restricted to the material class of interest is becoming a common approach. Even beyond dealing with data scarcity, chemical intuition obtained from other domains can enhance the performance of the model in the target domain. Therefore, as an intense amount of data continues to accumulate in the future, appropriately utilizing knowledge and data from other domains will become an increasingly important strategy.

5.3. Data Extraction from Literature

The quality of the data is crucial for developing reliable machine learning models. In this section, we focus on the discrepancies between real-world data and data obtained from simulations.

https://doi.org/10.1021/jacsau.4c00618
JACS Au 2024, 4, 3727−3743

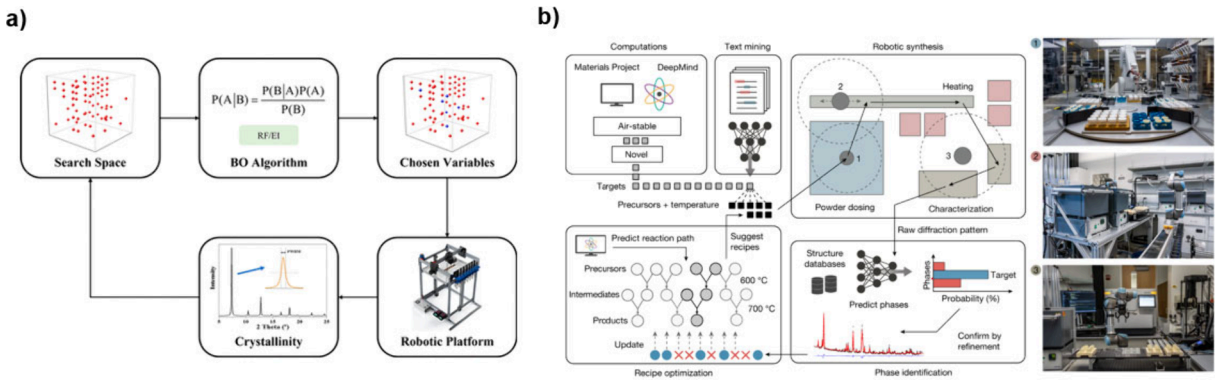


Figure 7. Autonomous materials discovery. (a) Schematic showing a process of synthesizing ZIF-67 with improved crystallinity via the integration of a robotic platform with a Bayesian optimization algorithm. Reprinted with permission from ref 21. Copyright 2021 American Chemical Society. (b) Autonomous materials discovery with the autonomous lab. Powder dosing, sample heating, and product characterization were automated while the transfer between these stations is performed using robotic arms. Reprinted with permission from ref 172. Copyright 2023 Springer Nature.

Most existing MOF databases are generated through simulations, which allow them to cover vast chemical spaces with high time efficiency. However, simulation methods are not perfectly accurate and can introduce small or large errors. Additionally, simulations are often conducted under assumptions such as perfect crystallinity, which can lead to discrepancies from real materials. To overcome these limitations, researchers have attempted to extract data from the experimental literature (often referred to as data mining) so that experimental data can be directly utilized for model training.

In the early stages of data mining for MOFs, rule-based extraction methodologies and relatively simple natural language processing (NLP) techniques were employed. Park et al.[164] developed a text mining algorithm that captures units for surface area and pore volume from HTML format inputs. Nandy et al.[165] used NLP and image analysis techniques to mine solvent-removal stability and thermal degradation temperatures of MOFs from experimental literature. The scope of data mining later expanded to include synthetic information on MOFs, where synthesis conditions of MOFs were extracted from experimental procedures and further processed for deeper understanding on their synthesis.[166,167] More recently, large language models have emerged as an irreplaceable tool in data mining. Compared with rule-based algorithms, the use of large language models has made the mining process more flexible and accurate. These models have also impacted the MOF community, with the Yaghi group employing ChatGPT for the data mining of MOF data.[168,169]

Data mining is an evolving field and is expected to play a crucial role in constructing high-quality experimental databases. The MOF domain is particularly suited to active mining given the large volume of ongoing experimental work. One factor that could drive rapid progress in this field is the establishment of standardized practices in data reporting and management. If researchers adopt a uniform format for data reporting in the scientific literature, the accuracy of data mining could be significantly enhanced. Additionally, since researchers typically report only the top (or best) experimental results, the volume of experimental data could be dramatically increased if they also publish their failed or partially successful outcomes.

5.4. Automation of Experiments and Autonomous Laboratories

Lastly, in this section, we highlight promising and futuristic approaches of data acquisition. Automation of the experimental process is an emerging concept in materials discovery that has gained huge traction in recent days.[170,171] This is made possible by using robotic systems, where robotic arms and other mechanical systems handle tasks such as sample preparation, mixing chemicals, and operating instruments. Such automation possesses huge advantages compared with experiments carried out by humans. Automated experiments can run continuously (i.e., 24 h per day/7 days a week) with higher efficiency, thereby allowing researchers to obtain more experimental data in an efficient manner. Additionally, it minimizes the risk of human error in experimental procedures, leading to more reliable and reproducible results. Due to these benefits, automation of experimental processes is an attractive method for obtaining large volumes of high-quality data.

The implementation of autonomous systems has also been applied to the field of MOFs. In research carried out by Moosavi et al.,[20] they utilized an automated system to acquire experimental data on the synthesis of HKUST-1. Their system was able to carry out 30 reactions per cycle, where one cycle was completed within a single day. Through repeated experiments, they verified that the robotic platform provides a consistent synthesis protocol with high reproducibility and good control over the synthesis variables. They represented the synthesis conditions as 9-dimensional synthesis vectors and implemented genetic algorithm to find the optimal synthesis conditions. The synthesis served as the chromosome, and through selection, crossover, and mutation, the synthesis conditions were optimized generation by generation. As a result, they achieved the HKUST-1 sample with the highest surface area reported to date.

In addition, Xie et al.[21] used a robotic platform for the rapid synthesis of ZIF-67 via the Joule heating method (Figure 7a). Their robotic platform controlled two chemical parameters (a molar ratio of Co ions to 2-methylimidazole and a total volume of precursors) and two processing parameters (applied DC voltage and reaction time). Using the data provided by the robotic system, they applied a Bayesian optimization algorithm to optimize the synthesis conditions of ZIF-67. Bayesian optimization suggests new synthesis conditions based on past evaluations (crystallinity of the samples in this work). It was observed that as iterations proceeded, structures with gradually higher crystallinity were obtained. Through this process and further analysis, they were able to retrieve chemical insights into the crystallinity of ZIF-67.

https://doi.org/10.1021/jacsau.4c00618
JACS Au 2024, 4, 3727−3743

An autonomous lab (also called a self-driving lab) is a more advanced and futuristic concept, in which laboratory processes and experiments are conducted without human intervention (Figure 7b).[172,173] The facility typically features robotics, artificial intelligence (AI), and multiple sensors, where sophisticated AI algorithms manage and optimize experimental parameters, analyze data in real-time, and autonomously make decisions based on predefined objectives. Active learning is often referred to in this context, as it helps in selecting the next set of experiments, thereby making the whole process more efficient. Few materials domains, including organic semiconductor lasers (OSLs), are being tested for the introduction of autonomous lab, yet several obstacles remain, including both mechanical and software challenges.[171] To the best of our knowledge, MOF synthesis in a completely autonomous lab with zero human intervention has not been reported yet, but it is our opinion that the efforts of researchers will make it possible in the near future.

6. CONCLUSIONS

In this Perspective, we introduced the recent trends of machine learning studies in MOFs and the future directions pursued by the researchers. Regression models for elucidating structure−property relationships were first covered, from conventional methods to transformer-based models, which have recently shown exceptional performance. Subsequently, we covered the recent trends in machine learning studies with a particular focus on MLPs, which have immense potential across a wide range of applications. MLPs are now gaining attention in the field of MOFs, with the anticipation that they will pave the way for the complicated chemistry of MOFs to be handled in an effective manner. Generative models are another emerging branch offering a new paradigm in materials design. We shared recent attempts to implement them in MOFs, along with some aspects of materials design using generative models. Lastly, several challenges encountered using machine learning methodologies, and efforts to address them have been discussed.

Machine learning has played a critical role in shifting the paradigm of research in the materials field. MOFs are a domain where their achievements are particularly pronounced, and significant advancements have been witnessed through the introduction of machine learning-based methodologies. Through continued research efforts, machine learning is poised to unlock the full potential of MOFs, driving further progress and innovation in materials science.

AUTHOR INFORMATION

Corresponding Author

Jihan Kim − *Department of Chemical and Biomolecular Engineering, Korea Advanced Institute of Science and Technology (KAIST), Daejeon 34141, Republic of Korea;* orcid.org/0000-0002-3844-8789; Email: jihankim@kaist.ac.kr

Authors

Junkil Park − *Department of Chemical and Biomolecular Engineering, Korea Advanced Institute of Science and Technology (KAIST), Daejeon 34141, Republic of Korea;* orcid.org/0000-0003-1770-7291

Honghui Kim − *Department of Chemical and Biomolecular Engineering, Korea Advanced Institute of Science and Technology (KAIST), Daejeon 34141, Republic of Korea;* orcid.org/0000-0002-1527-8641

Yeonghun Kang − *Department of Chemical and Biomolecular Engineering, Korea Advanced Institute of Science and Technology (KAIST), Daejeon 34141, Republic of Korea;* orcid.org/0000-0001-5191-5735

Yunsung Lim − *Department of Chemical and Biomolecular Engineering, Korea Advanced Institute of Science and Technology (KAIST), Daejeon 34141, Republic of Korea;* orcid.org/0000-0002-6327-9286

Complete contact information is available at: https://pubs.acs.org/10.1021/jacsau.4c00618

Author Contributions

The manuscript was written through contributions of all authors. All authors have given approval to the final version of the manuscript.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

This project was supported by National Research Foundation of Korea (NRF) under grant No. RS-2024-00337004.

REFERENCES

(1) Zhang, X.; Lin, R. B.; Wang, J.; Wang, B.; Liang, B.; Yildirim, T.; Zhang, J.; Zhou, W.; Chen, B. Optimization of the pore structures of MOFs for record high hydrogen volumetric working capacity. *Advanced materials* **2020**, *32* (17), 1907995.

(2) Nguyen, T.; Shimizu, G.; Rajendran, A. Post-Combustion CO2 capture by vacuum swing adsorption using a hydrophobic metal-organic framework (MOF), CALF-20: Multi-objective optimization and experimental validation. *ChemRxiv*, 2022.

(3) Li, B.; Wen, H.-M.; Zhou, W.; Xu, J. Q.; Chen, B. Porous metal-organic frameworks: promising materials for methane storage. *Chem.* **2016**, *1* (4), 557−580.

(4) Park, J.; Suh, B. L.; Kim, J. Computational design of a photoresponsive metal−organic framework for post combustion carbon capture. *J. Phys. Chem. C* **2020**, *124* (24), 13162−13167.

(5) Wang, Q.; Astruc, D. State of the art and prospects in metal−organic framework (MOF)-based and MOF-derived nanocatalysis. *Chem. Rev.* **2020**, *120* (2), 1438−1511.

(6) Doonan, C. J.; Sumby, C. J. Metal−organic framework catalysis. *CrystEngComm* **2017**, *19* (29), 4044−4048.

(7) Shan, Y.; Zhang, G.; Shi, Y.; Pang, H. Synthesis and catalytic application of defective MOF materials. *Cell Reports Physical Science* **2023**, *4* (3), 101301.

(8) Jo, Y. M.; Jo, Y. K.; Lee, J. H.; Jang, H. W.; Hwang, I. S.; Yoo, D. J. MOF-based chemiresistive gas sensors: toward new functionalities. *Adv. Mater.* **2023**, *35* (43), 2206842.

(9) Small, L. J.; Henkelis, S. E.; Rademacher, D. X.; Schindelholz, M. E.; Krumhansl, J. L.; Vogel, D. J.; Nenoff, T. M. Near-zero power MOF-based sensors for NO2 detection. *Adv. Funct. Mater.* **2020**, *30* (50), 2006598.

(10) Mallakpour, S.; Nikkhoo, E.; Hussain, C. M. Application of MOF materials as drug delivery systems for cancer therapy and dermal treatment. *Coord. Chem. Rev.* **2022**, *451*, 214262.

(11) Suresh, K.; Matzger, A. J. Enhanced drug delivery by dissolution of amorphous drug encapsulated in a water unstable metal−organic framework (MOF). *Angew. Chem., Int. Ed.* **2019**, *58* (47), 16790−16794.

(12) Colón, Y. J.; Gómez-Gualdrón, D. A.; Snurr, R. Q. Topologically guided, automated construction of metal−organic frameworks and their evaluation for energy-related applications. *Cryst. Growth Des.* **2017**, *17* (11), 5801−5810.

(13) 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* (20), 23647−23654.
(14) Chung, Y. G.; Haldoupis, E.; Bucior, B. J.; Haranczyk, M.; Lee, S.; Zhang, H.; Vogiatzis, K. D.; Milisavljevic, M.; Ling, S.; Camp, J. S.; et al. Advances, updates, and analytics for the computation-ready, experimental metal−organic framework database: CoRE MOF 2019. *Journal of Chemical & Engineering Data* **2019**, *64* (12), 5985−5998.
(15) Moghadam, P. Z.; Li, A.; Wiggin, S. B.; Tao, A.; Maloney, A. G.; Wood, P. A.; Ward, S. C.; Fairen-Jimenez, D. Development of a Cambridge Structural Database subset: a collection of metal−organic frameworks for past, present, and future. *Chem. Mater.* **2017**, *29* (7), 2618−2625.
(16) Ahmed, A.; Seth, S.; Purewal, J.; Wong-Foy, A. G.; Veenstra, M.; Matzger, A. J.; Siegel, D. J. Exceptional hydrogen storage achieved by screening nearly half a million metal-organic frameworks. *Nat. Commun.* **2019**, *10* (1), 1568.
(17) Colón, Y. J.; Snurr, R. Q. High-throughput computational screening of metal−organic frameworks. *Chem. Soc. Rev.* **2014**, *43* (16), 5735−5749.
(18) Chong, S.; Lee, S.; Kim, B.; Kim, J. Applications of machine learning in metal-organic frameworks. *Coord. Chem. Rev.* **2020**, *423*, 213487.
(19) Wei, J.; Chu, X.; Sun, X. Y.; Xu, K.; Deng, H. X.; Chen, J.; Wei, Z.; Lei, M. Machine learning in materials science. *InfoMat* **2019**, *1* (3), 338−358.
(20) 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* (1), 539.
(21) Xie, Y.; Zhang, C.; Deng, H.; Zheng, B.; 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* (45), 53485−53491.
(22) Zuluaga, M.; Krause, A. e-pal: An active learning approach to the multi-objective optimization problem. *Journal of Machine Learning Research* **2016**, *17* (104), 1−32.
(23) Fuhr, A. S.; Sumpter, B. G. Deep generative models for materials discovery and machine learning-accelerated innovation. *Frontiers in Materials* **2022**, *9*, 865270.
(24) Sanchez-Lengeling, B.; Aspuru-Guzik, A. Inverse molecular design using machine learning: Generative models for matter engineering. *Science* **2018**, *361* (6400), 360−365.
(25) Behler, J. Perspective: Machine learning potentials for atomistic simulations. *J. Chem. Phys.* **2016**, *145* (17), 170901.
(26) Castel, N.; André, D.; Edwards, C.; Evans, J. D.; Coudert, F.-X. Machine learning interatomic potentials for amorphous zeolitic imidazolate frameworks. *Digital Discovery* **2024**, *3*, 355−368.
(27) Sun, Y.; DeJaco, R. F.; Li, Z.; Tang, D.; Glante, S.; Sholl, D. S.; Colina, C. M.; Snurr, R. Q.; Thommes, M.; Hartmann, M.; et al. Fingerprinting diverse nanoporous materials for optimal hydrogen storage conditions using meta-learning. *Science Advances* **2021**, *7* (30), eabg3983.
(28) Lakshmanaprabu, S.; Shankar, K.; Ilayaraja, M.; Nasir, A. W.; Vijayakumar, V.; Chilamkurti, N. Random forest for big data classification in the internet of things using optimal features. *International journal of machine learning and cybernetics* **2019**, *10* (10), 2609−2618.
(29) Corso, G.; Stark, H.; Jegelka, S.; Jaakkola, T.; Barzilay, R. Graph neural networks. *Nature Reviews Methods Primers* **2024**, *4* (1), 17.
(30) Pinheiro Cinelli, L.; Araújo Marins, M.; Barros da Silva, E. A.; Lima Netto, S. Variational autoencoder. In *Variational Methods for Machine Learning with Applications to Deep Networks*; Springer, 2021; pp 111−149.
(31) Aurpa, T. T.; Sadik, R.; Ahmed, M. S. Abusive Bangla comments detection on Facebook using transformer-based deep learning models. *Social Network Analysis and Mining* **2022**, *12* (1), 24.
(32) Lyu, Z.; Xu, X.; Yang, C.; Lin, D.; Dai, B. Accelerating diffusion models via early stop of the diffusion process. *arXiv*, 2205.12524, 2022.
(33) Yao, Z.; Sánchez-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.; et al. Inverse design of nanoporous crystalline reticular materials with deep generative models. *Nature Machine Intelligence* **2021**, *3* (1), 76−86.
(34) Batzner, S.; Musaelian, A.; Sun, L.; Geiger, M.; Mailoa, J. P.; Kornbluth, M.; Molinari, N.; Smidt, T. E.; Kozinsky, B. E (3)-equivariant graph neural networks for data-efficient and accurate interatomic potentials. *Nat. Commun.* **2022**, *13* (1), 2453.
(35) Kang, Y.; Park, H.; Smit, B.; Kim, J. A multi-modal pre-training transformer for universal transfer learning in metal−organic frameworks. *Nature Machine Intelligence* **2023**, *5* (3), 309−318.
(36) Park, J.; Lee, Y.; Kim, J. Multi-modal conditioning for metal-organic frameworks generation using 3D modeling techniques. *ChemRxiv*, 2024. DOI: 10.26434/chemrxiv-2024-w8fps.
(37) Canivet, J.; Bonnefoy, J.; Daniel, C.; Legrand, A.; Coasne, B.; Farrusseng, D. Structure−property relationships of water adsorption in metal−organic frameworks. *New J. Chem.* **2014**, *38* (7), 3102−3111.
(38) Wu, D.; Yang, Q.; Zhong, C.; Liu, D.; Huang, H.; Zhang, W.; Maurin, G. Revealing the structure−property relationships of metal−organic frameworks for CO2 capture from flue gas. *Langmuir* **2012**, *28* (33), 12094−12099.
(39) Fahrmeir, L.; Kneib, T.; Lang, S.; Marx, B.; Fahrmeir, L.; Kneib, T.; Lang, S.; Marx, B. *Regression models*; Springer, 2013.
(40) Hearst, M. A.; Dumais, S. T.; Osuna, E.; Platt, J.; Scholkopf, B. Support vector machines. *IEEE Intelligent Systems and their applications* **1998**, *13* (4), 18−28.
(41) Chen, T.; Guestrin, C. Xgboost: A scalable tree boosting system. *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* **2016**, 785−794.
(42) Breiman, L. Random forests. *Machine learning* **2001**, *45*, 5−32.
(43) Fernandez, M.; Woo, T. K.; Wilmer, C. E.; Snurr, R. Q. Large-scale quantitative structure−property relationship (QSPR) analysis of methane storage in metal−organic frameworks. *J. Phys. Chem. C* **2013**, *117* (15), 7681−7689.
(44) Fanourgakis, G. S.; Gkagkas, K.; Tylianakis, E.; Klontzas, E.; Froudakis, G. A robust machine learning algorithm for the prediction of methane adsorption in nanoporous materials. *J. Phys. Chem. A* **2019**, *123* (28), 6080−6087.
(45) 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. *Molecular Systems Design & Engineering* **2019**, *4* (1), 162−174.
(46) Wu, X.; Xiang, S.; Su, J.; Cai, W. Understanding quantitative relationship between methane storage capacities and characteristic properties of metal−organic frameworks based on machine learning. *J. Phys. Chem. C* **2019**, *123* (14), 8550−8559.
(47) Burner, J.; Luo, J.; White, A.; Mirmiran, A.; Kwon, O.; Boyd, P. G.; Maley, S.; Gibaldi, M.; Simrod, S.; Ogden, V.; et al. ARC−MOF: a diverse database of metal-organic frameworks with DFT-derived partial atomic charges and descriptors for machine learning. *Chem. Mater.* **2023**, *35* (3), 900−916.
(48) Yuan, X.; Li, L.; Shi, Z.; Liang, H.; Li, S.; Qiao, Z. Molecular-fingerprint machine-learning-assisted design and prediction for high-performance MOFs for capture of NMHCs from air. *Advanced Powder Materials* **2022**, *1* (3), 100026.
(49) Zhou, J.; Cui, G.; Hu, S.; Zhang, Z.; Yang, C.; Liu, Z.; Wang, L.; Li, C.; Sun, M. Graph neural networks: A review of methods and applications. *AI open* **2020**, *1*, 57−81.
(50) Xie, T.; Grossman, J. C. Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties. *Physical review letters* **2018**, *120* (14), 145301.
(51) Rosen, A. S.; Iyer, S. M.; Ray, D.; Yao, Z.; 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* (5), 1578−1597.
(52) Rosen, A. S.; Fung, V.; Huck, P.; O'Donnell, C. T.; Horton, M. K.; Truhlar, D. G.; Persson, K. A.; Notestein, J. M.; Snurr, R. Q. High-throughput predictions of metal−organic framework electronic proper-

https://doi.org/10.1021/jacsau.4c00618
JACS Au 2024, 4, 3727−3743

ties: theoretical challenges, graph neural networks, and data exploration. *npj Computational Materials* **2022**, *8* (1), 1−10.
(53) Chen, C.; Ye, W.; Zuo, Y.; Zheng, C.; Ong, S. P. Graph networks as a universal machine learning framework for molecules and crystals. *Chem. Mater.* **2019**, *31* (9), 3564−3572.
(54) Shoghi, N.; Kolluru, A.; Kitchin, J. R.; Ulissi, Z. W.; Zitnick, C. L.; Wood, B. M. From molecules to materials: Pre-training large generalizable models for atomic property prediction. *arXiv*, 2310.16802, 2023.
(55) Gasteiger, J.; Shuaibi, M.; Sriram, A.; Günnemann, S.; Ulissi, Z.; Zitnick, C. L.; Das, A. Gemnet-oc: developing graph neural networks for large and diverse molecular simulation datasets. *arXiv*, 2204.02782, 2022.
(56) Wang, J.; Liu, J.; Wang, H.; Zhou, M.; Ke, G.; Zhang, L.; Wu, J.; Gao, Z.; Lu, D. A comprehensive transformer-based approach for high-accuracy gas adsorption predictions in metal-organic frameworks. *Nat. Commun.* **2024**, *15* (1), 1904.
(57) Cui, J.; Wu, F.; Zhang, W.; Yang, L.; Hu, J.; Fang, Y.; Ye, P.; Zhang, Q.; Suo, X.; Mo, Y.; et al. Direct prediction of gas adsorption via spatial atom interaction learning. *Nat. Commun.* **2023**, *14* (1), 7043.
(58) 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* (5), 2958−2967.
(59) Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. *Advances in neural information processing systems* **2017**, *30*, 6000−6010.
(60) Meredig, B.; Agrawal, A.; Kirklin, S.; Saal, J. E.; Doak, J. W.; Thompson, A.; Zhang, K.; Choudhary, A.; Wolverton, C. Combinatorial screening for new materials in unconstrained composition space with machine learning. *Phys. Rev. B* **2014**, *89* (9), No. 094104.
(61) 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* (1), 1−10.
(62) 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* (21), 6185−6192.
(63) Lim, Y.; Park, J.; Lee, S.; Kim, J. Finely tuned inverse design of metal−organic frameworks with user-desired Xe/Kr selectivity. *Journal of Materials Chemistry A* **2021**, *9* (37), 21175−21183.
(64) Park, J.; Lim, Y.; Lee, S.; Kim, J. Computational design of metal−organic frameworks with unprecedented high hydrogen working capacity and high synthesizability. *Chem. Mater.* **2023**, *35* (1), 9−16.
(65) Jose, A.; Devijver, E.; Jakse, N.; Poloni, R. Informative Training Data for Efficient Property Prediction in Metal−Organic Frameworks by Active Learning. *J. Am. Chem. Soc.* **2024**, *146* (9), 6134−6144.
(66) Park, H.; Majumdar, S.; Zhang, X.; Kim, J.; Smit, B. Inverse design of metal−organic frameworks for direct air capture of CO 2 via deep reinforcement learning. *Digital Discovery* **2024**, *3* (4), 728−741.
(67) Coulomb, C. A. Premier mémoire sur l'électricité et le magnétisme. *Histoire de l'Academie royale des sciences* **1785**, 569.
(68) Lennard-Jones, J. E. Cohesion. *Proceedings of the Physical Society* **1931**, *43* (5), 461.
(69) Van Duin, A. C.; Dasgupta, S.; Lorant, F.; Goddard, W. A. ReaxFF: a reactive force field for hydrocarbons. *J. Phys. Chem. A* **2001**, *105* (41), 9396−9409.
(70) Deringer, V. L.; Caro, M. A.; Csányi, G. Machine learning interatomic potentials as emerging tools for materials science. *Adv. Mater.* **2019**, *31* (46), 1902765.
(71) Mueller, T.; Hernandez, A.; Wang, C. Machine learning for interatomic potential models. *J. Chem. Phys.* **2020**, *152* (5), No. 050902.
(72) Musil, F.; Grisafi, A.; Bartók, A. P.; Ortner, C.; Csányi, G.; Ceriotti, M. Physics-inspired structural representations for molecules and materials. *Chem. Rev.* **2021**, *121* (16), 9759−9815.
(73) Drautz, R. Atomic cluster expansion for accurate and transferable interatomic potentials. *Phys. Rev. B* **2019**, *99* (1), No. 014104.
(74) Bartók, A. P.; Kondor, R.; Csányi, G. On representing chemical environments. *Phys. Rev. B* **2013**, *87* (18), 184115.
(75) Rupp, M.; Tkatchenko, A.; Müller, K.-R.; Von Lilienfeld, O. A. Fast and accurate modeling of molecular atomization energies with machine learning. *Physical review letters* **2012**, *108* (5), No. 058301.
(76) Huo, H.; Rupp, M. Unified representation for machine learning of molecules and crystals. *arXiv*, 1704.06439, 2017.
(77) Braams, B. J.; Bowman, J. M. Permutationally invariant potential energy surfaces in high dimensionality. *Int. Rev. Phys. Chem.* **2009**, *28* (4), 577−606.
(78) Wang, H.; Zhang, L.; Han, J.; Weinan, E. DeePMD-kit: A deep learning package for many-body potential energy representation and molecular dynamics. *Comput. Phys. Commun.* **2018**, *228*, 178−184.
(79) Batatia, I.; Batzner, S.; Kovács, D. P.; Musaelian, A.; Simm, G. N.; Drautz, R.; Ortner, C.; Kozinsky, B.; Csányi, G. The design space of e (3)-equivariant atom-centered interatomic potentials. *arXiv*, 2205.06643, 2022.
(80) Lei, X.; Medford, A. J. A universal framework for featurization of atomistic systems. *J. Phys. Chem. Lett.* **2022**, *13* (34), 7911−7919.
(81) Deng, B.; Zhong, P.; Jun, K.; Riebesell, J.; Han, K.; Bartel, C. J.; Ceder, G. CHGNet as a pretrained universal neural network potential for charge-informed atomistic modelling. *Nature Machine Intelligence* **2023**, *5* (9), 1031−1041.
(82) Chen, C.; Ong, S. P. A universal graph deep learning interatomic potential for the periodic table. *Nature Computational Science* **2022**, *2* (11), 718−728.
(83) Batatia, I.; Kovacs, D. P.; Simm, G.; Ortner, C.; Csányi, G. MACE: Higher order equivariant message passing neural networks for fast and accurate force fields. *Advances in Neural Information Processing Systems* **2022**, *35*, 11423−11436.
(84) Batatia, I.; Benner, P.; Chiang, Y.; Elena, A. M.; Kovács, D. P.; Riebesell, J.; Advincula, X. R.; Asta, M.; Baldwin, W. J.; Bernstein, N. A foundation model for atomistic materials chemistry. *arXiv*, 2401.00096, 2023.
(85) Riebesell, J.; Goodall, R. E.; Jain, A.; Benner, P.; Persson, K. A.; Lee, A. A. Matbench Discovery--An evaluation framework for machine learning crystal stability prediction. *arXiv*, 2308.14920, 2023.
(86) Gilmer, J.; Schoenholz, S. S.; Riley, P. F.; Vinyals, O.; Dahl, G. E. Neural message passing for quantum chemistry. In *International conference on machine learning*, 2017; PMLR: pp 1263−1272.
(87) Bronstein, M. M.; Bruna, J.; Cohen, T.; Veličković, P. Geometric deep learning: Grids, groups, graphs, geodesics, and gauges. *arXiv*, 2104.13478, 2021.
(88) Jain, A.; Hautier, G.; Ong, S. P.; Moore, C. J.; Fischer, C. C.; Persson, K. A.; Ceder, G. Formation enthalpies by mixing GGA and GGA+ U calculations. *Phys. Rev. B* **2011**, *84* (4), No. 045115.
(89) Kingsbury, R. S.; Rosen, A. S.; Gupta, A. S.; Munro, J. M.; Ong, S. P.; Jain, A.; Dwaraknath, S.; Horton, M. K.; Persson, K. A. A flexible and scalable scheme for mixing computed formation energies from different levels of theory. *npj Computational Materials* **2022**, *8* (1), 195.
(90) 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 CO2's Chemisorption and Diffusion in Mg-MOF-74. *ACS Nano* **2023**, *17* (6), 5579−5587.
(91) Goeminne, R.; Vanduyfhuys, L.; Van Speybroeck, V.; Verstraelen, T. DFT-Quality adsorption simulations in metal−organic frameworks enabled by machine learning Potentials. *J. Chem. Theory Comput.* **2023**, *19* (18), 6313−6325.
(92) Liu, S.; Dupuis, R.; Fan, D.; Benzaria, S.; Bonneau, M.; Bhatt, P.; Eddaoudi, M.; Maurin, G. Machine learning potential for modelling H 2 adsorption/diffusion in MOFs with open metal sites. *Chemical Science* **2024**, *15* (14), 5294−5302.
(93) Yu, Y.; Zhang, W.; Mei, D. Artificial neural network potential for encapsulated platinum clusters in mof-808. *J. Phys. Chem. C* **2022**, *126* (2), 1204−1214.
(94) 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* (6), 3793−3809.

https://doi.org/10.1021/jacsau.4c00618
JACS Au 2024, 4, 3727−3743

(95) Tayfuroglu, O.; Kocak, A.; Zorlu, Y. A neural network potential for the IRMOF series and its application for thermal and mechanical behaviors. *Phys. Chem. Chem. Phys.* **2022**, *24* (19), 11882−11897.
(96) 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 Computational Materials* **2023**, *9* (1), 1−8.
(97) Fan, D.; Ozcan, A.; Lyu, P.; Maurin, G. Unravelling abnormal in-plane stretchability of two-dimensional metal−organic frameworks by machine learning potential molecular dynamics. *Nanoscale* **2024**, *16* (7), 3438−3447.
(98) Fan, D.; Naskar, S.; Maurin, G. Unconventional mechanical and thermal behaviours of MOF CALF-20. *Nat. Commun.* **2024**, *15* (1), 3251.
(99) Balyakin, I.; Rempel, S.; Ryltsev, R.; Rempel, A. Deep machine learning interatomic potential for liquid silica. *Phys. Rev. E* **2020**, *102* (5), No. 052125.
(100) Quaranta, V.; Behler, J. r.; Hellström, M. Structure and dynamics of the liquid−water/zinc-oxide interface from machine learning potential simulations. *J. Phys. Chem. C* **2019**, *123* (2), 1293−1304.
(101) Schran, C.; Thiemann, F. L.; Rowe, P.; Müller, E. A.; Marsalek, O.; Michaelides, A. Machine learning potentials for complex aqueous systems made simple. *Proc. Natl. Acad. Sci. U. S. A.* **2021**, *118* (38), e2110077118.
(102) Yao, N.; Chen, X.; Fu, Z.-H.; Zhang, Q. Applying classical, ab initio, and machine-learning molecular dynamics simulations to the liquid electrolyte for rechargeable batteries. *Chem. Rev.* **2022**, *122* (12), 10970−11021.
(103) Goldsmith, B. R.; Esterhuizen, J.; Liu, J. X.; Bartel, C. J.; Sutton, C. Machine learning for heterogeneous catalyst design and discovery. *AIChE J.* **2018**, 2311.
(104) Mou, T.; Pillai, H. S.; Wang, S.; Wan, M.; Han, X.; Schweitzer, N. M.; Che, F.; Xin, H. Bridging the complexity gap in computational heterogeneous catalysis with machine learning. *Nature Catalysis* **2023**, *6* (2), 122−136.
(105) Stocker, S.; Jung, H.; Csányi, G.; Goldsmith, C. F.; Reuter, K.; Margraf, J. T. Estimating free energy barriers for heterogeneous catalytic reactions with machine learning potentials and umbrella integration. *J. Chem. Theory Comput.* **2023**, *19* (19), 6796−6804.
(106) Chen, D.; Shang, C.; Liu, Z.-P. Machine-learning atomic simulation for heterogeneous catalysis. *npj Computational Materials* **2023**, *9* (1), 2.
(107) Chanussot, L.; Das, A.; Goyal, S.; Lavril, T.; Shuaibi, M.; Riviere, M.; Tran, K.; Heras-Domingo, J.; Ho, C.; Hu, W.; et al. Open catalyst 2020 (OC20) dataset and community challenges. *ACS Catal.* **2021**, *11* (10), 6059−6072.
(108) Tran, R.; Lan, J.; Shuaibi, M.; Wood, B. M.; Goyal, S.; Das, A.; Heras-Domingo, J.; Kolluru, A.; Rizvi, A.; Shoghi, N.; et al. The Open Catalyst 2022 (OC22) dataset and challenges for oxide electrocatalysts. *ACS Catal.* **2023**, *13* (5), 3066−3084.
(109) Sivaraman, G.; Krishnamoorthy, A. N.; Baur, M.; Holm, C.; Stan, M.; Csányi, G.; Benmore, C.; Vázquez-Mayagoitia, Á. Machine-learned interatomic potentials by active learning: amorphous and liquid hafnium dioxide. *npj Computational Materials* **2020**, *6* (1), 104.
(110) Deringer, V. L.; Csányi, G. Machine learning based interatomic potential for amorphous carbon. *Phys. Rev. B* **2017**, *95* (9), No. 094203.
(111) Xu, N.; Shi, Y.; He, Y.; Shao, Q. A deep-learning potential for crystalline and amorphous Li−Si alloys. *J. Phys. Chem. C* **2020**, *124* (30), 16278−16288.
(112) Freitas, R.; Cao, Y. Machine-learning potentials for crystal defects. *MRS Commun.* **2022**, *12* (5), 510−520.
(113) Sriram, A.; Choi, S.; Yu, X.; Brabson, L. M.; Das, A.; Ulissi, Z.; Uyttendaele, M.; Medford, A. J.; Sholl, D. S. The Open DAC 2023 Dataset and Challenges for Sorbent Discovery in Direct Air Capture. *ACS Central Science* **2024**, *10* (5), 923−941.
(114) Kingma, D. P.; Welling, M. Auto-encoding variational bayes. *arXiv*, 1312.6114, 2013.
(115) Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative adversarial networks. *Communications of the ACM* **2020**, *63* (11), 139−144.
(116) Ho, J.; Jain, A.; Abbeel, P. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems* **2020**, *33*, 6840−6851.
(117) Wang, Z.; She, Q.; Ward, T. E. Generative adversarial networks in computer vision: A survey and taxonomy. *ACM Computing Surveys (CSUR)* **2022**, *54* (2), 1−38.
(118) Dhariwal, P.; Nichol, A. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems* **2021**, *34*, 8780−8794.
(119) Chowdhary, K.; Chowdhary, K. Natural language processing. *Fundamentals of artificial intelligence* **2020**, 603−649.
(120) Miao, Y. *Deep generative models for natural language processing*. Ph.D. Thesis, University of Oxford, 2017.
(121) Ramesh, A.; Pavlov, M.; Goh, G.; Gray, S.; Voss, C.; Radford, A.; Chen, M.; Sutskever, I. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, 2021; PMLR: pp 8821−8831.
(122) Ramesh, A.; Dhariwal, P.; Nichol, A.; Chu, C.; Chen, M. Hierarchical text-conditional image generation with clip latents. *arXiv*, 2204.06125, 2022.
(123) *Midjourney*. 2022. https://www.midjourney.com (accessed 28 Aug, 2024).
(124) Bubeck, S.; Chandrasekaran, V.; Eldan, R.; Gehrke, J.; Horvitz, E.; Kamar, E.; Lee, P.; Lee, Y. T.; Li, Y.; Lundberg, S. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv*, 2303.12712, 2023.
(125) Zhang, L.; Rao, A.; Agrawala, M. Adding conditional control to text-to-image diffusion models. *Proceedings of the IEEE/CVF International Conference on Computer Vision* **2023**, 3836−3847.
(126) Kang, S.; Cho, K. Conditional molecular design with deep generative models. *J. Chem. Inf. Model.* **2019**, *59* (1), 43−52.
(127) Hoogeboom, E.; Satorras, V. G.; Vignac, C.; Welling, M. Equivariant diffusion for molecule generation in 3d. In *International Conference on Machine Learning*, 2022; PMLR: pp 8867−8887.
(128) Igashov, I.; Stärk, H.; Vignac, C.; Schneuing, A.; Satorras, V. G.; Frossard, P.; Welling, M.; Bronstein, M.; Correia, B. Equivariant 3D-conditional diffusion model for molecular linker design. *Nature Machine Intelligence* **2024**, *6*, 417.
(129) Wu, K. E.; Yang, K. K.; Berg, R. v. d.; Zou, J. Y.; Lu, A. X.; Amini, A. P. Protein structure generation via folding diffusion. *arXiv*, 2209.15611, 2022.
(130) Hawkins-Hooker, A.; Depardieu, F.; Baur, S.; Couairon, G.; Chen, A.; Bikard, D. Generating functional protein variants with variational autoencoders. *PLoS computational biology* **2021**, *17* (2), e1008736.
(131) Xie, T.; Fu, X.; Ganea, O.-E.; Barzilay, R.; Jaakkola, T. Crystal diffusion variational autoencoder for periodic material generation. *arXiv*, 2110.06197, 2021.
(132) Freund, R.; Zaremba, O.; Arnauts, G.; Ameloot, R.; Skorupskii, G.; Dincă, M.; Bavykina, A.; Gascon, J.; Ejsmont, A.; Goscianska, J.; et al. The current status of MOF and COF applications. *Angew. Chem., Int. Ed.* **2021**, *60* (45), 23975−24001.
(133) Kim, B.; Lee, S.; Kim, J. Inverse design of porous materials using artificial neural networks. *Science advances* **2020**, *6* (1), eaax9324.
(134) Park, J.; Gill, A. P. S.; Moosavi, S. M.; Kim, J. Inverse design of porous materials: a diffusion model approach. *Journal of Materials Chemistry A* **2024**, *12*, 6507.
(135) Rappé, A. K.; Casewit, C. J.; Colwell, K.; Goddard, W. A., III; Skiff, W. M. UFF, a full periodic table force field for molecular mechanics and molecular dynamics simulations. *Journal of the American chemical society* **1992**, *114* (25), 10024−10035.
(136) *IZA Database*. https://www.iza-structure.org/databases/ (accessed 28 Aug. 2024).
(137) Park, H.; Yan, X.; Zhu, R.; Huerta, E.; Chaudhuri, S.; Cooper, D.; Foster, I.; Tajkhorshid, E. Ghp-mofassemble: Diffusion modeling, high throughput screening, and molecular dynamics for rational

discovery of novel metal-organic frameworks for carbon capture at scale. *arXiv*, 2306.08695, 2023.
(138) Fu, X.; Xie, T.; Rosen, A. S.; Jaakkola, T.; Smith, J. MOFDiff: Coarse-grained Diffusion for Metal-Organic Framework Design. *arXiv*, 2310.10732, 2023.
(139) M. Bran, A.; Cox, S.; Schilter, O.; Baldassari, C.; White, A. D.; Schwaller, P. Augmenting large language models with chemistry tools. *Nature Machine Intelligence* **2024**, *6*, 525.
(140) Kang, Y.; Kim, J. Chatmof: An autonomous ai system for predicting and generating metal-organic frameworks. *arXiv*, 2308.01423, 2023.
(141) Liu, S.; Nie, W.; Wang, C.; Lu, J.; Qiao, Z.; Liu, L.; Tang, J.; Xiao, C.; Anandkumar, A. Multi-modal molecule structure−text model for text-based retrieval and editing. *Nature Machine Intelligence* **2023**, *5* (12), 1447−1457.
(142) Besserve, M.; Sun, R.; Janzing, D.; Schölkopf, B. A theory of independent mechanisms for extrapolation in generative models. *In Proceedings of the AAAI Conference on Artificial Intelligence* **2021**, *35*, 6741−6749.
(143) Thompson, N. C.; Greenewald, K.; Lee, K.; Manso, G. F. The computational limits of deep learning. *arXiv*, 2007.05558, 2020.
(144) Villalobos, P.; Sevilla, J.; Besiroglu, T.; Heim, L.; Ho, A.; Hobbhahn, M. Machine learning model sizes and the parameter gap. *arXiv*, 2207.02852, 2022.
(145) Weiss, K.; Khoshgoftaar, T. M.; Wang, D. A survey of transfer learning. *Journal of Big data* **2016**, *3*, 1−40.
(146) Lim, Y.; Kim, J. Application of transfer learning to predict diffusion properties in metal−organic frameworks. *Molecular Systems Design & Engineering* **2022**, *7* (9), 1056−1064.
(147) Ma, R.; Colon, Y. J.; Luo, T. Transfer learning study of gas adsorption in metal−organic frameworks. *ACS Appl. Mater. Interfaces* **2020**, *12* (30), 34041−34048.
(148) Wang, C.-H.; Huang, K.-Y.; Yao, Y.; Chen, J.-C.; Shuai, H.-H.; Cheng, W.-H. Lightweight deep learning: An overview. *IEEE consumer electronics magazine* **2024**, *13*, 51.
(149) Yang, H.; Hu, C.; Zhou, Y.; Liu, X.; Shi, Y.; Li, J.; Li, G.; Chen, Z.; Chen, S.; Zeni, C. MatterSim: A Deep Learning Atomistic Model Across Elements, Temperatures and Pressures. *arXiv*, 2405.04967, 2024.
(150) Jiao, X.; Yin, Y.; Shang, L.; Jiang, X.; Chen, X.; Li, L.; Wang, F.; Liu, Q. Tinybert: Distilling bert for natural language understanding. *arXiv*, 1909.10351, 2019.
(151) Sanh, V.; Debut, L.; Chaumond, J.; Wolf, T. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv*, 1910.01108, 2019.
(152) Lan, Z.; Chen, M.; Goodman, S.; Gimpel, K.; Sharma, P.; Soricut, R. Albert: A lite bert for self-supervised learning of language representations. *arXiv*, 1909.11942, 2019.
(153) Sun, Z.; Yu, H.; Song, X.; Liu, R.; Yang, Y.; Zhou, D. Mobilebert: a compact task-agnostic bert for resource-limited devices. *arXiv*, 2004.02984, 2020.
(154) Xu, C.; Zhou, W.; Ge, T.; Wei, F.; Zhou, M. Bert-of-theseus: Compressing bert by progressive module replacing. *arXiv*, 2002.02925, 2020.
(155) de Wynter, A.; Perry, D. J. Optimal subarchitecture extraction for bert. *arXiv*, 2010.10499, 2020.
(156) Nandy, A.; Duan, C.; Kulik, H. J. Audacity of huge: overcoming challenges of data scarcity and data quality for machine learning in computational materials discovery. *Current Opinion in Chemical Engineering* **2022**, *36*, 100778.
(157) Lim, D.-W.; Kitagawa, H. Rational strategies for proton-conductive metal−organic frameworks. *Chem. Soc. Rev.* **2021**, *50* (11), 6349−6368.
(158) Sarango-Ramírez, M. K.; Park, J.; Kim, J.; Yoshida, Y.; Lim, D. W.; Kitagawa, H. Void space versus surface functionalization for proton conduction in metal−organic frameworks. *Angew. Chem., Int. Ed.* **2021**, *60* (37), 20173−20177.
(159) Mancuso, J. L.; Mroz, A. M.; Le, K. N.; Hendon, C. H. Electronic structure modeling of metal−organic frameworks. *Chem. Rev.* **2020**, *120* (16), 8641−8715.
(160) He, Y.; 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. *journal of physical chemistry letters* **2018**, *9* (16), 4562−4569.
(161) Kirklin, S.; Saal, J. E.; Meredig, B.; Thompson, A.; Doak, J. W.; Aykol, M.; Rühl, S.; Wolverton, C. The Open Quantum Materials Database (OQMD): assessing the accuracy of DFT formation energies. *npj Computational Materials* **2015**, *1* (1), 1−15.
(162) Cai, Z.; Li, W.; Chung, Y. G.; Li, S.; Liang, T.; Wu, T. Transfer learning-assisted computational screening of metal-organic frameworks and covalent-organic frameworks for the separation of Xe/Kr noble gas. *Sep. Purif. Technol.* **2024**, *348*, 127752.
(163) Park, H.; Kang, Y.; Kim, J. Enhancing Structure−Property Relationships in Porous Materials through Transfer Learning and Cross-Material Few-Shot Learning. *ACS Appl. Mater. Interfaces* **2023**, *15* (48), 56375−56385.
(164) Park, S.; Kim, B.; Choi, S.; Boyd, P. G.; Smit, B.; Kim, J. Text mining metal−organic framework papers. *J. Chem. Inf. Model.* **2018**, *58* (2), 244−251.
(165) Nandy, A.; Duan, C.; Kulik, H. J. Using Machine Learning and Data Mining to Leverage Community Knowledge for the Engineering of Stable Metal−Organic Frameworks. *J. Am. Chem. Soc.* **2021**, *143* (42), 17535−17547.
(166) 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* (11), 4510−4524.
(167) Park, H.; Kang, Y.; Choe, W.; Kim, J. Mining Insights on Metal−Organic Framework Synthesis from Scientific Literature Texts. *J. Chem. Inf. Model.* **2022**, *62* (5), 1190−1198.
(168) Zheng, Z.; Zhang, O.; Borgs, C.; Chayes, J. T.; Yaghi, O. M. ChatGPT Chemistry Assistant for Text Mining and the Prediction of MOF Synthesis. *J. Am. Chem. Soc.* **2023**, *145* (32), 18048−18062.
(169) Zheng, Z.; He, Z.; Khattab, O.; Rampal, N.; Zaharia, M. A.; Borgs, C.; Chayes, J. T.; Yaghi, O. M. Image and data mining in reticular chemistry powered by GPT-4V. *Digital Discovery* **2024**, *3* (3), 491−501.
(170) Xie, Y.; Sattari, K.; Zhang, C.; Lin, J. Toward autonomous laboratories: Convergence of artificial intelligence and experimental automation. *Prog. Mater. Sci.* **2023**, *132*, 101043.
(171) Seifrid, M.; Pollice, R.; Aguilar-Granda, A.; Morgan Chan, Z.; Hotta, K.; Ser, C. T.; Vestfrid, J.; Wu, T. C.; Aspuru-Guzik, A. Autonomous chemical experiments: Challenges and perspectives on establishing a self-driving lab. *Acc. Chem. Res.* **2022**, *55* (17), 2454−2466.
(172) Szymanski, N. J.; Rendy, B.; Fei, Y.; Kumar, R. E.; He, T.; Milsted, D.; McDermott, M. J.; Gallant, M.; Cubuk, E. D.; Merchant, A.; et al. An autonomous laboratory for the accelerated synthesis of novel materials. *Nature* **2023**, *624* (7990), 86−91.
(173) Vriza, A.; Chan, H.; Xu, J. Self-driving laboratory for polymer electronics. *Chem. Mater.* **2023**, *35* (8), 3046−3056.

https://doi.org/10.1021/jacsau.4c00618
JACS Au 2024, 4, 3727−3743