Check for updates

Received: 18 May 2022 | Revised: 15 July 2022 | Accepted: 31 July 2022

DOI: 10.1002/aic.17856

AIChE JOURNAL

RESEARCH ARTICLE

Thermodynamics and Molecular-Scale Phenomena

Active learning boosted computational discovery of covalent–organic frameworks for ultrahigh CH_4 storage

Hongjian Tang[1,2] | **Jianwen Jiang**[1]

[1]Department of Chemical and Biomolecular Engineering, National University of Singapore, Singapore, Singapore

[2]Key Laboratory of Energy Thermal Conversion and Control of Ministry of Education, School of Energy & Environment, Southeast University, Nanjing, People's Republic of China

Correspondence
Jianwen Jiang, Department of Chemical and Biomolecular Engineering, National University of Singapore, Singapore 117576, Singapore.
Email: chejj@nus.edu.sg

Funding information
Agency for Science, Technology and Research, Grant/Award Number: LCERFI01-0015 U2102d2004; Ministry of Education of Singapore, Grant/Award Number: C-261-000-207-532/C-261-000-777-532; National University of Singapore, Grant/Award Numbers: R-279-000-574-114, R-279-000-598-114, R-279-000-578-112

Abstract

As an environmental-benign fuel, methane (CH_4) has received considerable interest for developing high-capacity energy storage systems. Herein, we aim to rapidly discover covalent–organic frameworks (COFs) for ultrahigh CH_4 storage among 530,000+ COFs, including one experimental (Curated) and two hypothetical (Berkeley and Genomic) databases. First, the feature space of all the three COF databases is projected by t-Distributed Stochastic Neighbor Embedding (t-SNE) technique, which reveals a potential but unexplored regime in Genomic COFs. Subsequently, an active learning (AL) approach is developed by integrating parallel acquisition with molecular simulation to efficiently explore Genomic COFs. The parallel AL model demonstrates remarkable screening efficiency and shortlists top COFs by evaluating only 50 out of 445,845 Genomic COFs. A record-breaking Genomic COF is identified with CH_4 deliverable capacity of 222.2 v/v, surpassing the current world record (208.0 v/v from experiment and 217.9 v/v from simulation). Our AL approach is significantly faster than brute-force simulation and conventional machine learning, it would accelerate the discovery of advanced porous materials for broad applications.

KEYWORDS
active learning, covalent–organic frameworks, methane storage, molecular simulation, parallel acquisition

1 | INTRODUCTION

Combustion of fossil fuels (e.g., coal and oil) has produced over 35 gigatons of CO_2 annually, which adversely affects global climate and environment.[1] There is an urgent need to utilize alternative fuels with low carbon footprint. Methane (CH_4) is considered as a clean and cheap fuel source. To use CH_4 in mobile applications (e.g., transportation), it is crucial to develop safe and high-capacity storage systems for CH_4. The U.S. Department of Energy (DoE) outlined in 1993 the on-board CH_4 storage target as 150 v (STP)/v at 298 K and 35 bar, then revised it to 180 v (STP)/v in 2000. In 2012, the target was further revised to an ambitious value of 263 v (STP)/v at 298 K and 65 bar, which is equivalent to the density of compressed CH_4 at 298 K and 250 bar.[2] Unless otherwise mentioned, the volumetric capacity (v/v) in this study is at the STP, that is, the standard temperature and pressure (273.15 K and 1 bar).

In the ongoing quest of advanced materials for CH_4 storage, metal–organic frameworks (MOFs)[3] and covalent–organic frameworks (COFs)[4] have emerged as promising candidates. They can be synthesized from almost unlimited number of possible building blocks, thus possess a wide range of pore sizes, volumes, and shapes. In addition, the judicious selection of building blocks allows their pores and functionalities to be rationally tailed. From experimental measurements, high CH_4 storage capacities have been reported in a handful of MOFs (Table S1). Notably, HKUST-1,[5] UTSA-76a,[6] MOF-159,[7] and NJU-Bai 43[8] exhibit adsorption capacities ≈ 260 v/v at 298 K and 65 bar, as

© 2022 American Institute of Chemical Engineers.

well as deliverable capacities ≈ 200 v/v between 65 and 5 bar. Meanwhile, molecular simulation studies have been conducted to screen MOFs, COFs, porous polymer networks (PPNs), and other materials for CH_4 storage. Based on 137,953 hypothetical MOFs, ~300 MOFs were predicted to have storage capacities over 230 v/v at 35 bar.[9] Among 650,000 porous materials, the highest CH_4 deliverable capacity of 196 v (STP)/v was predicted in a hypothetical PPN between 65 and 5.8 bar.[10] A COF of tbd topology was shortlisted from 69,840 Berkeley COFs with a deliverable capacity of 216 v/v.[11] From trillions of MOFs, 96 were predicted to have deliverable capacities > 208 v/v with one about 217.9 v/v, which is the current world record from simulation.[12]

Over the last few years, machine learning (ML) has been increasingly applied to investigate CH_4 storage in MOFs and COFs.[13–19] Majority of these studies first generated a large volume of CH_4 adsorption data in a MOF or COF database by intensive simulations, then trained and interpreted ML models. For instance, CH_4 uptakes in ~13,000 MOFs were simulated and used to train a ML model, which was combined with a genetic algorithm to evolve MOFs *in silico* for CH_4 adsorption.[13] In addition to the large volume of data required for training, conventional ML models are largely trained using tree-based algorithms, which are subject to extrapolation issue. In other words, they are unable to make reliable predictions for unknown materials with features not used for training.[20] Such a drawback can be potentially overcome by emerging few-shot learning methods like transfer learning (TL)[21,22] and active learning (AL).[23–25] Essentially, TL is to pre-train a model on a known dataset to gain prior knowledge, which is transferred to the learning of new properties by fine-tuning the pre-trained model. This allows the TL model to achieve accurate predictions even with limited known data. However, TL requires latent representations rather domain-specific descriptors (i.e., pore size, pore volume, and framework chemistry of MOFs/COFs), thus causing the resultant model short of interpretability. Among different AL techniques, Bayesian optimization (BO)[26,27] is one of the most popular. It incorporates a probabilistic surrogate model and an acquisition function, thus allowing the model self-improving by dynamically acquiring most valuable training data. Recently, Deshewal et al. proposed a sequential BO screening approach and demonstrated that only hundreds of data were iteratively required to identify best structures from ~70,000 hypothetical COFs for CH_4 storage.[28] However, as only one COF was sampled in each BO iteration, their one-step screening approach was not highly efficient because the sequential validation on every acquired COF would be time-consuming by simulations of CH_4 uptakes at 5.8 bar and 65 bar, respectively. When handing a large and unknown database, a more rational strategy is to acquire multiple structures jointly in each BO iteration, which allows acquired candidates to be evaluated in batch, thereby improving BO efficiency.

Another important issue in ML is the cross-database generalizability and transferability of ML models. Based on ML-predicted CH_4 uptakes separately in MOFs and COFs, Fanourgakis et al. showed that their fingerprinting method for MOFs was generalizable to COFs, though the transferability of their ML models from MOFs to COFs was not examined.[15] By extending a deep learning model for CH_4 adsorption in a MOF database, Gurnani et al. found that their ML model was generalizable to MOFs beyond the feature scope of trained MOFs.[18] In this context, an interesting question is whether an unknown database deserves in-depth exploration for possible materials with targeted performance. Our recent study demonstrated that the transferability of ML models varied significantly in different MOF databases for propane/propylene (C3) separation; by analyzing the diversity and similarity of these databases in a projected feature space, we could intuitively and rapidly identify potential new MOFs for C3 separation.[29] Hence, prior to the cross-database application of ML models, it is crucial to assess the diversity and similarity among different databases.[30]

With the above issues in our mind, in this study we apply an interpretable and parallel AL approach, along with analysis of database diversity and similarity, to rapidly screen COFs for ultrahigh CH_4 storage. As illustrated in Scheme 1, the approach comprises four steps. (1) Three COF databases (Curated,[31,32] Berkeley,[11] and Genomic[33]) were collected and analyzed. The Curated and Berkeley databases were used to train and validate a parallel AL model, which was subsequently applied to screen Genomic COFs. (2) COFs were featurized by considering both geometric (pore size and geometry) and chemical (atom and bond type) features. (3) COF feature elimination was conducted using Random Forest (RF) regression. The downsized COF features were then visualized using t-SNE projection and correlated with CH_4 storage capacity. Furthermore, the efficiencies of three surrogate models including Gaussian process (GP) regression, Random Forest with variance (RFV), and Bayesian Neural Network (BNN), as well as different acquisition strategies, were compared and selected. (4) Finally, an interpretable and parallel AL model was integrated with molecular simulation to rapidly screen Genomic COFs for CH_4 storage and record-breaking COFs were identified.

2 | METHODOLOGY

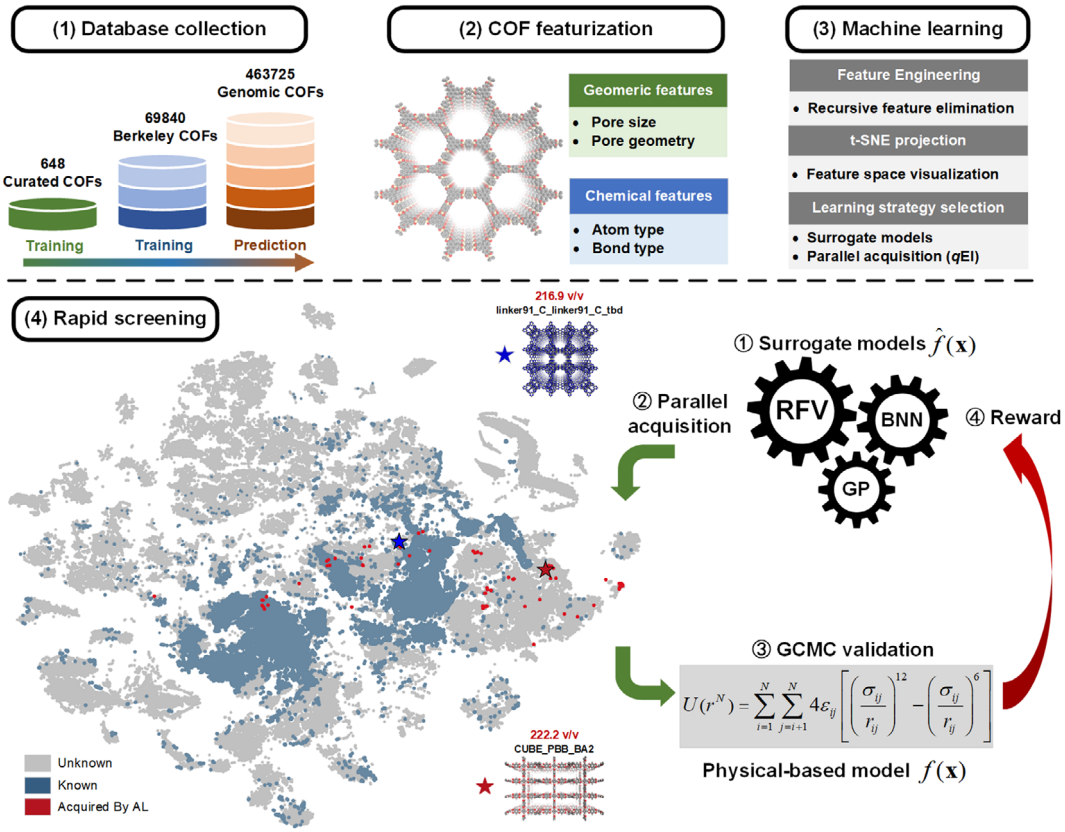
2.1 | COF databases

One experimental COF database (Curated) and two hypothetical COF databases (Berkeley and Genomic) were collected and examined in this work (Table 1). The dimensionalities of COF structures were also analyzed. Specifically, the dimensionalities of Berkeley COFs and Genomic 2D/3D COFs were collected from the source papers; while the dimensionalities of Curated COFs and Genomic Functional COFs were identified via Pymatgen[34] using a method proposed by Gorai et al.[35] Among 648 Curated COFs, 545 are 2D and 103 are 3D. By contrast, Berkeley COFs are dominated by 3D structures (61199), significantly more than 2D (8641); and the majority of Genomic COFs are also 3D.

2.2 | COF featurization

Two feature domains were considered for COFs: (1) geometric features and (2) chemical features as listed in Table 2. The geometric

15475905, 2022, 11, Downloaded from https://aiche.onlinelibrary.wiley.com/doi/10.1002/aic.17856 by Xian Jiaotong University, Wiley Online Library on [19/04/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License



SCHEME 1 Workflow. (1) Covalent–organic framework (COF) database collection. (2) COF featurization of geometric and chemical features. (3) Machine learning initiation including feature elimination, t-SNE projection, and learning strategy selection. (4) Parallel active learning integrated with molecular simulation for rapid screening of Genomic COFs.

TABLE 1 COF databases in this study

Database	Source	Number of structures	Dimensionality	Counts	Purpose	Ref.
Curated	Experimental	648	2D	545	To develop AL model	31, 32
			3D	103		
Berkeley	Hypothetical	69,840	2D	8641		11
			3D	61,199		
Genomic 2D	Hypothetical	166,681	2D	166,681	To screen new COFs	33
Genomic 3D	Hypothetical	129,753	3D	129,753		
Genomic functional	Hypothetical	167,291	2D	23,406		
			3D	143,885		

Abbreviations: AL, active learning; COFs, covalent–organic frameworks.

features including pore size and pore geometry were calculated using Zeo++.[36] The geometric features were shown as effective descriptors, especially for ML predictions of gas adsorption and separation in porous COFs and MOFs.[29] Our recent study demonstrated that atom types were important to capture MOF framework chemistry, thus significantly improving ML predictions for C3 separation.[29] Thus, these atom types were also adopted here to describe COF framework chemistry on the basis of hybridization and connectivity of atoms. For instance, we could uniquely and intuitively represent carbon atoms as C_1, C_2, C_3, and C_R respectively, indicating the linear, trigonal, tetrahedral, and resonant hybridization manner. In addition, bond types were introduced here to featurize COF framework chemistry in a more detailed way, as illustrated in Figure S1. The atom and bond types in each COF were enumerated by combining the lammps_interface with an in-house python script under the UFF4MOF nomenclature.[37] Figures S2–S11 show the counts of COFs with different atom and bond types. For being invariant to crystal cell dimensions, atom and bond types were quantified per unit volume (i.e., volumetric). A number of COFs were unfeaturizable and excluded in this study (Table S2). As listed in Table S3, COF

15475905, 2022, 11, Downloaded from https://aiche.onlinelibrary.wiley.com/doi/10.1002/aic.17856 by Xian Jiaotong University, Wiley Online Library on [19/04/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

TABLE 2 List of descriptors

Feature domain	Descriptor
Geometric features	Pore size: largest cavity diameter (LCD), pore limited diameter (PLD), largest free path diameter (LFPD)
	Pore geometry: volumetric surface area (VSA), gravimetric surface area (GSA), void fraction (VF), probe occupiable void fraction (VF_PO), pore volume (PV), probe occupiable pore volume (PV_PO), framework density (ρ)
Chemical features	Atom type density: H_, C_R, C_3, C_2, N_R, N_3, N_2, O_R, O_3, …
	Bond type density: C_2-C_R, C_2-H_, C_2-N_2, C_3-C_3, C_3-C_R, C_3-H_, C_R-C_R, C_R-H_, C_R-N_R, N_R-H_, N_R-O_R, …

features from different domains were grouped into five descriptor sets for further investigation.

2.3 | Molecular simulation

CH_4 uptake data in Berkeley COFs were directly extracted from the original literature.[11] The data in Curated and shortlisted Genomic COFs were estimated via grand canonical Monte Carlo (GCMC) simulations. Before simulations, the surface areas of Curated and Genomic COFs were evaluated using Zeo++[36] with a probe of 3.64 Å in diameter (representing N_2), and the inaccessible surface areas were blocked. Consistent with the study in Berkeley COFs,[11] CH_4 was mimicked as a united-atom model with the TraPPE force field,[38] and COFs were described by the Dreiding force field.[39] As exemplified in Figure S12, the simulated adsorption isotherms in two typical COFs agree well with experimental data. For CH_4 storage, adsorption was simulated at 298 K and 65 bar, while desorption was simulated at 298 K and 5.8 bar. The deliverable capacity ΔN_{CH_4} was calculated by CH_4 uptakes between adsorption (65 bar) and desorption (5.8 bar). A typical GCMC simulation in Curated COFs was run for 10^4 cycles (5000 cycles for equilibration and remaining 5000 cycles for ensemble average). In shortlisted Genomic COFs, a longer simulation with 10^5 cycles was run (2.5×10^4 cycles for equilibration and remaining 7.5×10^4 cycles for ensemble average). All the simulations were conducted using the RASPA package.[40] The Lennard-Jones interactions were truncated at 12.8 Å with tail corrections. The lengths of each COF structure were enlarged to at least 25.6 Å along three dimensions with the periodic boundary conditions.

2.4 | Machine learning

With the simulation data of ΔN_{CH_4} in Curated and Berkeley COFs, RF regression models were constructed to interpret feature importance and make feature selection. To train these models, Berkeley COFs were randomly split into training (90%) and test set (10%); subsequently, Curated COFs were used for model validation. There were 86 descriptors altogether for both geometric and chemical features in Berkeley COFs. To avoid overfitting, recursive feature elimination (RFE) was applied to downsize the feature dimensionality to 30. The RF regression was implemented via *scikit-optimize* toolkit (https://scikit-optimize.github.io/stable/). As mentioned earlier, analysis of database diversity and similarity is useful to examine the cross-database generalizability and transferability of ML models. Thus, t-Distributed Stochastic Neighbor Embedding (t-SNE) technique[41] was adopted here to visualize three COF databases. On this basis, high-dimensional COF features were reduced and projected into a 2D space. As COFs with similar features tend to aggregate into clusters, such a method is insightful to locate the feature regime of top-performing COFs and explore the underlying performance of unknown COFs. The implementation of t-SNE projection is briefly described in the Supporting Information.

For the AL, BO as the most popular AL technique was applied in this study.[26] BO involves a stochastic process comprising (1) a probabilistic surrogate model to cheaply approximate target $f(\mathbf{x})$ by a model prediction $\hat{f}(\mathbf{x})$ and to quantify the corresponding variance at each $\hat{f}(\mathbf{x})$. Here, the target $f(\mathbf{x})$ is deliverable capacity ΔN_{CH_4}; (2) an acquisition function to score and rank unknown COFs based on the prediction value and variance. For a specific COF, $\mathbf{x}$ indicates a vector representation of its features. During the acquisition in COF feature space, trade-off between the exploitation (high prediction value) and exploration (high prediction variance) is balanced. The exploitation can sample COFs near the optimum with a maximal $f(\mathbf{x})$; while the exploration can reduce prediction variance of the surrogate model. As illustrated in Figure S13a and Scheme 1, the AL under this study consisted of four steps in each loop: (1) training a surrogate model actively to approximate the ground truth $f(\mathbf{x})$ by a prediction value $\hat{f}(\mathbf{x})$; (2) acquiring samples from COF feature space; (3) simulating CH_4 uptakes in sampled COFs by GCMC simulations as observations to $f(\mathbf{x})$; (4) rewarding the surrogate model using the observations to improve $\hat{f}(\mathbf{x})$. As the most popular surrogate model for BO-based AL, GP regression[42] was used (see Figure S13). However, due to increasing complexity on the inversion of $n \times n$ kernel matrix along with increased training samples (n), training a GP model takes $O(n^3)$ time for each learning step of hyperparameter optimization (10 steps used here), which would be computationally expensive when handling a large volume of training data. Consequently, two alternative models namely Random Forest with conditional variance[42] (RFV, see Figure S14) and BNN (see Figure S15)[43] were also examined. The subtle difference of RFV from regular RF is that it enables variance estimation.

In addition to examining the three surrogated models (GP, RFV, and BNN) in AL, different acquisition methods were compared. Most acquisition functions like expected improvement (EI) are specified to acquire samples in sequence; that is, each acquisition step acquires only one structure ($\boldsymbol{q} = 1$) from an unknown COF feature space with the maximal EI score for post GCMC validation. Apparently, such serial acquisition is inefficient because (1) retraining of a surrogate model is sophisticated and time-consuming; (2) multicore computation

15475905, 2022, 11, Downloaded from https://aiche.onlinelibrary.wiley.com/doi/10.1002/aic.17856 by Xian Jiaotong University, Wiley Online Library on [19/04/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

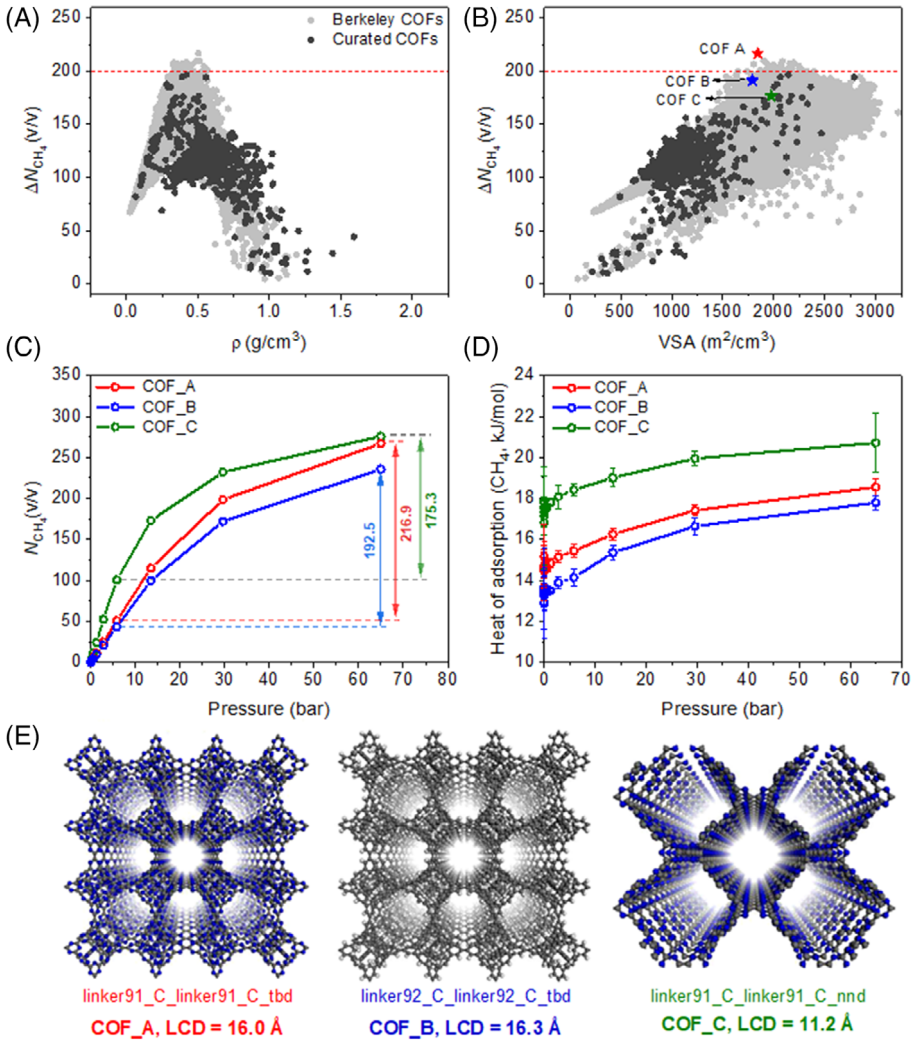


FIGURE 1 Relationships (A) $\Delta N_{CH_4} \sim \rho$ and (B) $\Delta N_{CH_4} \sim$ volumetric surface area (VSA) in Berkeley and Curated covalent–organic frameworks (COFs). The dashed line indicates $\Delta N_{CH_4} = 200$ v/v. (C) Adsorption isotherms in COF_A, B, and C. (D) Isosteric heats of CH_4 adsorption in COF_A, B, and C. (E) Structures of COF_A, B, and C.

allows GCMC validation in batch. Hence, parallel acquisition in COFs is highly desired to reduce model retraining and GCMC validation time. Instead of sequential EI, we implemented the parallel EI (***q***EI)[44] based on Kriging Believer (KB) method[45,46] to score multiple COFs jointly ($\boldsymbol{q} = 10$). The ***q***EI-based AL model was first initialized by ΔN_{CH_4} data in Curated COFs, acquiring ΔN_{CH_4} data in Berkeley COFs, and finally applied to rapidly screen Genomic COFs. More details regarding the surrogate models and ***q***EI acquisition method can be found in the Supporting Information.

3 | RESULTS AND DISCUSSION

3.1 | CH_4 storage in Curated and Berkeley COFs

Figure 1A shows the relationship between CH_4 deliverable capacity ΔN_{CH_4} and framework density ρ in Berkeley and Curated COFs. The ΔN_{CH_4} in Berkeley COFs appears to center at $\rho = 0.4$ g/cm^3. For the same ΔN_{CH_4}, Curated COFs generally have higher ρ than Berkeley counterparts. Such a difference between the two COF databases is primarily attributed to their distinct chemical constitutions. As illustrated in Figures S2 and S3, many heavy atom types (e.g., Cl, Br, Cu, Zn) that constitute Curated COFs are not in Berkeley COFs. All Curated COFs exhibit $\Delta N_{CH_4} < 200$ v/v. By contrast, a handful of Berkeley COFs possess $\Delta N_{CH_4} > 200$ v/v with the highest ΔN_{CH_4} of 216.9 v/v in COF_A (linker91_C_linker91_C_tbd). Apparently, there is a spacious space in hypothetical Berkeley COFs for ultrahigh CH_4 storage, which has not been experimentally reported or included in experimentally Curated COFs. Figure 1B shows the relationship $\Delta N_{CH_4} \sim$ volumetric surface area (VSA). In general, ΔN_{CH_4} is positively correlated with VSA. However, the largest VSA does not definitely suggest the highest ΔN_{CH_4}. Top-performing COFs with $\Delta N_{CH_4} > 200$ v/v are in a broad range of VSA from 1800 to 2700 m^2/cm^3. While the best Berkeley (COF_A) exhibits a VSA of 1844 m^2/cm^3, many COFs with VSA close to 1800 exhibit very low ΔN_{CH_4} of ~60 v/v.

Apparently, VSA is not the only factor governing ΔN_{CH_4} in COFs, other factors such as framework chemistry and pore geometry also play an indispensable role. This is further demonstrated by comparing CH_4 adsorption isotherms (Figure 1C) and heats of adsorption (Figure 1D) in COF_A and its two counterparts COF_B (linker91_C_linker91_C_nnd) and COF_C (inker92_C_linker92_C_tbd).

15475905, 2022, 11, Downloaded from https://aiche.onlinelibrary.wiley.com/doi/10.1002/aic.17856 by Xian Jiaotong University, Wiley Online Library on [19/04/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

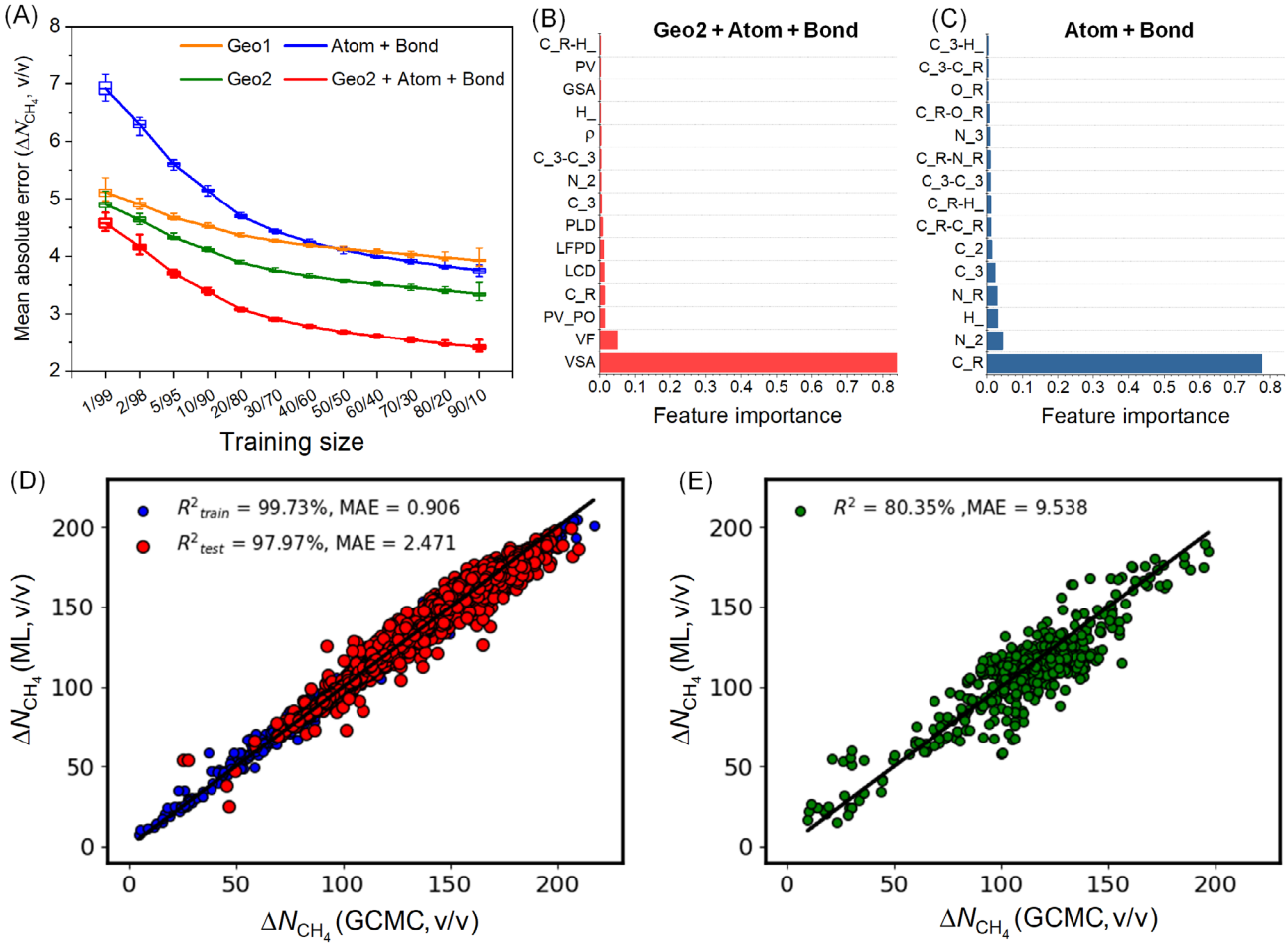


FIGURE 2 (A) Learning curves of Random Forest (RF) models to predict ΔN_{CH_4} with different split ratios of training/test sets in Berkeley covalent–organic frameworks (COFs). Top 15 important features using descriptor sets (B) "Geo2 + Atom + Bond" and (C) "Atom + Bond". (D) Training/test of the RF model trained with the downsized descriptor set "Geo2 + Atom + Bond" in Berkeley COFs. (E) Validation of the RF model trained with the downsized descriptor set "Geo2 + Atom + Bond" in Curated COFs.

Compared with COF_A, COF_B possesses a similar N_{CH4} at 5.8 bar but a significantly lower N_{CH4} at 65 bar, resulting in a lower ΔN_{CH_4} of 192.5 v/v. Although COF_C exhibits a slightly higher N_{CH4} at 65 bar than COF_A, its N_{CH4} at 5.8 bar is higher, thus causing the lowest ΔN_{CH_4} among the three COFs. The heats of adsorption in Figure 1D reveal that CH_4-framework interaction in COF_C is remarkably stronger than in COF_A and COF_B. As illustrated in Figure 1E, COF_B is a topological analogue to COF_A by substituting its linker91 (triazine) with linker92 (phenyl). Such linker substitution renders COF_B with different framework chemistry and decreased VSA, respectively, accounting for its weaker heat of adsorption and lower N_{CH4} at 65 bar. COF_C is a polymorph of COF_A with the same linker, but its topology, pore size, and shape are different from COF_A. The pores in COF_A are cylindrical with LCD of 16.0 Å, whereas they are squared with LCD of 11.2 Å in COF_C. The smaller pores in COF_C intensify interaction with CH_4 particularly at a low pressure, thus leading to higher N_{CH4} at 5.8 bar and largely lowering ΔN_{CH_4}. These observations are consistent with our previous studies that both pore geometry and framework chemistry are critical in governing the adsorption of C2 and C3 hydrocarbons in MOFs,[29,47] and further underpin the rationality of COF featurization in this work that incorporates both geometric and chemical features.

3.2 | Feature importance and selection

Prior to directly applying all the geometric and chemical descriptors (Table 2) to train AL models, the performance of different descriptor sets with a partial or full combination of COF features (Table S3) is benchmarked in Berkeley COFs via RF regression. Figure 2A shows the learning curves of RF models trained with different descriptor sets to predict ΔN_{CH_4} in Berkeley COFs. The RF model with descriptor set "Geo1" performs relatively poor in the test set to predict ΔN_{CH_4}. By including pore size (LPFD) and pore geometry (GSA + PV_PO +- VF_PO) in descriptor set "Geo2," a remarkable decrease in mean absolute error (MAE) is observed. It is interesting to note the RF model solely using chemical descriptor "Atom + Bond" performs better than the one using "Geo1" at a large split ratio (>50%). Similar

15475905, 2022, 11, Downloaded from https://aiche.onlinelibrary.wiley.com/doi/10.1002/aic.17856 by Xian Jiaotong University, Wiley Online Library on [19/04/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

performance is also observed in Figure S16 in the RF model to predict N_{CH_4} at 65 and 5.8 bar, respectively. Particularly for N_{CH_4} at 5.8 bar, "Atom + Bond" set outperforms "Geo2" when the training set ratio exceeds 50%, thus indicating the important role of framework chemistry in CH_4 adsorption at a low pressure. After the inclusion of framework chemistry, the MAE is clearly reduced in the regression using descriptor set "Geo2 + Atom + Bond." These observations corroborate that the framework chemistry of COFs is also important in governing CH_4 storage.

The RF models trained using "Geo2 + Atom + Bond" and "Atom + Bond" are further interpreted by counting the contributions of different features. The top 15 important features are ranked in Figure 2B,C. Based on descriptor set "Geo2 + Atom + Bond," VSA is found to be the most important feature and accounts for over 80% contribution. This coincides well with $\Delta N_{CH_4} \sim$ VSA relationship in Figure 1B and recent simulation studies of CH_4 storage in COFs and MOFs.[11,18] Chemical features are less important but not negligible. Among them, the contributions of carbon-related features like C_R (carbon atom with resonant hybridization) and C_3—C_3 (carbon–carbon bond with tetrahedral hybridization) are clearly higher than other atom and bond types. Difference in the importance of chemical features is amplified in Figure 2C based on descriptor set "Atom + Bond." In this case, C_R is substantially more important than others and contributes ~70%. Experimental studies have shown that aromatic rings in MOFs[48] and porous organic polymers[49] are favorable for CH_4 storage. Moreover, 10 out of 15 top important features are carbon-related. The reason is that C—C bonds are lighter than other bonds (e.g., C—N) in COFs.[11] Hence COFs with more C—C bonds especially C_R—C_R bonds tend to exhibit higher ΔN_{CH_4}.

Among the various descriptor sets, the 86-dimension "Geo2 + Atom + Bond" exhibits the best performance. With such a high-dimension descriptor set, overfitting might occur during model training in a small COF dataset (e.g., AL model initialization in 614 Curated COFs hereinafter). To examine this issue, RFE and cross-validated selection were applied to the set "Geo2 + Atom + Bond." As shown in Figure S17, upon decreasing feature dimension from 86 to 30, the MAE in ΔN_{CH_4} prediction remains nearly identical, but rises upon further decreasing dimension. Thus, we re-trained the RF model based on the downsized "Geo2 + Atom + Bond" descriptor set (Table S3) in Berkeley COFs with 90/10 split ratio of training/test sets. The MAE was measured as a loss function and minimized via random search hyperparameter grids (Table S4). As shown in Figure 2D, a desirable accuracy in ΔN_{CH_4} prediction is obtained with coefficient of determination (R^2) of 97.97% and MAE of 2.471 v/v in the test set. Furthermore, as demonstrated in Figure 2E, the out-of-sample validation in Curated COFs exhibits R^2 of 80.35% and MAE of 9.538 v/v. This reveals the good transferability of the re-trained RF model and the reliability of the downsized descriptors. Later, this RF model will be used as a conventional ML benchmark to compare with the parallel AL model hereinafter for screening of Genomic COFs.

Based on the downsized descriptor set "Geo2 + atom + bond" (Table S3), similarity and diversity among the three COF databases were compared by using t-SNE projection.[41] From the t-SNE maps in Figure 3, there are several interesting observations. (1) Despite the

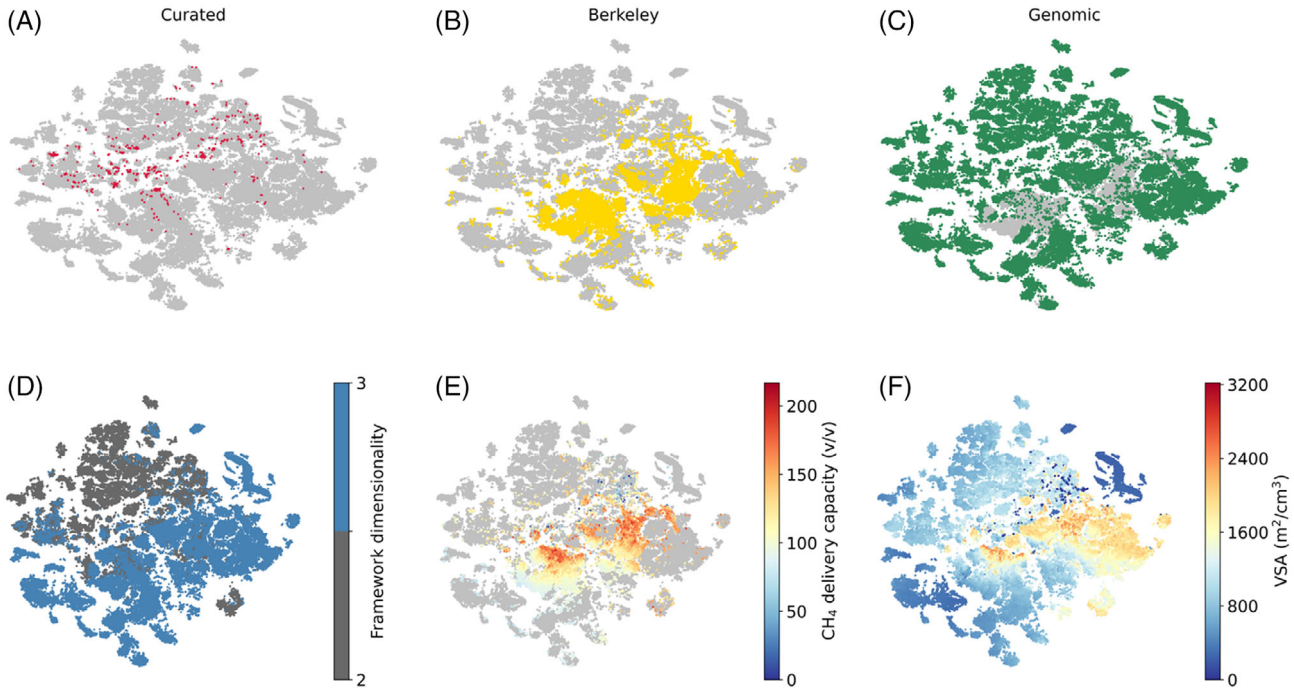


FIGURE 3 t-SNE maps based on the downsized descriptor set "Geo2 + Atom + Bond." The gray background indicates the overall feature space by combining three covalent–organic framework (COF) databases; each dot represents one COF. Distributions of (A) Curated COFs, (B) Berkeley COFs, and (C) Genomic COFs. The reduced feature space based on (D) framework dimensionality, (E) ΔN_{CH_4}, and (F) volumetric surface area. (D) and (F) are for all the COFs, while (E) is for Curated and Berkeley COFs.

15475905, 2022, 11, Downloaded from https://aiche.onlinelibrary.wiley.com/doi/10.1002/aic.17856 by Xian Jiaotong University, Wiley Online Library on [19/04/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

limited number (614 featurizable, Table S2) of structures, Curated COFs span a scattered regime (mainly the upper-left regime in Figure 3A), indicating the structural diversity of these experimentally synthesized COFs. Comprising 65,711 featurizable structures, Berkeley COFs are predominantly populated in the right-bottom regime (Figure 3B). By contrast, Genomic COFs with 445,845 featurizable structures exhibit a much broader distribution covering almost the whole feature space (Figure 3C). (2) An observable boundary exists between 2D- and 3D-COFs (Figure 3D), demonstrating the validity of our descriptors in capturing COF geometric features. This observation can be further used to interpret the dissimilarity among the three COF databases. Specifically, Curated COFs are majorly constituted of 2D structures, thus scarcely overlapping with Berkeley COFs that are dominated by 3D structures (Table 1). Detailed comparison across the three COF databases in terms of framework dimensionality is shown in Figure S18, where the spread of each COF database with a framework dimensionality is clearly projected in t-SNE maps. (3) Intriguingly, top-performing Curated and Berkeley COFs with $\Delta N_{CH_4} > 200$ v/v (red regime in Figure 3E) are primarily 3D instead of 2D. This is because 3D-COFs are more readily functionalizable than 2D counterparts, thus providing favorable adsorption sites (e.g., branched linkers and functional groups). In addition, 3D-COFs generally possess significantly larger VSA than 2D (Figure 3F). From earlier discussion, ΔN_{CH_4} is positively correlated with VSA (Figure 1B). It is thus highly anticipated to discover superior structures from the unexplored Genomic COF database for CH_4 storage, where massive 3D structures with the optimal VSA from 1800 to 2700 m^2/cm^3 are clearly seen in Figure 3F.

3.3 | Learning strategies

From the above t-SNE analysis on diversity and similarity among the three COF databases, we infer that Genomic COFs may contain candidates with high CH_4 storage capacities. The challenge is how to efficiently identify these candidates. In this regard, BO-based AL was combined with parallel acquisition to accelerate the screening of Genomic COFs for CH_4 storage. First, to determine which of the three surrogate models (GP, RFV, and BNN) should be adopted, we benchmarked their efficiencies in search for top Berkeley COFs. This is because the simulated ΔN_{CH_4} data were already available in Berkeley COFs, thus AL loops could be operated on the fly; otherwise, BO loop after each acquisition would be interrupted by GCMC simulations to evaluate ΔN_{CH_4} in acquired COF candidates. Initially, the simulated ΔN_{CH_4} data in Curated COFs were used as a training set. Subsequently, the single-point EI ($\boldsymbol{q} = 1$) was used as the acquisition function to score and rank COFs in Berkeley database. The candidate with the maximal EI score was added into the training set as a reward to improve the surrogate model for the next loop.

As shown in Figure 4A, GP regression exhibits the best search efficiency among the three surrogated models after combining with the sequential EI acquisition. GP_EI combination acquires the Berkeley COF (linker91_C_linker91_C_tbd) with the highest ΔN_{CH_4} (216.9 v/v) after 25 AL loops, which implies the best COF in Berkeley database for CH_4 storage can be identified by simulating only 25 structures. RFV_EI has a slightly lower efficiency than GP_EI and it identifies the best COF by 34 AL loops, significantly fewer than that of BNN_EI (145 loops). However, if taking the identified counts of top COFs with $\Delta N_{CH_4} > 200$ v/v as a criterion, as shown in Figure 4B, BNN_EI shows an efficiency comparable to GP_EI and faster than RFV_EI. The reasons are (1) High predication uncertainty in Berkeley COFs with $\Delta N_{CH_4} < 100$ v/v and $100 \sim 150$ v/v are seen respectively for GP and RFV (Figure S19). Exploration on COFs with high prediction uncertainty can well reward the surrogate model, thus enabling GP_EI and RFV_EI to identify the best COF rapidly. (2) The low and rigid dropout probability ($\delta = 0.05$) in BNN may lead to an overall low prediction certainty in Berkeley COFs (Figure S19). This makes BNN_EI more exploitation on COFs with high prediction value of ΔN_{CH_4}, thus resulting in more COFs with $\Delta N_{CH_4} > 200$ v/v. BNN-EI is expected to be comparable to GP_EI if its hyperparameters (i.e., dropout rate, hidden neurons, etc.) are optimized during each AL loop. Nevertheless, training and optimizing BNN would be generally sophisticated and time-consuming. The running time of 300 AL loops in Berkeley COFs based on the three surrogate models are compared and listed in Table S5, which indicates the remarkably less computational complexity of RFV_EI than that of GP_EI and BNN_EI. Hence, considering search efficiency and computational complexity, RFV was selected as the surrogate model for subsequent parallel AL.

Then, five parallel acquisition strategies were compared on their search efficiencies for top Berkeley COFs, including EI_greedy and four KB-related ***q***EI derivatives (KB, KBLB, KBUB, and KBR, see Table S6). In each AL loop, EI_greedy sampled 10 structures greedily from unknown COFs with the highest ranking EI score. Typically, this strategy estimated ΔN_{CH_4} in COFs independently. By contrast, in ***q***EI strategy, ΔN_{CH_4} in COFs were estimated dependently by virtually looping single-point EI acquisition for 10 times to approximate a joint estimation of ΔN_{CH_4} in 10 COFs. More details are provided in the Supporting Information. As corroborated in Figure 4C, the efficiency of ***q***EI in identifying the best Berkeley COF decreases: KBLB (4) > KBR (7) > EI_greedy (13) > KBUB (18) > KB (25). That is, KBLB_***q***EI finds the best Berkeley COF in 4 AL loops and greatly outperforms other ***q***EI strategies. Because each AL loop acquires 10 COFs in parallel, 4 AL loops means GCMC simulations are required to conduct in only 40 COFs. There are also additional important observations. (1) In search for the best COF, EI_greedy can surpass KBUB and KB strategies after 10 AL loops (Figure 4C). (2) If counting the number of top COFs with $\Delta N_{CH_4} > 200$ v/v (Figure 4D), EI_greedy has the poorest performance, significantly less than ***q***EI based on KB and derivatives. The former indicates that EI_greedy can obtain fewer top COFs in each batch to reward RFV model, especially the one with the highest EI; while the latter suggests that KB-related ***q***EI strategies are able to rank top COFs more accurately in each batch than EI_greedy. Hence, KB-related ***q***EI strategies, especially KBLB_***q***EI, are expected to have greater advantage when searching for top candidates in a larger and more diverse database like Genomic COFs.

15475905, 2022, 11, Downloaded from https://aiche.onlinelibrary.wiley.com/doi/10.1002/aic.17856 by Xian Jiaotong University, Wiley Online Library on [19/04/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

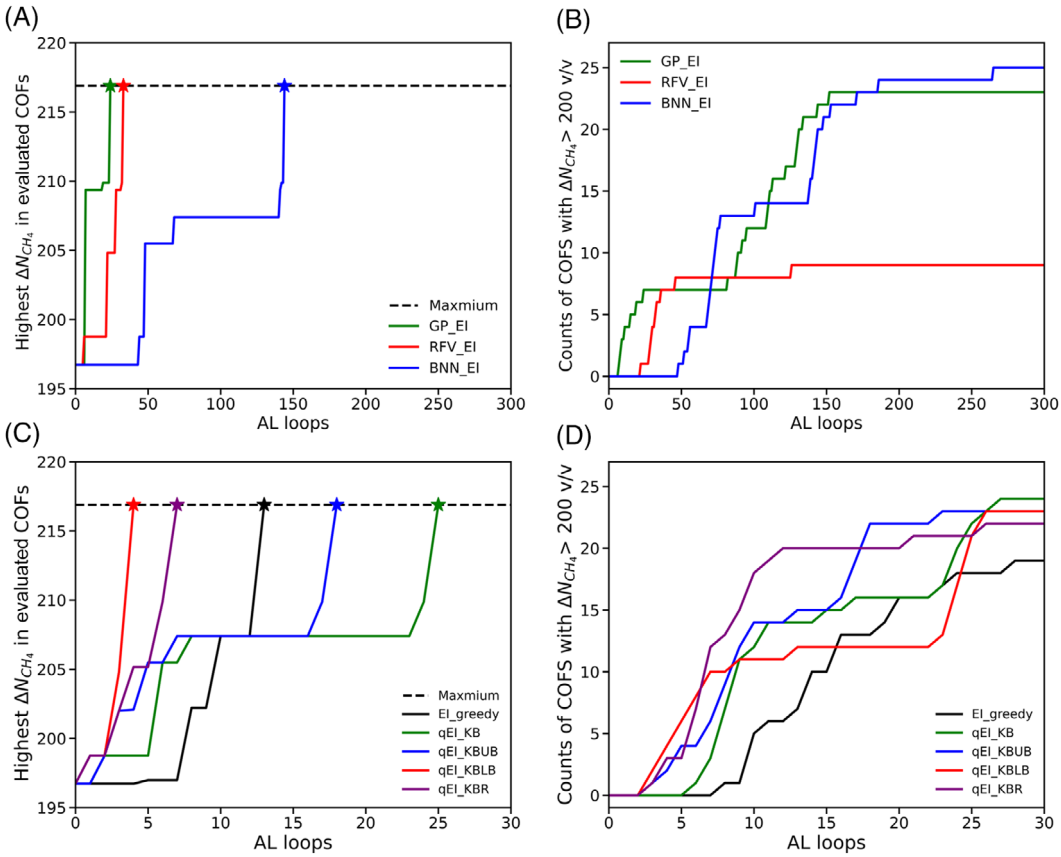


FIGURE 4 Performance comparison. The search efficiencies of three surrogate models combined with the sequential expected improvement (EI) acquisition for (A) highest ΔN_{CH_4} and (B) cumulative counts of covalent–organic frameworks (COFs) with ΔN_{CH_4} > 200 v/v in Berkeley COFs. The search efficiencies of different acquisition strategies combined with Random Forest with variance surrogate model for (C) highest ΔN_{CH_4} and (D) cumulative counts of top COFs with ΔN_{CH_4} > 200 v/v in Berkeley COFs.

3.4 | Rapid screening of Genomic COFs

From the performance benchmarked in Figure 4, the parallel AL model combining RFV model and KBLB_*q*EI acquisition stands out the best and was applied to rapidly screen unknown Genomic COFs for CH_4 storage. The screening is vividly animated by two dynamic images in the SI, where Videos S1 and S2 demonstrate the screening process in t-SNE map and $\Delta N_{CH_4} \sim$ VSA plot, respectively. The parallel AL model was initially trained by using CH_4 storage data in 614 Curated COFs (the black points in Figure 5A), then updated by acquiring 300 Berkeley COFs after 30 AL loops (the red line in Figure 4C,D, and the blue points in Figure 5A). Finally, the parallel AL model identified top Genomic COFs by acquisition in 10 AL loops (the red points in Figure 5A). Genomic COFs with ΔN_{CH_4} higher than the best Berkeley COF (216.9 v/v) were found in loop #2, #3, and #5, as also demonstrated in Figure 5B. Remarkably, 5 Genomic COFs with ΔN_{CH_4} > 216.9 v/v were found only in 5 loops (i.e., only simulating 50 structures). Apparently, the parallel AL model is highly efficient to screen top Genomic COFs. When AL loop continued to loop #10, no better COFs were found in the last 5 loops. Figure 5C shows the cumulative counts of top COFs with ΔN_{CH_4} > 200 v/v in each loop. There were 23 structures with ΔN_{CH_4} > 200 v/v in Berkeley COFs (the blue column). Upon acquiring Genomic COFs, 21 structures were found in the early 5 loops but only 6 more structures in the late 5 loops. Table S1 lists the top 10 Genomic COFs identified from this work. The best Genomic COF (CUBE_KET2_BA2) found in loop #5 possesses the highest ΔN_{CH_4} of 222.2 v/v, and 5 Genomic COFs surpass the best Berkeley COF with ΔN_{CH_4} > 216.9 v/v.

As illustrated by the dynamic image (Video S1) and the t-SNE maps in Figure S20, with the parallel AL model, we can efficiently explore the desired feature space of top Genomic COFs and avoid exhaustively examining undesirable structures. Moreover, such a screening approach based on the parallel AL model is several orders of magnitude faster than brute-force simulation-based screening. It is also intriguing to compare the parallel AL model with the conventional ML model. We take the RF regression model trained in the Berkeley COFs (see Figure 2D,E) as the conventional ML model. Figure S21 shows top 100 Genomic COFs predicted by each model. After GCMC validation, as listed in Table S7, only 4 out of 100 Genomic COFs predicted by the conventional ML model are confirmed to posses ΔN_{CH_4} > 200 v/v, with the highest ΔN_{CH_4} of 214.2 v/v. By contrast, 27 out of 100 Genomic COFs predicted by the parallel AL model are confirmed and the highest ΔN_{CH_4} is 222.2 v/v. Considering the fact that CH_4 storage data in 66,307 Berkeley COFs were used to train

15475905, 2022, 11, Downloaded from https://aiche.onlinelibrary.wiley.com/doi/10.1002/aic.17856 by Xian Jiaotong University, Wiley Online Library on [19/04/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

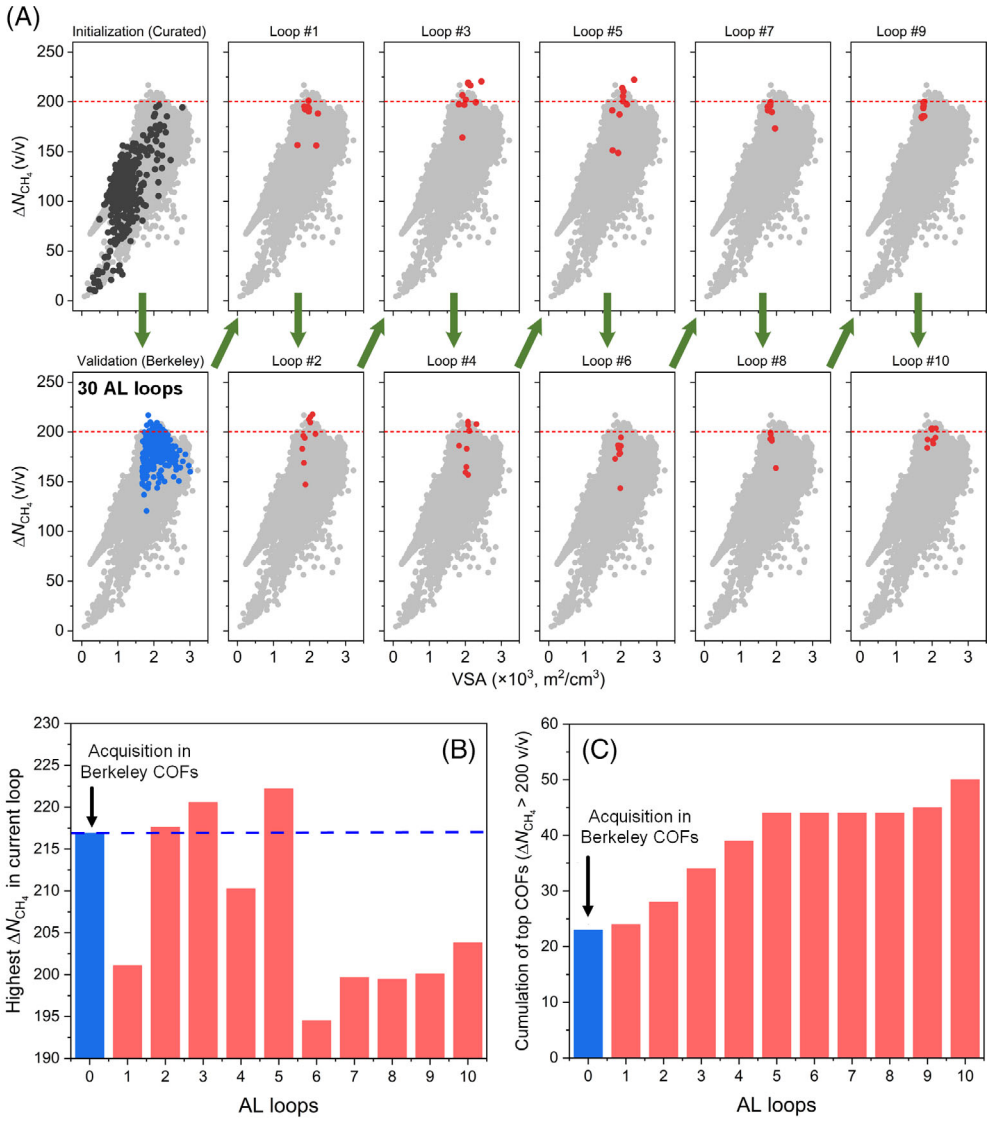


FIGURE 5 (A) Rapid screening of Genomic covalent–organic frameworks (COFs) via parallel active learning (AL). The gray points are all the COFs in the three databases. The back points are the Curated COFs, the blue points are 300 Berkeley COFs acquired after 30 AL loops, and red points are top Genomic COFs acquired after each AL loop. The dash line indicates $\Delta N_{CH_4} = 200$ v/v. (B) Highest ΔN_{CH_4} in each AL loop. The dashed line indicates $\Delta N_{CH_4} = 216.9$ v/v in the best Berkeley COF. (C) Cumulative counts of top COFs (with $\Delta N_{CH_4} > 200$ v/v) in each AL loop.

the conventional ML model, which is overwhelmingly larger than the number (614 Curated COFs for initialization and 300 Berkeley COFs for acquisition) used to develop the parallel AL model, the performance of the latter is significant and much more accurate than the former. Similar drawback of conventional ML model is also clearly seen in a recent preprint aiming to predict CH_4 storage in COFs.[19] Therein, a training set containing 84,800 COFs was utilized to derive ML-based equations, followed by intensive GCMC simulations in 10,000 top-performing COFs to validate ML predictions. In addition to high computational cost, the ML-based equations lack physical meaning and are subject to extrapolation issue. One of the top Genomic COFs (silicon_105-biqin-cellopt) identified by our parallel AL model was not predicted by the ML-based equations.[19]

The 5 top Genomic COFs (with $\Delta N_{CH_4} > 216.9$ v/v) identified from this work are further compared with top porous materials reported in the literature. As shown in Table S1 and Figure 6A, all the 5 Genomic COFs surpass the current world record capacity of 208.0 v/v from experiment.[7] Moreover, 4 of them have capacity higher than the world record capacity of 217.9 v/v from simulation.[12] CUBE_PBB_BA2 stands out with the highest ΔN_{CH_4} of 222.2 v/v. The atomic structures of the 5 top Genomic COFs are displayed in Table S8 (silicon_105-mi-noopt is a duplicate of silicon_105-mi-cellopt and thus not displayed). All the 5 Genomic COFs have 3D dimensionality and can be classified into two categories based on their topologies: cds and ctn. To the best of our knowledge, COFs with cds topology have not been experimentally reported; while well-known 3D COFs in the Curated COF database like COF-102, -103, -105, and -202 typically possess ctn topology. In this regard, the top Genomic COF with ctn topology (silicon_105-mi-cellopt) is further compared with its topological analogues COF-102, -103, -105, and -202. As shown in Figure 6B, silicon_105-mi-cellopt exhibits ΔN_{CH_4} higher than the four Curated COFs. Among the four Curated COFs, ΔN_{CH_4} in COF-102 and -103 is higher than in COF-105 and -202.

To elucidate the difference in ΔN_{CH_4} among the top COFs with ctn topology (Figure 6B), the Shapley Additive Explanations (SHAP)[50] were used to generate SHAP force plots for three COFs

15475905, 2022, 11, Downloaded from https://aiche.onlinelibrary.wiley.com/doi/10.1002/aic.17856 by Xian Jiaotong University, Wiley Online Library on [19/04/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

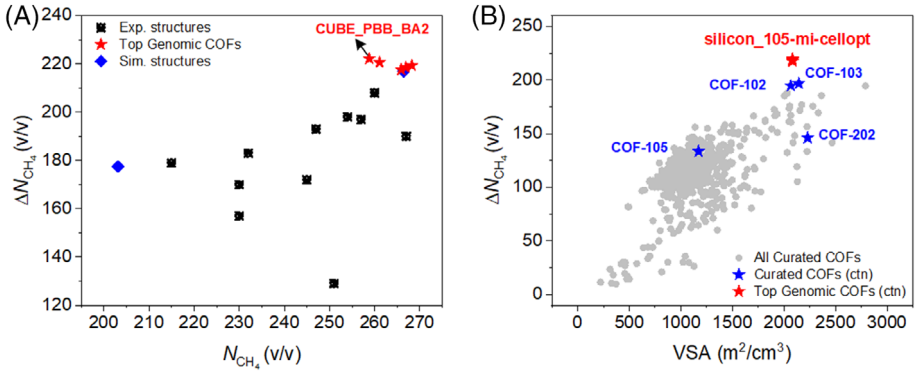


FIGURE 6 (A) Comparison of top five Genomic covalent–organic frameworks (COFs) with reported porous materials in the literature. (B) Comparison of top Genomic COF (silicon_105-mi-cellopt) with four Curated COFs (COF-102, -103, -105, and -202) with ctn topology.

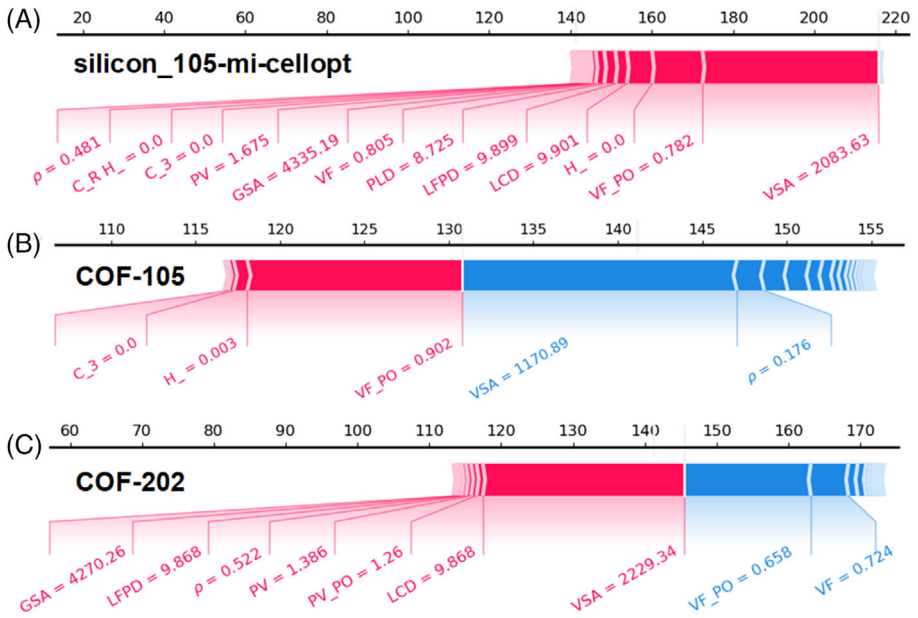


FIGURE 7 SHAP force plots of (A) silicon_105-mi-cellopt, (B) COF-105, and (C) COF-202. The features in red indicate positive impact on ΔN_{CH_4}, while the features in blue indicate negative impact on .ΔN_{CH_4}

(silicon_105-mi-cellopt, COF-105, and COF-202). As shown in Figure 7, high VSA and high VF_PO in silicon_105-mi-cellopt account for the most positive impact on ΔN_{CH_4}. In COF-105, high VF_PO promotes ΔN_{CH_4}, while low VSA is extremely adverse to ΔN_{CH_4}. In contrast to COF-105, COF-202 possesses high VSA but low VF_PO, thus substantially inhibiting ΔN_{CH_4}. These observations suggest: (1) hypothetical Genomic COFs with ctn topology like silicon_105-mi-cellopt and silicon_105-biqin-cellopt are potentially synthesizable; (2) COF-105 and COF-202 can act as starting structures to improve CH_4 storage capacity by concurrently constructing high VSA and high void fraction.

4 | CONCLUSIONS

We develop a remarkably efficient screening approach through integrating parallel AL and molecular simulation, which accelerates the discovery of top COFs for CH_4 storage by several orders of magnitude compared with brute-force simulation and conventional supervised ML method. The VSA is found to be a governing factor and positively correlated with CH_4 deliverable capacity. The t-SNE analysis in feature space and feature interpretation corroborate that pore geometry and framework chemistry also play an important role. The database analysis reveals a potential but unexplored performance space in Genomic COFs toward high CH_4 storage. For AL, different surrogate models and parallel acquisition strategies are examined in Berkeley COFs. The AL model combining RFV and KBLB-based parallel acquisition is demonstrated to be highly efficient in screening Genomic COFs. By simulating only 50 structures, 5 Genomic COFs are identified from 445,845 with CH_4 deliverable capacity higher than that in the best Berkeley COF, also surpassing the current world record. The parallel AL approach developed in this study would facilitate the development of COFs and other porous materials for CH_4 storage and other practical applications.

AUTHOR CONTRIBUTIONS

Hongjian Tang: Conceptualization (equal); data curation (equal); formal analysis (equal); investigation (equal); methodology (equal); resources (equal); software (equal); validation (equal); visualization (equal); writing – original draft (equal). **Jianwen Jiang:** Funding acquisition (equal); supervision (equal); validation (equal); visualization (equal); writing – review and editing (equal).

15475905, 2022, 11, Downloaded from https://aiche.onlinelibrary.wiley.com/doi/10.1002/aic.17856 by Xian Jiaotong University, Wiley Online Library on [19/04/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

ACKNOWLEDGMENTS

The authors gratefully acknowledge the A*STAR LCER-FI project (LCERFI01-0015 U2102d2004), the Ministry of Education of Singapore, and the National University of Singapore (R-279-000-578-112, R-279-000-598-114, R-279-000-574-114) for financial support.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

DATA AVAILABILITY STATEMENT

Data are available on https://github.com/davidtangGT/COF_Active_learning

ORCID

Hongjian Tang https://orcid.org/0000-0003-0484-3665

Jianwen Jiang https://orcid.org/0000-0003-1310-9024

REFERENCES

1. Carbon Dioxide Information Analysis Center, Oak Ridge National Laboratory, Tennessee, USA, CO_2 emissions. Accessed April 1, 2022. https://data.worldbank.org/indicator/EN.ATM.CO2E.KT
2. Methane opportunities for vehicular energy (MOVE), Advanced Research Project Agency-Energy, U.S. Department of Energy; 2012.
3. He Y, Chen F, Li B, Qian G, Zhou W, Chen B. Porous metal–organic frameworks for fuel storage. *Coord Chem Rev*. 2018;373:167-198.
4. Geng K, He T, Liu R, et al. Covalent–organic frameworks: design, synthesis, and functions. *Chem Rev*. 2020;120:8814-8933.
5. Peng Y, Krungleviciute V, Eryazici I, Hupp JT, Farha OK, Yildirim T. Methane storage in metal–organic frameworks: current records, surprise findings, and challenges. *J Am Chem Soc*. 2013;135:11887-11894.
6. Li B, Wen H-M, Wang H, et al. A porous metal–organic framework with dynamic pyrimidine groups exhibiting record high methane storage working capacity. *J Am Chem Soc*. 2014;136:6207-6210.
7. Gándara F, Furukawa H, Lee S, Yaghi OM. High methane storage capacity in aluminum metal–organic frameworks. *J Am Chem Soc*. 2014;136:5271-5274.
8. Zhang M, Zhou W, Pham T, et al. Fine tuning of mof-505 analogues to reduce low-pressure methane uptake and enhance methane working capacity. *Angew Chem Int Ed*. 2017;56:11426-11430.
9. Wilmer CE, Leaf M, Lee CY, et al. Large-scale screening of hypothetical metal–organic frameworks. *Nat Chem*. 2012;4:83-89.
10. Simon CM, Kim J, Gomez-Gualdron DA, et al. The materials genome in action: identifying the performance limits for methane storage. *Energy Environ Sci*. 2015;8:1190-1199.
11. Mercado R, Fu R-S, Yakutovich AV, Talirz L, Haranczyk M, Smit B. In silico design of 2D and 3D covalent–organic frameworks for methane storage applications. *Chem Mater*. 2018;30:5069-5086.
12. Lee S, Kim B, Cho H, et al. Computational screening of trillions of metal–organic frameworks for high-performance methane storage. *ACS Appl Mater Interfaces*. 2021;13:23647-23654.
13. Beauregard N, Pardakhti M, Srivastava R. In silico evolution of high-performing metal organic frameworks for methane adsorption. *J Chem Inform Model*. 2021;61:3232-3239.
14. Fanourgakis GS, Gkagkas K, Tylianakis E, Froudakis GE. A universal machine learning algorithm for large-scale screening of materials. *J Am Chem Soc*. 2020;142:3814-3822.
15. Fanourgakis GS, 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;123:6080-6087.
16. Fernandez M, Woo TK, Wilmer CE, Snurr RQ. Large-scale quantitative structure–property relationship (QSPR) analysis of methane storage in metal–organic frameworks. *J Phys Chem C*. 2013;117:7681-7689.
17. Kim S-Y, Kim S-I, Bae Y-S. Machine-learning-based prediction of methane adsorption isotherms at varied temperatures for experimental adsorbents. *J Phys Chem C*. 2020;124:19538-19547.
18. Gurnani R, Yu Z, Kim C, Sholl DS, Ramprasad R. Interpretable machine learning-based predictions of methane uptake isotherms in metal–organic frameworks. *Chem Mater*. 2021;33:3543-3552.
19. Ahmed A. Machine learning-guided equations for super-fast prediction of methane storage capacities of COFs. ChemRxiv. 2020.
20. Fanourgakis GS, Gkagkas K, Froudakis GE. Introducing artificial MOFs for improved machine learning predictions: identification of top-performing materials for methane storage. *J Chem Phys*. 2022;156:054103.
21. Sun Y, DeJaco RF, Siepmann JI. Deep neural network learning of complex binary sorption equilibria from molecular simulation data. *Chem Sci*. 2019;10:4377-4388.
22. Sun Y, DeJaco RF, Li Z, et al. Fingerprinting diverse nanoporous materials for optimal hydrogen storage conditions using meta-learning. *Sci Adv*. 2021;7:eabg3983.
23. Lookman T, Balachandran PV, Xue D, Yuan R. Active learning in materials science with emphasis on adaptive sampling using uncertainties for targeted design. *npj Comput Mater*. 2019;5:21.
24. Jablonka KM, Jothiappan GM, Wang S, Smit B, Yoo B. Bias free multiobjective active learning for materials design and discovery. *Nat Commun*. 2021;12:1-10.
25. Schmidt J, Marques MR, Botti S, Marques MA. Recent advances and applications of machine learning in solid-state materials science. *npj Comput Mater*. 2019;5:83.
26. Snoek J, Larochelle H, Adams RP. Practical Bayesian optimization of machine learning algorithms. *Adv Neural Inform Process Syst*. 2012;25:2951-2959.
27. Wang K, Dowling AW. Bayesian optimization for chemical products and functional materials. *Curr Opin Chem Eng*. 2022;36:100728.
28. Deshwal A, Simon CM, Doppa JR. Bayesian optimization of nanoporous materials. *Mol Syst Des Eng*. 2021;6:1066-1086.
29. Tang H, Xu Q, Wang M, Jiang J. Rapid screening of metal–organic frameworks for propane/propylene separation by synergizing molecular simulation and machine learning. *ACS Appl Mater Interfaces*. 2021;13:53454-53467.
30. Moosavi SM, Nandy A, Jablonka KM, et al. Understanding the diversity of the metal-organic framework ecosystem. *Nat Commun*. 2020;11:1-10.
31. Ongari D, Yakutovich AV, Talirz L, Smit B. Building a consistent and reproducible database for adsorption evaluation in covalent–organic frameworks. *ACS Cent Sci*. 2019;5:1663-1675.
32. Ongari D, Talirz L, Smit B. Too many materials and too many applications: an experimental problem waiting for a computational solution. *ACS Cent Sci*. 2020;6:1890-1900.
33. Lan Y, Han X, Tong M, et al. Materials genomics methods for high-throughput construction of COFs and targeted synthesis. *Nat Commun*. 2018;9:1-10.
34. Ong SP, Richards WD, Jain A, et al. Python materials genomics: a robust, open-source python library for materials analysis. *Comput Mater Sci*. 2013;68:314-319.
35. Gorai P, Toberer ES, Stevanović V. Computational identification of promising thermoelectric materials among known quasi-2D binary compounds. *J Mater Chem A*. 2016;4:11110-11116.
36. Willems TF, Rycroft CH, Kazi M, Meza JC, Haranczyk M. Algorithms and tools for high-throughput geometry-based analysis of crystalline porous materials. *Microporous Mesoporous Mater*. 2012;149:134-141.
37. Boyd PG, Moosavi SM, Witman M, Smit B. Force-field prediction of materials properties in metal-organic frameworks. *J Phys Chem Lett*. 2017;8:357-363.

15475905, 2022, 11, Downloaded from https://aiche.onlinelibrary.wiley.com/doi/10.1002/aic.17856 by Xian Jiaotong University, Wiley Online Library on [19/04/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

38. Martin MG, Siepmann JI. Transferable potentials for phase equilibria. 1. United-atom description of n-alkanes. *J Phys Chem B*. 1998;102:2569-2577.
39. Mayo SL, Olafson BD, Goddard WA. Dreiding: a generic force field for molecular simulations. *J Phys Chem*. 1990;94:8897-8909.
40. Dubbeldam D, Calero S, Ellis DE, Snurr RQ. RASPA: molecular simulation software for adsorption and diffusion in flexible nanoporous materials. *Mol Simul*. 2016;42:81-101.
41. Van der Maaten L, Hinton G. Visualizing data using t-SNE. *J Mach Learn Res*. 2008;9:2579-2605.
42. Hutter F, Xu L, Hoos HH, Leyton-Brown K. Algorithm runtime prediction: methods & evaluation. *Artif Intell*. 2014;206:79-111.
43. Gal Y, Ghahramani Z. Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning, International Conference on Machine Learning; 2016:1050-1059.
44. Ginsbourger D, Le Riche R, Carraro L. A multi-points criterion for deterministic parallel global optimization based on gaussian processes, NCP07; 2008.
45. Ginsbourger D, Le Riche R, Carraro L. Kriging is well-suited to parallelize optimization. *Computational Intelligence in Expensive Optimization Problems*. Springer; 2010:131-162.
46. Roux É, Tillier Y, Kraria S, Bouchard PO. An efficient parallel global optimization strategy based on kriging properties suitable for material parameters identification. *Arch Mech Eng*. 2020;67:169-195.
47. Tang H, Jiang J. In silico screening and design strategies of ethane-selective metal–organic frameworks for ethane/ethylene separation. *AIChE J*. 2021;67:e17025.
48. He Y, Zhou W, Qian G, Chen B. Methane storage in metal–organic frameworks. *Chem Soc Rev*. 2014;43:5657-5678.
49. Rozyyev V, Thirion D, Ullah R, et al. High-capacity methane storage in flexible alkane-linked porous aromatic network polymers. *Nat Energy*. 2019;4:604-611.
50. Lundberg S, Lee SI. A unified approach to interpreting model predictions. The 31st Conference on Neural Information Processing Systems; 2017.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Tang H, Jiang J. Active learning boosted computational discovery of covalent–organic frameworks for ultrahigh CH_4 storage. *AIChE J*. 2022;68(11):e17856. doi:10.1002/aic.17856

15475905, 2022, 11, Downloaded from https://aiche.onlinelibrary.wiley.com/doi/10.1002/aic.17856 by Xian Jiaotong University, Wiley Online Library on [19/04/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License