Journal Name

ARTICLE TYPE

Cite this: DOI: 00.0000/xxxxxxxxxx

Sequential Design of Adsorption Simulations in Metal-Organic Frameworks†

Krishnendu Mukherjee, Alexander W. Dowling, and Yamil J. Colón*[a]

Received Date
Accepted Date

DOI: 00.0000/xxxxxxxxxx

The large number of possible structures of metal-organic frameworks (MOFs) and their limitless potential applications has motivated molecular modelers and researchers to develop methods and models to efficiently assess MOF performance. Some of the techniques include large-scale high-throughput molecular simulations and machine learning models. Despite those advances, the number of possible materials and the potential conditions that could be used still pose a formidable challenge for model development requiring large data sets. Therefore, there is a clear need for algorithms that can efficiently explore the spaces while balancing the number of simulations with prediction accuracy. Here, we present how active learning can sequentially select simulation conditions for gas adsorption, ultimately resulting in accurate adsorption predictions with an order of magnitude less number of simulations. We model adsorption of pure components methane and carbon dioxide in Cu-BTC. We employ Gaussian process regression (GPR) and use the resulting uncertainties in the predictions to guide the next sampling point for molecular simulation. We outline the procedure and demonstrate how this model can emulate adsorption isotherms at 300 K from 10^{-6} to 300 bar (methane)/100 bar (carbon dioxide). We also show how this procedure can be used for predicting adsorption on a temperature-pressure phase space for a temperature range of 100 to 300 K, and pressure range of 10^{-6} to 300 bar (methane)/100 bar (carbon dioxide).

1 Introduction

Metal-organic frameworks (MOFs) are crystalline nanoporous materials comprised of inorganic nodes connected by organic linkers.[1] The chemical versatility of the building blocks provides a unique opportunity to tailor and design these materials with desired textural and chemical properties. The design flexibility of MOFs has resulted in their deployment for energy storage, catalysis, drug delivery, photonics, sensors, etc.[2–10] Despite the potential of these materials and their increasing numbers in experimental and synthetic studies, there is a challenge to determine which are the best materials and what are the conditions (e.g., temperature, pressure) that maximize their performance.

Molecular simulations have played an important role in the design and discovery of MOFs in a variety of applications.[11] Molecular models that describe the interactions between the materials and adsorbents of interest have been used to provide important physical insights and guide experiments towards promising candidates. The number of MOFs has kept increasing and so new algorithms and techniques have been introduced to enhance computational screening capabilities.[12–15] Some of these algorithms include those for crystal generation and enumeration, characterization of porous structures, and performance evaluation.[16–18] The use of these large-scale, high-throughput computational screening techniques on databases of MOF structures (experimental or computationally generated) has revealed structure-property relationships and identified top performing materials for many applications.[19–22] These studies can produce large amounts of data relating the physical and textural properties of MOFs (void fraction, surface area, pore volume, etc.) to their performance.

The deluge of data has allowed researchers to employ machine learning (ML) algorithms in a multitude of settings, with emphasis on gas adsorption and separations.[23–27] Some examples include hydrogen storage, methane storage, and Xe/Kr separations.[28–32] These ML models have resulted in important physical insight through the development of new descriptors capable of capturing important factors for applications of interest.[33–36] ML studies have also resulted in surrogate models capable of calculations that are orders of magnitude faster than the molecular simulations they rely upon for data.[37] Therein lies a challenge and bottleneck for workflows that rely on ML for predictions: large datasets are needed for the proper training and use of many ML algorithms. In cases where obtaining data is difficult or time consuming, the potential of ML algorithms and workflows is severely limited. As such, recent efforts have focused on using data that is already available in the literature or can be easily obtained using

*[a] Department of Chemical and Biomolecular Engineering, University of Notre Dame, IN 46556, USA. E-mail: ycolon@nd.edu

† Electronic Supplementary Information (ESI) available: [details of any supplementary information available should be included here]. See DOI: 00.0000/00000000.

simulations to make predictions for new systems. We recently demonstrated this type of approach using transfer learning.[38] Transfer learning leverages information used to train a model to produce a new model applied in a novel context using significantly less data. We trained deep neural networks (DNNs) for hydrogen adsorption at 243 K and 100 bar. We then used it as a source task where all the layers of the DNN remain fixed and only the last layer is fit for a new target task. New target tasks included hydrogen and methane adsorption at different temperatures. Interestingly, although the transfer learning model used an order of magnitude less data, we found higher accuracy compared with direct training. However, transferring the learning from pure component adsorption of hydrogen or methane to separations of Xe/Kr proved challenging because the underlying features that account for the behaviors are different.

Another approach involves training a multilayer perceptron (MLP) on alchemical species; these are modeled using arbitrary forcefield parameters that do not necessarily correspond to real molecules. With enough sampling in the alchemical space, the parameters that correspond to the real molecules will be included. Anderson and coworkers successfully demonstrated this type of approach, training an MLP using isotherms of alchemical species and making accurate isotherm predictions of real and simple molecules.[39] Extrapolations to other molecules not included in the training set showed reasonable accuracy. Most recently, Sturluson and coworkers implemented an algorithm to complete missing adsorption and physical property data in covalent organic frameworks (COFs) based on available data.[40] They trained a low rank model of adsorption-property matrices which makes "recommendations" in places where there is missing data. Through this the researchers were able to make predictions of missing values and group materials by their adsorption performance.

Alternative approaches employ an active learning (AL) approach — also known as sequential design — to help balance the accuracy of the predictive models with the number of data points to be acquired. This can be particularly attractive in situations where the feature space is very large (adsorption while varying temperature and pressure conditions) and/or time-consuming or resource-intensive experiments or simulations are needed. These approaches are increasing in popularity in the molecular simulation space. Uteva and coworkers recently implemented AL for intermolecular potential energy surfaces, showing improvement over grid-based approaches.[41] Similarly, Vandermause used AL to balance the use of quantum mechanical calculations to produce force fields.[42] In the context of porous materials, Santos and coworkers present a recent example where they seek to connect different length and time scales.[43] To do so, they require expensive molecular dynamics (MD) simulations. They used AL where the simulations were chosen based on model uncertainty through a query-by-committee approach. They show they require an order of magnitude less simulations to build their data set.

Herein we present an AL approach to balance model prediction accuracy with the number of simulations required to build a reasonable data set. The method relies on Gaussian process regression (GPR) where a data prior is fit.[44] The GPR model returns a prediction mean and prediction standard derivation (uncertainty), the later of which is used to determine the next individual simulation to be performed. We demonstrate this approach modeling adsorption of pure components methane and carbon dioxide in Cu-BTC for single isotherms and the temperature-pressure space. We outline the algorithm and show an order of magnitude saving on the number of simulations required to accurately assess the adsorption landscape.

2 Methods

2.1 Active Learning

The procedure outlined in this work intelligently selects the next adsorption simulation to be performed to facilitate training an accurate Gaussian process (GP) surrogate model. A GP is a non-parametric ML model that describes a real process $f(x)$ with a distribution over functions which have a joint Gaussian distribution, shown as N in equation below, described by a mean $\mu(x)$ and covariance $K(x,x^{'})$ function[44]:

$$f(x) \sim N(\mu(x), K(x,x^{'})). \tag{1}$$

There are many potential choices for $K(x,x^{'})$. We chose the rational quadratic kernel as it has been used before to describe adsorption loading in MOFs[45]:

$$K(x,x^{'}) = \left(1 + \frac{d(x,x^{'})^2}{2\alpha l^2}\right)^{-\alpha}, \tag{2}$$

where $d(x,x^{'})$ is the Euclidean distance between x and $x^{'}$, l is the length scale of the kernel, and α is the scale mixture parameter. The hyperparameters of the kernel, l and α, are found by maximizing the log-marginal-likelihood; the L-BGFS-B optimization algorithm implemented in scikit-learn was used in this work.[46,47] Importantly, to ensure the GP is fit appropriately, we take the log (base 10) of all the data (pressure, temperature, adsorption loading) and standardize the input variables (pressure and temperature), before it is run through the GP workflow.

Another important aspect of AL procedures is the acquisition function: how to choose the next simulation. For this purpose, there are many available choices of acquisition functions such as expected improvement (EI), upper confidence bound (UCB), probability of improvement (PI) etc.[48] Each of these functions have been developed for either more exploitative purposes (searching close to the prior input) or exploratory (far from prior data input). The purpose of this study is to explore the pressure and temperature conditions quickly and accurately for a given adsorption process. To achieve this, we settled on a "greedy" approach or one that simply "explores" the space. So, for iteration $n+1$, we choose conditions (pressure or temperature and pressure) x_{n+1} that maximizes the GPR prediction variance σ_n^2 constrained by bounds represented as the set $\mathscr{X}$:

$$x_{n+1} = \underset{x \in \mathscr{X}}{\operatorname{argmax}} \quad \sigma_n^2(x). \tag{3}$$

This is known as active learning MacKay, which was originally proposed in the context of neural networks.[49] Seo and coworkers implemented the idea for GPs.[50] After the new simulation at

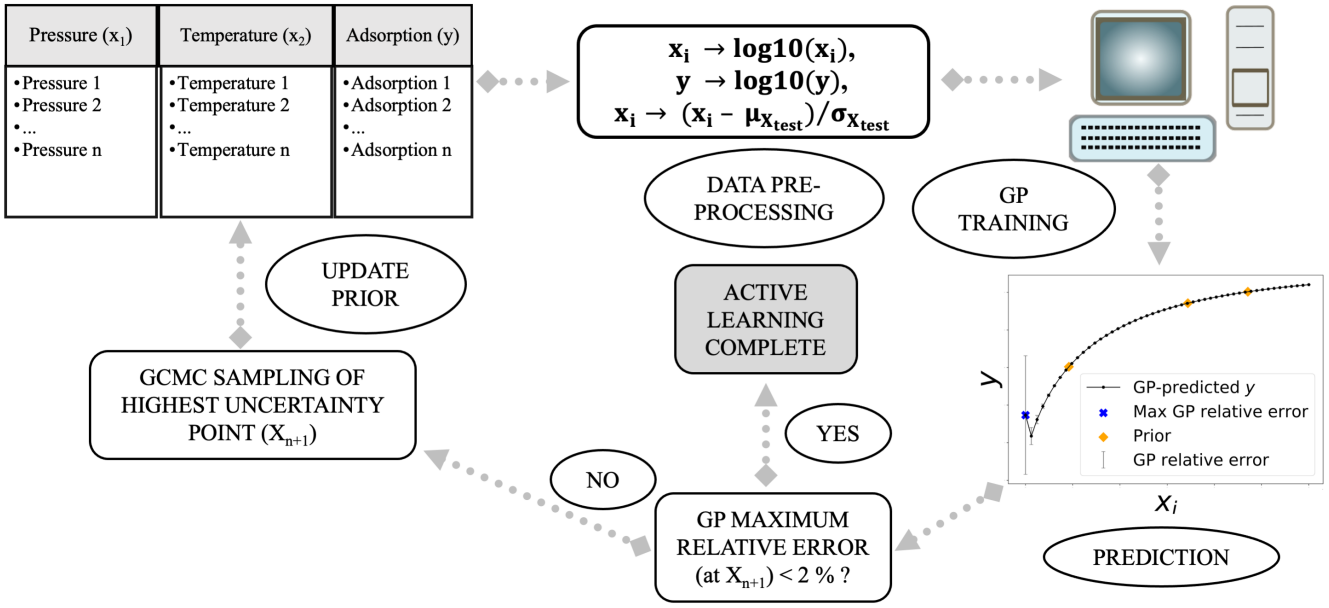

Fig. 1 A simple AL workflow for predicting adsorption isotherm in MOFs. The first step is generating prior data, as shown in the top left table with input variables, x_1, x_2 as pressure and temperature, and adsorption (output) as y. In the next step, data pre-processing is done by taking the log (base 10) for all the x_i and y, followed by standardizing the input variables x_i. Also, the input variables are standardized with respect to the mean and standard deviation of the test set X_{test}. This is followed by training the pre-processed data with a Gaussian process (GP) regression. After training is complete, adsorption predictions are made for the test set. The GP predicted relative error is then calculated for all the test points on the isotherm, and then maximum GP relative error is extracted. If the value of this error is less than 2 % (our convergence limit for the AL, except section 3.4 where it is set to 3 %) then learning is complete. If not, then the point with maximum relative error is sampled using another GCMC simulation, and the prior is updated with this data. After prior updating, next cycle of AL begins and it goes on until the maximum GP relative error goes below threshold.

x_{n+1} is performed and the data is gathered, the GP is refit and the procedure is repeated until $\sigma_n^2(x)$ is below some threshold. We picked this threshold as 2 % for our methane and carbon dioxide adsorption (section 3.1, 3.2, and 3.3) while a 3 % limit was chosen for carbon dioxide adsorption with two features (section 3.4). This limit is user defined and can be chosen as per the level of accuracy desired for an application. For adsorption, 2% to 3% prediction accuracy of a surrogate model is acceptable for most purposes. After this step, at the beginning of the procedure a GP prior is usually fit using data spread in x. Figure 1 summarizes the AL procedure in this work. Also, irrespective of the performance of the AL at the first iteration, we forced the algorithm to complete the first cycle. This was done because for some specific choice of priors, the GP can become overconfident and the GP predicted relative error might be too low. The test set for AL was linearly spaced between the pressure limits. For methane this range was 10^{-6} to 300 bar, while for carbon dioxide the limit was set to 10^{-6} to 100 bar. We used this as a test set, denoted by X_{test}, which was an array consisting of 50 grid points. The next point x_{n+1} for AL was determined from this set only. For the case of two features (section 3.3 and 3.4), we added a temperature grid as well. The temperature test set was also linearly spaced from 100 to 300 K for both methane and carbon dioxide, and it consisted of 40 points. While we used this linearly spaced grid criteria for testing and building the AL model, we also did an interpolation test at the low pressure region (10^{-6} to 1 bar) for both methane and carbon dioxide. For this test, we had 50 grid points spaced in the natural log-scale to test the performance of the final GP regression after AL has finished. We only did this interpolation test for low pressure region and we kept the temperature range same as the X_{test}. Also for the interpolation test, the input variables were standardized against X_{test}. This was done to create an environment in which a user can test the power of a final AL fit model which is completely blind to the interpolation test information. The AL performance for both the AL initial test set (i.e. on X_{test}) and low pressure interpolation test are reported in results section. Also, a set of GCMC simulation were done for both tests to generate the ground truth data. The details of GCMC simulation uncertainty (σ_{GCMC}) for both the X_{test} and low pressure interpolation test are given in the respective tables in the next sections.

2.2 Molecular Simulations

Adsorption loading at various temperatures (100 to 300 K) and pressures (10^{-6} to 300 bar) were calculated using grand canonical Monte Carlo (GCMC) simulations in RASPA.[51–53] MC moves employed were insertions, deletions, reinsertions at a random point in the space, rotations, and translations with equal probability; 2,000 initialization and 20,000 production cycles were used for methane and 2,000 initialization cycles and 20,000 production cycles for carbon dioxide. Methane and carbon dioxide were modeled using TraPPE.[54] Nonbonded interactions for methane and carbon dioxide were modeled as a Lennard-Jones (LJ) or LJ + Coulomb, respectively with a cutoff for van der Waals interactions of 12.5 Å.[55] MOF Cu-BTC (BTC stands for

benzene-1,3,5-tricarboxylate) was chosen for this study.[56] The charge for Cu-BTC was taken from Castillo et. al from 2008, where they obtained the partial charges from fitting different set of charges to reproduce water adsorption isotherms.[57] Cu-BTC is chosen in this work since it is a popular MOF in the domain for separating hydrocarbons and has been studied extensively for molecules such as O_2, N_2, CH_4 and CO_2. In this work, Cu-BTC was modeled as rigid and parameters for nonbonded interactions were taken from the Universal Force Field (UFF).[58] Lorentz-Berthelot mixing rules were used for cross-term interactions.[59] This methodology for modeling adsorption in Cu-BTC has been validated against experiments. For example, Yang and coworkers used GCMC simulation approach (using UFF parameters for Cu-BTC and TraPPE for adsorbates) to predict mixture adsorption of CO_2, CH_4, and H_2 as well as CO_2 separation from flue gases.[60,61] Their predictions matched very closely with experimental isotherms at the same operating conditions. The GCMC methodology has also been adopted by Wang and team for simulating adsorption for hydrocarbon mixtures in Cu-BTC (CO_2/CO, CO_2/C_2H_4, and C_2H_4/C_2H_6) and they found close agreement with experiments.[62]

2.3 Prior Generation Strategy

We selected 3 schemes for generating the prior dataset for adsorption isotherms. The first two were based of Latin hypercube sampling (LHS) in the input feature space.[63] LHS sampling scheme was chosen since it's a 'space-filling' method which utilizes the entire range of model input. Since we do not have a prior knowledge of how input probabilities are distributed (pressure/temperature or both), a 'space-filling' design is a better choice since it distributes the input equally in the design space. In this work, we adopted two different implementations of the LHS scheme. For the first LHS-based sampling, pressure was sampled linearly from the pressure range. For the second LHS-based prior, pressure was sampled in a log (base 10) scale in the respective pressure range. Temperature was fixed at 300 K for section 3.1 and 3.2. For performing AL with 2 features (section 3.3 and 3.4), temperature was sampled linearly for both the LHS-based priors. In the third prior, named 'boundary-informed prior', samples were hand-picked at the limits of the test range (for two features, this would be a meshgrid of pressure and temperature points). For example, in case of methane adsorption with two features (section 3.3) we choose the pressure and temperature points as shown in table 1. The boundary-informed prior thus has 50 points for this section (five temperature points for each of the ten pressure mark). For section 3.4 (carbon dioxide adsorption with two features), boundary-informed prior had 40 points (8 pressure points with five temperature for each pressure mark). The pressure points of 200 and 300 bar (from table 1) were missing for carbon dioxide since the high pressure limit is 100 bar.

Table 1 Boundary-informed prior grid points for CH_4 adsorption in Cu-BTC MOF for two features

Pressure (in bar)	Temperature (in K)
10^{-6}	
10^{-5}	100
10^{-4}	
10^{-3}	150
10^{-2}	
10^{-1}	200
1	
10	250
100	
200	300
300	

2.4 Error Calculation

Three error metrics are used in the AL framework:

1. GP-predicted Relative Error — This is the ratio of GP-predicted uncertainty at a point by the GP-predicted adsorption. Please note the aim of the AL procedure is to constrain the GP-predicted relative error within a threshold limit (refer to figure 1 and 2).

$$\text{GP relative Error in \% (at a point)} = \frac{\sigma_{\text{GP-predict}}(x)}{Y_{\text{GP-predict}}(x)} \times 100 \quad (4)$$

2. Relative Error — This is ratio of the difference between GP-predicted adsorption and the ground truth adsorption calculated by GCMC simulation.

$$\text{Relative Error in \% (at a point)} = \left| \frac{Y_{\text{GP-predict}}(x) - Y_{\text{GCMC}}(x)}{Y_{\text{GCMC}}(x)} \right| \times 100 \quad (5)$$

3. Mean Relative Error (MRE) — This is calculated as a mean of the relative error for an entire AL iteration. We compare this error with the maximum GP relative error to check for speed of convergence of the AL protocol. Also, since MRE compares GP-predicted adsorption and ground truth based off of GCMC simulations, it serves as a parameter to gauge the performance of the AL model.

$$\text{MRE in \%} = \left(\sum_{i=1}^{n} \left| \frac{Y_{\text{GP-predict}}(x_i) - Y_{\text{GCMC}}(x_i)}{Y_{\text{GCMC}}(x_i)} \right| \right) \times \frac{100}{n} \quad (6)$$

3 Results and Discussions

3.1 Methane Isotherms

Methane adsorption is Type I and is relatively simple to model as a single sphere without electrostatic interactions. Figure 2 shows the evolution of the GP fit through all the iterations of the AL procedure for a methane isotherm at 300 K. Starting from 8 data points selected using boundary-informed prior scheme, only 2 iterations are needed to decrease the relative error of the GP fit (equation 4) to under 2 %. For the LHS-based priors, four points were chosen for building the prior dataset. For boundary-informed prior, we find a good agreement between the GP predictions and the simulation results, and we show this case in figure 2. Panel a (in figure 2) shows the GP fit using the simulation data selected through boundary-informed prior. The first GP fit clearly struggles at high pressures where it under predicts methane loading and this is also reflected in high GP relative error (shown as grey bars in the plot). The highest relative standard deviation was at 300 bar and panel b shows the resulting GP fit with the new

simulation result added to the data (blue marker). The GP fit now resembles more of what is expected of an adsorption isotherm.

Qualitatively the fit does not change very much after the first iteration. Panel c shows the final GP fit compared to a full simulated isotherm with a good agreement between the GP fit and the GCMC simulation results. The last panel d shows the comparison of GP fit with GCMC simulations at low pressure range (10^{-6} to 1 bar). The final adsorption isotherm GP fits along with the GCMC simulations for both linear-spaced and log-spaced LHS have been included in the Supporting Information (figure S1).

The performance of other priors, linear-LHS and log-LHS, along with boundary-informed are tabulated in table 2. As we can observe, the overall MRE (for (X_{test})) is only 1.15 % for linear-spaced prior while it is slightly higher for boundary-informed prior at 2.14 %. The log-spaced prior has a very high MRE of 7.80 % for X_{test} compared to the other priors. For the low pressure interpolation test, the log-spaced LHS finished after one iteration and performed the best among all the prior with an MRE of 13.61 %. However, the MRE of the log-based prior was high for the X_{test} range. Overall the GP fits obtained using the previous protocols shows poor performance in the low pressure regime (10^{-6} to 1 bar) for all the priors, especially for linear-spaced one. In some cases the predicted adsorption can differ by an order of magnitude. Hence, despite the perceived agreement of the GP fit, the errors in the low pressure region are high when comparing with simulation results. This can present significant challenges in analysis for separations where ideal adsorbed solution theory (IAST) is used and it is particularly sensitive to the results in the low pressure regime.[64]

Table 2 Performance of different priors for predicting CH_4 uptake in Cu-BTC MOF (all errors are expressed in %)

Prior type	Iterations	MRE (X_{test})	MRE (Low Pressure)
Boundary-informed	2	2.14	15.79
Linear-spaced LHS	1	1.15	37.84
Log-spaced LHS	1	7.80	13.61
GCMC	-	σ_{GCMC} (X_{test})	σ_{GCMC} (Low Pressure)
Ground truth	-	0.74	73.97

These three AL approaches for adsorption isotherms show good agreement with the GCMC simulations despite not having to physically simulate all the points in the test set. The first approach only used eight data points for the prior, including one at each boundary of the isotherm in pressure, and the AL procedure converges with only two additional iterations. For the linear-spaced LHS approach, the low pressure regimes of the isotherm are problematic and resulted in a high MRE. The log-spaced prior, shows a better MRE at low pressure than the boundary-informed approach, but its performance was poor for pressure range in the X_{test}. The code for the AL along with the data are publicly available and can be accessed through this link: https://github.com/mukherjee07/Sequential_Design_of_Adsorption_simulations_in_MOFs.

3.2 Carbon Dioxide Isotherms

Carbon dioxide adsorption isotherms present an interesting contrast to methane adsorption. Namely, the electrostatic interactions of the molecule induce a much sharper transition in the isotherm. Despite this, we see very similar behavior and results for the AL procedures for carbon dioxide when compared to methane. We carried out AL with three priors, as was done for methane. The first difference in this study was the total pressure range, which was 10^{-6} to 100 bar. Another difference is 9 points were chosen for the boundary-informed prior since the transition is sharper for carbon dioxide adsorption. For each of the log-spaced and linear-spaced priors, four points in the isotherm were generated in an automated fashion similar to methane adsorption.

The AL converged to a 2 % GP relative error in a similar number of iterations as for methane, except for log-spaced prior where it took 6 iterations. Table 3 shows the final GP fit results compared to the simulated isotherm. Boundary-informed prior GP fit had the best MRE at both high pressure (X_{test}) and low pressure range.

Table 3 Performance of different priors for predicting CO_2 uptake in Cu-BTC MOF (all errors are expressed in %)

Prior type	Iterations	MRE (X_{test})	MRE (Low Pressure)
Boundary-informed	3	1.52	20.39
Linear-spaced LHS	1	3.66	1011.21
Log-spaced LHS	6	2.07	73.37
GCMC	-	σ_{GCMC} (X_{test})	σ_{GCMC} (Low Pressure)
Ground truth	-	1.04	51.95

The GP fit carbon dioxide isotherm are shown in the Supporting Information figure S2. We find the final GP fit for boundary-informed prior performs excellently in the X_{test} pressure range as well in the low pressure region. The log-spaced one performs well at the high pressure region (X_{test}) while the linear-spaced one has high error at low pressure as well as at the tail of the pressure range. This also becomes evident from the MRE for both the LHS-based prior schemes in table 3. The linear-spaced prior based GP fit has a MRE of 3.66 % for the X_{test} isotherm with a very high MRE of 1011.21 % for the low pressure range. The log-spaced GP fit performs better than the linear one but still has a higher MRE compared to boundary-informed for both the pressure ranges. Through this comparison, it is evident that boundary-informed prior outperforms both LHS schemes when using pressure as a single feature.

3.3 Temperature-Pressure diagrams for methane adsorption

We performed adsorption simulation in the temperature and pressure phase-space (two features for AL) with priors based on boundary-informed and LHS sampling schemes. The pressure and temperature range for this study was 10^{-6} bar to 300 bar, and 100 K to 300 K respectively. The boundary-informed one, similar to the previous methane and carbon dioxide isotherms, was curated to bias the training data with hand-picked pressure and temperature points. For each pressure point as reported in table 1 five temperature points (100 K, 150 K, 200 K, 250 K, and 300 K) were

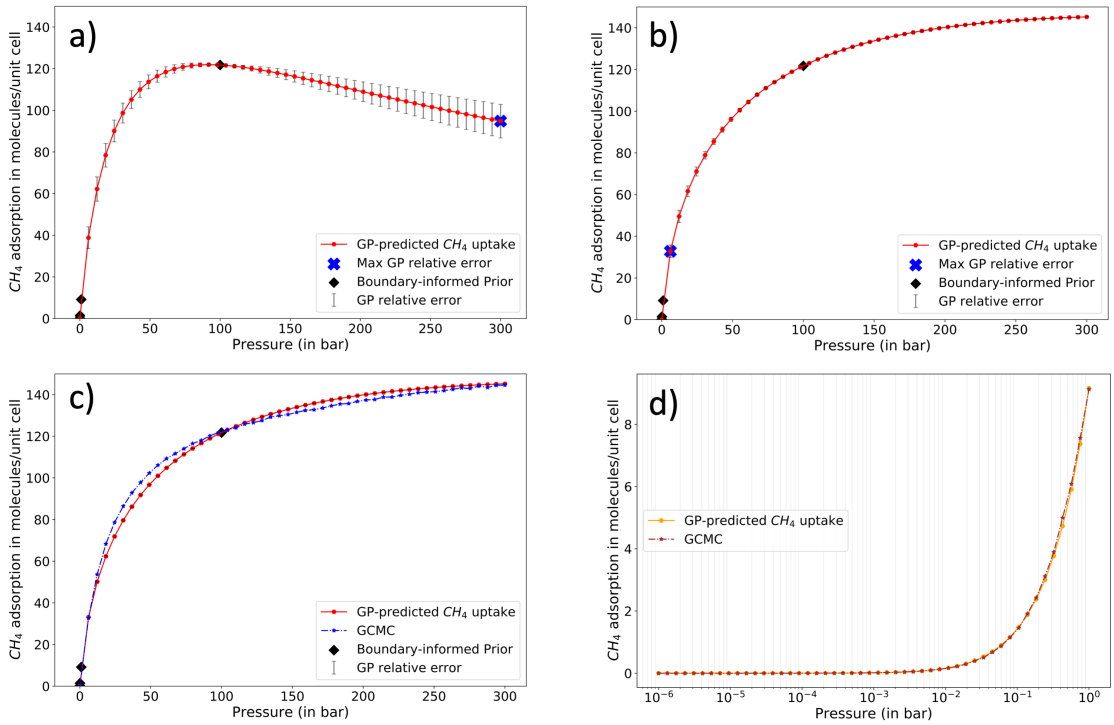

Fig. 2 Evolution of GP fit to GCMC simulations of a methane isotherm at 300 K. Line in red represents the GP fit. Panel a shows the GP fit to the data resulting from boundary-informed prior selection. Boundary-informed prior points are shown in black diamonds. Notice that the model is erroneously predicting maximum adsorption at ∼100 bar and it is decreasing as we are moving beyond the 100 bar range. This is happening since the model does not have any information on adsorption at high pressure and hence, the maximum GP relative error (shown as blue ex mark) is at 300 bar. Panels b show the subsequent iteration of the GP and the next point that needs to be added to the data fit is at low pressure region (blue ex marker at ∼10 bar). Panel c shows the final GP fit along with results from GCMC simulations (ground truth) for a full methane isotherm. Panel d shows the GP fit along with GCMC simulations for the pressure range of 10^{-6} to 1 bar

chosen to form a 50 point prior. The other priors were LHS based, linear and log-spaced, sampled along the temperature and pressure phase space. Both the LHS and log-based prior had also 50 points for a fair comparison with the boundary-informed prior.

The ground truth dataset was created using GCMC simulations for two separate tests as explained previously. Like, adsorption for a single feature, X_{test} was linearly spaced with 50 points between 10^{-6} to 300 bar, which was biased for the high pressure region. This same dataset had 40 temperature points divided linearly from 100 K to 300 K for each pressure. Thus X_{test} had 2000 points. For the low-pressure interpolation ground truth, the pressure range was from 10^{-6} to 1 bar, with 50 pressure points linearly distributed in the log-space of this range. The temperature points was distributed linearly as X_{test}, with 40 points in temperature for each pressure. Thus the low-pressure ground truth also had 2000 points. The AL fit was done with X_{test}, and so the AL protocol had zero knowledge of the adsorption in low-pressure region. This was purposefully done to test the power of the method for interpolating to low-pressure region similar to the one for single feature in section 3.1 and 3.2.

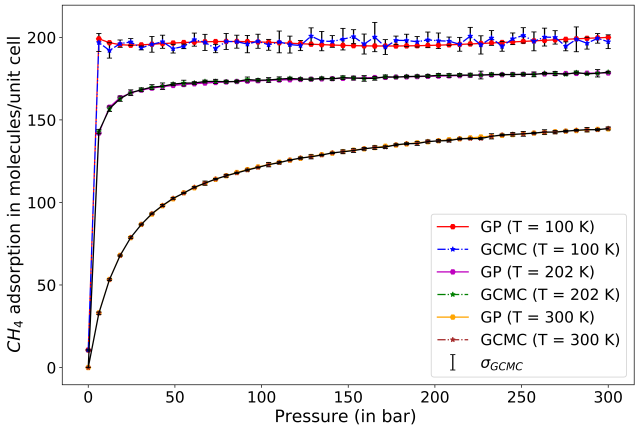

Fig. 3 Comparison of GP-predicted CH_4 uptake with GCMC simulation predicted for pressure range from 10^{-6} to 300 bar, at temperature of 100 K, 202 K and 300 K for boundary-informed prior.

The best performing prior for this study was boundary-informed and the final GP fit with GCMC simulation is shown in figure 3. The GP fit predicts the uptake very close to the GCMC and the MRE is only 0.86 % for X_{test} as reported in table 4. With only a total of 33 iterations, it can predict the uptake for a phase space of 2000 points with a very low MRE (less than 1 %). The log-spaced prior based GP fit had a slightly higher MRE of 7.99 % followed by linearly-spaced GP fit with 8.62 %. The number of iterations for the log-spaced was 19, comparable with that of boundary-informed while for the linear-spaced it was only 6.

Though each of these priors performed reasonably well in the X_{test} range, the low-pressure test set revealed appreciable differences in their performance. Observing the MRE for low pressure interpolation, we find that log-spaced prior is at 13.43 %, which is the best among the three. This was followed by boundary-informed at 18.30 %, and then we had the linear-spaced LHS with a very poor MRE of 85.74 %. The methane uptake at low pressure indicates that the performance of these models are comparable. However observing the methane uptake at this range, as shown in figure 4, we find a substantial difference in their predictions especially at the lowest temperature of 100 K. In figure 4a, the boundary-informed prior based uptake performs reasonably well at 100 K while in 4b and 4c, the GP fits from linear and log-spaced LHS prior are very far off from ground truth. The log-spaced prior first over predicts then returns to the GCMC simulation range while the linear-based prior under predicts the GCMC ground truth. The situation improves for both the LHS schemes at higher temperature of 202 K and 300 K since here the GP fit starts to match the ground truth.

Now, as we observe in the low-pressure interpolation test, we find the pressure range (10^{-6} to 10^{-1} bar) and temperature (100 K) where GP is failing for certain priors. A simple explanation would be that the GP was 'not' built or trained on this interpolation test set (please note the AL model was built on X_{test} which was biased towards high pressure region). Hence, it didn't have the knowledge that so many pressure points exists from 10^{-6} to 10^{-1} bar. In another scenario, if we had used the interpolation test set for building the AL model, then all the priors had performed quite well (since during training, AL model would run as long as the GP predicted uncertainties in all the points at the low-pressure interpolation set goes below the threshold). This also explains why boundary-informed prior is so good in these low-pressure regions which is because boundary-informed was seeded with points in the low pressure as well as points in the high-pressure zone. Hence, boundary-informed has balance of both the low- and high-pressure points, thus it is a better choice of prior than the LHS ones. This phenomenon becomes very important for prior selection, especially at low temperature regions since the adsorption isotherm shifts to the left and hence a lack of low-pressure points can make the model fail at lower temperatures.

This behaviour is also reflected in relative error isotherm plot at the low pressure range in figure 5. We see for boundary-informed prior, the highest relative error is 140 % at a single point and the rest of the errors are less than 100 % at these temperatures. The relative errors, when compared to boundary-informed, are very high for linear-spaced and log-spaced prior schemes. Though it can be pointed out that error range of 50 % for the boundary-informed prior is still high for predictions, we should observe that the uncertainty of the GCMC simulations is also very high in this range (table 4). In figure S10 in the Supporting Information, we have shown the ratio of standard deviation to methane uptake at low pressure for the GCMC simulation, and we can find that this ratio is well above 1.0 (more than 100 % error) for a major portion of this space and in the extremely low-pressure region it is as high as 4.0 (this is true for pressure range 10^{-6} to 10^{-4} bar). The uncertainty is large here due to the reason that methane adsorption at many pressure points at this range is 0 or $\sim$0, which also explains why the relative errors are so high as they have the adsorption term in the denominator (a value close to 0 in uptake can shoot this error easily). Thus, given the high uncertainty in GCMC simulation in this region as well as near zero methane uptake lev-

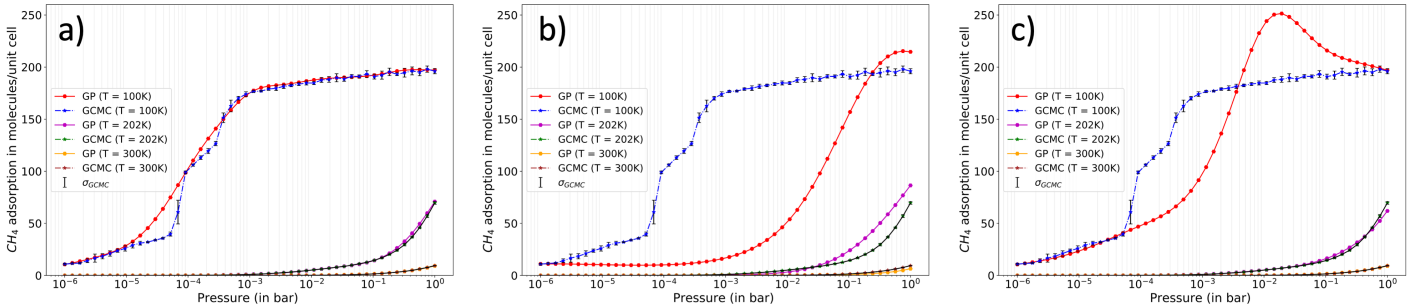

Fig. 4 Methane uptake comparison between GP and GCMC simulation in Cu-BTC at low pressure range for different priors, a) Boundary-informed prior, b) Linear-spaced LHS prior, and c) Log-spaced LHS prior

els, the boundary-informed prior uptake results can be accepted. Also, in figure 5 and 4, we find that boundary-informed prior has low relative error than log-spaced prior but in table 4, we see the MRE of log-spaced prior is lower than boundary-informed. This can be explained from our choice of temperature, which were 100 K, 202 K and 300 K, for figures 4, and 5, points which corresponded to the prior points in boundary-informed. The log-spaced prior was sampled in LHS and hence the temperature had a wide distribution and hence log-spaced prior would overall outperform boundary-informed prior if we take the complete temperature range into consideration. However, at the temperature boundaries (figures 4, and 5), the boundary-informed prior would have lower errors than log-spaced ones.

Table 4 Performance of different priors for predicting CH_4 uptake in Cu-BTC MOF with two features (all errors are expressed in %)

Prior type	Iterations	MRE (X_{test})	MRE (Low Pressure)
Boundary-informed	33	0.86	18.30
Linear-spaced LHS	6	8.62	85.74
Log-spaced LHS	19	7.99	13.43
GCMC	-	σ_{GCMC} (X_{test})	σ_{GCMC} (Low Pressure)
Ground truth	-	3.18	23.53

Another aspect of this study is the convergence of AL with iterations. Figure 6 presents AL based on boundary-informed prior convergence in terms of maximum GP-predicted relative error and MRE. Since the AL continues until the maximum GP relative error is less than 2 %, it takes a number of iterations before the protocol converges. In figure 6 we can observe that the GP maximum error quickly goes to a very low point (say 3 %). However to reach 2 % maximum error for the GP, it takes a large number of iterations. For boundary-informed prior it took 33 iterations to converge. While 33 iterations of AL was quite fast for methane adsorption, a molecule which doesn't have electrostatic interactions, this aspect can play an important role for more complex molecules. We will address this issue further for carbon dioxide adsorption in the next section and examine how fast the boundary-informed prior errors are converging with respect to iterations. Further, although not covered in this study, features such as molecule flexibility, chain length, and different configurations can also influence the AL convergence rate.

3.4 Temperature-Pressure diagrams for carbon dioxide adsorption

As mentioned earlier, carbon dioxide adsorption on Cu-BTC is more complex than methane adsorption due to electrostatic interactions. For carbon dioxide adsorption, the boundary-informed prior performs the best. However, AL converges very slowly for carbon dioxide and hence for this case, we changed the limit of maximum GP relative error (which was 2 % for all cases before) convergence limit to 3 %. In table 5, the MRE reported were based on prior convergence of maximum GP relative error of 3 %. One interesting observation is that boundary-informed MRE at low pressure for carbon dioxide adsorption with a 3 % cut-off is closer to that of methane at the threshold of 2 %. This might be due to a high value of maximum uncertainty in the low pressure region for the case of carbon dioxide adsorption, and so to obtain a flat GP relative error, AL needs more iterations. However, since MRE presents a mean property of the relative error, the majority of the points for carbon dioxide adsorption had a lower error for this low pressure region and hence the MRE was also smaller. We also observed that the linear and log-LHS priors took more iterations, 10 and 50 respectively, in case of carbon dioxide to get a maximum GP relative error of 3 %, than methane, which was only 6 and 19 to a achieve a 2 % maximum GP relative error.

In figure 7, we have shown the final GP fit based on boundary-informed prior compared with GCMC simulations (ground truth). We find a very close agreement between the GP fit and GCMC calculations. The uncertainty (shown as σ_{GCMC}), however, is very high for temperature of 100 K and here the GP under predicts the carbon dioxide uptake in the mid-pressure range. However this error is very small and is close to the 2 % relative error limit. In figure S8 and S9 of Supporting Information, we have the carbon dioxide uptake and relative error plots for the low-pressure interpolation test region, for all the three priors. Here, similar to figures 4 and 5, we find boundary-informed prior performs the best compared to the LHS-based schemes. The reasons for this performance are the same as discussed in the section 3.3 for methane, that boundary-informed prior had the most information in the prior corresponding to the low-pressure region and thus, performs the best compared to other schemes. The failures in case of boundary-informed priors (which happens at adsorption rise at 100 K) and LHS-based schemes (10^{-6} to 10^{-3} bar for low temperatures) can be explained by the lack of knowledge by

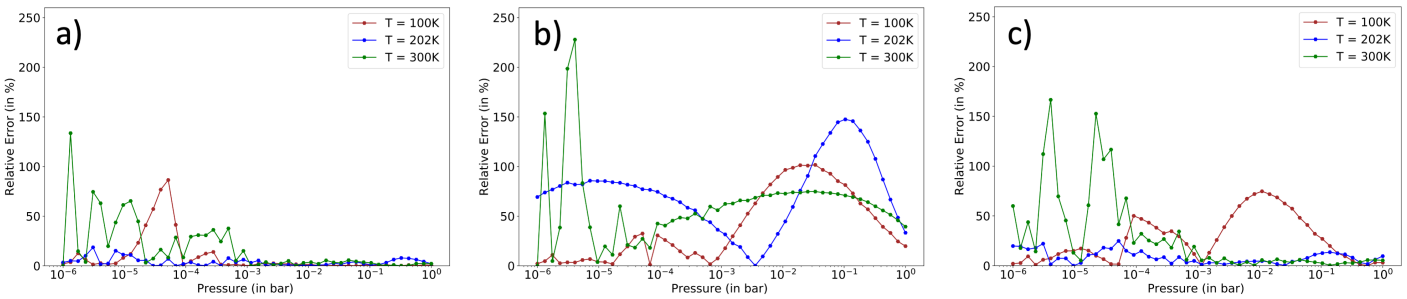

Fig. 5 Relative error (in %) comparison between GP and GCMC simulation in Cu-BTC at low pressure range for different priors, a) Boundary-informed prior, b) Linear-spaced LHS prior, and c) Log-spaced LHS prior

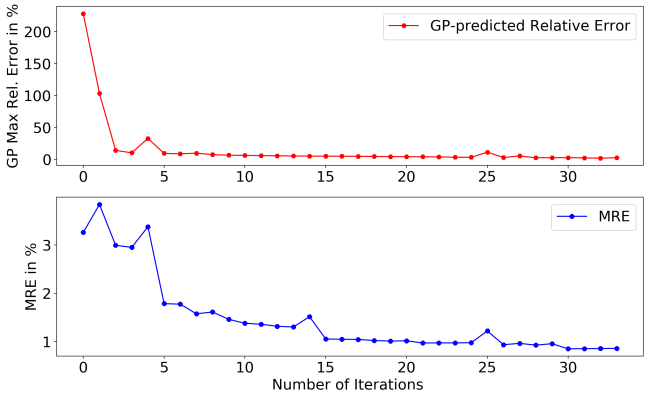

Fig. 6 Maximum GP-predicted relative error and MRE (mean relative error) with AL for methane adsorption in Cu-BTC with iterations for boundary-informed prior.

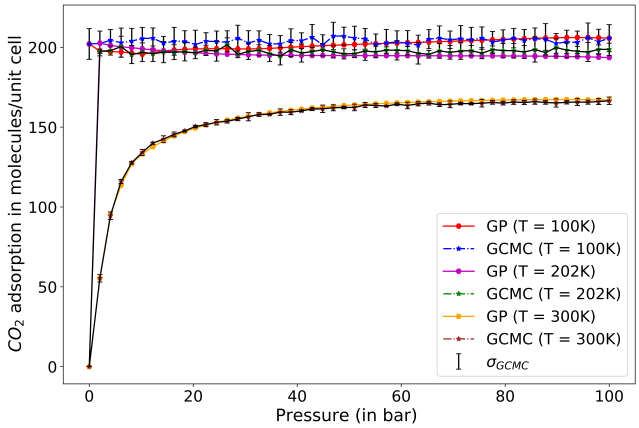

Fig. 7 Comparison of GP-predicted CO_2 uptake with GCMC simulation predicted for pressure range from 10^{-6} to 100 bar, at temperature of 100 K, 202 K and 300 K for boundary-informed prior.

the AL model, or due to extremely high uncertainties in GCMC simulations due to the near zero adsorption of carbon dioxide at these pressure points.

As discussed before, convergence of maximum GP relative error with iterations is very slow for carbon dioxide (shown in figure 8 for the boundary-informed prior). It took 33 iterations for the maximum GP relative error to reach 3 %, however it takes 129 iterations to reach the limit of 2 %. Still, the performance of a 3 % convergence is very good and comparable to methane adsorption at the 2 % threshold. The MRE, as shown in table 5 at both the full pressure and low pressure ranges are comparable, if not lower, than that of methane.

Apart from the slow convergence and encountering higher uncertainties at low temperature, AL does manage to predict carbon dioxide uptake with comparable accuracy with that of methane for two features. This further proves that the method is transferable to complex molecules and we can also effectively explore the adsorption conditions of temperature and pressure (including the low pressure region) for these complex molecules with a limited number of simulations dictated by AL.

4 Conclusions

Based on the methane and carbon dioxide adsorption proof-of-concept case studies, we can conclude that the AL framework is a promising method to efficiently collect data from molecular simulations, and the trained GPR surrogate models can replace GCMC simulation for emulating adsorption isotherm. For the case of pressure and temperature adsorption space for methane and carbon dioxide (section 3.3 and 3.4), we showed that with only 33 iterations of AL iterations, the algorithm can predict 4000 data points in temperature and the pressure range. This includes the low pressure region which is important for separation predictions (IAST). We can recognize here that with less than 2 % of the data AL can accurately estimate the full isotherms for a large temperature and pressure range. Having a protocol like AL to sequentially select adsorption simulations for surrogate models can save orders of magnitude in terms of computational cost in designing cheap and reliable surrogate model for adsorption prediction.

AL is also much faster than GCMC simulations and a GPR sur-

Table 5 Performance of different priors for predicting CO_2 uptake in Cu-BTC MOF with two features (all errors are expressed in %)

Prior type	Iterations	MRE (X_{test})	MRE (Low Pressure)
Boundary-informed	33	2.43	18.10
Linear-spaced LHS	9	2.79	43.11
Log-spaced LHS	49	2.64	18.07
GCMC	-	σ_{GCMC} (X_{test})	σ_{GCMC} (Low Pressure)
Ground truth	-	3.47	15.64

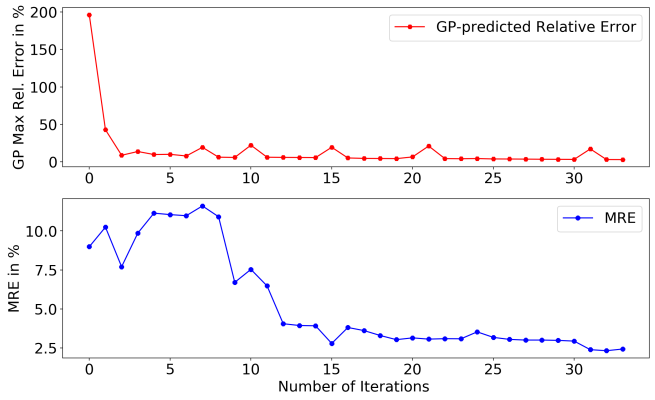

Fig. 8 Maximum GP-predicted relative error and MRE (mean relative error) with AL for CO_2 adsorption in Cu-BTC with iterations for boundary-informed prior.

rogate model only takes a few seconds to a few minutes to predict the whole isotherm. If we take the complete pressure and temperature space, the computational cost of the GPR remains very low, and the prediction is finished within minutes. However a single GCMC adsorption simulation at a fixed pressure and temperature can take from a few minutes to a few hours (can also go beyond a day depending on molecule complexity and number of production runs). Thus, predicting a full isotherm (with 50 points) can take a day or longer for complex molecules, while performing a pressure-temperature phase space simulation can take between a week to a month in terms of computational cost. In essence, AL is order of magnitudes faster than conventional GCMC simulation for predicting adsorption simulation in MOFs. These features of AL carry immense potential for material discovery especially for high-throughput simulations. An example would be material discovery for adsorption/regeneration in a pressure/temperature swing fashion. Discovery of an ideal material for such applications would require adsorption isotherm predictions throughout the pressure-temperature phase space and would be very expensive with conventional GCMC. However, with the AL approach this material discovery time would be much shorter. This would assist the search immensely as one can potentially test an order of magnitude more candidate material with the same computational resource.

Among the priors we tested, the boundary-informed one performed best considering both the X_{test} and low pressure interpolation dataset. We also found the log-spaced LHS prior can outperform boundary-informed prior in the low pressure range but has large relative errors at the high pressure region. Similarly, the linear-spaced LHS prior generally performs well at high pressures but is very poor in the low pressure range. In contrast, the boundary-informed prior has a good balance of both the low and high pressure points, and thus comes across as a better choice for building priors for AL. Moreover, the boundary-informed prior has points that are distributed in orders of magnitude at the low-pressure region which is essential for capturing the adsorption rise (the area with the highest uncertainty). This becomes evident as we compare the prior performance in the low-pressure region for methane and carbon dioxide, we observe boundary-informed prior captures the trend very closely. However, it must be clarified that there are many ways to construct a prior and we have only tested and compared three strategies in this work. In this context more novel prior models can be explored including schemes like orthogonal arrays and composite designs.[65]

Alternative AL approaches can also be explored, including the addition of multiple sampling points in a parallel fashion during the building of the GP model. In each iteration, we can select multiple points for sampling which have a GP predicted relative errors above a set uncertainty threshold. While, this study presents a simple application of AL for relatively simple molecules (methane and carbon dioxide), further studies on the number of features and other aspects of AL are needed to comprehensively understand the usefulness of AL for adsorption in MOFs.

Conflicts of interest

There are no conflicts to declare.

Acknowledgements

KM would like to thank the ND Energy's Eilers Family Graduate fellowship. YJC acknowledges support from the University of Notre Dame and AWD acknowledges support from the U.S. National Science Foundation under award CBET-1941596. All authors thank the ND Center for Research Computing for computation resources and technical support.

Notes and references

1 M. Kondo, T. Yoshitomi, H. Matsuzaka, S. Kitagawa and K. Seki, *Angewandte Chemie*, 1997, **36**, 1725–1727.
2 H. W. Langmi, J. Ren, B. North, M. Mathe and D. Bessarabov, *Electrochimica Acta*, 2014, **128**, 368–392.
3 P. Boyd, A. Chidambaram, E. García-Díez, C. Ireland, T. Daff, R. Bounds, A. Gładysiak, P. Schouwink, S. Moosavi, M. Maroto-Valer, J. Reimer, J. Navarro, T. Woo, S. Garcia, K. Stylianou and B. Smit, *Nature*, 2019, **576**, 253–256.
4 Z. Hu, Y. Wang, B. B. Shah and D. Zhao, *Advanced Sustainable Systems*, 2019, **3**, 1800080.
5 V. Pascanu, G. González Miera, A. K. Inge and B. Martín-Matute, *Journal of the American Chemical Society*, 2019, **141**, 7223–7234.
6 R.-B. Lin, S. Xiang, H. Xing, W. Zhou and B. Chen, *Coordination Chemistry Reviews*, 2019, 87–103.
7 L. E. Kreno, K. Leong, O. K. Farha, M. Allendorf, R. P. Van Duyne and J. T. Hupp, *Chemical Reviews*, 2012, **112**, 1105–1125.
8 L. Wang, M. Zheng and Z. Xie, *J. Mater. Chem. B*, 2018, **6**, 707–717.
9 L. Wang, M. Zheng and Z. Xie, *J. Mater. Chem. B*, 2018, **6**, 707–717.
10 R. A. Fritz, Y. J. Colón and F. Herrera, *Chem. Sci.*, 2021, **12**, 3475–3482.
11 A. Sturluson, M. T. Huynh, A. R. Kaija, C. Laird, S. Yoon, F. Hou, Z. Feng, C. E. Wilmer, Y. J. Colón, Y. G. Chung, D. W.

Siderius and C. M. Simon, *Molecular Simulation*, 2019, **45**, 1082–1121.

12 P. Z. Moghadam, A. Li, X.-W. Liu, R. Bueno-Perez, S.-D. Wang, S. B. Wiggin, P. A. Wood and D. Fairen-Jimenez, *Chem. Sci.*, 2020, **11**, 8373–8387.

13 N. Rampal, A. Ajenifuja, A. Tao, C. Balzer, M. S. Cummings, A. Evans, R. Bueno-Perez, D. J. Law, L. W. Bolton, C. Petit, F. Siperstein, M. P. Attfield, M. Jobson, P. Z. Moghadam and D. Fairen-Jimenez, *Chem. Sci.*, 2021, **12**, 12068–12081.

14 R. B. Getman, Y.-S. Bae, C. E. Wilmer and R. Q. Snurr, *Chemical Reviews*, 2012, **112**, 703–723.

15 Q. Yang and C. Zhong, *The Journal of Physical Chemistry B*, 2006, **110**, 17776–17783.

16 Y. G. Chung, J. Camp, M. Haranczyk, B. J. Sikora, W. Bury, V. Krungleviciute, T. Yildirim, O. K. Farha, D. S. Sholl and R. Q. Snurr, *Chemistry of Materials*, 2014, **26**, 6185–6192.

17 F.-X. Coudert and A. H. Fuchs, *Coordination Chemistry Reviews*, 2016, **307**, 211–236.

18 P. Li, N. A. Vermeulen, C. D. Malliakas, D. A. Gómez-Gualdrón, A. J. Howarth, B. L. Mehdi, A. Dohnalkova, N. D. Browning, M. O'Keeffe and O. K. Farha, *Science*, 2017, **356**, 624–627.

19 C. Wilmer, M. Leaf, C. Lee, O. Farha, B. Hauser, J. Hupp and R. Snurr, *Nature Chemistry*, 2012, **4**, 83–89.

20 Y. Bao, R. Martin, C. Simon, M. Haranczyk, B. Smit and M. Deem, *The Journal of Physical Chemistry C*, 2015, **119**, 186–195.

21 S. Li, Y. G. Chung and R. Q. Snurr, *Langmuir*, 2016, **32**, 10368–10376.

22 P. Wollmann, M. Leistner, U. Stoeck, R. Grünker, K. Gedrich, N. Klein, O. Throl, W. Grählert, I. Senkovska, F. Dreisbach and S. Kaskel, *Chem. Commun.*, 2011, **47**, 5151–5153.

23 K. Mukherjee and Y. J. Colón, *Molecular Simulation*, 2021, **47**, 857–877.

24 Z. Shi, W. Yang, X. Deng, C. Cai, Y. Yan, H. Liang, Z. Liu and Z. Qiao, *Mol. Syst. Des. Eng.*, 2020, **5**, 725–742.

25 M. Fernandez, P. G. Boyd, T. D. Daff, M. Z. Aghaji and T. K. Woo, *The Journal of Physical Chemistry Letters*, 2014, **5**, 3056–3060.

26 M. Z. Aghaji, M. Fernandez, P. G. Boyd, T. D. Daff and T. K. Woo, *European Journal of Inorganic Chemistry*, 2016, **2016**, 4505–4511.

27 Y. G. Chung, D. A. Gómez-Gualdrón, P. Li, K. T. Leperi, P. Deria, H. Zhang, N. A. Vermeulen, J. F. Stoddart, F. You, J. T. Hupp, O. K. Farha and R. Q. Snurr, *Science Advances*, 2016, **2**, e1600909.

28 A. W. Thornton, C. M. Simon, J. Kim, O. Kwon, K. S. Deeg, K. Konstas, S. J. Pas, M. R. Hill, D. A. Winkler, M. Haranczyk and B. Smit, *Chemistry of Materials*, 2017, **29**, 2844–2854.

29 N. S. Bobbitt and R. Q. Snurr, *Molecular Simulation*, 2019, **45**, 1069–1081.

30 M. Pardakhti, E. Moharreri, D. Wanik, S. L. Suib and R. Srivastava, *ACS Combinatorial Science*, 2017, **19**, 640–645.

31 G. S. Fanourgakis, K. Gkagkas, E. Tylianakis, E. Klontzas and G. Froudakis, *The Journal of Physical Chemistry A*, 2019, **123**, 6080–6087.

32 C. M. Simon, R. Mercado, S. K. Schnell, B. Smit and M. Haranczyk, *Chemistry of Materials*, 2015, **27**, 4459–4475.

33 M. Fernandez and A. S. Barnard, *ACS Combinatorial Science*, 2016, **18**, 243–252.

34 M. Fernandez, P. G. Boyd, T. D. Daff, M. Z. Aghaji and T. K. Woo, *The Journal of Physical Chemistry Letters*, 2014, **5**, 3056–3060.

35 B. J. Bucior, N. S. Bobbitt, T. Islamoglu, S. Goswami, A. Gopalan, T. Yildirim, O. K. Farha, N. Bagheri and R. Q. Snurr, *Mol. Syst. Des. Eng.*, 2019, **4**, 162–174.

36 A. Sturluson, M. T. Huynh, A. H. P. York and C. M. Simon, *ACS Central Science*, 2018, **4**, 1663–1676.

37 B. J. Befort, R. S. DeFever, G. M. Tow, A. W. Dowling and E. J. Maginn, *Machine Learning Directed Optimization of Classical Molecular Modeling Force Fields*, 2021.

38 R. Ma, Y. J. Colón and T. Luo, *ACS Applied Materials & Interfaces*, 2020, **12**, 34041–34048.

39 R. Anderson, A. Biong and D. A. Gómez-Gualdrón, *Journal of Chemical Theory and Computation*, 2020, **16**, 1271–1283.

40 A. Sturluson, A. Raza, G. D. McConachie, D. Siderius, X. Fern and C. Simon, *ChemRxiv*, 2021, 1–25.

41 E. Uteva, R. S. Graham, R. D. Wilkinson and R. J. Wheatley, *The Journal of Chemical Physics*, 2018, **149**, 174114.

42 J. Vandermause, S. B. Torrisi, S. Batzner, Y. Xie, L. Sun, A. M. Kolpak and B. Kozinsky, *On-the-Fly Active Learning of Interpretable Bayesian Force Fields for Atomistic Rare Events*, 2019.

43 J. E. Santos, M. Mehana, H. Wu, M. Prodanović, Q. Kang, N. Lubbers, H. Viswanathan and M. J. Pyrcz, *The Journal of Physical Chemistry C*, 2020, **124**, 22200–22211.

44 C. E. Rasmussen and C. K. I. Williams, *Gaussian Processes for Machine Learning*, MIT Press, 2006.

45 A. Gopalan, B. Bucior, N. Bobbitt and R. Snurr, *Molecular Physics*, 2019, **117**, 3683–3694.

46 D. Liu and J. Nocedal, *Mathematical Programming*, 1989, **45**, 503–528.

47 F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot and E. Duchesnay, *Journal of Machine Learning Research*, 2011, **12**, 2825–2830.

48 B. Shahriari, K. Swersky, Z. Wang, R. P. Adams and N. de Freitas, *Proceedings of the IEEE*, 2016, **104**, 148–175.

49 D. J. C. MacKay, *Neural Computation*, 1992, **4**, 590–604.

50 S. Seo, M. Wallat, T. Graepel and K. Obermayer, Proceedings of the International Joint Conference on Neural Networks, 2000, pp. 241 – 246 vol.3.

51 G. Maurin, P. L. Llewellyn and R. G. Bell, *The Journal of Physical Chemistry B*, 2005, **109**, 16084–16091.

52 R. Q. Snurr, A. T. Bell and D. N. Theodorou, *The Journal of Physical Chemistry*, 1993, **97**, 13742–13752.

53 D. Dubbeldam, S. Calero, D. E. Ellis and R. Q. Snurr, *Molecular Simulation*, 2016, **42**, 81–101.

54 B. Eggimann, A. Sunnarborg, H. Stern, A. Bliss and J. Siepmann, *Molecular Simulation*, 2014, **40**, 101–105.
55 J. E. Lennard-Jones, *Proceedings of the Physical Society*, 1931, **43**, 461–482.
56 S. S.-Y. Chui, S. M.-F. Lo, J. P. H. Charmant, A. G. Orpen and I. D. Williams, *Science*, 1999, **283**, 1148–1150.
57 J. M. Castillo, T. J. H. Vlugt and S. Calero, *The Journal of Physical Chemistry C*, 2008, **112**, 15934–15939.
58 A. K. Rappe, C. J. Casewit, K. S. Colwell, W. A. Goddard and W. M. Skiff, *Journal of the American Chemical Society*, 1992, **114**, 10024–10035.
59 H. A. Lorentz, *Annalen der Physik*, 1881, **248**, 127–136.
60 Q. Yang and C. Zhong, *The Journal of Physical Chemistry B*, 2006, **110**, 17776–17783.
61 Q. Yang, C. Xue, C. Zhong and J.-F. Chen, *AIChE Journal*, 2007, **53**, 2832–2840.
62 S. Wang, Q. Yang and C. Zhong, *Separation and Purification Technology*, 2008, **60**, 30–35.
63 M. D. McKay, R. J. Beckman and W. J. Conover, *Technometrics*, 1979, **21**, 239–245.
64 K. S. Walton and D. S. Sholl, *AIChE Journal*, 2015, **61**, 2757–2762.
65 T. Simpson, J. Peplinski, P. Koch and J. Allen, *Engineering with Computers*, 2001, **17**, 129–150.