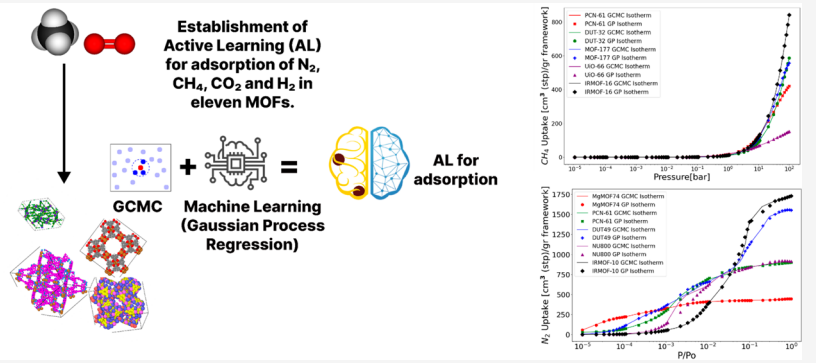
I&EC research
Industrial & Engineering Chemistry Research
pubs.acs.org/IECR
Article

Active Learning for Adsorption Simulations: Evaluation, Criteria Analysis, and Recommendations for Metal−Organic Frameworks

Etinosa Osaro, Krishnendu Mukherjee, and Yamil J. Colón*

Cite This: *Ind. Eng. Chem. Res.* 2023, 62, 13009−13024
Read Online



ABSTRACT: High-throughput molecular simulations and machine learning (ML) have been implemented to adequately screen a large number of metal−organic frameworks (MOFs) for applications involving adsorption. Grand canonical Monte Carlo (GCMC) simulations have proven effective in calculating the adsorption capacity at given pressures and temperatures, but they can require expensive computational resources. While they can be resource-efficient, ML models can require large datasets, creating a need for algorithms that can efficiently characterize adsorption; active learning (AL) can play a very important role in this regard. In this work, we make use of Gaussian process regression (GPR) to model pure component adsorption of nitrogen at 77 K from 10^{-5} to 1 bar, methane at 298 K from 10^{-5} to 100 bar, carbon dioxide at 298 K from 10^{-5} to 100 bar, and hydrogen at 77 K from 10^{-5} to 100 bar on PCN-61, MgMOF-74, DUT-32, DUT-49, MOF-177, NU-800, UiO-66, ZIF-8, IRMOF-1, IRMOF-10, and IRMOF-16. The GPR model requires an initial training of the model with an initial dataset, the prior one, and, in this study of evaluating AL, we make use of three different prior selection schemes. Each prior scheme is updated with a sampling point resulting from the GP model uncertainties. This protocol continues until a maximum GPR relative error of 2% is attained. We make a recommendation on the best prior selection scheme for the total 44 adsorbate−adsorbent pairs primarily making use of the mean absolute error and the total amount of points required for convergence of the model. To further evaluate the AL framework, we apply the BET consistency criteria on the simulated and GP nitrogen isotherms and compare the resulting surface areas.

INTRODUCTION

Metal−organic frameworks (MOFs) are nanoporous hybrid solids composed of organic ligands linked together by metal ions or clusters (nodes). Their modular design allows for extensive synthetic tunability and, thus, precise chemical and structural control. Porosity, stability, particle shape, and conductivity are just a few of the qualities that may be customized for specific purposes via an innovative synthetic design. The tuneability makes these materials attractive to meet the demands of energy storage technologies, catalysis, drug delivery, lithium-ion batteries, etc.[1−8]

Over the years, there have been tens of thousands of MOFs synthesized and over hundreds of thousands of MOFs predicted to be synthesized since the studies that gave rise to the synthesis of MOF-5 and others.[9−11] Because of the control over the pore chemistry and diversity, MOFs have been studied in applications that involve adsorption. Adsorption loading (uptake) can be calculated using Monte Carlo simulations. Grand canonical Monte Carlo (GCMC) has shown great effectiveness in the simulation of adsorption isotherms of several materials.[12−16]

Machine learning (ML) techniques are being used to estimate the adsorption loading for many MOFs in response to the researchers' need for faster and efficient data access. Over the past years, ML has been used in material science research[17] for a variety of reasons, including polymer property

Received: May 11, 2023
Revised: July 22, 2023
Accepted: July 24, 2023
Published: August 8, 2023

https://doi.org/10.1021/acs.iecr.3c01589
Ind. Eng. Chem. Res. 2023, 62, 13009−13024

prediction,[18] zeolite structure categorization,[19] crystal structure prediction,[20] etc. The estimation of gas storage capacity has been predicted using ML,[21−23] as well as predicting the oxidation state of MOFs, optimizing the swing adsorption process conditions with MOFs, and assigning partial charges to MOF atoms.[17,24] The high-throughput screening of MOFs for hydrogen separation have also been studied using ML.[25,26]

The rule of thumb is that ML requires a large set of data for the proper training of the model, but given the high cost of running many molecular simulations, there is need for an alternative. Several authors have developed ML models to predict isotherms;[27−29] however, a large dataset is required to train these models. Built on the principle of ML is a method referred to as "active learning" (AL). As a subset of the ML, the AL algorithm actively chooses the subset of instances from the pool of unlabeled data that will be labeled next. The basic tenet of the active learner algorithm notion is that, if given the freedom to select the labels it wants to learn, an ML algorithm could be able to achieve a greater degree of accuracy while utilizing fewer training points. Santos and his co-workers investigated as a dataset construction method because big databases of molecular dynamics (MD) calculations are expensive to produce. They demonstrate that the method requires a tenth of the MD calculations, compared to the production of random datasets and produces accurate models that generalize to actual scanning electron microscopy geometries. Their method is to create and train deep neural network models based on physics by learning from a database of MD computations. By foreseeing the statistical distribution of gas inside nanopores, the model explains the adsorption process.[30]

In the context of adsorption in MOFs, Mukherjee and co-workers[31] worked on using AL to predict the isotherms of methane and carbon dioxide in HKUST-1, while navigating the pressure and temperature phase space simultaneously. In this paper, we also present the use of AL (otherwise known as sequential design) to predict the adsorption isotherms of various gaseous molecules across several MOFs while not relying on a large dataset. MOFs of diverse surface areas,[32−34] topologies,[35−39] and pore size distributions[40−42] were selected for this work. Our choice of structures is derived from previous works that seek to estimate the surface area of MOFs across a diverse set of structures.[39]

As AL continues to become an important tool, the purpose of this study is to develop and evaluate AL frameworks to generate the adsorption of gas molecules for numerous MOFs at different temperatures and pressures. In this study, we evaluated different AL approaches for predicting the pure component adsorption isotherms for nitrogen (N_2), carbon dioxide (CO_2), methane (CH_4), and hydrogen (H_2) in a diverse set of MOFs. We then calculated the surface area of those MOFs from the AL predicted uptake and compared it to the BET surface area from the GCMC N_2 isotherms. As part of our analysis of the AL procedure for the chosen molecules in several MOFs, we take a deep look to observe any consistency in the posterior points added generated from the AL method for every adsorbate−adsorbent pair. We also analyze the use of different kernels for the GP and make recommendations for different priors, depending on the structure under study.

2.0. METHODS

2.1. Grand Canonical Monte Carlo. Nitrogen isotherms at 77 K, carbon dioxide isotherms at 298 K, methane isotherms at 298 K, and hydrogen isotherms at 77 K were all simulated using the RASPA code.[43] The pressure ranges used in this work were 10^{-5} to 1 bar for nitrogen and 10^{-5} to 100 bar for methane, carbon dioxide, and hydrogen. All adsorbents were modeled using the Universal Force Field (UFF)[44] and the adsorbates were modeled using TraPPE,[44] except for hydrogen. Hydrogen was modeled using the Feynman−Hibbs corrections to account for quantum effects.[45−50] It was modeled using a Lennard-Jones parameter at the center of mass and charges at the center of mass and nuclei.[51−53] Lorentz−Berthelot mixing rules were employed for cross-term interactions.[54] 50 000 production cycles and 10 000 equilibration cycles were employed in these GCMC simulations. GCMC simulations moves included the random movement of a randomly chosen molecule, the insertion of a solute particle at a randomly determined site, the deletion of a randomly chosen molecule, and rotation.[12] The adsorbents of interest in this study were IRMOF-1,[55] IRMOF-10,[55] IRMOF-16,[55] NU-800,[56] UiO-66,[57] ZIF-8,[58] DUT-32,[59] DUT-49,[60] MOF-177,[61] PCN-61,[62] and Mg-MOF74.[63] MOF atoms were held fixed at crystallographic positions. The charges on the MOFs were not considered.

2.2. Active Learning. In this study, the AL process is configured to train and fit certain historical data (prior) to a model. The posterior point in this approach that has the biggest associated relative error in the predictions updates the prior, which is used to refit the model. This procedure is repeated until the predicted relative error is <2%.

This AL protocol makes use of the Gaussian process regressor (GPR). The mean and covariance functions of a Gaussian process (GP) provide a complete description of the process.[64] This is mathematically described as

$$f \approx \mathrm{GP}(m(x), K(x, x')) \tag{1}$$

In eq 1, the function f has a GP distribution with a mean (m) function and covariance function (K).

The GP model was implemented using several GPR kernels which include the rational quadratic (RQ) kernel,[65−68] radial basis function (RBF),[69,70] and the Matern kernel.[71,72]

The RBF kernel also known as the squared exponential kernel has the form of

$$K(x, x') = \exp\left[-\frac{(x - x')^2}{2l^2}\right] \tag{2}$$

The RBF kernel is characterized by l, which is the length scale that accounts for the variance, while $x - x'$ is the Euclidean distance between x and x'.

The RQ kernel is of the form

$$K(x, x') = \left[1 + \frac{(x - x')^2}{2\alpha l^2}\right]^{-\alpha} \tag{3}$$

The RQ kernel is equivalent to the summation of many RBF kernels, and it is characterized by $x - x'$, which is the Euclidean distance between x and x' ; l is the length scale, and α controls the weighting of large-scale and small-scale fluctuations.

The Matern kernel has the following mathematical form:

$$K(x, x') = \frac{1}{\Gamma(\nu)2^{\nu-1}}\left(\frac{\sqrt{2\nu}}{L}\mathrm{d}(x, x')\right)^{\nu} k_{\nu}\left(\frac{\sqrt{2\nu}}{l}\mathrm{d}(x, x')\right) \tag{4}$$

https://doi.org/10.1021/acs.iecr.3c01589
Ind. Eng. Chem. Res. 2023, 62, 13009−13024

In eq 4, $d(x, x')$ represents the Euclidean distance, k_ν represents the modified Bessel function, and Γ represents the gamma function. The symbol ν is an additional parameter that represents the smoothness of the function. As ν tends to infinity, this kernel becomes equivalent to the RBF kernel, and when ν is 0.5, it becomes like the exponential kernel.

Based on the work done by Mukherjee and colleagues,[31] prior to running the data through the GP process, we take the logarithm (base 10) of all the data (pressure and adsorption loading) and standardize the input feature (pressure). For this work, we select the posterior point from the unseen array of test-data points, which has the highest corresponding GPR relative error. Once the prior has been updated with this posterior pressure point and its corresponding GCMC simulation uptake, the GP is then retrained, and the isotherms are predicted. This process continues until the GPR relative error of ≤2% is achieved. Scikit-learn library is used for the GP implementation.[73]

For this work, the AL method is used to determine the next point of GCMC simulation based on the GP uncertainty in the pressure test array data and thus predict the full isotherm within the pressure range. This eliminates the need to simulate many computationally intensive data points in that pressure but rather few points (referred to as the prior) and the next points for GCMC simulations, until the AL convergence policy is met. This saves computational resources. However, to evaluate the best prior schemes, we conducted 46 GCMC simulations for N_2 for the pressure range, and 64 GCMC simulations for CH_4, CO_2, and H_2 for their pressure ranges. The priors were selected from these points, as explained in section 2.2.1. Also, we compared the isotherms from the GP model to the resulting isotherms from these GCMC data points, as seen in the Results and Discussion section.

It is important to note that this AL technique is applicable to any MOF, gas type, and mixtures, as well as thermodynamic conditions (pressure and temperature).

2.2.1. Prior Selection. Three prior selection schemes were employed in this research: the boundary-informed, log-spaced, and two-data prior selection for the AL on all the molecules of interest. The prior points for the different adsorbates are shown in Table S1 in the Supporting Information (SI). Briefly, the boundary-informed prior contains points at low and high pressures, the log-spaced is logarithmically spaced in pressure, and the two-data prior contains only the lowest and highest pressures.

2.2.2. Error Calculations. In this AL framework, we calculate the following errors:

(1) GPR Predicted Relative Error: This is the ratio of the GP predicted uncertainty (σ_{GP}) to the predicted adsorption ($Y_{GP-predict}$) from the chosen model. This is the relative error we set to decrease below 2%.

This can also be described mathematically using eq 5:

$$\text{GPR relative error} = \frac{\sigma_{\text{GP}}}{Y_{\text{GP-predict}}} \tag{5}$$

(2) Mean Relative Error (MRE): The MRE is used to check the agreement between the AL-based model and the ground truth. As shown below, MRE compares the GP predicted adsorption to the ground truth GCMC results.

The mathematical formula is

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

(3) Mean Absolute Error (MAE): The MAE measures the average absolute difference between the GP-predicted value and actual output from ground truth. The mathematical formula is of the form

$$\text{MAE} = \frac{1}{n} \sum_{j=1}^{n} |Y_{\text{GP-predict}} - Y_{\text{GCMC}}| \tag{7}$$

In eqs 6 and 7, n represents the total data points.

(4) Surface Area True Relative Error: This is the true relative error between the GP results' computed surface area and the GCMC results' computed surface areas. The mathematical form is

$$\text{TRE (in \%)} = \frac{|\text{SA}_{\text{GP}} - \text{SA}_{\text{GCMC}}|}{\text{SA}_{\text{GCMC}}} \times 100 \tag{8}$$

2.3. BET Surface Area Calculations from N_2 GCMC and AL Results. The BET theory is based on the physical adsorption of gas molecules on a solid surface to form multilayers,[74,75] and it is typically used to characterize MOF surface area using N_2 isotherms at 77 K. In this work, we apply the Rouquerol et al. consistency theory to the BET theory[76,77] to calculate the surface area of 11 MOFs for both the GCMC and AL results using the BETSI code.[78]

The AL algorithm produced its isotherms of the test dataset, which is passed into the BETSI algorithm to calculate the surface area. The BETSI algorithm is responsible for selecting the data points that match Rouquerol et al. consistency theory.

3.0. RESULTS AND DISCUSSION

We first performed GCMC simulations for all 11 MOFs of varying surface area values (PCN-61, MgMOF-74, DUT-32, DUT-49, MOF-177, NU-800, UiO-66, ZIF-8, IRMOF-1, IRMOF-10, and IRMOF-16) for nitrogen, methane, carbon dioxide, and hydrogen adsorption. These results will be considered the ground truth when compared against the GPR from the AL. We first use nitrogen adsorption to test the various kernels and choose the best performer. The chosen kernel will then be used for the rest of the adsorbates. We then discuss the AL results for all molecules by selecting the results of two adsorbents for each adsorbate. Other MOFs' results for each molecule are shown in the Supporting Information (SI). The AL protocols for the various adsorbate−adsorbent combinations were evaluated based on the MAE values, and the prior scheme with the lowest MAE was selected as the best performing prior scheme. In the case of similar performance, we looked at the number of iterations the AL took to reach a 2% predicted error. We also report the true mean relative error to show the observed agreement between the AL and GCMC simulations.

3.1. Nitrogen Isotherms. We performed GCMC simulations using the RASPA code[43] as described in the methodology. The isotherms generated are shown in Figure S1 in the SI. The simulation was performed at 77 K with pressure values ranging from 10^{-5} to 1 bar, consisting of 46 data points. These isotherms are later compared to the GPR results from the AL procedure. We applied the AL protocol

https://doi.org/10.1021/acs.iecr.3c01589
Ind. Eng. Chem. Res. 2023, 62, 13009−13024

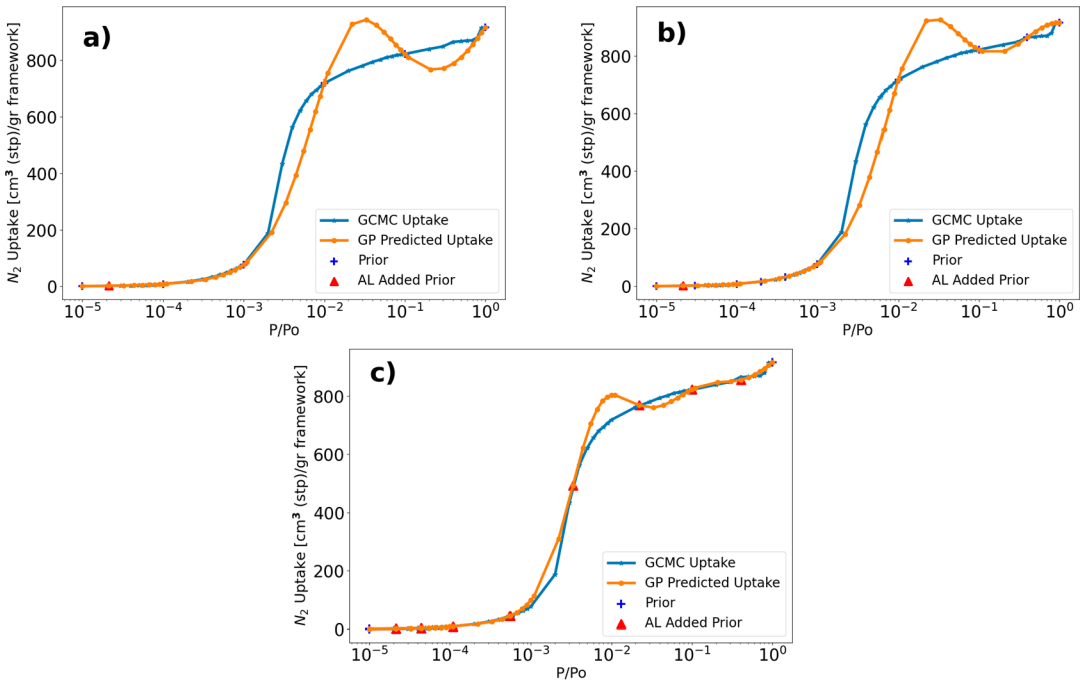


Figure 1. Nitrogen uptake comparison between GCMC simulation and GP in IRMOF 1 using the RQ kernel for three different priors: (a) boundary-informed prior, (b) log-spaced prior, and (c) two-data prior.

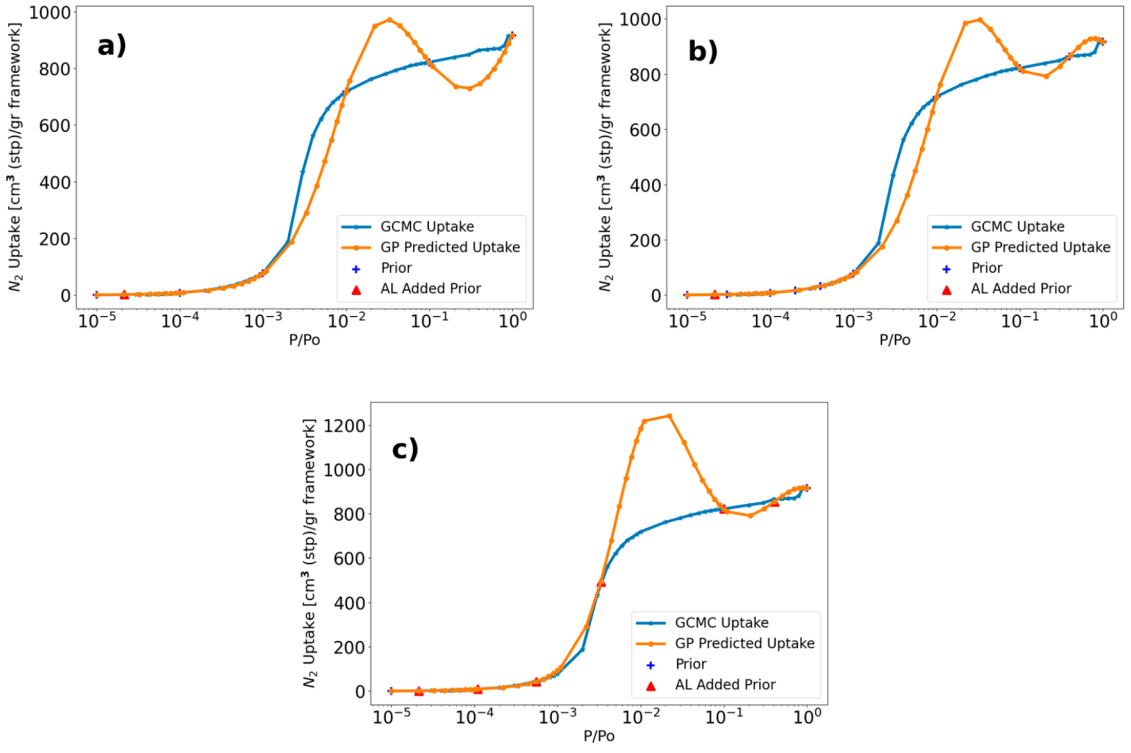


Figure 2. Nitrogen uptake comparison between GCMC simulation and GP in IRMOF 1 using the RBF kernel for three different priors: (a) boundary-informed prior, (b) log-spaced prior, and (c) two-data prior.

using the three prior selection schemes described in section 2.1: boundary-informed, log-spaced, and two-data. Various kernels of interest and approaches as described in section 2.2 were used to evaluate the AL protocol and its predictions. On determining the GCMC generated isotherms, we apply the BET SI algorithm[78] to calculate the surface area of each MOF. After determining the predicted loadings from various prior selection schemes and the best kernel choice, we also calculate the surface areas using the same algorithm and compare the results.

3.1.1. N_2 AL in IRMOF-1. Figure 1 shows the AL procedure and the resulting fits with three different prior strategies using the RQ kernel for nitrogen adsorption in IRMOF-1. The results show that the AL with the boundary-informed and log-

https://doi.org/10.1021/acs.iecr.3c01589
Ind. Eng. Chem. Res. 2023, 62, 13009−13024

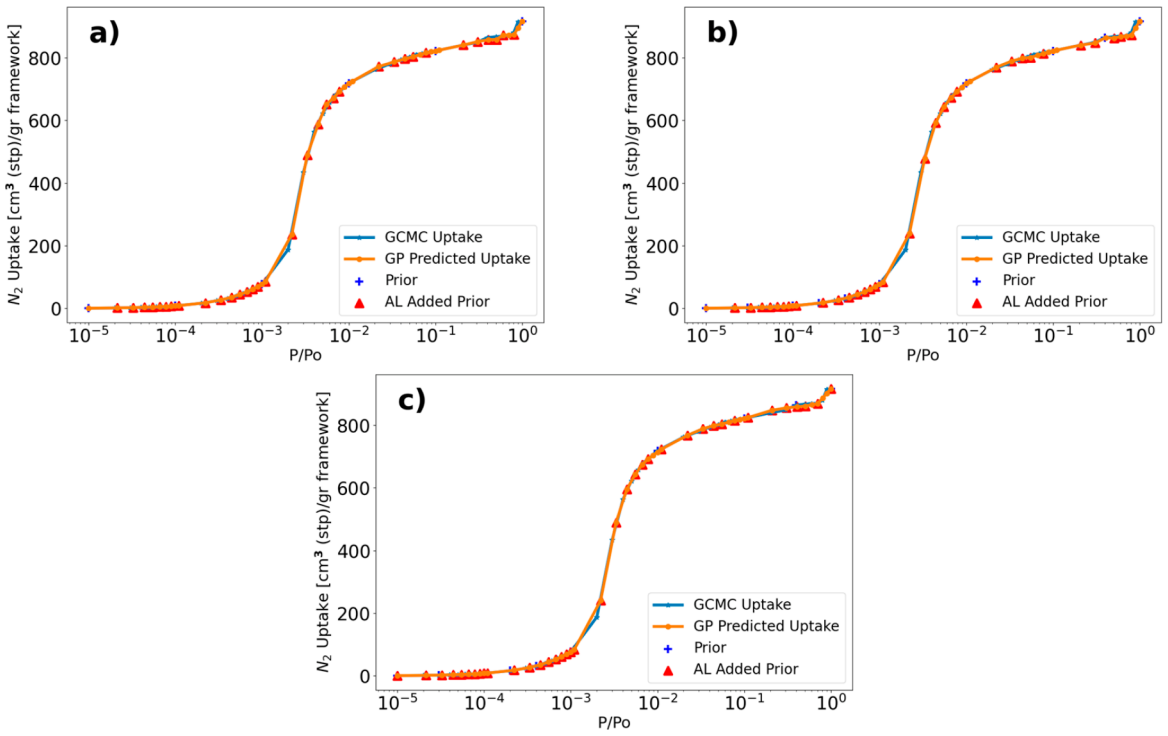


Figure 3. Nitrogen uptake comparison between GCMC simulation and GP in IRMOF 1 using the Matern kernel for three different priors: (a) boundary-informed prior, (b) log-spaced prior, and (c) two-data prior.

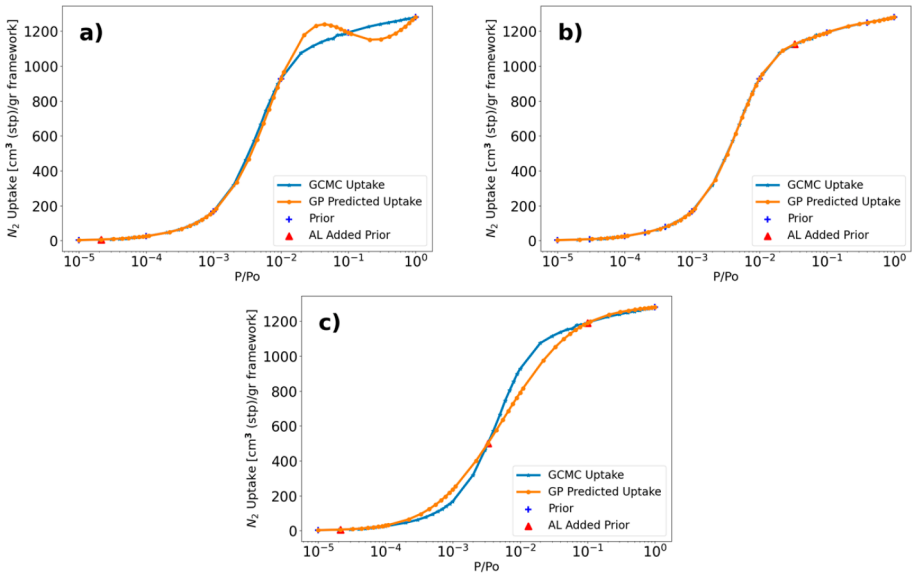


Figure 4. Nitrogen uptake comparison between GCMC simulation and GP in MOF-177 using the RQ for three different priors: (a) boundary-informed prior, (b) log-spaced prior, and (c) two-data prior.

spaced priors sampled the same point and converged to within 2% relative error in one iteration. The two-data prior, however, sampled eight additional points taking the total training data points to ten. The MREs of all of the prior schemes are reported in the SI. The GP predicted isotherm for the boundary-informed and the log-spaced prior shows a bad fit from relative pressures of $\sim 10^{-3}$ to 1, while the AL added priors for the two-data prior allowed some fitting of the isotherm between those same relative pressure ranges and generally gave better predictions from the isotherms as seen in Figure 1c. In these figures, the blue marker line represents the GCMC ground truth, the orange marker line is the final GP model, the blue crosses are the initial data or prior, and the red triangles are the added data to the GP during the AL procedure.

Using the RQ kernel, the boundary-informed and log-spaced prior performed underwhelmingly as compared with the two-data prior selection scheme, as shown in Figure 1. We start to see significant deviations at relative pressure of 10^{-3} to 1 for both the boundary-informed and log-spaced prior schemes, and lesser deviations were observed for the two-data prior.

https://doi.org/10.1021/acs.iecr.3c01589
Ind. Eng. Chem. Res. 2023, 62, 13009−13024

To further analyze the choice of kernel, we decided to evaluate the RBF kernel on this system, and the results are shown in Figure 2. The boundary-informed, log-spaced, and the two-data prior all show significant deviations in the isotherms, as seen in the figures. The two-data prior performs badly at relative pressure between 10^{-3} and 10^{-1}, as a result of the AL algorithm failing to pick and sample pressure points in that region. The GP model performs poorly in this region due to insufficient starting prior data points in the high-pressure region. The boundary-informed and log-spaced prior performed better than the two-data prior but resulted in significant deviations in the isotherms because the AL algorithm did not sample more points.

Generally, these results show that the RQ kernel performed better than the RBF kernel.

Figure 3 shows the implementation of the Matern kernel; the isotherms generated were better compared to the RQ and RBF kernels, but a lot of data points were required due to the underconfidence of the matern kernel. Oversampling of many points to update the prior was one of the situations we looked to avoid in this work. The boundary-informed prior sampled 34 more points, making a total of 40 points to completely train the model from a total of 46 data points. The log-spaced prior resulted in a total of 42 data points and the two-data prior required a total of 39 data-points.

Based on the analysis of the isotherms, MREs (see the SI), MAE and the total amount of points needed to converge the model, we concluded that the RQ kernel meets all our criteria. Hence, this kernel was chosen as the default kernel for other MOFs and adsorbents in this paper. The RQ kernel has also been used for other ML studies of MOFs.[31] To further support the use of the RQ kernel, we show better performance of the RQ to the RBF kernel for IRMOF-10 in Figures S9i and S9ii, respectively, in the SI. The isotherms from the RQ kernel were better than those from the RBF kernel. We generally observed good fits from this decision, as seen in the next sections and in the SI.

3.1.2. N_2 AL in MOF-177. For the RQ kernel, the boundary-informed converged within a 2% GP relative error with one iteration for the boundary-informed and log-spaced prior schemes. The two-data prior scheme required three iterations. Figure 4 shows the comparison between the predicted and actual isotherms.

For the boundary-informed prior, we observed deviations in the predictions from ~10^{-2} to 1 bar (high-pressure region). These deviations are experienced due to no sampling of the high-pressure regions by the GP model. None of these deviations were observed for the log-spaced prior scheme due to the sampling of more points in the high-pressure region as required by this prior scheme. The two-data prior resulted in poor predictions, as shown in Figure 4c.

The fits of the three prior schemes for the other MOFs (using the RQ kernel) are all shown in the SI. Based on the best performing prior, a comparison between the ground truth and GP predicted isotherm is shown for some select MOFs in Figure 5. We see a good comparison between the GP model (symbols) and the ground truth data (lines). Noticeable deviations in the predictions are observed at pressure ranges of 10^{-3} to 10^{-2} bar. However, overall, the fits show that the AL algorithm produces good results, despite the limited number of simulations.

The total sampling points (required training data), prior schemes, resulting MAE, and recommendation for all 11 MOFs of interest are provided in Table 1. From Table 1, we can see that the log-space prior selection scheme is most popular among the recommendations due to a lower MAE value than others. Thus, the log-spaced prior is the recommended prior for N_2 studies at 77 K.

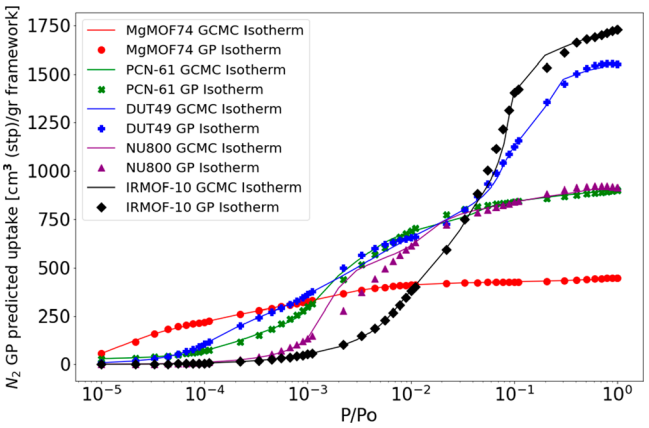


Figure 5. Comparison of GP and GCMC nitrogen isotherms for some select MOFs.

Table 1. Nitrogen AL Summary Showing Errors, MAE, and Recommended Prior Selection Scheme for All 11 MOFs[a]

	Total data points for different priors			MAE [cm^3/gr framework] for different priors		
MOF	BI	LS	TD	BI	LS	TD
PCN-61	7	**11**	6	11.98	**5.88**	11.54
MgMOF-74	11	**12**	11	4.36	**3.48**	4.62
DUT 32	7	**12**	4	27.6	**16.76**	26.88
DUT 49	7	**11**	5	34.14	**22**	27.54
MOF 177	7	**11**	5	26.88	**10.7**	32.58
NU-800	7	11	8	10.26	13.98	43.76
UiO-66	7	**11**	4	2.08	**1.26**	5.28
ZIF-8	11	12	**15**	3.9	3.4	**1.9**
IRMOF-1	7	11	**10**	38.3	33.34	**23.96**
IRMOF-10	7	**11**	5	55.12	**43.68**	63.3
IRMOF-16	7	**11**	5	105.3	**53.18**	89.56

[a]"LS" represents log-spaced, "BI" represents boundary-informed, and "TD" represents two-data priors. Bolded values represent the recommended prior for the relevant MOF.

3.2. Surface Area Calculations from N_2 GCMC and GP Isotherms. We computed N_2 isotherms at 77 K to calculate the surface area (SA) in m^2/g of several MOFs. On using the BET SI algorithm,[78] and to further evaluate the AL results, we obtained the surface areas from the GCMC and the AL results for the three prior schemes. These results are reported in Table 2, and the values in parentheses represent the percentage surface area true relative error (TRE). Based on the TREs, we see a consistent lower TRE for the prior scheme selected in Table 1. This further proves that recommendations of the prior for selection are majorly dependent on the MAE, followed by the number of data points required to converge the model to the set AL protocol of 2% GP relative error.

3.3. Methane Isotherms. CH_4 isotherms across the same set of MOFs described in section 3.1 were calculated using the RASPA code.[43] The isotherms for the several MOFs are shown in Figure S11 in the SI. For the sake of visualization, we show the isotherms in the order of low-pressure ranges, high-

https://doi.org/10.1021/acs.iecr.3c01589
Ind. Eng. Chem. Res. 2023, 62, 13009−13024

Table 2. Computed Surface Areas from Ground Truth and Results from All Prior Schemes[a]

MOF	GCMC SA (m^2/g)	BI prior SA (m^2/g)	LS prior SA (m^2/g)	TD Prior SA (m^2/g)
PCN-61	3416	3552 (3.98)	3383 (0.97)	3536 (3.51)
MgMOF-74	1806	1797 (0.95)	1807 (0.06)	1809 (0.17)
DUT 32	4725	4719 (0.13)	4829 (2.2)	4572 (3.24)
DUT 49	4847	4656 (3.94)	4754 (1.92)	4210 (13.14)
MOF 177	5011	5294 (5.65)	4982 (0.58)	4966 (0.9)
NU-800	3491	3478 (0.37)	3394 (2.78)	3441 (1.43)
UiO-66	1313	1305 (0.61)	1305 (0.61)	1297 (1.22)
ZIF-8	1402	1405 (0.21)	1403 (0.07)	1403 (0.07)
IRMOF-1	3455	3588 (3.85)	3450 (0.14)	3454 (0.03)
IRMOF-10	6333	6588 (4.03)	6230 (1.63)	4729 (25.33)
IRMOF-16	5376	5196 (3.35)	5499 (2.29)	6533 (21.52)

[a]The values in parentheses represent the surface area true relative error relative to the GCMC surface area.

pressure ranges, and altogether. The simulations were done at 298 K with pressure values ranging from 10^{-5} to 10^{2} bar, consisting of 64 data points.

The boundary-informed prior, log-spaced prior, and two-point prior selection were also used for all the MOFs using the RQ kernel as described in section 2.2. The results for the adsorbate−adsorbent pair as shown in the next few subsections were analyzed in the same manner as the nitrogen predictions; the MAE, and in the case of equal or significantly close MAE then we use the total amount of data points required to fully train the model within a convergence of 2% GP relative error. In the case of methane, the total amount of data points in the ground-truth was 64. The boundary-informed prior contains 8 pressure points while that of the log-spaced prior contains 12 data points out of the sixty-four data points total points. The two-data prior still has two points which are the lowest and highest pressures from the GCMC simulation. In the subsequent sections, we show and discuss two adsorbents (others are shown in the SI).

3.3.1. CH_4 in MgMOF-74. With five iterations, the lowest MAE across all the prior schemes was the logarithmically spaced prior with a value of 1.56 (cm^3(stp)/gr framework). The boundary informed prior with two iterations resulted in a MAE of 3.44 (cm^3(stp)/gr framework), and the two-data prior finished in four iterations with a MAE of 2.1 (cm^3(stp)/gr framework). Figure 6 shows that the logarithmically spaced prior gave the best fit compared to the boundary-informed and two-data priors. From the fits, we see that the boundary-informed prior performed fits were good, until high pressures of 10−100 bar where fluctuations are observed. The log-spaced prior gave good fits at all pressure ranges while the two-data prior performed well up until pressures of 1−100 bar. Based on the performance of all the fits (MAE), we recommend the use of the logarithmically spaced prior regardless of having the most sampled points in addition to the initial prior data points.

3.3.2. CH_4 in IRMOF-16. All fits as seen in Figures 7 were good using all of the prior selection schemes. A total of nine data points gave a MAE of 3.6 (cm^3(stp)/gr framework) using the boundary-informed prior, while with 20 data points to train the model, the MAE of the log-spaced prior was 1.92 (cm^3(stp)/gr framework). The two-data prior with six total points resulted in a MAE of 2.14 (cm^3(stp)/ gr framework). We immediately noted that the log-spaced prior sampled a lot of points (∼41% of the available dataset), and this is something we would like to avoid. The two-data prior required ∼9% of the total data, and with a very close MAE to that of the log-spaced prior, we recommend the use of the two-data prior scheme for this absorbent.

The fits of the three prior schemes for the other MOFs are all shown in the SI. The results for some MOFs are listed in Figure 8. On comparing these results, we can see that the AL

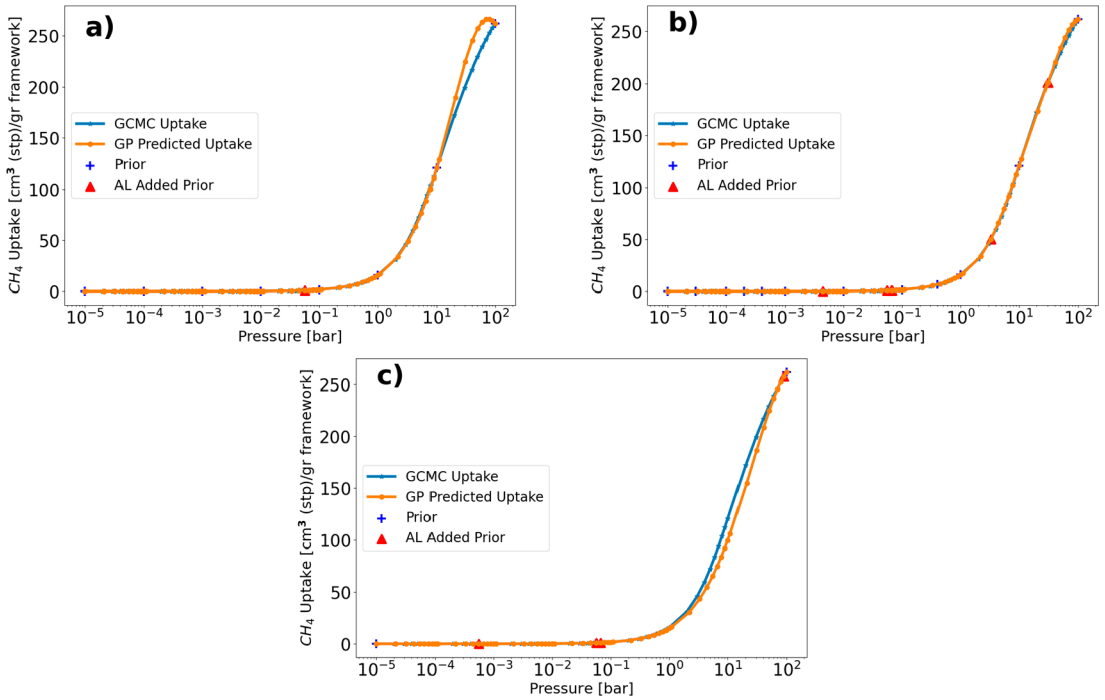


Figure 6. Methane uptake comparison between GCMC simulation and GP in MgMOF-74 using the RQ kernel for three different priors: (a) boundary-informed prior, (b) log-spaced prior, and (c) two-data prior.

https://doi.org/10.1021/acs.iecr.3c01589
Ind. Eng. Chem. Res. 2023, 62, 13009−13024

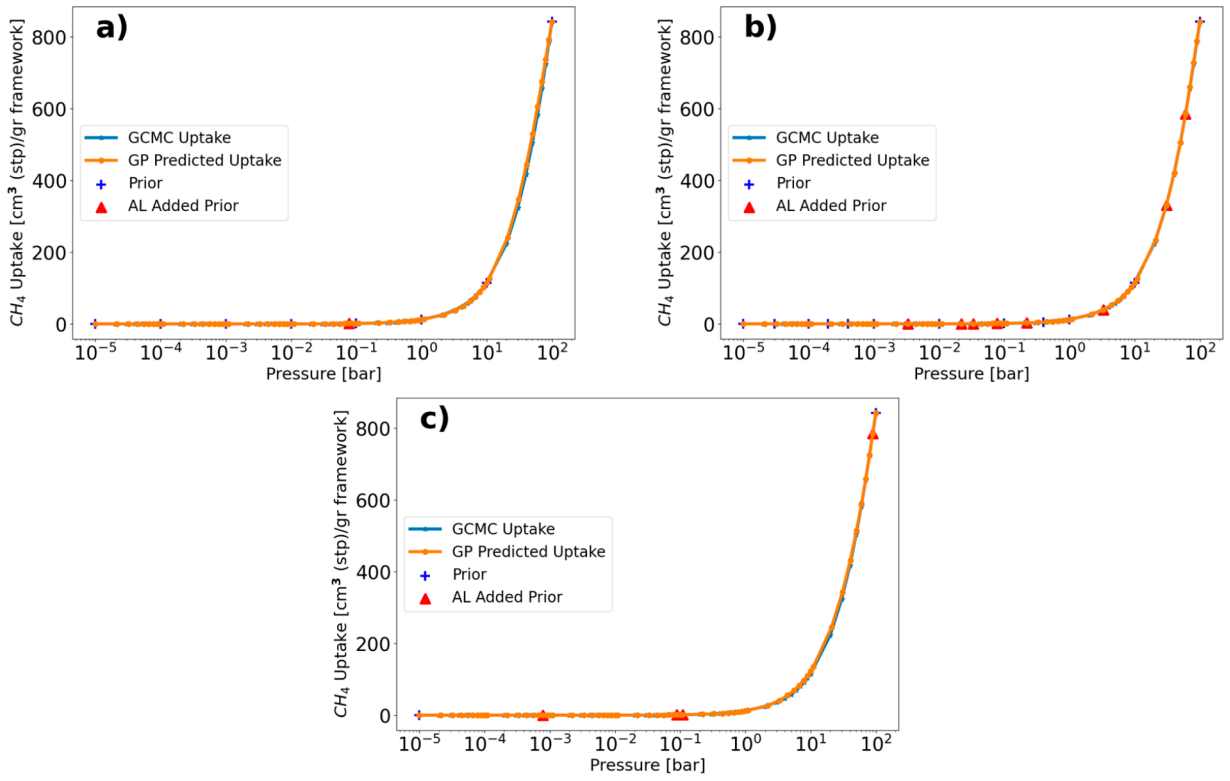


Figure 7. Methane uptake comparison between GCMC simulation and GP in IRMOF-16 using the RQ kernel for three different priors: (a) boundary-informed prior, (b) log-spaced prior, and (c) two-data prior.

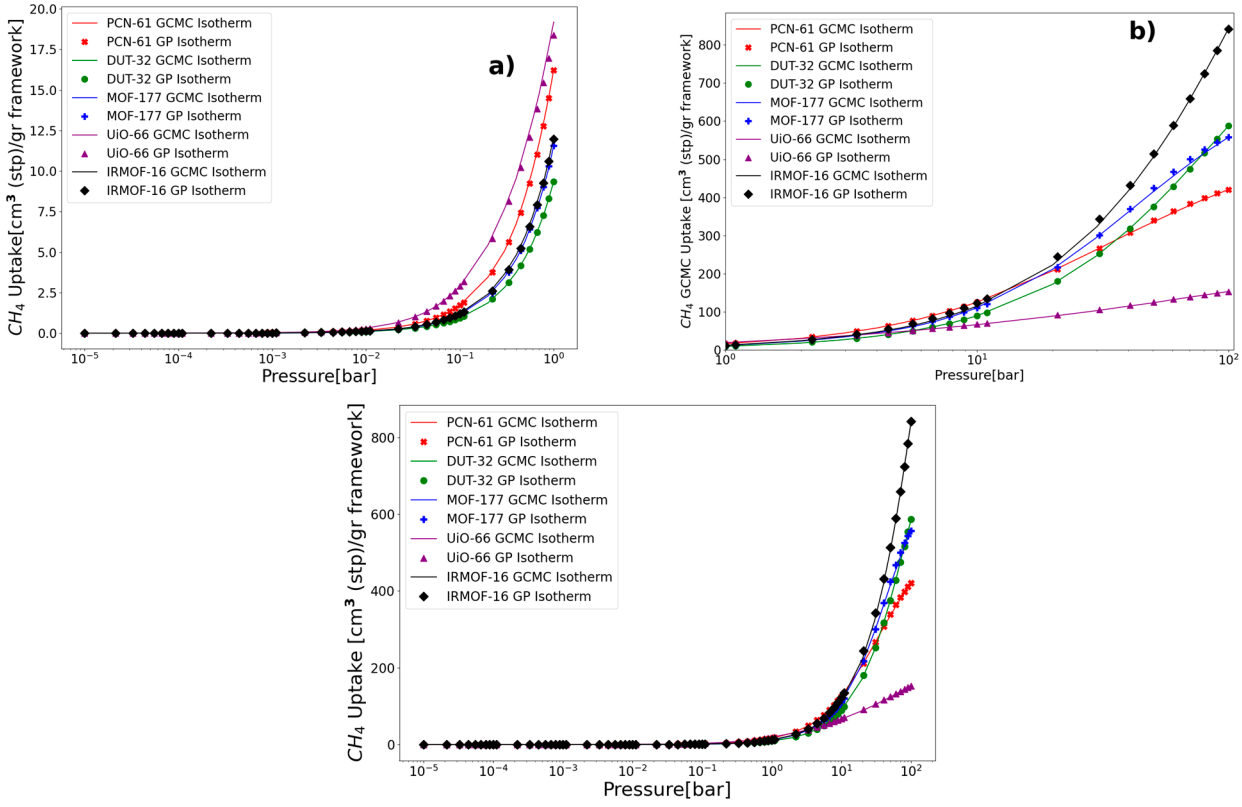


Figure 8. Comparison of GP and GCMC methane isotherms for some select MOFs. For visualization purposes, we show three different pressure ranges for the isotherms: (a) the low-pressure region, (b) the high-pressure regions, and (c) all the pressure ranges used for methane.

protocol works well for methane using a small amount of data points required for simulations. Here, we see an excellent

https://doi.org/10.1021/acs.iecr.3c01589
Ind. Eng. Chem. Res. 2023, 62, 13009−13024

comparison between the GP model (symbols) and the ground-truth (lines) with limited simulation data. No noticeable deviations were present in the comparisons.

Table 3 shows a summary of all of the MOFs discussed for the methane AL study. We generally see that the log-spaced prior resulted in more iterations than boundary-informed or the two-data priors. Based on the resulting MAEs, four of the MOFs were better modeled with the boundary-informed prior and four of the MOFs were also better modeled with the log-spaced prior. This similarity makes inferring a generic prior for other MOFs being studied in this work; however, from Table 3, it is obvious that the boundary-informed prior all resulted in a lesser number of iterations than the log-spaced prior, and thus picking the boundary-informed prior will be recommended as the generic prior.

Table 3. Methane AL Summary Showing Errors, MAE and Recommended Prior Selection Scheme for All 11 MOFs[s]

	Total Data Points for different priors			MAE [cm^3/gr framework] for different priors		
MOF	BI	LS	TD	BI	LS	TD
PCN-61	**9**	18	6	**1.7**	1.76	2.3
MgMOF-74	9	**17**	6	3.44	**1.56**	2.1
DUT 32	9	**13**	5	3.04	**1.16**	5.18
DUT 49	9	17	6	2.56	2.18	**2.02**
MOF 177	9	**16**	26	5.1	**2.18**	1.96
NU-800	**9**	21	10	**2.02**	1.92	2.12
UiO-66	9	14	5	0.94	0.72	**0.74**
ZIF-8	9	**18**	6	1.62	**0.88**	1.56
IRMOF-1	**9**	18	6	1	1.44	2.3
IRMOF-10	**9**	17	8	**4.52**	2.42	3.3
IRMOF-16	9	20	**6**	3.6	1.92	**2.04**

[s]"LS" represents log-spaced, "BI" represents boundary-informed, and "TD" represents two-data priors. Values shown in bold font represent the recommended prior for the relevant MOF.

The preference for the selection of the best prior scheme has always been the prior scheme with the lowest MAE; however, it was observed that, for MOFs such as the MOF-177, NU-800, UiO-66, and IROF-16, the MAE value between the lowest and the next lowest prior scheme MAEs were close. There were significant differences in the total amount of points required to converge to the 2% maximum GP relative error; hence, for these MOFs, we selected the best prior scheme using the total amount of data points required.

3.4. Carbon Dioxide Isotherms. Like methane isotherms, the CO_2 isotherms were generated via GCMC using RASPA in the same set of MOFs at 298 K. Figure S43 in the SI shows adsorption from experiments and GCMC simulations in Mg-MOF-74, showing good agreement. The pressure points of interest in the CO_2 isotherms and AL range from 10^{-5} to 10^{2} bar, consisting of 64 data points. The isotherms generated are shown in Figure S21 in the SI. On applying the same sets of prior schemes are applied to test for the best one when carrying out AL on CO_2 in the various MOFs in the next subsections. Once again, we evaluate their performance using the same metrics as described for AL in methane and thus recommend the best prior selection per MOF using the RQ kernel.

3.4.1. CO_2 in DUT-32. The two-data prior was the chosen prior for this adsorbate−adsorbent pair with the lowest MAE of 6.3 (cm^3(stp)/gr framework) while requiring a total of 10 data points to satisfy the AL protocol. The next lowest prior scheme was the log-spaced prior with 6.38 (cm^3(stp)/gr

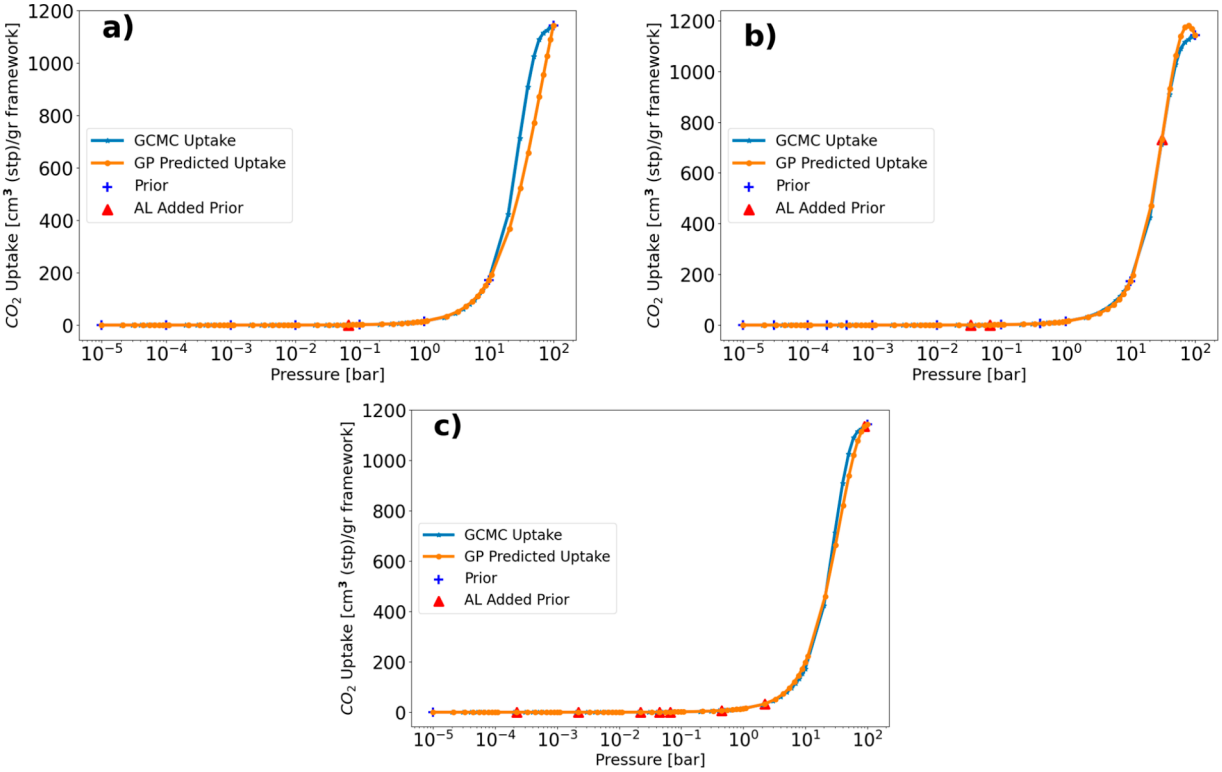


Figure 9. Carbon dioxide uptake comparison between GCMC simulation and GP in DUT-32 using the RQ kernel, for three different priors: (a) boundary-informed prior, (b) log-spaced prior, and (c) two-data prior.

https://doi.org/10.1021/acs.iecr.3c01589
Ind. Eng. Chem. Res. 2023, 62, 13009−13024

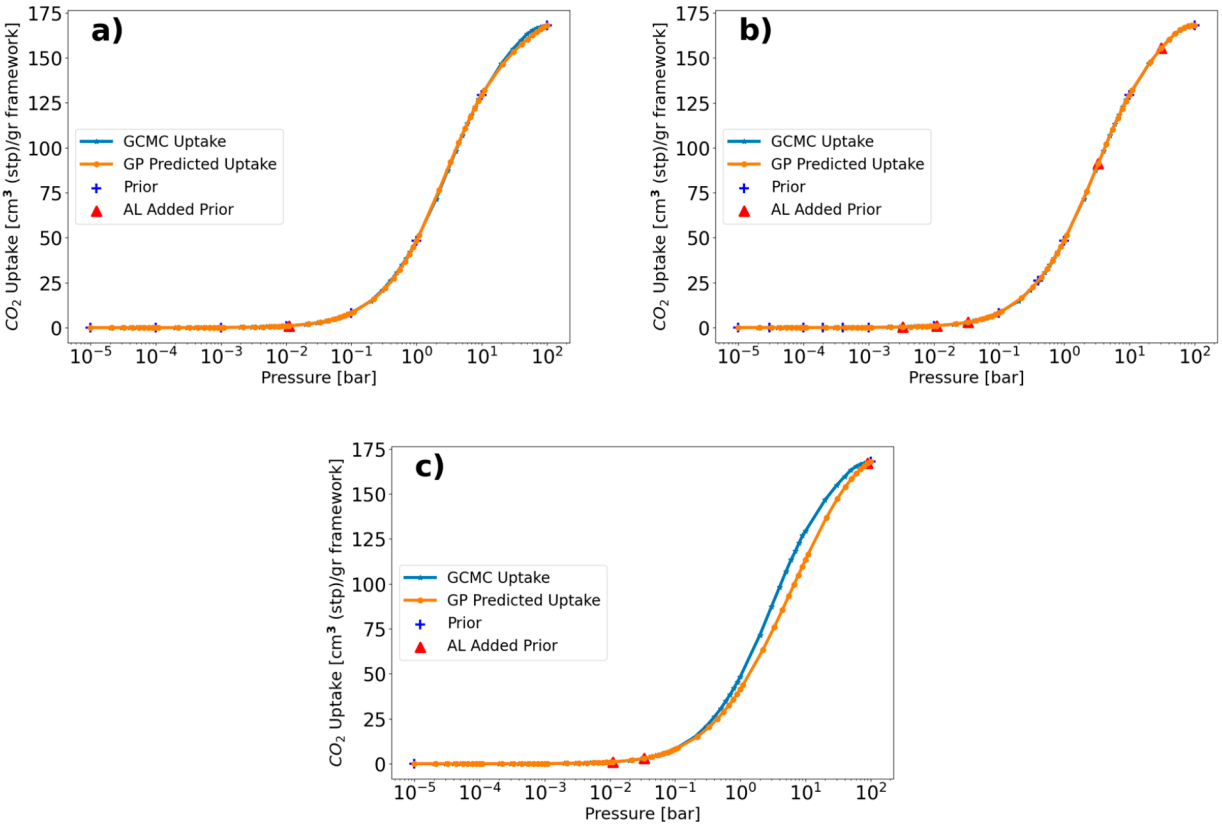


Figure 10. Carbon dioxide uptake comparison between GCMC simulation and GP in UiO-66 using the RQ kernel for three different priors: (a) boundary-informed prior, (b) log-spaced prior, and (c) two-data prior.

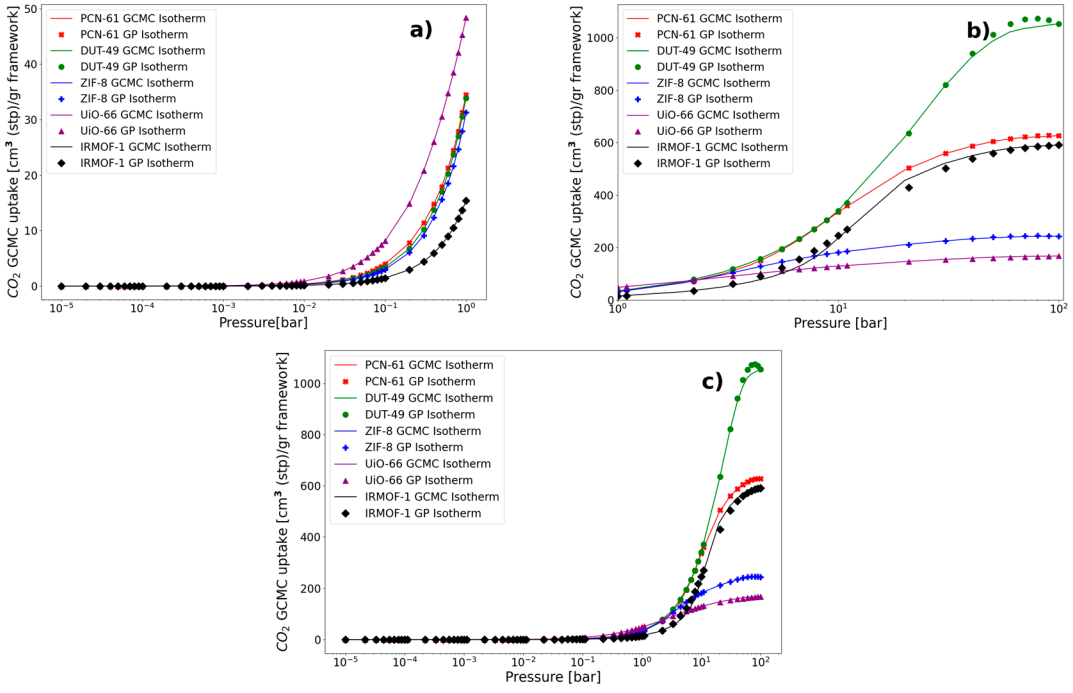


Figure 11. Comparison of the GP and GCMC carbon dioxide isotherms for some select MOFs. For visualization purposes, we show three different pressure ranges for the isotherms: (a) the low-pressure region, (b) the high-pressure regions, and (c) all the pressure ranges used for carbon dioxide.

framework), and the highest was the boundary-informed prior with a MAE of 21.9 (cm^3(stp)/gr framework. Figure 9a shows a significant deviation in the isotherms from pressures of 10 to 100 bar. The isotherms for the log-spaced prior in Figures 9b and 9c show less deviations.

https://doi.org/10.1021/acs.iecr.3c01589
Ind. Eng. Chem. Res. 2023, 62, 13009−13024

3.4.2. CO_2 in UiO-66. For UiO-66, the preferred prior scheme will be the boundary-informed prior because we observed a MAE of 1 (cm^3(stp)/gr framework using just nine total data points, as opposed to the log-spaced prior which produced a MAE of 0.96 (cm^3(stp)/gr framework), using 17 total data points. The MAEs as seen are close, hence resulting in picking the best prior scheme based on the total number of data points. The isotherms in Figure 10b using the log-spaced prior are almost perfect, while that of the two-data prior has deviations at high pressures, with a MAE of 2.7 (cm^3(stp)/gr framework.

The fits of the three prior schemes for the other MOFs are all shown in the SI. In Figure 11, we show the resulting GP predicted isotherms using the best prior (symbols) against the GCMC isotherms (lines) for some select MOF. The comparison between the GP model and the ground-truth is excellent, especially considering the limited number of simulations required for AL. However, for the DUT-49 MOF, we see noticeable deviations in the high-pressure range.

Table 4 shows a summary of all of the MOFs discussed for the carbon dioxide AL study. From Table 4, the most common recommendation for prior scheme was the LS prior with five of the MOFs best modeled with this prior. It is also important to note the LS prior resulted in more iterations than the boundary-informed prior, but the primary choice for recommendation of prior scheme remains the MAE. Despite this, the results suggest the LS prior is a good choice for other MOFs studied that may be studied for CO_2 adsorption not discussed in this work.

Table 4. Carbon Dioxide AL Summary Showing Errors, MAE, and Recommended Prior Selection Scheme for All 11 MOFs[a]

	Total Data Points for different priors			MAE [cm^3/gr framework] for different priors		
MOF	BI	LS	TD	BI	LS	TD
PCN-61	9	**17**	6	12.26	**3.76**	7.5
MgMOF-74	9	**13**	6	8.18	**3.22**	9.62
DUT 32	9	15	**10**	21.9	6.38	**6.3**
DUT 49	**9**	13	6	**6.18**	6.56	15.18
MOF 177	9	**15**	5	17.7	**8.06**	12
NU-800	9	**18**	8	7.86	**4.74**	7.88
UiO-66	9	17	5	**1**	0.96	2.7
ZIF-8	9	**17**	6	4.22	**1.64**	5.66
IRMOF-1	9	17	**13**	10.58	4.86	**4.46**
IRMOF-10	9	18	**9**	26.2	13.44	**13.28**
IRMOF-16	9	17	**14**	43.2	17.04	**10.6**

[a]"LS" represents log-spaced, "BI" represents boundary-informed, and "TD" represents two-data priors. Values shown in bold font represent the recommended prior for the relevant MOF.

3.5. Hydrogen Isotherms. The hydrogen isotherms were generated from the GCMC simulations by using the RASPA code. These simulations were done at 77 K. The pressure range used in the simulations and both the AL protocol were from 10^{-5} to 10^2 bar, consisting of 64 data points. The same moves and cycles as described in the methodology apply here. Due to the cryogenic temperature, we use the Feynman–Hibbs correction.[49] The isotherms generated from the Monte Carlo simulations are shown in Figure S31 in the SI, in the order of low, high, and all pressure ranges. Like the AL work done on previous adsorbates in the various MOFs, we evaluate their performance using the same metrics as described for AL in CO_2 and thus recommend the best prior selection per MOF. Given the good fits experienced in the previous sections, we also stick to the use of the RQ kernel for all of the MOFs and prior schemes.

3.5.1. H_2 in NU-800. On sampling one additional point to make seven total data points by the boundary-informed prior AL protocol, the resulting MAE was 5.58 (cm^3(stp)/gr framework). The two-data prior required a total of five data points to give a MAE of 30.16 (cm^3(stp)/gr framework); this is backed up from the very significant deviations in the isotherms at high pressures. The log-spaced prior resulted in a MAE of 6.1 (cm^3(stp)/gr framework) requiring 17 total data points. Figure 12 shows good fits for the boundary-informed and the log-spaced priors and a large deviation at pressures of 1–100 bar for the two-data prior.

3.5.2. H_2 in ZIF-8. The isotherms generated from the GP model are shown in Figure 13. Figure 13a has the boundary-informed prior showing a significant deviation in the isotherms from 10 to 100 bar, resulting in a MAE of 5.54 (cm^3(stp)/gr framework). With a MAE of 3.72 (cm^3(stp)/gr framework), the isotherms for the log-spaced prior in Figure 13b were much better. Very high deviations occur in Figure 13c, and this is supported by the MAE of 11.58 (cm^3(stp)/gr framework). For this adsorbate–adsorbent pair, we recommend the use of the log-spaced prior.

We also show the GP predicted uptake, compared with the GCMC isotherms for some select MOFs in Figure 14. Like our findings for the GP models for N_2, CH_4, and CO_2, we observe accurate predictions from the GP models for H_2 when compared with the ground-truth data. The recommendations and outcomes for all the MOFs are discussed in this paper and in the SI are presented in Table 5. We observe that the boundary-informed prior was the most frequently recommended selection scheme. Thus, this is a recommended prerequisite for H_2 adsorption in MOFs.

Table 5 shows a summary of all of the MOFs discussed for the hydrogen AL study. The isotherms for 10 other MOFs are shown in the SI. From Table 5, the most common recommendation for the prior scheme was the BI prior and thus is a safe prior choice for all other MOFs studied for hydrogen adsorption not discussed in this study.

4. CONCLUSIONS

As shown by the results from the adsorbate–adsorbent pairs discussed in this paper, the AL method proves to be an efficient way of selecting which GCMC simulations to perform to construct surrogate models using GPs. For nitrogen, we found that the highest number of data points required to accurately describe and predict isotherms across a range of MOFs is 15 out of 46 total data points, using the two data points prior in ZIF-8. Similarly, for methane, the highest number of data points was 26 out of 64, using the two-data prior in MOF-177; however, the recommended prior selection scheme for this MOF only required 16 total data points. The highest number of data points for carbon dioxide isotherms was 18 out of 64. The isotherms prove that with ~20%–30% of GCMC simulations as intelligently selected by the AL criteria, we can generate almost perfect isotherms compared to the ground-truth data of larger GCMC simulation data-points generated (46 in the case of N_2 and 64 in the case of CH_4, CO_2, and H_2).

https://doi.org/10.1021/acs.iecr.3c01589
Ind. Eng. Chem. Res. 2023, 62, 13009–13024

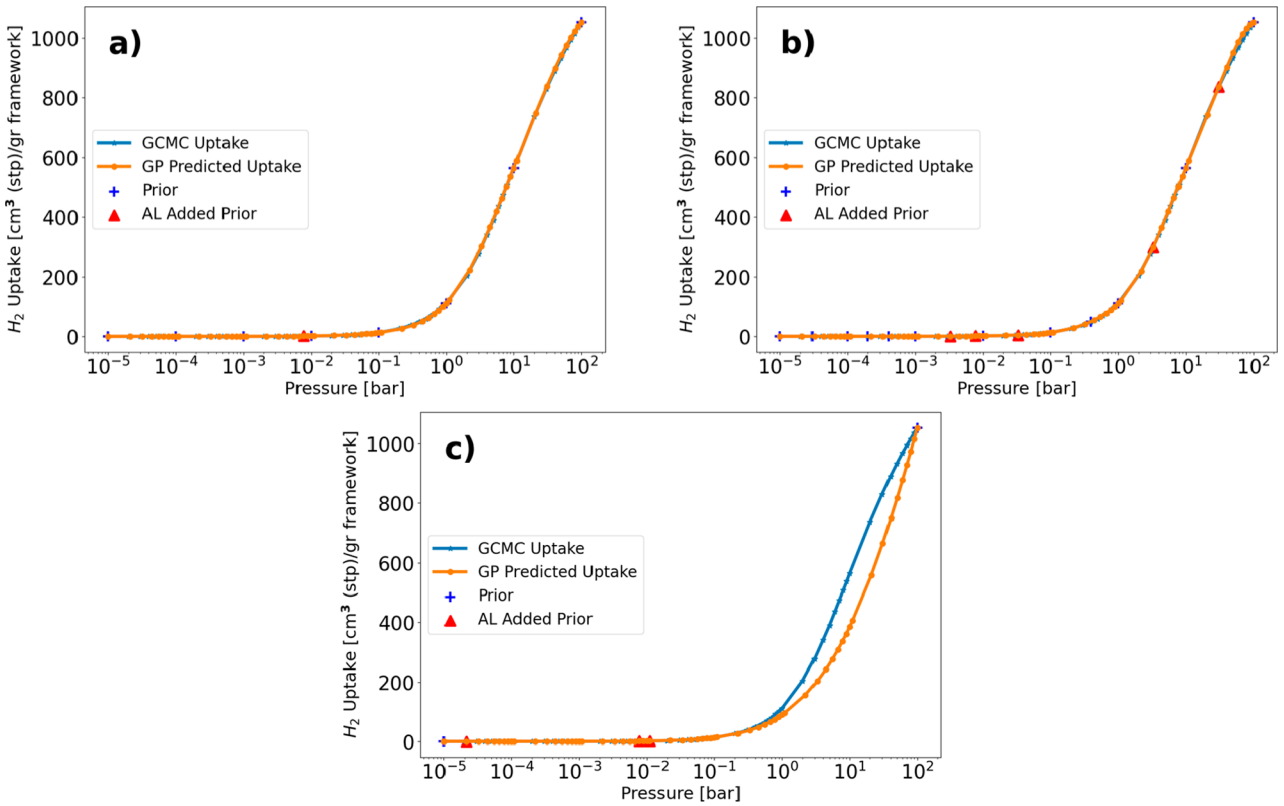


Figure 12. Hydrogen uptake comparison between GCMC simulation and GP in NU-800 using the RQ kernel for three different priors: (a) boundary-informed prior, (b) log-spaced prior, and (c) two-data prior.

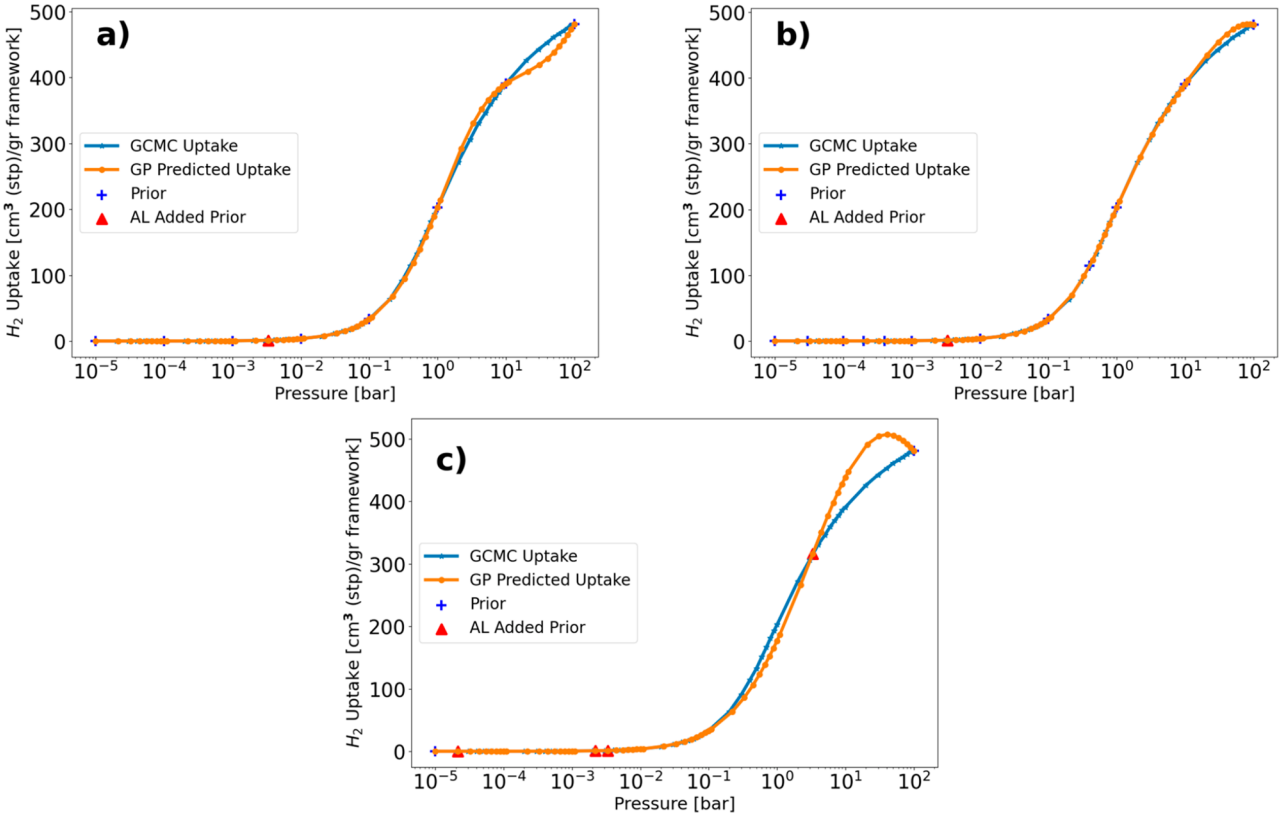


Figure 13. Hydrogen uptake comparison between GCMC simulation and GP in ZIF-8 using the RQ kernel, for three different priors: (a) boundary-informed prior, (b) log-spaced prior, and (c) two-data prior.

https://doi.org/10.1021/acs.iecr.3c01589
Ind. Eng. Chem. Res. 2023, 62, 13009−13024

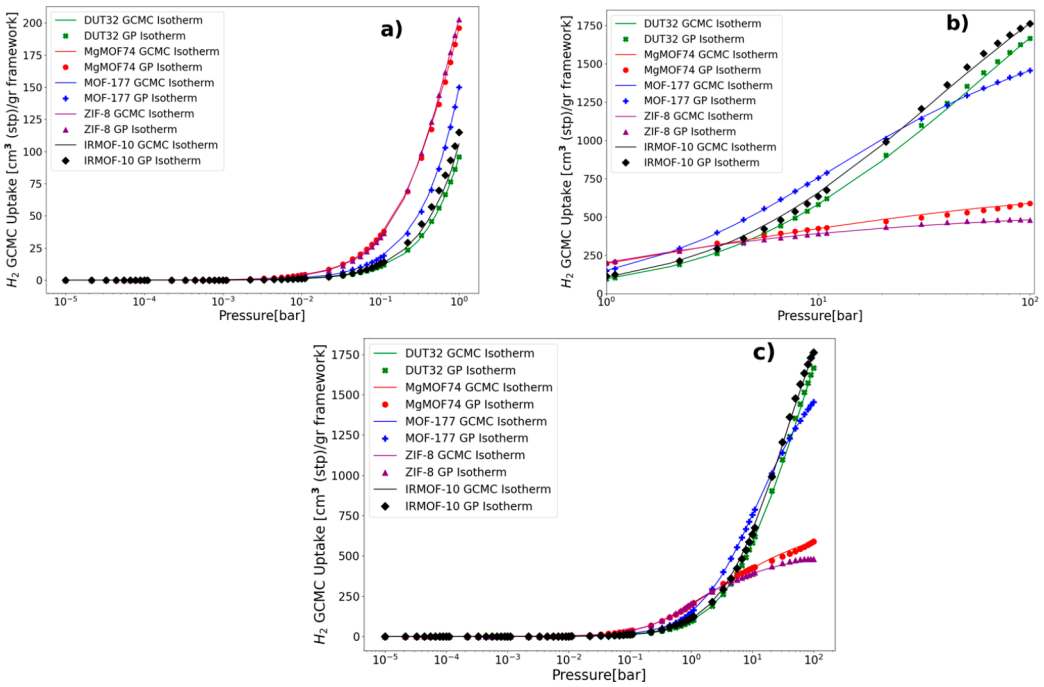


Figure 14. Comparison of the GP and GCMC hydrogen isotherms for some select MOFs. For visualization purposes, we show three different pressure ranges for the isotherms: (a) the low-pressure region, (b) the high-pressure regions, and (c) all the pressure ranges used for hydrogen.

Table 5. Hydrogen AL Summary Showing Errors, MAE, and Recommended Prior Selection Scheme for All 11 MOFs[a]

	Total Data Points for different priors			MAE [cm³/gr framework] for different priors		
MOF	BI	LS	TD	BI	LS	TD
PCN-61	9	**15**	5	9.34	**5.04**	17.84
MgMOF-74	**9**	13	5	**5.32**	5.6	25.2
DUT 32	**9**	14	5	**6.18**	7.02	15.22
DUT 49	9	14	7	10.02	6.62	**6.52**
MOF 177	**9**	14	11	**6.96**	7.32	7.64
NU-800	**9**	17	5	**5.58**	6.3	30.16
UiO-66	**9**	27	6	**2.04**	2.08	20.48
ZIF-8	9	**13**	6	5.54	**3.72**	11.58
IRMOF-1	9	**18**	9	11.78	**6.62**	21.64
IRMOF-10	9	15	**9**	7.52	6.22	**6.08**
IRMOF-16	9	13	**9**	16.46	14.18	**9.84**

[a]"LS" represents log-spaced, "BI" represents boundary-informed, and "TD" represents two-data priors. Values shown in bold font represent the recommended prior for the relevant MOF.

With AL, computational resources can be effectively saved as the GP model assists in identifying the specific pressure points necessary to carry out the GCMC simulation. AL enables the selection of informative data points that contribute the most to reducing uncertainty in the predictions. By intelligent selection of these pressure points, the need for extensive and computationally expensive GCMC simulations across the entire pressure range can be minimized. This targeted approach optimizes computational resources by focusing on the most informative regions of the pressure space, resulting in more efficient and accurate predictions of the adsorption behavior in MOFs.

Our results for nitrogen indicate that the AL schemes are effective and can be applied to determine the surface areas of various types of MOFs. It is important to note that only three prior selection schemes were used in this study; thus, further exploration of different schemes and AL kernels is recommended. Future research will also focus on other sets of MOFs, and isotherm types not examined in this study. However, from this study, we can infer that the boundary informed, or log space prior scheme, will be a good starting point for these cases.

■ ASSOCIATED CONTENT

Data Availability Statement

The code, algorithms, and adsorption data can be accessed at the following GitHub link: https://github.com/theOsaroJ/ActivelearningforMOFs. In addition to the MOF AL algorithm process, the repository also contains data from simulations that were conducted as well as the data generated by the AL algorithm for the recommended prior scheme.

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.iecr.3c01589.

Active learning results of other adsorbate−adsorbent pairs not reported in the manuscript. GP mean relative errors for all prior schemes generated for all adsorbates and adsorbents. Comparison between the GCMC and experimental isotherms for chosen structures (PDF)

■ AUTHOR INFORMATION

Corresponding Author

Yamil J. Colón − *Department of Chemical and Biomolecular Engineering, University of Notre Dame, Notre Dame, Indiana 46556, United States;* orcid.org/0000-0001-5316-9692; Email: ycolon@nd.edu

Authors

Etinosa Osaro − *Department of Chemical and Biomolecular Engineering, University of Notre Dame, Notre Dame,*

Indiana 46556, United States; orcid.org/0000-0003-2744-339X

Krishnendu Mukherjee − *Department of Chemical and Biomolecular Engineering, University of Notre Dame, Notre Dame, Indiana 46556, United States;* orcid.org/0000-0002-7727-1461

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.iecr.3c01589

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

Authors gratefully acknowledge NSF CAREER Award No. CBET-2143346. We also thank the Center for Research Computing at the University of Notre Dame for computational resources.

■ REFERENCES

(1) Baumann, A. E.; Burns, D. A.; Liu, B.; Thoi, V. S. Metal-Organic Framework Functionalization and Design Strategies for Advanced Electrochemical Energy Storage Devices. *Commun. Chem.* **2019**, *2* (1), 1−14.

(2) Langmi, H. W.; Ren, J.; North, B.; Mathe, M.; Bessarabov, D. Hydrogen Storage in Metal-Organic Frameworks: A Review. *Electrochim. Acta* **2014**, *128* (2013), 368−392.

(3) Mao, H.; Tang, J.; Day, G. S.; Peng, Y.; Wang, H.; Xiao, X.; Yang, Y.; Jiang, Y.; Chen, S.; Halat, D. M.; et al. A Scalable Solid-State Nanoporous Network with Atomic-Level Interaction Design for Carbon Dioxide Capture. *Sci. Adv.* **2022**, *8* (31), 6849.

(4) Hu, Z.; Wang, Y.; Shah, B. B.; Zhao, D. CO 2 Capture in Metal-Organic Framework Adsorbents: An Engineering Perspective. *Adv. Sustain. Syst.* **2019**, *3* (1), No. 1800080.

(5) Pascanu, V.; González Miera, G.; Inge, A. K.; Martín-Matute, B. Metal-Organic Frameworks as Catalysts for Organic Synthesis: A Critical Perspective. *J. Am. Chem. Soc.* **2019**, *141* (18), 7223−7234.

(6) Lin, R. B.; Xiang, S.; Zhou, W.; Chen, B. Microporous Metal-Organic Framework Materials for Gas Separation. *Chem.* **2020**, *6* (2), 337−363.

(7) Kreno, L. E.; Leong, K.; Farha, O. K.; Allendorf, M.; Van Duyne, R. P.; Hupp, J. T. Metal-Organic Framework Materials as Chemical Sensors. *Chem. Rev.* **2012**, *112* (2), 1105−1125.

(8) Maranescu, B.; Visa, A. Applications of Metal-Organic Frameworks as Drug Delivery Systems. *Int. J. Mol. Sci.* **2022**, *23* (8), 4458.

(9) Moosavi, S. M.; Nandy, A.; Jablonka, K. M.; Ongari, D.; Janet, J. P.; Boyd, P. G.; Lee, Y.; Smit, B.; Kulik, H. J. Understanding the Diversity of the Metal-Organic Framework Ecosystem. *Nat. Commun.* **2020**, *11* (1), 1−10.

(10) Li, H.; Eddaoudi, M.; O'Keeffe, M.; Yaghi, O. M. Design and Synthesis of an Exceptionally Stable and Highly porous metal-organic framework. *Nature* **1999**, *402* (6759), 276−279.

(11) Fujita, M.; Washizu, S.; Ogura, K.; Kwon, Y. J. Preparation, Clathration Ability, and Catalysis of a Two-Dimensional Square Network Material Composed of Cadmium(II) and 4, 4′-Bipyridine. *J. Am. Chem. Soc.* **1994**, *116* (3), 1151−1152.

(12) Tao, Y. R.; Zhang, G. H.; Xu, H. J. Grand Canonical Monte Carlo (GCMC) Study on Adsorption Performance of Metal Organic Frameworks (MOFs) for Carbon Capture. *Sustain. Mater. Technol.* **2022**, *32*, e00383.

(13) Ohba, T.; Inaguma, Y.; Kondo, A.; Kanoh, H.; Noguchi, H.; Gubbins, K. E.; Kajiro, H.; Kaneko, K. GCMC Simulations of Dynamic Structural Change of Cu−Organic Crystals with N2 Adsorption. *J. Exp. Nanosci.* **2006**, *1* (1), 91−95.

(14) Rogge, S. M. J.; Goeminne, R.; Demuynck, R.; Gutiérrez-Sevillano, J. J.; Vandenbrande, S.; Vanduyfhuys, L.; Waroquier, M.; Verstraelen, T.; Van Speybroeck, V. Modeling Gas Adsorption in Flexible Metal−Organic Frameworks via Hybrid Monte Carlo/Molecular Dynamics Schemes. *Adv. Theory Simulations* **2019**, *2* (4), 1800177−15.

(15) Peng, X.; Cheng, X.; Cao, D. Computer Simulations for the Adsorption and Separation of CO 2/CH4/H2/N2 Gases by UMCM-1 and UMCM-2 Metal Organic Frameworks. *J. Mater. Chem.* **2011**, *21* (30), 11259−11270.

(16) Erucar, I.; Keskin, S. Unlocking the Effect of H2O on CO2 Separation Performance of Promising MOFs Using Atomically Detailed Simulations. *Ind. Eng. Chem. Res.* **2020**, *59* (7), 3141−3152.

(17) Jain, A.; Hautier, G.; Ong, S. P.; Persson, K. New Opportunities for Materials Informatics: Resources and Data Mining Techniques for Uncovering Hidden Relationships. *J. Mater. Res.* **2016**, *31* (8), 977−994.

(18) Sumpter, B. G.; Noid, D. W. Neural Networks as Tools for Predicting Materials Properties. *Annu. Technol. Conf. - ANTEC, Conf. Proc.* **1995**, *2*, 2556−2560.

(19) Carr, D. A.; Lach-hab, M.; Yang, S.; Vaisman, I. I.; Blaisten-Barojas, E. Machine Learning Approach for Structure-Based Zeolite Classification. *Microporous Mesoporous Mater.* **2009**, *117* (1−2), 339−349.

(20) Schmidt, J.; Marques, M. R. G.; Botti, S.; Marques, M. A. L. Recent Advances and Applications of Machine Learning in Solid-State Materials Science. *npj Comput. Mater.* **2019**, *5* (1), DOI: 10.1038/s41524-019-0221-0.

(21) Aghaji, M. Z.; Fernandez, M.; Boyd, P. G.; Daff, T. D.; Woo, T. K. Quantitative Structure−Property Relationship Models for Recognizing Metal Organic Frameworks (MOFs) with High CO2 Working Capacity and CO2/CH4 Selectivity for Methane Purification. *Eur. J. Inorg. Chem.* **2016**, *2016* (27), 4505−4511.

(22) Shi, Z.; Yang, W.; Deng, X.; Cai, C.; Yan, Y.; Liang, H.; Liu, Z.; Qiao, Z. Machine-Learning-Assisted High-Throughput Computational Screening of High Performance Metal-Organic Frameworks. *Mol. Syst. Des. Eng.* **2020**, *5* (4), 725−742.

(23) Ahmed, A.; Siegel, D. J. Predicting Hydrogen Storage in MOFs via Machine Learning. *Patterns* **2021**, *2* (7), No. 100291.

(24) Altintas, C.; Altundal, O. F.; Keskin, S.; Yildirim, R. Machine Learning Meets with Metal Organic Frameworks for Gas Storage and Separation. *J. Chem. Inf. Model.* **2021**, *61* (5), 2131−2146.

(25) Bai, X.; Shi, Z.; Xia, H.; Li, S.; Liu, Z.; Liang, H.; Liu, Z.; Wang, B.; Qiao, Z. Machine-Learning-Assisted High-Throughput Computational Screening of Metal−Organic Framework Membranes for Hydrogen Separation. *Chem. Eng. J.* **2022**, *446* (P2), No. 136783.

(26) Li, H.; Wang, C.; Zeng, Y.; Li, D.; Yan, Y.; Zhu, X.; Qiao, Z. Combining Computational Screening and Machine Learning to Predict Metal−Organic Framework Adsorbents and Membranes for Removing CH4 or H2 from Air. *Membranes (Basel)* **2022**, *12* (9), 830.

(27) Wu, X.; Xiang, S.; Su, J.; Cai, W. Understanding Quantitative Relationship between Methane Storage Capacities and Characteristic Properties of Metal-Organic Frameworks Based on Machine Learning. *J. Phys. Chem. C* **2019**, *123* (14), 8550−8559.

(28) Fernandez, M.; Boyd, P. G.; Daff, T. D.; Aghaji, M. Z.; Woo, T. K. Rapid and Accurate Machine Learning Recognition of High Performing Metal Organic Frameworks for CO2 Capture. *J. Phys. Chem. Lett.* **2014**, *5* (17), 3056−3060.

(29) Anderson, R.; Biong, A.; Gómez-Gualdrón, D. A. Adsorption Isotherm Predictions for Multiple Molecules in MOFs Using the Same Deep Learning Model. *J. Chem. Theory Comput.* **2020**, *16* (2), 1271−1283.

(30) Santos, J. E.; Mehana, M.; Wu, H.; Prodanović, M.; Kang, Q.; Lubbers, N.; Viswanathan, H.; Pyrcz, M. J. Modeling Nanoconfinement Effects Using Active Learning. *J. Phys. Chem. C* **2020**, *124* (40), 22200−22211.

(31) Mukherjee, K.; Dowling, A. W.; Colón, Y. J. Sequential Design of Adsorption Simulations in Metal Organic Frameworks. *Mol. Syst. Des. Eng.* **2022**, *7* (3), 248−259.

(32) Streppel, B.; Hirscher, M. BET Specific Surface Area and Pore Structure of MOFs Determined by Hydrogen Adsorption at 20 K. *Phys. Chem. Chem. Phys.* **2011**, *13* (8), 3220−3222.
(33) Farha, O. K.; Eryazici, I.; Jeong, N. C.; Hauser, B. G.; Wilmer, C. E.; Sarjeant, A. A.; Snurr, R. Q.; Nguyen, S. T.; Yazaydin, A. Ö.; Hupp, J. T. Metal-Organic Framework Materials with Ultrahigh Surface Areas: Is the Sky the Limit? *J. Am. Chem. Soc.* **2012**, *134* (36), 15016−15021.
(34) Furukawa, H.; Ko, N.; Go, Y. B.; Aratani, N.; Choi, S. B.; Choi, E.; Yazaydin, A. Ö.; Snurr, R. Q.; O'Keeffe, M.; Kim, J.; Yaghi, O. M. Ultrahigh Porosity in Metal-Organic Frameworks. *Science (80-.)* **2010**, *329* (5990), 424−428.
(35) Alexandrov, E. V.; Shevchenko, A. P.; Blatov, V. A. Topological Databases: Why Do We Need Them for Design of Coordination Polymers? *Cryst. Growth Des.* **2019**, *19* (5), 2604−2614.
(36) Alexandrov, E. V.; Blatov, V. A.; Kochetkov, A. V.; Proserpio, D. M. Underlying Nets in Three-Periodic Coordination Polymers: Topology, Taxonomy and Prediction from a Computer-Aided Analysis of the Cambridge Structural Database. *CrystEngComm* **2011**, *13* (12), 3947−3958.
(37) O'Keeffe, M.; Peskov, M. A.; Ramsden, S. J.; Yaghi, O. M. (RCSR) Database of, and Symbols for, Crystal. *Acc. Chem. Res.* **2008**, *41* (12), 1782−1789.
(38) Blatov, V. A.; Shevchenko, A. P.; Proserpio, D. M. Applied Topological Analysis of Crystal Structures with the Program Package Topospro. *Cryst. Growth Des.* **2014**, *14* (7), 3576−3586.
(39) Gómez-Gualdrón, D. A.; Moghadam, P. Z.; Hupp, J. T.; Farha, O. K.; Snurr, R. Q. Application of Consistency Criteria to Calculate BET Areas of Micro- and Mesoporous Metal-Organic Frameworks. *J. Am. Chem. Soc.* **2016**, *138* (1), 215−224.
(40) Yuan, S.; Zou, L.; Qin, J. S.; Li, J.; Huang, L.; Feng, L.; Wang, X.; Bosch, M.; Alsalme, A.; Cagin, T.; Zhou, H. C. Construction of Hierarchically Porous Metal-Organic Frameworks through Linker Labilization. *Nat. Commun.* **2017**, *8* (May), 1−10.
(41) Trepte, K.; Schwalbe, S. Systematic Analysis of Porosities in Metal-Organic Framework. *ChemRxiv* **2019**, 1−9.
(42) Feng, L.; Wang, K. Y.; Lv, X. L.; Yan, T. H.; Zhou, H. C. Hierarchically Porous Metal-Organic Frameworks: Synthetic Strategies and Applications. *Natl. Sci. Rev.* **2020**, *7* (11), 1743−1758.
(43) Dubbeldam, D.; Calero, S.; Ellis, D. E.; Snurr, R. Q. RASPA: Molecular Simulation Software for Adsorption and Diffusion in Flexible Nanoporous Materials. *Mol. Simul.* **2016**, *42* (2), 81−101.
(44) Potoff, J. J.; Siepmann, J. I. Vapor-Liquid Equilibria of Mixtures Containing Alkanes, Carbon Dioxide, and Nitrogen. *AIChE J.* **2001**, *47* (7), 1676−1682.
(45) Ahmed, A.; Liu, Y.; Purewal, J.; Tran, L. D.; Wong-Foy, A. G.; Veenstra, M.; Matzger, A. J.; Siegel, D. J. Balancing Gravimetric and Volumetric Hydrogen Density in MOFs. *Energy Environ. Sci.* **2017**, *10* (11), 2459−2471.
(46) Fischer, M.; Hoffmann, F.; Fröba, M. Preferred Hydrogen Adsorption Sites in Various MOFs-A Comparative Computational Study. *ChemPhysChem* **2009**, *10* (15), 2647−2657.
(47) Basdogan, Y.; Keskin, S. Simulation and Modelling of MOFs for Hydrogen Storage. *CrystEngComm* **2015**, *17* (2), 261−275.
(48) Bureekaew, S.; Amirjalayer, S.; Tafipolsky, M.; Spickermann, C.; Roy, T. K.; Schmid, R. MOF-FF - A Flexible First-Principles Derived Force Field for Metal-Organic Frameworks. *Phys. Status Solidi Basic Res.* **2013**, *250* (6), 1128−1141.
(49) Wahiduzzaman, M.; Walther, C. F. J.; Heine, T. Hydrogen Adsorption in Metal-Organic Frameworks: The Role of Nuclear Quantum Effects. *J. Chem. Phys.* **2014**, *141* (6), DOI: 10.1063/1.4892670.
(50) Liu, J.; Johnson, K.; Culp, J.; Natesakhawat, S.; Bockrath, B.; Sankar, S. G.; Zande, B.; Garberoglio, G. Experimental and Theoretical Studies of Gas Adsorption in Cu3(Btc)2 : An Effective Activation. *J. Phyc. Chem. C* **2007**, *111* (26), 9305−9313.
(51) Fairen-Jimenez, D.; Colón, Y. J.; Farha, O. K.; Bae, Y. S.; Hupp, J. T.; Snurr, R. Q. Understanding Excess Uptake Maxima for Hydrogen Adsorption Isotherms in Frameworks with Rht Topology. *Chem. Commun.* **2012**, *48* (85), 10496−10498.
(52) Hu, N.; Sun, X.; Hsu, A. Monte Carlo Simulations of Hydrogen Adsorption in Alkali-Doped Single-Walled Carbon Nanotubes. *J. Chem. Phys.* **2005**, *123* (4), DOI: 10.1063/1.1954727.
(53) Michels, A.; de Graaff, W.; Ten Seldam, C. A. Virial Coefficients of Hydrogen and Deuterium at Temperatures between −175°C and + 150°C. Conclusions from the Second Virial Coefficient with Regards to the Intermolecular Potential. *Physica* **1960**, *26* (6), 393−408.
(54) Lorentz, H. A. Ueber Die Anwendung Des Satzes Vom Virial in Der Kinetischen Theorie Der Gase. *Ann. Phys.* **1881**, *248* (1), 127−136.
(55) Wang, X.-S.; Ma, S.; Sun, D.; Parkin, S.; Zhou, H. C. A Mesoporous Metal-Organic Framework with Permanent Porosity. *J. Am. Chem. Soc.* **2006**, *128* (51), 16474−16475.
(56) Li, B.; Wen, H. M.; Zhou, W.; Xu, J. Q.; Chen, B. Porous Metal-Organic Frameworks: Promising Materials for Methane Storage. *Chem.* **2016**, *1* (4), 557−580.
(57) Cavka, J. H.; Jakobsen, S.; Olsbye, U.; Guillou, N.; Lamberti, C.; Bordiga, S.; Lillerud, K. P. A New Zirconium Inorganic Building Brick Forming Metal Organic Frameworks with Exceptional Stability. *J. Am. Chem. Soc.* **2008**, *130* (42), 13850−13851.
(58) Bergaoui, M.; Khalfaoui, M.; Awadallah-F, A.; Al-Muhtaseb, S. A Review of the Features and Applications of ZIF-8 and Its Derivatives for Separating CO2 and Isomers of C3- and C4-Hydrocarbons. *J. Nat. Gas Sci. Eng.* **2021**, *96* (June), No. 104289.
(59) Grünker, R.; Bon, V.; Müller, P.; Stoeck, U.; Krause, S.; Mueller, U.; Senkovska, I.; Kaskel, S. A New Metal-Organic Framework with Ultra-High Surface Area. *Chem. Commun.* **2014**, *50* (26), 3450−3452.
(60) Stoeck, U.; Krause, S.; Bon, V.; Senkovska, I.; Kaskel, S. A Highly Porous Metal−Organic Framework, Constructed from a Cuboctahedral Super-Molecular Building Block, with Exceptionally High Methane Uptake. *Chem. Commun.* **2012**, *48* (88), 10841−10843.
(61) Alezi, D.; Belmabkhout, Y.; Suyetin, M.; Bhatt, P. M.; Weseliński, L. J.; Solovyeva, V.; Adil, K.; Spanopoulos, I.; Trikalitis, P. N.; Emwas, A. H.; Eddaoudi, M. MOF Crystal Chemistry Paving the Way to Gas Storage Needs: Aluminum-Based Soc -MOF for CH4, O2, and CO2 Storage. *J. Am. Chem. Soc.* **2015**, *137* (41), 13308−13318.
(62) Zhao, D.; Yuan, D.; Sun, D.; Zhou, H. C. Stabilization of Metal-Organic Frameworks with High Surface Areas by the Incorporation of Mesocavities with Microwindows. *J. Am. Chem. Soc.* **2009**, *131* (26), 9186−9188.
(63) Su, X.; Bromberg, L.; Martis, V.; Simeon, F.; Huq, A.; Hatton, T. A. Postsynthetic Functionalization of Mg-MOF-74 with Tetraethylenepentamine: Structural Characterization and Enhanced CO2 Adsorption. *ACS Appl. Mater. Interfaces* **2017**, *9* (12), 11299−11306.
(64) Rasmussen, C. E. Gaussian Processes in Machine Learning. *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)* **2004**, *3176*, 63−71.
(65) Gheytanzadeh, M.; Baghban, A.; Habibzadeh, S.; Esmaeili, A.; Abida, O.; Mohaddespour, A.; Munir, M. T. Towards Estimation of CO2 Adsorption on Highly Porous MOF-Based Adsorbents Using Gaussian Process Regression Approach. *Sci. Rep.* **2021**, *11* (1), 1−13.
(66) Deringer, V. L.; Bartók, A. P.; Bernstein, N.; Wilkins, D. M.; Ceriotti, M.; Csányi, G. Gaussian Process Regression for Materials and Molecules. *Chem. Rev.* **2021**, *121* (16), 10073−10141.
(67) Dudek, A.; Baranowski, J. Gaussian Processes for Signal Processing and Representation in Control Engineering. *Appl. Sci.* **2022**, *12* (10), 4946.
(68) Wilson, A. G.; Adams, R. P. Gaussian Process Kernels for Pattern Discovery and Extrapolation. *30th Int. Conf. Mach. Learn. ICML 2013* **2013**, *28* (PART3), 2104−2112.

(69) Pilario, K. E.; Shafiee, M.; Cao, Y.; Lao, L.; Yang, S. H. A Review of Kernel Methods for Feature Extraction in Nonlinear Process Monitoring. *Processes* **2020**, *8* (1), 24−47.
(70) Khan, S.; Naseem, I.; Togneri, R.; Bennamoun, M. A Novel Adaptive Kernel for the RBF Neural Networks. *Circuits, Syst. Signal Process* **2017**, *36* (4), 1639−1653.
(71) Melkumyan, A.; Ramos, F. Multi-Kernel Gaussian Processes. *IJCAI Int. Jt. Conf. Artif. Intell.* **2011**, 1408−1413.
(72) Deshwal, A.; Doppa, J. R. Combining Latent Space and Structured Kernels for Bayesian Optimization over Combinatorial Spaces. *Adv. Neural Inf. Process. Syst.* **2021**, *10* (NeurIPS), 8185−8200.
(73) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; Duchesnay, E. Scikit-Learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12* (January), 2825−2830.
(74) van der Walt, S.; Colbert, S C.; Varoquaux, G. The NumPy Array: A Structure for Efficient Numerical Computation, Computing in Science & Engineering. *Comput. Sci. Eng.* **2011**, *13* (2), 22−30.
(75) Harris, C. R.; Millman, K. J.; van der Walt, S. J.; Gommers, R.; Virtanen, P.; Cournapeau, D.; Wieser, E.; Taylor, J.; Berg, S.; Smith, N. J.; Kern, R.; Picus, M.; Hoyer, S.; van Kerkwijk, M. H.; Brett, M.; Haldane, A.; del Río, J. F.; Wiebe, M.; Peterson, P.; Gérard-Marchant, P.; Sheppard, K.; Reddy, T.; Weckesser, W.; Abbasi, H.; Gohlke, C.; Oliphant, T. E. Array Programming with NumPy. *Nature* **2020**, *585* (7825), 357−362.
(76) Walton, K. S.; Snurr, R. Q. Applicability of the BET Method for Determining Surface Areas of Microporous Metal-Organic Frameworks. *J. Am. Chem. Soc.* **2007**, *129* (27), 8552−8556.
(77) Rodríguez-Reinoso, F. Characterization of Porous Solids VI: Preface. *Stud. Surf. Sci. Catal.* **2002**, *144*, xv.
(78) Osterrieth, J. W. M.; Rampersad, J.; Madden, D.; Rampal, N.; Skoric, L.; Connolly, B.; Allendorf, M. D.; Stavila, V.; Snider, J. L.; Ameloot, R.; et al. How Reproducible Are Surface Areas Calculated from the BET Equation? *Adv. Mater.* **2022**, *34*, No. 2201502.

https://doi.org/10.1021/acs.iecr.3c01589
Ind. Eng. Chem. Res. 2023, 62, 13009−13024