Active learning of alchemical adsorption simulations; towards a universal adsorption model.

Etinosa Osaro[a], Fernando Fajardo-Rojas[b], Gregory M. Cooper[a], Diego Gómez-Gualdrón[b], and Yamil J. Colón*[a].

a. Department of Chemical and Biomolecular Engineering, University of Notre Dame, IN 46556, USA
b. Department of Chemical and Biological Engineering, Colorado School of Mines, 1500 Illinois St, Golden, CO 80401, USA

*Corresponding author: Yamil J. Colón (ycolon@nd.edu)

Abstract

Adsorption is a fundamental process studied in materials science and engineering because it plays a critical role in various applications, including gas storage and separation. Understanding and predicting gas adsorption within porous materials demands comprehensive computational simulations that are often resource intensive, limiting the identification of promising materials. Active learning (AL) methods offer an effective strategy to reduce the computational burden by selectively acquiring critical data for model training. Metal-organic frameworks (MOFs) exhibit immense potential across various adsorption applications due to their porous structure and their modular nature, leading to diverse pore sizes and chemistry that serve as an ideal platform to develop adsorption models. Here, we demonstrate the efficacy of AL in predicting gas adsorption within MOFs using “alchemical” molecules and their interactions as surrogates for real molecules. We first applied AL separately to each MOF, reducing the training dataset size by 57.5% while retaining predictive accuracy. Subsequently, we amalgamated the refined datasets across 1800 MOFs to train a multilayer perceptron (MLP) model, successfully predicting adsorption of real molecules. Furthermore, by integrating MOF features into the AL framework using principal component analysis (PCA), we navigated MOF space effectively, achieving high predictive accuracy with only a subset of MOFs. Our results highlight AL's efficiency in reducing dataset size, enhancing model performance, and offering insights into adsorption phenomenon in large datasets of MOFs. This study underscores AL's crucial role in advancing computational material science and developing more accurate and less data intensive models for gas adsorption in porous materials.

Main

Metal-organic frameworks (MOFs) stand as versatile porous materials with exquisitely tunable structures, and tremendous potential for numerous applications across various fields.[1–6] A large fraction of these applications seek to exploit the adsorption properties of these materials, which are composed of interconnected building blocks (i.e., metallic nodes and organic linkers).[7–10] For instance, numerous MOFs could be imparted with adsorption properties to substitute ca. 80% of heat-based chemical separations processes with adsorption-based ones.[11] Therefore, harnessing the full potential of MOFs, and accelerating their development by anticipating MOF designs that embody desired properties through computation requires reliable predictions of adsorption behavior within these intricate frameworks.

The vast MOF space, spanning countless unique structures formed from different combinations of constitutive building blocks, poses an immense challenge to predicting gas adsorption behaviors across this expansive space of materials sufficiently fast. Depending on the complexity of the molecule model, predicting the adsorption loading of a molecule within a MOF through classical techniques such as grand canonical Monte Carlo (GCMC) simulations[12–16] may require substantial and specialized computational resources. Each GCMC simulation involves comprehensive exploration of the configurational space of gas molecules within MOF pores, calculating interaction energies, and sampling numerous adsorption states. Regardless of the complexity of the molecule model, the computational expense escalates significantly as the number of MOFs, adsorbate molecules, and adsorption conditions under consideration increase.

Machine learning (ML) seems poised to be an important tool to predict adsorption in MOFs.[17–22] However, developing ML that can comprehensively navigate the immense space formed by different MOF and molecule pairings demands a high volume of training data to achieve reliable predictions. Acquiring such large datasets can be an arduous, time-consuming, and computationally expensive task. Several ML adsorption models documented in the literature demand an extensive dataset ranging from thousands to millions of data points. Our solution, active learning (AL), circumvents this necessity.

To circumvent the above issue, AL could be used as a strategic approach to optimize the data acquisition process. AL, a subfield of ML, reduces the data burden to train a model through an iterative effort that guides the collection of training data only towards the most informative data points, while simultaneously using these data points to train a surrogate model to predict the quantity of interest and the uncertainty associated with the prediction.[23–27] In this work, we will add data to the training using the points for which the prediction uncertainty is highest.

We envision AL to play a crucial role in the development of ML adsorption models by guiding the selection of adsorption scenarios that offer surrogate models the most information about adsorption behavior, thereby reducing the computational expense associated with conducting GCMC simulations to generate training data. We select specific combinations of MOFs, molecules, and conditions that contribute to the surrogate model's predictions of adsorption.

AL has been demonstrated to reduce the data burden to train models that predict adsorption of specific molecules. For instance, in a previous study, Osaro and coworkers[28] demonstrated the development of a model a few predict full pure gas isotherms for methane, nitrogen, hydrogen and carbon dioxide using few training datapoints across eleven MOFs. In another instance, Mukherjee and coworkers[29] used AL to develop a model to predict full isotherms for methane and carbon dioxide in HKUST-1 at different temperatures. AL has additionally been employed to train a model capable of predicting adsorption behaviors for various gas pairs, including xenon-krypton, carbon dioxide-methane, and hydrogen sulfide-carbon dioxide, within a single MOF.[30] However, the use of AL towards the development of universal adsorption models, and the extent to which it could reduce the data burden in such context, has not been explored.

Training of a universal adsorption model implies presenting the model with adsorption data for different molecules, along with some representation of said molecules. As molecules can be modeled using some combinations of values for parameters in intermolecular (e.g., Lennard Jones and Coulombic parameters) and intramolecular potentials, Gómez-Gualdrón and coworkers showed that to "teach" a model about adsorption, one does not need to limit the adsorption data to real molecules. Specifically, they created 200 alchemical molecules using arbitrary combinations of potential parameters,[31] obtained adsorption data for them using molecular simulation and used the data to train a multi-layer perceptron (MLP) model capable of predicting full isotherms for real, small, non-polar, near-spherical, rigid molecules.

The above MLP demonstrates the concept of a ML-based universal adsorption model to be sound. Yet, the feasibility of a truly universal model is contingent on the ability to incorporate sufficient adsorption data for molecules (real or alchemical) with more diverse sizes, shapes, polarity, and flexibility. However, the above MLP required approximately 5 million GCMC data points, encompassing adsorption data for 200 small, non-polar, near-spherical, rigid alchemical adsorbates, on a relatively small set of 1,800 topologically and chemically diverse ToBaCCo[32]-generated MOFs, at fugacities of 0.01, 0.025, 0.05, 0.075, 0.1, 0.25, 0.5, 0.75, 1, 5, 10, 50, 75, and 100 bar[33]. Overall, each MOF required about 2,800 adsorption data points for training.

In the work above, alchemical adsorbates were adequately represented by four features, which were explored in a somewhat exhaustive fashion. Expansion of the training data to include molecules with more diverse sizes, shapes, polarity, and flexibility would require representing the adsorbates with more features, increasing the dimensionality of the adsorbate space, whose exhaustive exploration would imply an intractable number of GCMC simulations. Thus, a critical bottleneck that needs to be overcome to truly open the path towards a universal adsorption model is to gain the ability to efficiently explore the adsorbate (plus adsorbent) space.

Crucially, in this work we demonstrate for the first time the ability of AL to cut down the size of training datasets that includes adsorption data of multiple molecules, using the adsorbate space explored earlier by Gómez-Gualdrón and coworkers as a testbed. We first approached this task on a per MOF strategy by using AL to generate the training data for each MOF (1800 MOFs in total), which resulted in 57.5% data savings. The resulting surrogate models from AL per MOF are used to generate training data for a new MLP model, which was shown to retain the original predictive performance of the original MLP by Gómez-Gualdrón and coworkers. Encouraged by the results, we then approach this AL task on a joint MOF-adsorbate basis (alchemical adsorbates and 3445 MOFs). Excitingly, this approach results in drastic data savings of 99.8%. Lastly, we analyze the AL process, focusing on its selected features as the model is developed, providing insights into AL campaigns for adsorption.

Results

AL on alchemical molecules

The MLP previously developed by Gómez-Gualdrón and coworkers[33] used approximately 5 million training data points derived from GCMC simulations involving 200 alchemical adsorbates across 1800 MOFs. It established strong correlations between the predictions generated by the MLP model and the adsorption results obtained from GCMC simulations and will be used as a surrogate for GCMC in this study section. In this section, we demonstrate the efficacy of AL in developing a similarly predictive model, while reducing the training dataset on a per MOF basis. We constructed a new MLP model capable of predicting the adsorption behavior of real molecules using training data originated from the surrogate GP models developed by AL for each MOF. Namely, with our first AL approach, we executed the AL process in each MOF separately, and subsequently amalgamated the training datapoints selected by AL for each of all 1800 MOFs into a unified dataset. The latter was then utilized to train a new MLP model, which was tested to predict adsorption of real molecules.

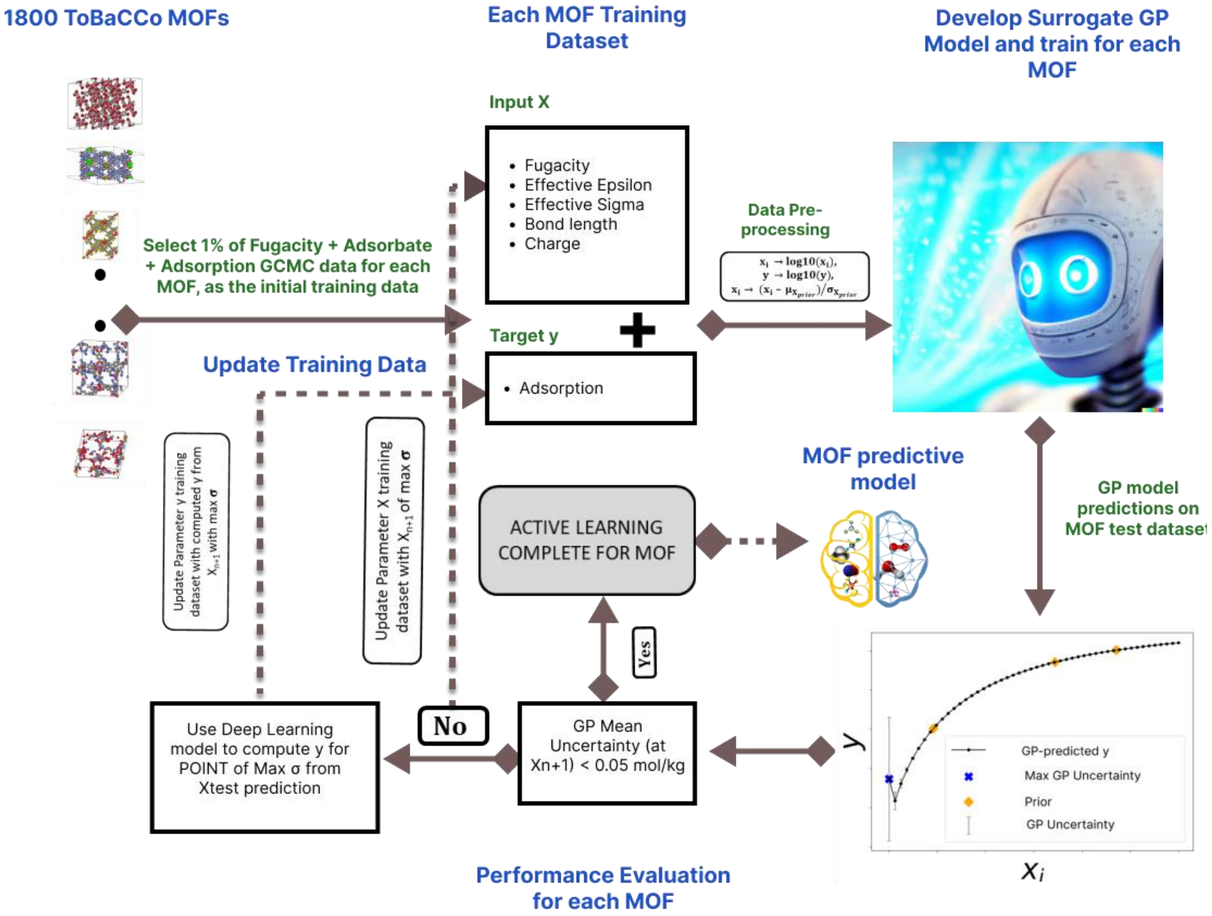
The AL algorithm for the above approach is illustrated in **Figure 1**. The adsorbate features that the Gaussian Process (GP) model (F) developed for each MOF uses to make adsorption predictions (adsorption loading N) are adsorbate surrogate Lennard-Jones parameters (effective epsilon (ε) and effective sigma (σ)), bond length (l), and charge (q). The adsorbate (which is defined by the combination of values of the aforementioned parameters) and fugacity (f) combinations to be iteratively added to the training data are automatically selected by the AL. To commence the AL algorithm for each MOF, we curated an initial set of twenty-six training data points encompassing fugacities, diverse alchemical adsorbates, and their respective adsorption values. These data points were chosen to represent a broad range derived from the training dataset. A sample of the initial training data for a single MOF is available in the project's GitHub repository.

The logarithms of the adsorbate features were used as input to the GP model, except in the case of charge and bond-length. All features underwent z-score standardization before being inputted into the GP model, which utilized a rational quadratic (RQ) kernel to perform the regression. At each AL iteration, the GP model trained (using the initial training dataset) for each MOF was used to compute the GP mean uncertainty for each prediction, which is a direct output of the GP model.

Following training, the model was utilized to predict adsorption based on randomly chosen values of alchemical parameters and fugacities, referred to as testing data, as detailed in **Table S1** of the supporting information (SI). Importantly, all features in the testing dataset fell within the bounds of the parameters of the training data.

The point in the testing dataset with the maximum GP uncertainty was identified and fed into the Gómez-Gualdrón and coworkers MLP model to compute the considered ground truth adsorption value, as it earlier proved to have accurate correlations with adsorption from GCMC simulations[33]. Adsorbate and fugacity combinations continued to be added iteratively to the training data until the mean predicted uncertainty of the GP was under 0.05 mol/kg. Once the threshold was met, the final GP model was utilized to predict the adsorption in the testing dataset. For each MOF, the entire testing dataset is inputted into the MLP model to generate the MLP ground truth adsorption values for comparisons with the GP predicted adsorption. We can use the MLP model to generate the ground truth data because of the high accuracy of the model when compared to GCMC simulations.[33]

Figure 1. AL framework on alchemical molecules.


1800 ToBaCCo MOFs
Each MOF Training Dataset
Develop Surrogate GP Model and train for each MOF
Input X
Fugacity
Effective Epsilon
Effective Sigma
Bond length
Charge
Select 1% of Fugacity + Adsorbate + Adsorption GCMC data for each MOF, as the initial training data
Data Pre-processing
$x_i \rightarrow \log10(x_i)$,
$y \rightarrow \log10(y)$,
$x_i \rightarrow (x_i - \mu_{x_{prior}})/\sigma_{x_{prior}}$
Target y
Adsorption
Update Training Data
Update Parameter y training dataset with computed y from X_{n+1} with max σ
Update Parameter X training dataset with X_{n+1} of max σ
MOF predictive model
GP model predictions on MOF test dataset
ACTIVE LEARNING COMPLETE FOR MOF
Yes
No
Use Deep Learning model to compute y for POINT of Max σ from Xtest prediction
GP Mean Uncertainty (at Xn+1) < 0.05 mol/kg
y
GP-predicted y
Max GP Uncertainty
Prior
GP Uncertainty
X_i
Performance Evaluation for each MOF

The GPR model with the input features of fugacity, effective epsilon, effective sigma, bond length, and charge is used to predict the target variable of adsorption loading, after the model has been trained with some initial training data. The test data set (the test input features) with the highest predicted GP uncertainty is fed into the MLP model and the ground-truth is computed. This point is added to the initial training data and the model is retrained and the predictions on the test dataset is done again. This is done iteratively until the GP mean uncertainty is less than 0.05 mol/kg. This procedure is conducted independently in each of the 1800 MOFs. Prior as used in this figure is the initial training data.

We assessed the GP model performance by calculating the R^2 between the "MLP ground truth" and the GP model predictions. **Figure 2** presents the evolution of R^2 and GP model uncertainty for the two most extreme cases within the 1800 MOFs. The one that required the most AL iterations to reach the 0.05 mol/kg (**Figure 2a**), and the one that required the fewest iterations (**Figure 2b**).

The MOF with the highest number of AL iterations of 1882, initially had a GP mean uncertainty of 0.142 mol/kg and an R^2 value of 0.68 in the first iteration. Over subsequent iterations, it achieved a final GP mean uncertainty of 0.049 mol/kg, accompanied by an R^2 value of 0.99. Conversely, the MOF with the fewest AL iterations, only 106, started with a higher initial GP mean uncertainty of 0.496 mol/kg and a lower R^2 value of 0.46. However, it also reached a final R^2 value of 0.99 at a GP mean uncertainty of 0.049 mol/kg. Notably, fluctuations in the GP mean uncertainty (represented by the blue line) closely corresponded to fluctuations in the R^2 value (represented by the orange line). These fluctuations highlight the correlation between GP mean uncertainty and R^2 values, emphasizing the impact of iterative data inclusion on model performance.

The two cases above show that the GP model can predict the adsorption of adsorbates with the prescribed ending threshold regardless of the starting quality of the GP model. Albeit the efficacy (i.e., number of iterations) with which AL achieves the desired goal clearly differs across MOFs. In **Figure S1**, we show how the textural properties of MOFs (largest pore diameter (LPD), pore limiting diameter (PLD), void fraction (VF), surface area (SA), pore size standard deviation (PSSD) and the inverse framework density (IFD)) influence the number of AL iterations where MOFs with LPD values < 30 Å, PLD values < 20 Å, void fractions > 0.6, PSSD values < 7.5, and IFD values < 4 demanded more AL iterations.

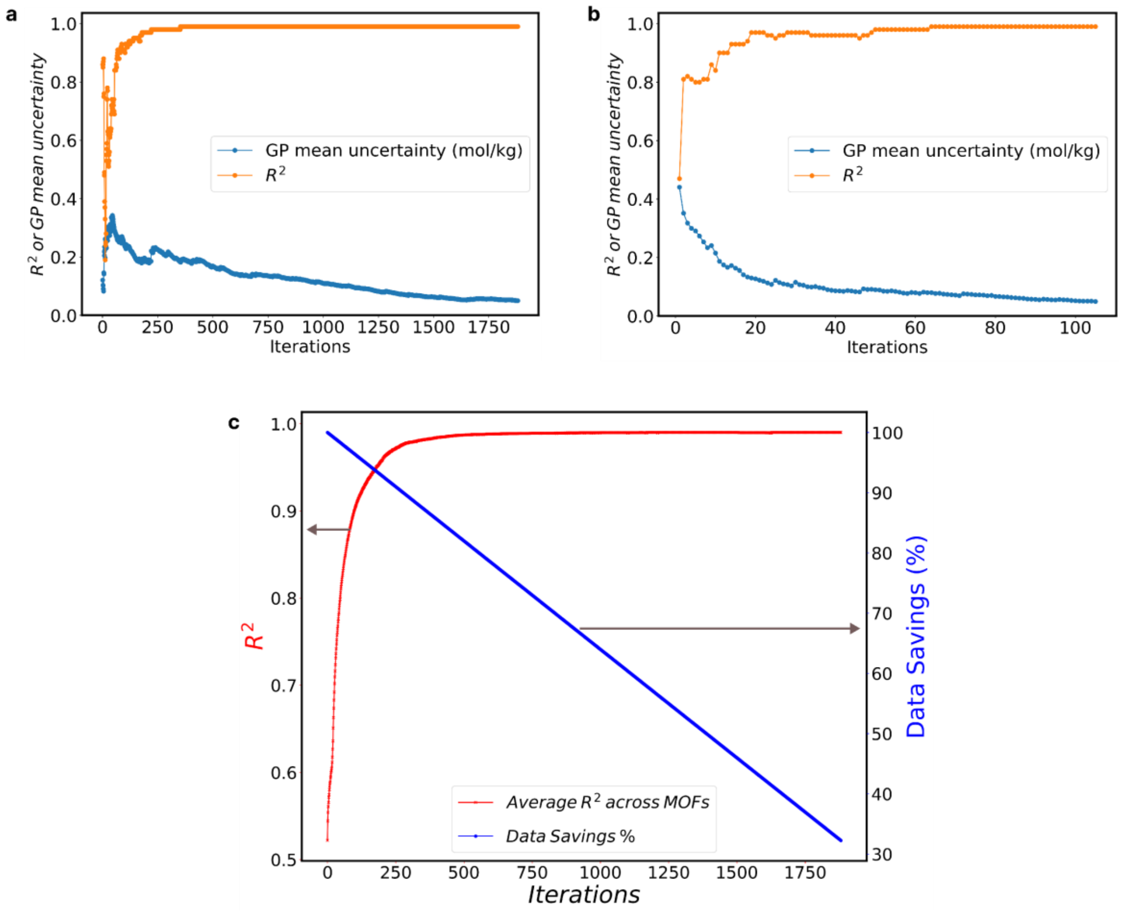
The percentage of data savings, as a function of AL iterations can be calculated by **Eq. 1:**

$$\%\, Data\ savings = 100 - 100 \times \left(\frac{(\#\ of\ AL\ iterations\ \times 1800) + \#\ of\ initial\ training\ data\ across\ all\ MOFs}{Total\ \#\ of\ training\ data\ points (original\ GCMC\ data)} \right) \quad \text{(Eq. 1).}$$

Based on the final GP models for all the MOFs, we achieved a data savings of 57.5% compared to the original MLP training data.

Figure 2c illustrates the collective impact of AL on enhancing GP predictions across all 1800 MOFs, depicting the average R^2 value (red line) at each iteration. In tandem with the improvement in R^2, we present the corresponding percentage of data savings (blue line), as calculated by Eq. 1. Notably, as AL iterations progress, we observe a consistent rise in the average R^2 values, indicative of the AL criterion's efficacy. Around the 500th AL iteration and beyond, the average R^2 across all GP models reaches 0.99, regardless of the GP mean uncertainty across MOFs. This trend underscores the potency of AL in optimizing dataset efficiency while upholding predictive accuracy, providing valuable insights for refining AL policies and strategies.

Figure 2. Perfomance of GPR model across MOFs and data savings.



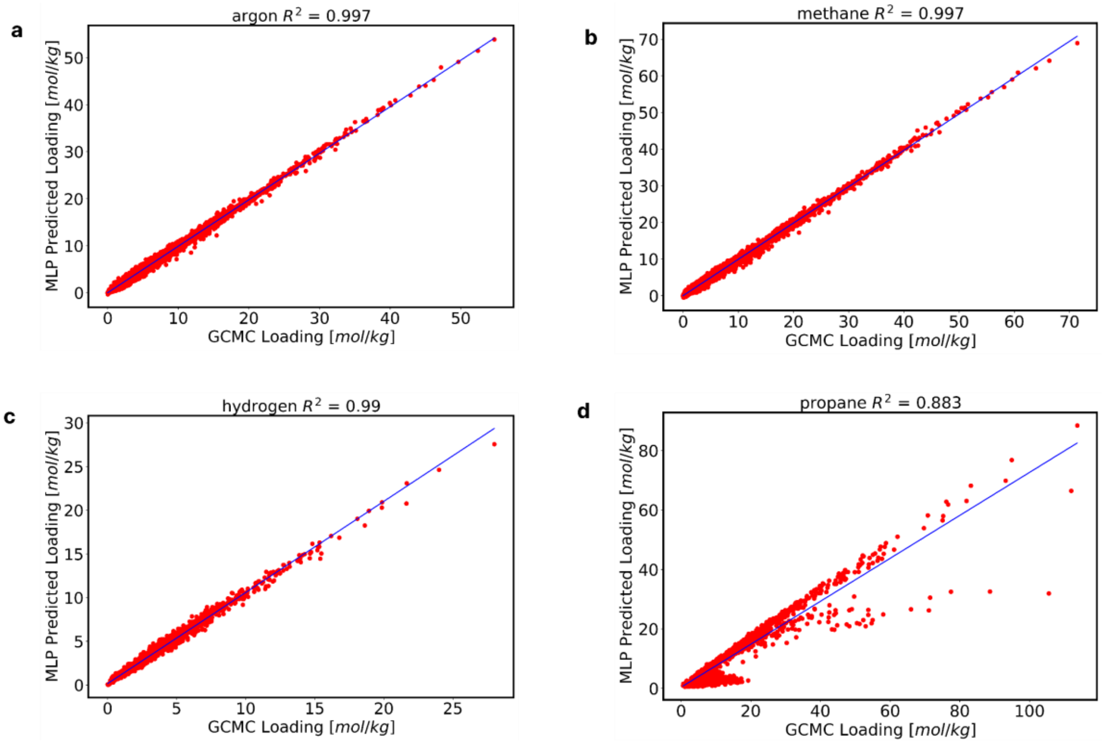
a) Evolution of the GP mean uncertainty and R^2 of the MOF with the most AL iterations; **b)** Evolution of the GP mean uncertainty and R^2 of the MOF with the fewest AL iterations; **c)** Average R^2 across all MOFs (left y-axis) and the corresponding % data savings (right y-axis) at various AL iterations.

We developed 1800 GP models for MOFs using AL, necessitating a separate prediction case for each MOF when making predictions for other alchemical molecules. We leverage the datapoints used to train the final GP models at the 0.05 mol/kg uncertainty level for each of the 1800 MOFs to produce a new MLP model. All these datapoints were collected into one single training data featuring 2.1 million data points (57.5% data savings relative to the 5 million used to train the original MLP). Next, we utilized TensorFlow[34] to train a new MLP model while optimizing its associated hyperparameters (details can be found in the **SI)**, which was used to predict the adsorption of real molecules within different MOFs, as done previously by Gómez-Gualdrón and coworkers[33]. **Figure 3** shows a comparison between our GP-informed MLP and the GCMC simulations.

Notice that the predictions in **Figure 3** correspond to real molecules despite the training data corresponding to alchemical molecules. Note that molecules such as argon and methane (**Figures 3a** and **3b**, respectively) can be considered interpolations between alchemical molecules used in the training. The new MLP, also retains the same limitations of the original MLP when predicting adsorption for real molecules that fall below and above the alchemical parameter "range" considered for training. While hydrogen predictions remain accurate (**Figure 3c**), predictions for larger molecules like propane prove to be less successful (**Figure 3d**). These results reemphasize the need to expand the training data to include, for instance, larger alchemical adsorbates, but we show that could be achieve efficiently using AL. Additional results and predictions can be found in the **SI**.

We further compared the predictions of the new MLP and the original MLP model. The results of these predictions of the adsorption of real molecules from both models are shown in **Table 1**. The new MLP model, trained on 57.5% less data than the original MLP model, exhibited a comparable performance to the original MLP model. A similar level of performance by both MLP models was maintained for molecules within and outside the alchemical range. These results show that AL is useful in scaling down on the training data required by MLP models.

Figure 3. Perfomance of the newly developed MLP model.



Performance of the new MLP model. The new MLP model was evaluated in the prediction of adsorption of real molecules among the range of alchemical training adsorbates **a)** argon, **b) m**ethane, and extrapolated outside the range of alchemical training adsorbates **c) h**ydrogen, and **d) p**ropane**.**

Table 1: R^2 comparison between the original MLP model and new MLP model trained on the GP model final training data. The red items refer to molecules extrapolated outside the range of the alchemical training adsorbates.

Adsorbate	*New MLP Model (R^2)*	*Original MLP Model (R^2)*
Argon	**0.997**	**0.997**
Ethane	**0.996**	**0.996**
Krypton	**0.997**	**0.997**
Methane	**0.997**	**0.997**
Nitrogen	**0.997**	**0.997**
Xenon	**0.996**	**0.996**
Butane	**0.638**	**0.616**
Helium	**0.981**	**0.982**
Hydrogen	**0.99**	**0.993**
Isobutane	**0.287**	**0.27**
Propane	**0.833**	**0.871**
Benzene	**0.783**	**0.775**

Simultaneous AL on adsorbate, fugacity, and MOF space

The approximate halving of the training data by applying AL to the adsorbate space was encouraging, but arguably represents insufficient data savings for the increase in adsorbate space dimensionality that would occur if one expanded the types of adsorbates included in the training dataset to account for higher adsorbate complexity. Additionally, developing individual GP models for each MOF to generate the training data is notably demanding and tasking, as it requires performing AL on each MOF. To address this challenge, we decided to adopt a strategy where AL operates simultaneously on the adsorbate and MOF space. The underlying hypothesis was that what a model learns about adsorption in one MOF may be applicable to other similar MOFs, making exhaustive training data generation for all MOFs in a database unnecessary.

To achieve the above, we sought to incorporate the MOF features as additional input to a GP surrogate model within an AL framework that also selects the most "informative" MOFs to be used in training data generation. However, as the MOF space is inherently high-dimensional—where each MOF can be described as combination of chemical and textural characteristics such as node and linker types, void fraction, surface area, pore sizes, and so forth—it is imperative to reduce the dimensionality of the MOF representation to make AL exploration more efficient.

To this end, we resorted to principal component analysis[35–39] (PCA), which is a widely employed technique to transform high-dimensional data into a lower-dimensional space, while retaining the essential patterns and structures inherent in the original data. In this study, PCA was applied to the textural properties of a larger number of MOFs (3445) than in our first approach. We used the following properties for PCA: largest pore diameter (LPD), diffusion-limiting pore diameter (DLPD), void fraction (VF), surface area (SA), standard deviation of the pore size distribution (PSSD), and the inverse framework density (IFD). These 3445 MOFs were the dataset MOFs from two previous studies.[31,33] These MOF properties were chosen due to their deemed importance in predicting adsorption.[40–43] The resulting principal components can be effectively interpreted as representative descriptors that capture the prevailing patterns and variabilities in the textural properties of the MOFs. **Figure S12** shows that the first two principal component directions (PC1 and PC2) account for 91% of the cumulative variance, suggesting they are sufficient for the AL to meaningfully navigate the MOF space. Formally, the inclusion of the MOF into the AL process makes it so that $N \sim F(f, \varepsilon, \sigma, l, q, PC1, PC2)$.

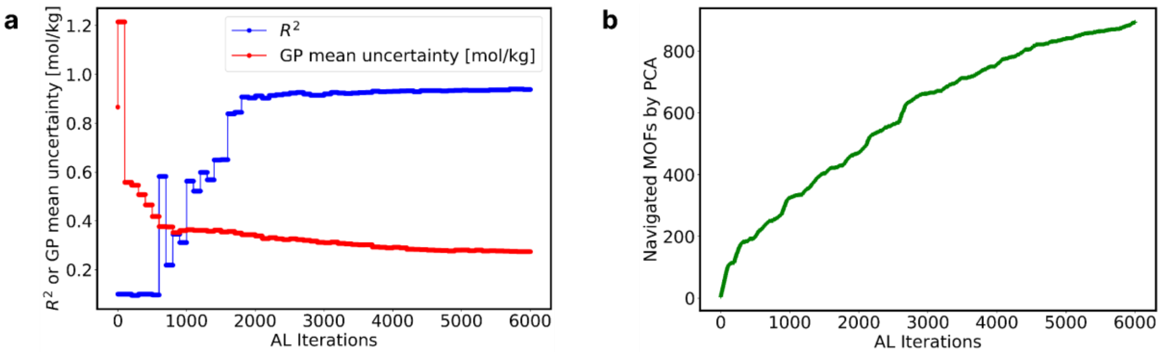
We selected four MOFs positioned at the boundaries of PC1 and PC2 to initiate the AL iterations. For each MOF, we provided adsorption data for a single alchemical adsorbate, chosen randomly, at a random fugacity. The initial training dataset comprises this data, while the remaining data from all the MOFs, fugacity, and adsorbates constitute the testing dataset in this context.

The bagging approach was applied to the full testing dataset. This approach involves the segmentation of the large dataset into bags of smaller datasets for predictions. The dataset, comprising 3445 MOFs and over 5 million simulation data points[31,33] obtained through GCMC simulations, was systematically organized into 100 bags. Each bag was designed to contain diverse data samples, systematically categorized by fugacities and adsorbates. This categorization also ensured variability and comprehensive coverage within each set while incorporating all 3445 MOFs (represented by their PCs) in every bag. The purpose of the bagging process was simply to parallelize the testing of the model.

Upon segmentation, the initial GP model is evaluated on each bag, where the uncertainties in each data point are collected across all the bags. Upon compilation of uncertainties from the bags, the maximum GP uncertainty across all bags was estimated. At this point, the test array corresponding to adsorption at that specific point of maximum uncertainty was retrieved. Subsequently, this array, along with its corresponding GCMC adsorption data, was used to update the training dataset. This process was repeated 6000 times (see **Figure S14**).

Upon 6,000 iterations, the final training dataset selected by AL consists of 6004 data points. The evolution of the average R^2 and GP mean uncertainty as a function of iteration is shown in **Figure 4a**. Initially, the model had a GP mean uncertainty of 0.87 mol/kg and a low R^2 of 0.1, which were substantially improved to a final average GP mean uncertainty of 0.27 mol/kg and an R^2 of 0.94. **Figure 4b** shows the number of MOFs used by the AL as the iterations increase, which by the 6000th iteration corresponds to 893 MOFs. Using just 0.11% of the data, encompassing only 26% of the MOFs in the database, we constructed a GP-PCA model with an R^2 of 0.94. This kind of data savings are extremely encouraging for future exploration of datasets including a larger variety of adsorbate types.

Figure 4. Analysis of the GP-PCA model.



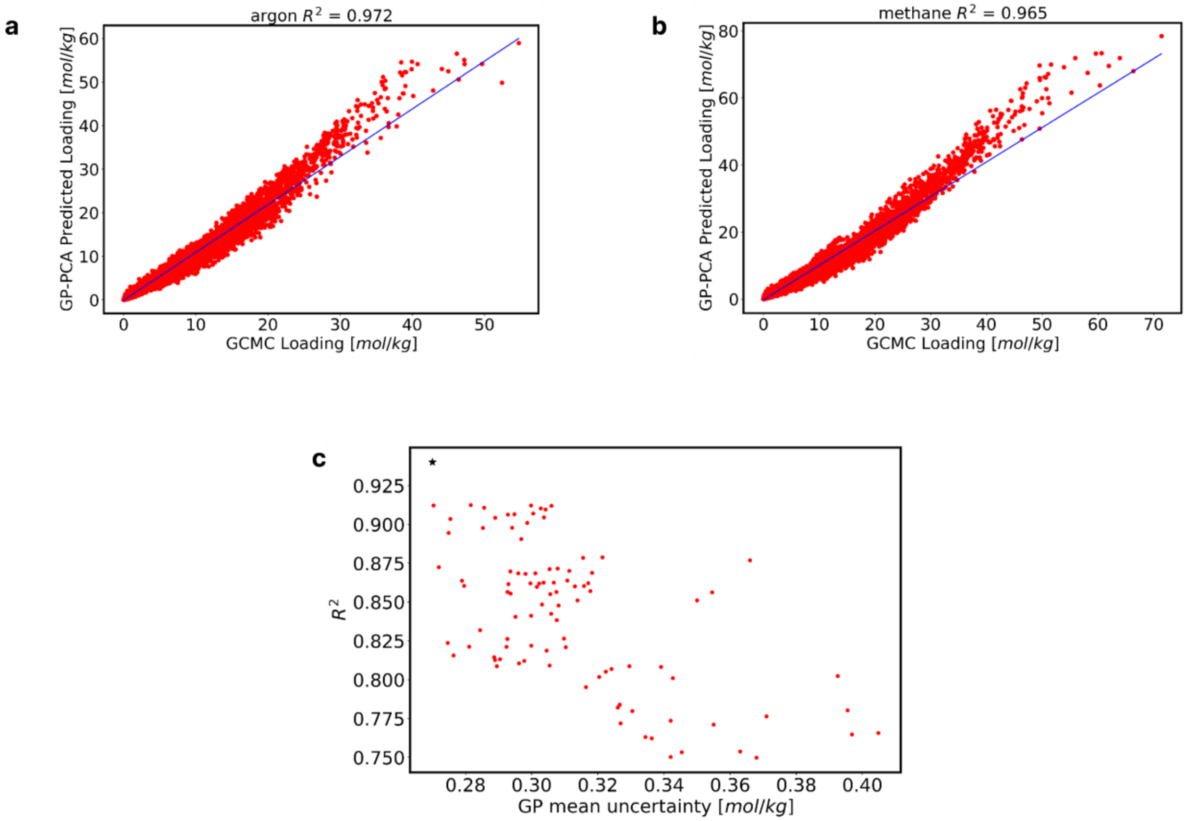
a) Evolution of average R^2, average GP MAE, and number of MOFs attained during 6000 iterations. **b)** Number of MOFs navigated as a function of the number of iterations.

Figures 5a through **5b** highlight predictions made for real molecules utilizing the newly developed GP model. **Figures 5a** and **5b** show adsorption predictions for argon and methane in the 3445 MOFs at a range of fugacities between 1e-2 and 100 bar. Results for other real molecules are shown in **Figure S15**. With these results, it is possible to say that the model does well in predicting the adsorption of real molecules.

Notice that while we showed the possibility of training a suitable adsorption model using only 6,004 datapoints, AL is directly responsible for that outcome. To illustrate this point, we fit GPs to 100 randomly selected training data sets, each with 6,004 datapoints. **Figure 5c** shows the significant variability in the average R^2 and GP mean uncertainty for these 100 randomly selected training datasets. The lowest achieved R^2 was 0.75, coupled with a GP mean uncertainty of 0.368 mol/kg. Conversely, the highest R^2 obtained was 0.91, accompanied by a GP mean uncertainty of 0.282 mol/kg. On average, across all 100 randomly selected training datasets, the average R^2 was 0.84, while the GP mean uncertainty averaged 0.312 mol/kg, which are worse than the R^2 of 0.94 and GP mean uncertainty of 0.27 mol/kg attained by the AL-selected dataset (star symbol in **Figure 5c)**.

Finally, we developed another MLP model trained using the 6004 data points from the GP-PCA model. Employing the same method and hyperparameters as previously applied, we fine-tuned the MLP model to ensure consistency and comparability with our earlier methodologies. Subsequently, we utilized this MLP model to predict the adsorption of real molecules across multiple MOFs. However, upon evaluation, our findings revealed a reduction in accuracy compared to the GP-PCA model. This divergence in predictive performance underscores the intricate challenges inherent in modeling gas adsorption phenomena within MOFs using traditional MLP approaches. For instance, employing the GP-PCA model yields an R^2 value of 0.972, whereas the MLP model achieves an R^2 value of 0.922. Other real molecules predictions can be found in the SI.

Figure 5. Performance of the GP-PCA model, and random sampled training data performance.

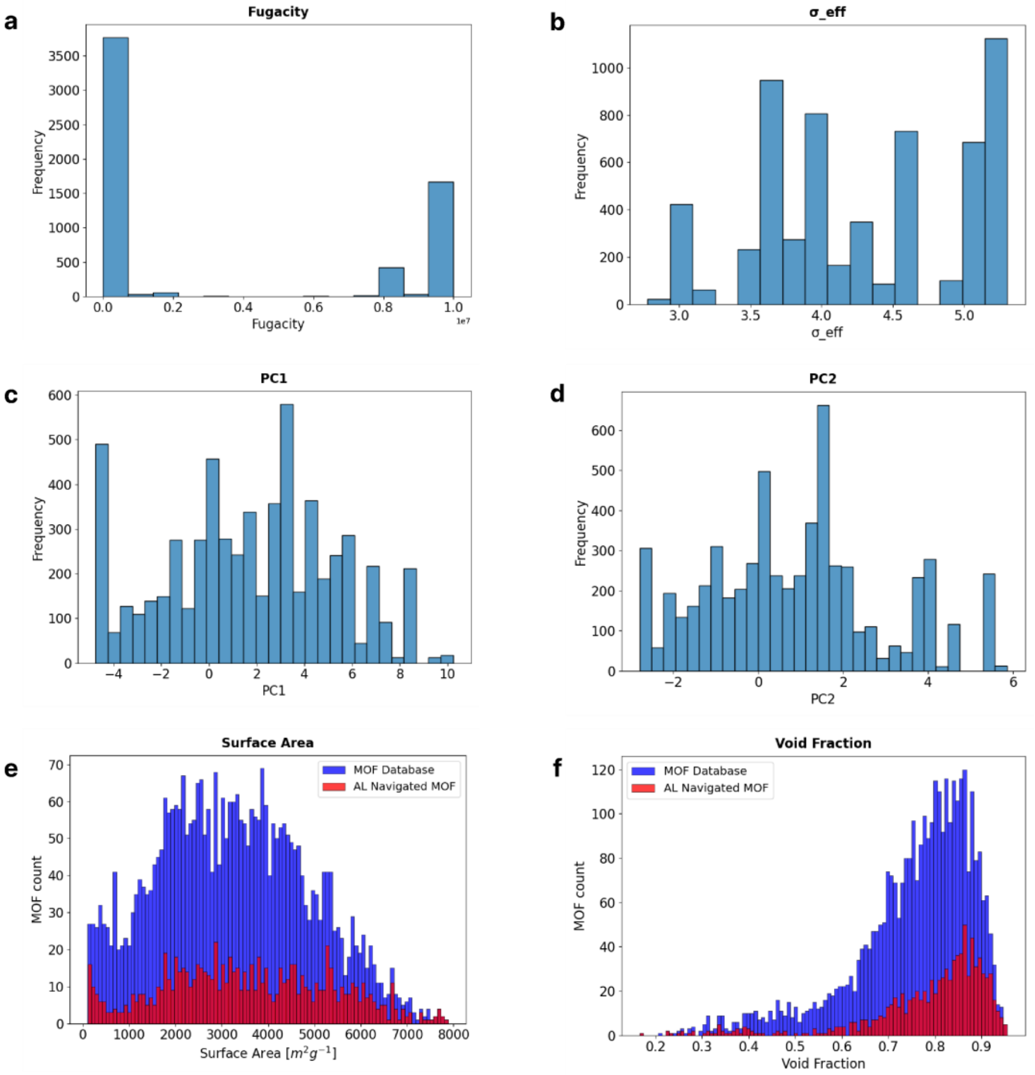


a) – b) Model prediction of real molecules (Argon, and Methane). **c)** R^2(s) and GP MAE(s) across all randomly selected training data. Each point on the graph corresponds to a unique training dataset, with its associated R^2 and GP MAE values.

Feature navigation and evolution of the GP-PCA model

Looking to understand how AL made the selection of fugacity and adsorbate and MOF features to be included in the training dataset**, Figures 6a** through **6d** show the analysis of the some of the feature regions (fugacity, σ_{eff}, PC1, and PC2). From the observations for fugacity, the model requires more training data at the boundaries of the feature. While it is not clear whether this is coincidental or not, it is worth noting that such fugacities tend to inform the model about adsorption at dilute (i.e., Henry's region) and near-pore saturation conditions. Contrastingly, the model required more evenly distributed training data along the σ_{eff}, PC1, and PC2 input features. From these results, we can see that AL intelligently selects what regions to sample and that it requires more diverse sampling for the adsorbates than any other input feature of the model. To get a more meaningful interpretation of the explored MOF space, we revert the PC1 and PC2 back to their textural properties, and in **Figure 6e-f**, we show the regions navigated by the AL in terms of the surface area and void fraction. Remarkably, the observed distribution of textural properties in the 893 MOFs picked by AL mimics closely the distribution of these properties in the complete 3445-MOF set. This suggests that the AL picks a representative sample of MOFs for each combination textural property values.

Figure 6. Training data sampled by the AL GP-PCA model.



a) – d) AL selected training data regions for fugacity, effective sigma, PC1 and PC2, respectively. **e) – f)** AL sampled regions of surface area and void fraction as a comparison to the distribution of the available MOF dataset.

Discussions

AL has emerged as a critical methodology to optimize the data selection and acquisition for gas adsorption in MOFs and therefore to understand the most important features to predict their adsorption isotherms. Initially in this work, through a systematic application of AL to a dataset of 1800 MOFs, significant strides were made in reducing the dataset while continuously enhancing model performance. This iterative process was marked by a noticeable correlation between the reduction in GP mean uncertainty and the increment in R^2 values through the iterative process, indicating a consistent improvement trend across multiple MOFs.

For the first approach with 1800 MOFs, we performed AL on each MOF by setting a convergence criterion of 0.05 mol/kg in the GP mean uncertainty, the resultant amalgamation of training data from diverse MOFs formed a comprehensive dataset encompassing approximately 2.1 million data points, which was used to train a new MLP model. This newly created MLP model showcased precision in predicting the adsorption behaviors of real molecules within the specified alchemical range. It was notable that the model's performance was comparable to the original one. This achievement highlights the efficiency of AL in selecting informative datasets while significantly saving computational resources.

As a second approach applied to 3445 MOFs, PCA was instrumental in identifying significant dimensions contributing to variance within MOF textural properties enhancing the MOF textural space exploration. Including PC1 and PC2 in the AL process contributed to robust model training and enhanced predictive accuracy. Furthermore, an in-depth analysis of textural properties within the navigated MOF subset shows the preservation of the overall distribution of textural properties compared with the available data. This diverse and extensive coverage across MOF textural properties is observed in the comparative histograms. This fact is proof of the navigation process's effectiveness in encapsulating and representing essential material characteristics within the navigated subset.

The scope of this work is primarily constrained by the limited range of considered adsorbates, delineated by the parameters of the alchemical model. However, this constraint mirrors that of previous models. Specifically, the study focuses on predicting adsorption isotherms for small, nearly spherical, nonpolar, monoatomic, and diatomic adsorbates across various fugacities, at 298 K, consistent with the conditions of GCMC simulations. Despite these limitations, the integration of AL in data generation represents a significant advancement toward establishing a more comprehensive and universally applicable adsorption model for gases within MOFs. This approach signifies progress in refining predictive models, particularly in terms of reducing data requirements. Importantly, it sets a clear path for expanding the model's versatility by incorporating new data, thereby enhancing its applicability across a broader range of scenarios.

In summary, these cumulative findings highlight the efficacy of AL in navigating complex MOF and adsorbate spaces, accurately predicting adsorption, and enriching our understanding of the phenomena, specifically in MOFs. This evidence solidifies AL as a valuable and necessary methodology in material science research, offering an effective way to overcome data-scarcity while paving the way for future advancements in this domain.

Methods

Gaussian process regression (GPR) is a probabilistic ML technique effectively used for non-linear regression tasks. It operates on the principles of Bayesian statistics and assumes a prior distribution over functions, defining a distribution over the entire space of functions that could describe the underlying data.

The fundamental concept behind GPR involves modeling the relationship between input features (predictors) and output variables (predictions) using Gaussian processes (GPs). GPs are defined by a mean function and a covariance function (also known as kernel function). The mean function represents the average trend of the data, while the covariance function captures the similarity between pairs of data points x and x'. The GPR is mathematically represented by $f \sim GP\big(m(x), K(x, x')\big)$, where the function f

has a GP distribution with mean function (m) and covariance function (K). Here, as K we use the rational quadratic (RQ) kernel, which takes the mathematical form:

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

The kernel is characterized by $x - x'$ which is the Euclidean distance between x and x' data points (the input data); l is a parameter that signifies the length scale, defining the characteristic length over which variations in the function occur, and α plays a pivotal role governing the balance between large-scale and small-scale fluctuations within the function.

In this study of gas adsorption in MOFs, GPR was utilized to model the relationship between various features associated with MOFs and the adsorption of specific molecules. On the navigation of fugacity and alchemical molecules across 1800 MOFs, the applied GP model trained individual GPR models for each MOF to predict the adsorption behavior of different molecules within that MOF. In the scenario involving the navigation of fugacity, alchemical molecules, and the 3445 MOFs represented by the principal component derived from their textural properties, a single GP model is trained to predict adsorption. The complete training process was carried out through AL iterations strategically selecting data points that improve the predictive accuracy of the model, ultimately reducing computational cost to generate the training data. The GPflow library[44] was used for implementing the AL workflow and the GP used the rational quadratic (RQ) kernel[45–48].

AL Algorithm Implementation on Fugacity and Adsorbates.

The study used an AL algorithm to navigate adsorption scenarios on 1800 MOFs. The features of the GP model (F) used for each MOF are fugacity (f), surrogate Lennard-Jones parameters epsilon effective (ε) and sigma effective (σ), bond length (l), and charges (q); these are the parameters that the AL algorithm automatically selects at each iteration. The GP model uses those features to make adsorption predictions in the MOF (N), mathematically $N{\sim}F(f, \varepsilon, \sigma, l, q)$. **Figure 1** illustrates the AL convergence criteria set at 0.05 mol/kg GP mean uncertainty and the algorithm workflow across iterations for all MOFs analyzing GP mean uncertainty and R^2 behavior.

Multilayer Perceptron (MLP) Model Training

A unified training data set comprising approximately 2.1 million data points from multiple MOFs was created using the AL process. A Multilayer Perceptron (MLP) model was trained using TensorFlow[49]. The selected model configuration included 500 epochs a batch size of 128, and a learning rate of 0.00001. The MLP architecture featured an input layer followed by three hidden layers, each employing Leaky ReLU[50] activation functions.

AL Algorithm Implementation on fugacity, adsorbates, and adsorbents.

Principal components (PCs) of the MOF textural properties were generated using scikit-learn[51,52]. This was applied to a dataset containing 3445 MOFs textural properties to identify primary dimensions significantly contributing to variance. The AL process was identical as before but adding PC1 and PC2 as input features to the AL process: $N{\sim}F(f, \varepsilon, \sigma, l, q, PC1, PC2)$.

Bagging Approach for Model Testing

The GCMC simulation data, exceeding 5 million points and taken from previous studies[31,33], was segmented into 100 bags to represent diverse adsorbates across 3445 MOFs. Ensuring that each bag encompassed the PCs of all MOFs for uniformity and representativity in each bag. These bags were structured to vary across fugacity and adsorbate types.

Data availability

The data used in this study was obtained from two previous studies.[31,33] However, the data was reorganized to be used for this research study. The reorganized data can be found on the GitHub page via: https://github.com/theOsaroJ/Active-Learning-of-alchemical-adsorption-simulations-towards-a-universal-adsorption-model.

Code availability.

The AL algorithms, along with examples of the GP model and the newly developed MLP model tailored for fugacity and adsorbate scenarios, are accessible on GitHub. Additionally, both the GP-PCA and MLP models, designed to encompass fugacity, adsorbate, and adsorbent space, are available on GitHub at https://github.com/theOsaroJ/Active-Learning-of-alchemical-adsorption-simulations-towards-a-universal-adsorption-model.

Acknowledgements.

E.O would like to thank the Lucy Family Institute for Data and Society as well as The Patrick and Jana Eilers Graduate Student Fellowship for Energy Related Research at the University of Notre Dame. D.A.G-.G. acknowledges funding from the Institute for Data-Driven Dynamical Design (ID4) funded through NSF grant OAC-2118201 and also thanks access to the Mio supercomputer at Colorado School of Mines. Y.J.C gratefully acknowledge NSF CAREER Award No. CBET-2143346. E.O and Y.J.C also thank the Center for Research Computing at the University of Notre Dame for computational resources.

References.

1. Li, H. *et al.* Porous metal-organic frameworks for gas storage and separation: Status and challenges. *EnergyChem* **1**, 100006 (2019).

2. Islamov, M. *et al.* High-throughput screening of hypothetical metal-organic frameworks for thermal conductivity. *NPJ Comput Mater* **9**, 1–12 (2023).

3. Feng, L., Wang, K. Y., Willman, J. & Zhou, H. C. Hierarchy in Metal-Organic Frameworks. *ACS Cent Sci* **6**, 359–367 (2020).

4. Raptopoulou, C. P. Metal-organic frameworks: Synthetic methods and potential applications. *Materials* **14**, 1–32 (2021).

5. Borboudakis, G. *et al.* Chemically intuited, large-scale screening of MOFs by machine learning techniques. *NPJ Comput Mater* **3**, 1–6 (2017).

6. Comlek, Y., Pham, T. D., Snurr, R. Q. & Chen, W. Rapid design of top-performing metal-organic frameworks with qualitative representations of building blocks. *NPJ Comput Mater* **9**, (2023).

7. 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* **2**, 1–14 (2019).

8. Langmi, H. W., Ren, J., North, B., Mathe, M. & Bessarabov, D. Hydrogen storage in metal-organic frameworks: A review. *Electrochim Acta* **128**, 368–392 (2014).

9. Mao, H. *et al.* A scalable solid-state nanoporous network with atomic-level interaction design for carbon dioxide capture. **6849**, 1–28 (2022).

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

11. Sholl, D. S. & Lively, R. P. Seven chemical separations to change the world. *Nature* **532**, 435–437 (2016).

12. Lin, S. *et al.* Machine-learning-assisted screening of pure-silica zeolites for effective removal of linear siloxanes and derivatives. *J Mater Chem A Mater* **8**, 3228–3237 (2020).

13. Ohba, T. *et al.* GCMC simulations of dynamic structural change of Cu–organic crystals with N2 adsorption. *J Exp Nanosci* **1**, 91–95 (2006).

14. Rogge, S. M. J. *et al.* Modeling Gas Adsorption in Flexible Metal–Organic Frameworks via Hybrid Monte Carlo/Molecular Dynamics Schemes. *Adv Theory Simul* **2**, 1–15 (2019).

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* **21**, 11259–11270 (2011).

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* **59**, 3141–3152 (2020).

17. Gómez-Gualdrón, D. A., Simon, C. M. & Colón, Y. Efficient Data Utilization in Training Machine Learning Models for Nanoporous Materials Screening. *AI-Guided Design and Property Prediction for Zeolites and Nanoporous Materials* 343–376 (2023) doi:10.1002/9781119819783.ch13.

18. Yang, C. *et al.* Application of machine learning in MOFs for gas adsorption and separation. *Mater Res Express* (2023) doi:10.1088/2053-1591/ad0c07.

19. Guo, W. *et al.* Deep Learning Models for Predicting Gas Adsorption Capacity of Nanomaterials. *Nanomaterials* **12**, (2022).

20. Yang, Z., Chen, B., Chen, H. & Li, H. A critical review on machine-learning-assisted screening and design of effective sorbents for carbon dioxide (CO2) capture. *Front Energy Res* **10**, 1–19 (2023).

21. Zhang, X. *et al.* High-throughput and machine learning approaches for the discovery of metal organic frameworks. *APL Mater* **11**, (2023).

22. Mukherjee, K. & Colón, Y. J. Machine learning and descriptor selection for the computational discovery of metal-organic frameworks. *Mol Simul* **47**, 857–877 (2021).

23. Shields, M. D. *et al.* Active learning applied to automated physical systems increases the rate of discovery. *Sci Rep* **13**, 1–9 (2023).

24. Sheng, Y. *et al.* Active learning for the power factor prediction in diamond-like thermoelectric materials. *NPJ Comput Mater* **6**, 1–7 (2020).

25. Vandermause, J. *et al.* On-the-fly active learning of interpretable Bayesian force fields for atomistic rare events. *NPJ Comput Mater* **6**, 1–11 (2020).

26. Bassman, L. *et al.* Active learning for accelerated design of layered materials. *NPJ Comput Mater* **4**, 1–9 (2018).

27. Lookman, T., Balachandran, P. V., Xue, D. & Yuan, R. Active learning in materials science with emphasis on adaptive sampling using uncertainties for targeted design. *NPJ Comput Mater* **5**, (2019).

28. Osaro, E., Mukherjee, K. & J. Colón, Y. Active Learning for Adsorption Simulations: Evaluation, Criteria Analysis, and Recommendations for Metal–Organic Frameworks. *Industrial &amp; Engineering Chemistry Research* **62**, 13009–13024 (2023).

29. Mukherjee, K., Dowling, A. W. & Colón, Y. J. Sequential design of adsorption simulations in metal organic frameworks. *Mol Syst Des Eng* **7**, 248–259 (2021).

30. Mukherjee, K., Osaro, E. & Colón, Y. J. Active learning for efficient navigation of multi-component gas adsorption landscapes in a MOF. *Digital Discovery* 1506–1521 (2023) doi:10.1039/d3dd00106g.

31. Anderson, R. & Gómez-Gualdrón, D. A. Deep learning combined with IAST to screen thermodynamically feasible MOFs for adsorption-based separation of multiple binary mixtures. *Journal of Chemical Physics* **154**, (2021).

32. Colón, Y. J., Gómez-Gualdrón, D. A. & Snurr, R. Q. Topologically Guided, Automated Construction of Metal-Organic Frameworks and Their Evaluation for Energy-Related Applications. *Cryst Growth Des* **17**, 5801–5810 (2017).

33. 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* **16**, 1271–1283 (2020).

34. Rucci, M. & Casile, A. Fixational instability and natural image statistics: Implications for early visual representations. *Network: Computation in Neural Systems* **16**, 121–138 (2005).

35. Kollias, L., Rousseau, R., Glezakou, V. A. & Salvalaglio, M. Understanding Metal-Organic Framework Nucleation from a Solution with Evolving Graphs. *J Am Chem Soc* **144**, 11099–11109 (2022).

36. Sapnik, A. F. *et al.* Multivariate analysis of disorder in metal–organic frameworks. *Nat Commun* **13**, 1–13 (2022).

37. Nandy, A., Duan, C. & Kulik, H. J. Using Machine Learning and Data Mining to Leverage Community Knowledge for the Engineering of Stable Metal-Organic Frameworks. *J Am Chem Soc* **143**, 17535–17547 (2021).

38. Lid, P. P. & Planning, S. *PRINCIPAL COMPONENTS ANALYSIS (PCA)* Xln 1*. vol. 19 (1993).

39. Jollife, I. T. & Cadima, J. Principal component analysis: A review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* **374**, (2016).

40. Yang, W. *et al.* Computational screening of metal-organic framework membranes for the separation of 15 gas mixtures. *Nanomaterials* **9**, (2019).

41. Pardakhti, M., Moharreri, E., Wanik, D., Suib, S. L. & Srivastava, R. Machine Learning Using Combined Structural and Chemical Descriptors for Prediction of Methane Adsorption Performance of Metal Organic Frameworks (MOFs). *ACS Comb Sci* **19**, 640–645 (2017).

42. Cooper, G. M. & Colón, Y. J. Metal-organic framework clustering through the lens of transfer learning. *Mol Syst Des Eng* **8**, 1049–1059 (2023).

43. Mukherjee, K. & Colón, Y. J. Machine learning and descriptor selection for the computational discovery of metal-organic frameworks. *Mol Simul* **47**, 857–877 (2021).

44. de Matthews, A. G. *et al.* GPflow: A Gaussian Process Library using TensorFlow Mark van der Wilk. *Journal of Machine Learning Research* **18**, 1–6 (2017).

45. Gheytanzadeh, M. *et al.* Towards estimation of CO2 adsorption on highly porous MOF-based adsorbents using gaussian process regression approach. *Sci Rep* **11**, 1–13 (2021).

46. Deringer, V. L. *et al.* Gaussian Process Regression for Materials and Molecules. *Chem Rev* **121**, 10073–10141 (2021).

47. Dudek, A. & Baranowski, J. Gaussian Processes for Signal Processing and Representation in Control Engineering. *Applied Sciences (Switzerland)* **12**, (2022).

48. Wilson, A. G. & Adams, R. P. Gaussian process kernels for pattern discovery and extrapolation. *30th International Conference on Machine Learning, ICML 2013* **28**, 2104–2112 (2013).

49. Abadi, M. *et al.* TensorFlow: A system for large-scale machine learning. (2016).

50. Xu, B., Wang, N., Chen, T. & Li, M. Empirical Evaluation of Rectified Activations in Convolutional Network. (2015).

51. Pedregosa, F. *et al.* Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* **12**, 2825–2830 (2011).

52. Deekshith, P. & Singh, R. P. Review on Advanced Machine Learning Model : Scikit-learn. *International Journal of Scientific Research and Engineering Development (IJSRED)* **3**, 526–529 (2020).