ACCELERATED DISCOVERY OF CH_4 UPTAKE CAPACITY MOFS USING BAYESIAN OPTIMIZATION

A PREPRINT

Eric Taw
Department of Chemical Engineering,
University of California, Berkeley;
Material Science Division,
Lawrence Berkeley National Laboratory
erictaw@berkeley.edu

Jeffrey B. Neaton
Department of Physics,
University of California, Berkeley;
Material Science Division,
Lawrence Berkeley National Laboratory;
Kavli Energy NanoScience Institute at Berkeley
jbneaton@lbl.gov

November 5, 2021

ABSTRACT

High-throughput computational studies for discovery of metal-organic frameworks (MOFs) for separations and storage applications are often limited by the costs of computing thermodynamic quantities, with recent studies reliant ab initio results for a narrow selection of MOFs and empirical force-field methods for larger selections. Here, we conduct a proof-of-concept study using Bayesian optimization on CH_4 uptake capacity of hypothetical MOFs for an existing dataset (Wilmer et al, Nature Chem. 2012, **4**, 83). We show that less than 0.1% of the database needs to be screened with our Bayesian optimization approach to recover the top candidate MOFs. This opens the possibility of efficient screening of MOF databases using accurate ab-initio calculations for future adsorption studies on a minimal subset of MOFs. Furthermore, Bayesian optimization and the surrogate model presented here can offer interpretable material design insights and our framework will be applicable in the context of other target properties.

1 Introduction

Metal-organic frameworks (MOFs) are a class of highly porous, crystalline materials known for their high surface areas and potential for separations applications. [1] MOFs consist of inorganic secondary building units (SBUs) and organic linkers in a particular topology. [2] The many possible SBUs, organic linkers, and topologies leads to a combinatorial number of hypothetical MOFs (hMOFs), and prior studies have developed systematic ways of constructing a sizable database of such hypothetical structures. [3, 4] Even a limited selection of building blocks derived from experimentally synthesized structures can lead to $> 10^5$ candidate MOFs. [5] Recently published databases consist of over 300,000 hypothetical MOF structures. [6]

Existing databases enable high-throughput computational screening of MOFs, allowing one to not only propose candidate MOFs for a given application, but also observe structural and chemical trends. Among the first studies to do so is from Wilmer et al. [5], which found methane storage MOFs with uptake capacities greater than the state-of-the-art at the time of publication. Furthermore, they obtained global insights across all 130,000 hMOFs, such as an optimal range of pore diameter and gravimetric surface area. [5] Similar studies have appeared for H_2 storage [7], flue gas carbon capture [8], and Xe/Kr separations. [9] Notably, Chung et al. used genetic algorithms to dramatically reduce the number of MOF candidates screened while still identifying top candidates for flue gas carbon capture. [10]

When deciding on methods for discovering new candidate materials with a given property, one must balance the trade-off between screening through a vast number of candidates and the accuracy of the method for predicting the property. For gaseous adsorption, one brute-force approach could be to screen through a large database of structures using computationally cheap methods relying on force fields, as is the case for the studies cited above. However, for some applications, force fields are insufficiently predictive [11]. On the flip side, more accurate methods like density

functional theory (DFT) are more computationally expensive, particularly for large unit cell materials like MOFs, due to the cubic scaling of DFT with the number of atoms. [12] Thus, DFT calculations are often reserved for smaller, more targeted datasets, such as in Rosen et al., where the binding energy of O_2 was evaluated for less than 100 MOFs. [13]

Recent studies have shown Bayesian optimization to be a promising method of finding new material candidates without screening through a large combinatorial database. Herbol et. al. used DFT to calculate the solvation enthalpy of various hybrid organic-inorganic perovskites, and a custom Bayesian optimization algorithm allowed them to discover new halide perovskite-solvent combinations without performing calculations for the entire search space. [14] Yamawaki et al. discovered new nanostructured graphene for thermoelectric applications. [15] In both cases, Bayesian optimization allowed the authors to identify candidate materials much more efficiently than random searching despite a large combinatorial search space.

To the best of our knowledge, no literature currently exists applying Bayesian optimization techniques for MOFs despite the prevalence of high throughput screening studies of MOF adsorbents. [5, 16, 17, 6, 18, 19] While genetic algorithms could be used to dramatically reduce the computational expense of high-throughput screening procedures, such methods can require extensive tuning by hand for good performance, and the tuning procedure can involve little physical or statistical intuition. [20]

In this work, we propose a simple Bayesian optimization method for identifying the most promising candidate MOF in a database for CH_4 uptake capacity at a given pressure. We show that this Bayesian optimization approach can reliably discover the top 10 candidates for CH_4 uptake in a set of 51,000 hypothetical MOFs by sampling less than 0.1% of the dataset. Furthermore, the model we use identifies the organic linker as the most important factor in estimating the methane uptake capacity. In what follows, we first briefly review the two key components of Bayesian optimization, the surrogate model and the acquisition function. Then, we discuss the specifics of our surrogate model.

2 Bayesian Optimization

Bayesian optimization (BO) is a well-known technique for sample-efficient, derivative-free global optimization of an objective function, in this case the CH_4 uptake capacity (see below), and has found applications in fields ranging from robotics [21] to hyperparameter tuning [22, 23] (and, more amusingly, chocolate chip cookie recipe optimization [24]). We summarize Bayesian optimization below and refer readers to Frazier's recent tutorial for further details. [25] All Gaussian processes were implemented with GPyTorch. [26]

2.1 Surrogate Model

BO relies on a surrogate model, which approximates an objective function such that one may obtain cheap estimates of the objective, i.e., the quantity of interest (CH_4 uptake capacity), as well as the uncertainty of that estimate. Here, we model the objective as a Gaussian process (GP) with a linear mean. Our model can be broken down into two components:

$$\hat{y}(\mathbf{x}_{struct}, \mathbf{x}_{ident}) = LR(\mathbf{x}_{struct}) + GP(\mathbf{x}_{ident}) \tag{1}$$

where a linear regression (LR) model correlates structural features $\mathbf{x}_{struct}$, here the dominant pore diameter (in Å), void fraction, gravimetric surface area (in m^2/g), and the volumetric surface area (in m^{-1}). The coefficients of our LR model are found by evaluating the normal equation; our GP further trains on the error of the LR model using identity features $\mathbf{x}_{ident}$ (here, interpenetration capacity, actual interpenetration, inorganic node, primary linker, and secondary linker). We adapt the feature set used by Chung et al. [27], and more in-depth discussion of this feature set can be found in their paper and their associated supplementary information.

GPs rely on a kernel function $k(\mathbf{x}_i, \mathbf{x}_j)$ which quantifies the similarity between two identity feature vectors $\mathbf{x}_i$ and $\mathbf{x}_j$. A naïve implementation of the identity features is to one-hot encode every feature, though this is not sufficiently informative for a GP to adequately approximate the objective function and will lead to limitations in Bayesian optimization. [28] In cheminformatics, one commonly uses similarity metrics like the cosine distance $k(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{x}_i \cdot \mathbf{x}_j / \|\mathbf{x}_i\| \|\mathbf{x}_j\|$ between molecular fingerprints, which are binary vector representations of the molecular graph. [29] Here, we represent the primary and secondary organic linkers as molecular fingerprints such that we can use the cosine kernel to measure the similarity between two sets of identity features. See Figure 1 for an example of how hMOFs are represented for the GP in this work. However, due to the length of the molecular fingerprint (here we use fingerprints of length 2048, but this can be set arbitrarily long by the user), the primary and secondary linker features will be heavily

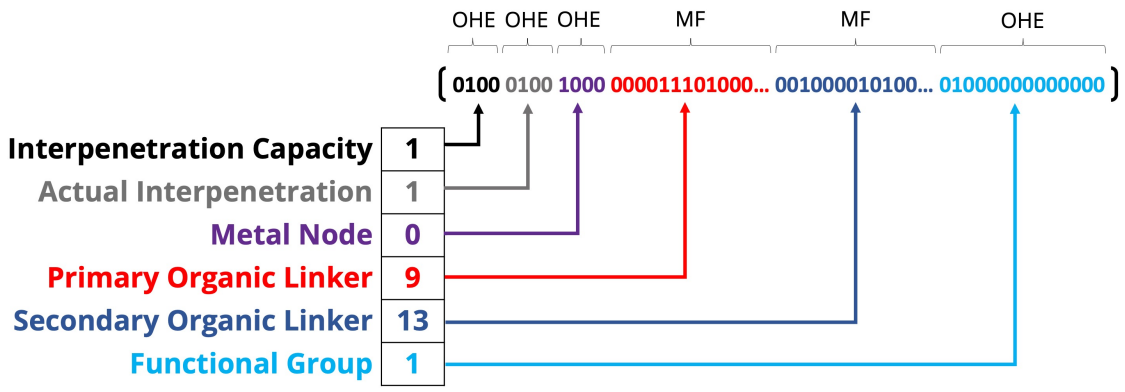

Figure 1: Schematic for how identity features of hMOFs are represented. OHE = one-hot encoding, MF = molecular fingerprint

weighed in the cosine kernel compared to the other identity features, which are represented by one-hot encoded vectors with lengths of less than 20. Thus, we use the weighted cosine kernel

$$k(\mathbf{x}_i, \mathbf{x}_j) = \sum_i \theta_i \frac{\mathbf{x}_i \cdot \mathbf{x}_j}{\|\mathbf{x}_i\|\|\mathbf{x}_j\|}, \tag{2}$$

where θ_i are weights for the i-th feature. Here, there are 6 such weights, each corresponding to one of the identity features listed above (also, see Table 1). Each θ_i are constrained to be positive and $\sum \theta_i = 1$. In doing so, each feature is not only length-independent, but also allows the user to interpret feature importance; higher weighted features are deemed more important than lesser weighted features.

2.2 Acquisition Function

Given the estimates of the objective function using the surrogate model on the test set, acquisition functions choose the next structure to evaluate. A key trade-off that any acquisition function must make is between exploration and exploitation. Exploration refers to evaluating MOF structures where the surrogate model is least confident to improve the model's overall accuracy across the search space. Exploitation refers to evaluating structures that the surrogate model believes will maximize/minimize the objective function. Throughout this study, we evaluate new points that maximize the upper confidence bound (UCB), which is expressed as

$$f_{UCB}(\mathbf{x}) = \mu(\mathbf{x}) + \beta\sigma(\mathbf{x}), \tag{3}$$

where μ is the prediction by the surrogate model, σ represents the uncertainty, and $\beta \geq 0$ is a constant that tunes the contributions of the uncertainty of the surrogate model. If $\beta = 0$, this amounts to a greedy search, in other words, fully exploitative, while $\beta \to \infty$ results in a fully explorative algorithm. Unless otherwise noted, any results found in the rest of this paper uses UCB with $\beta = 0$.

2.3 Dataset

Hypothetical MOF data are obtained mainly from Chung et al. [27], which contains a 51,000 hMOF subset of the original 130,000 as published by Wilmer et al. [5] Each hMOF is characterized by their interpenetration capacity, the actual interpenetration of the structure, the identity of the inorganic node, the organic linker, and the functional group additions to the organic linker, if any. Hereafter, these are referred to as *identity features*. These are extracted from the filenames of the reduced hMOF database at `https://github.com/snurr-group`. Wilmer et al. calculated the pore diameter, void fraction, gravimetric surface area, volumetric surface area, and density of each structure using Zeo++. [30] Hereafter, these features are referred to as *structural features*. Furthermore, they calculated the methane uptake capacities (in units of vol. STP CH_4 per vol. hMOF, hereafter shortened to v/v) ranging from 1 bar to 248 bar using grand canonical Monte Carlo (GCMC) as implemented in RASPA. [31] Here, we arbitrarily chose to optimize the CH_4 uptake capacity at 35 bar, as all structures used by Chung et al. had capacities computed at this pressure. The techniques presented here are not particular to this pressure of CH_4. Structural features and uptake capacities are

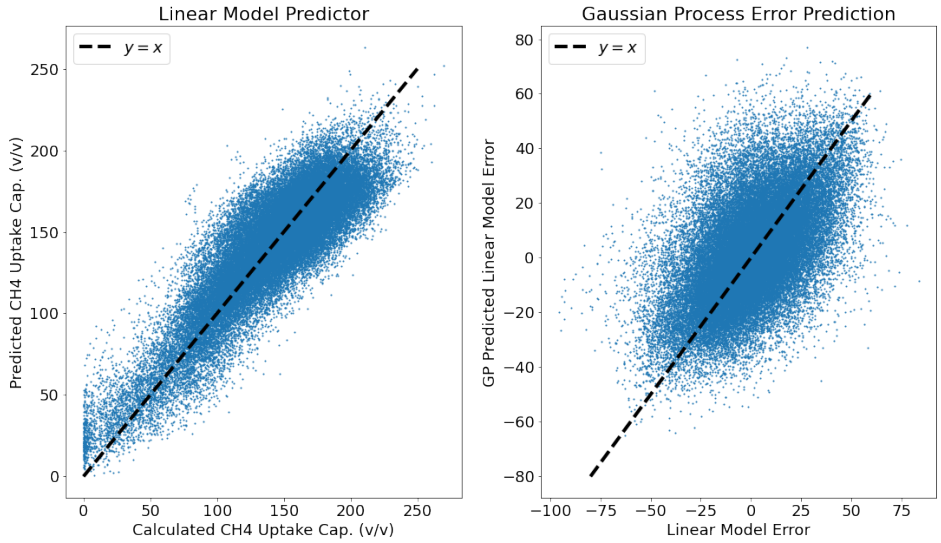

Figure 2: A combined linear model and error-correcting GP was trained on a random sample of 200 hMOFs. Parity plots of (left) the linear model regressing structural features against the methane uptake capacity at 35 bar, and (right) a GP with a weighted cosine kernel regressing the identity features against the error between the linear model and uptake capacity.)

incorporated by matching the hMOF ID in the CSV file for the reduced hMOF database with the hMOF ID found among the MOFs used by Chung et al. [27]

3 Results

3.1 Machine Learning

We initially perform linear regression using structural features to predict the CH_4 uptake capacity at 35 bar for a random sample of 200 hMOFs. As most methane adsorption processes rely on non-bonding interactions with the organic linkers involving van der Waals dispersion [32], the uptake capacity should scale well with surface area, and we expect such a basic model to perform reasonably well. From Figure 2 (left), a basic linear regression (LR) model based solely on structural features can predict the uptake capacity with reasonable accuracy (R^2 = 0.76, root mean squared error (RMSE) = 21.0 vol CH_4/vol hMOF). Regressing against larger pressure uptake data (for instance, at 100 bar or 248 bar) results in better fits, as exhibited by the increasing correlation coefficient R^2 with increasing pressure (see Figure 6). This suggests that structural features are more important in high-pressure regimes and less important at low-pressure regimes.

A GP trained using identity features is used to learn the error between predictions from the LR model on structural features and the uptake capacity as calculated by GCMC. As can be seen from Figure 2, the fit to test data is relatively poor (R^2 = 0.30, RMSE = 17.5 vol CH_4/vol hMOF). Combining this with the linear model, such that the final model is the sum of the linear model and the GP trained on the linear model error, results in a marginally improved model (R^2 = 0.82, RMSE = 17.5 vol CH_4/vol hMOF). As can be seen in Table 1, the weighted kernel reveals the importance of each identity feature. Both the primary and secondary linkers are weighted the most heavily, indicating higher importance. The functional group identity is important, but less so than the linkers, while the interpenetration features and the identity of the inorganic indicate the least importance. This matches chemical intuition, as the uptake capacity can be readily tuned by modifying the organic linker. [32]

The predictive performance of the combined LR and GP model using just 200 hMOFs as the training set (representing about 0.4% of the entire hMOF dataset) begs the question: how small of a training set can we use and still make usable predictions and extract chemical trends? We examine the performance on unseen test data using 20, 50, 100, 150, and 200 hMOFs as the training set and find that training set sizes below 150 show greater variability in predictive performance (see Figure 3, left). Moreover, the importance of the organic linker features is learned even with only

Pressure	Interpen. Capacity	Actual Interpen.	Inorganic Node	Functional Group	Primary Linker	Secondary Linker
35 bar	0.044 ± 0.013	0.046 ± 0.016	0.043 ± 0.011	0.159 ± 0.023	0.365 ± 0.087	0.342 ± 0.084
1 bar	0.020 ± 0.010	0.042 ± 0.024	0.039 ± 0.012	0.108 ± 0.027	0.424 ± 0.098	0.367 ± 0.087

Table 1: Normalized weights (θ_f) of the weighted cosine kernel (eq. 2) after training on random samples of 200 hMOFs and their corresponding 35 bar and 1 bar uptake capacity. 50 repeats were used to generate the mean, and intervals are given in $\pm$ 1 standard deviation. (Interpen. = Interpenetration)

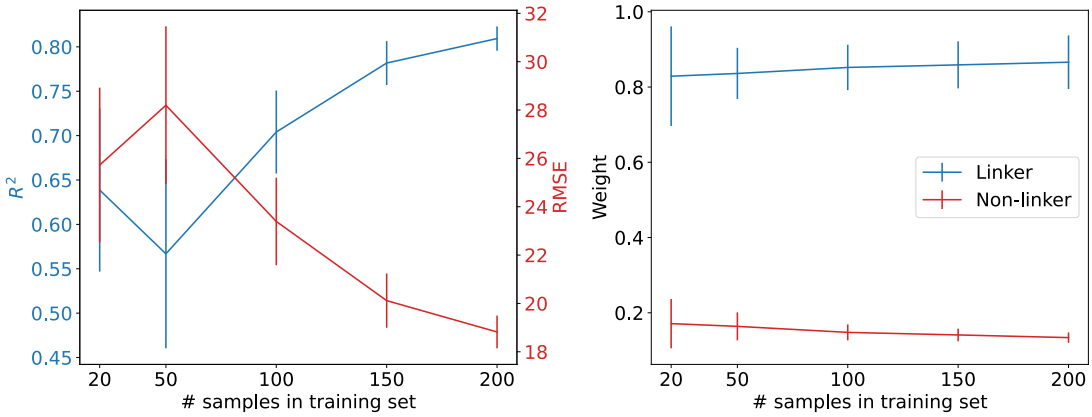

Figure 3: (left) R^2 and RMSE with increasing training size. (right) Sum of the weights of linker-related identity features (namely, the functional group, primary linker, and secondary linker) and non-linker identity features (interpenetration capacity, actual interpenetration, inorganic). Each training set was randomly sampled from the hMOF database. 50 repeats of each training set size was used to generate error bars, which represent $\pm$ 1 standard deviation.

20 samples (see Figure 3, right). This highlights the model's capability to make quantitative predictions and elucidate avenues of material design in the low-data regime.

3.2 Bayesian Optimization

Encouraged by the results even at 20 training examples, we use this model as the surrogate model in Bayesian optimization to find the highest uptake capacity hMOFs in the database and compared this to a LR model trained only on structural features. We begin with 50 random samples of 20 hMOFs, each of which act as starting points to Bayesian optimization. To provide a fair comparison, both surrogate models used, LR and LR+GP, are initiated with the same set of 20 hMOFs. Each iteration of BO adds another training example as specified by the acquisition function until a maximum of 50 training examples was evaluated.

35 bar As can be seen from Figure 4, both the LR model and LR + GP model are, on average, able to find high uptake capacity hMOFs while evaluating only 50 hMOFs (0.1% of the dataset). The LR + GP model converges to a top 10 candidate faster than the LR model, and all 50 repeats of BO obtain a top 10 candidate material. The LR model tends to converge somewhat slower, and a few trials of BO with LR were unable to find a candidate hMOF better than 220 v/v. However, $>$50% of the repeats with the LR model found the top candidate structure while the LR + GP model found one of the top 2 in 48/50 repeats. This can be attributed to overfitting of the LR + GP model in the high uptake capacity regime. Notably, both surrogate models find high-ranking candidates far more quickly than random sampling, though clearly the LR+GP model more reliably finds high uptake capacity hMOFs.

1 bar To show wider applicability of Bayesian optimization for gas uptake problems, we attempt to identify hMOFs with the highest CH_4 uptake capacity at 1 bar. This is a fundamentally different problem, as the gas uptake mechanism at higher pressures can be characterized by pore-filling for which structural features are likely to play a more important role in predicting uptake. [33] In contrast, low pressure gas uptake depends more heavily on the interactions between CH_4 and linker. [34]

Anomalies in the hMOF database requires further data processing prior to Bayesian optimization. Due to the lower uptake capacities, we remove any candidates with $>$33% error, a metric that is also provided by the database and is calculated based on the variance of prior Monte Carlo moves by Wilmer et al. [5] Moreover, some candidates have an

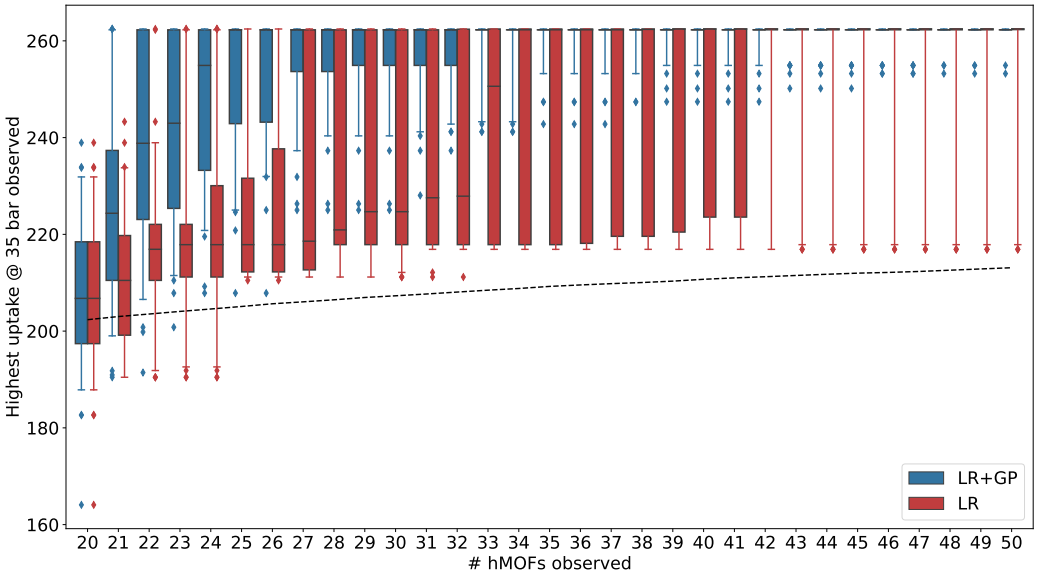

Figure 4: Highest uptake observed as Bayesian optimization progresses over 50 independent repeats with the LR surrogate model (red) and the LR+GP surrogate model (blue). Boxes encompass the 25% and 75% percentiles of uptake, while whiskers of the box plots indicate 5% and 95% percentiles. Diamond points represent outliers outside that range. The dashed line indicates the expected highest uptake of a random sample of the indicated size.

unexpectedly higher uptake at low pressure CH_4 than higher ones. Filtering out all candidates with $>$33% error and non-monotonically-increasing isotherms leaves 45,700 candidate hMOFs.

Starting with 20 randomly sampled hMOFs as its initial training set and ending with 50 training examples, as with the 35 bar case, Bayesian optimization was able to reliably find top 20 candidates with 43/50 repeats finding the top candidate in the entire dataset. Furthermore, the kernel weights for the organic linkers are higher than in the 35 bar case (Table 1), indicating greater sensitivity of the uptake capacity to the organic linkers at the low-pressure regime.

4 Conclusion and Future Work

We show in this study that

- Basic models such as linear regression are able to predict the uptake capacity of moderate-pressure CH_4 by MOFs using only structural features, such as pore diameter, surface area, and void fraction. The predictive accuracy of the linear model quickly drops as the bulk pressure decreases, implying that structural features matter less at lower pressures and a nonlinear model is necessary to estimate the uptake capacity in the low-pressure regime.
- Gaussian processes, with a cosine kernel, can enhance the linear model's performance using the identities of the inorganic node, the organic linkers, and functional groups on those linkers. Moreover, we observe that, even with an exceptionally small amount of training data, the surrogate model identifies the organic linkers as important features for predicting uptake capacities.
- Bayesian optimization performed with a combined linear model and Gaussian process results in quickly finding a top 10 CH_4 uptake candidate hMOF evaluating only 0.1% of the dataset that contains 51,000 possible candidates for a bulk pressure of 35 bar. The same model and hMOF representation is applicable for 1 bar bulk pressure data, indicating that this method still finds top candidate structures despite different relevant chemistry (ie high-pressure uptake typically correlates well with pore volume, while low-pressure adsorption correlates with adsorbate-adsorbent interactions).

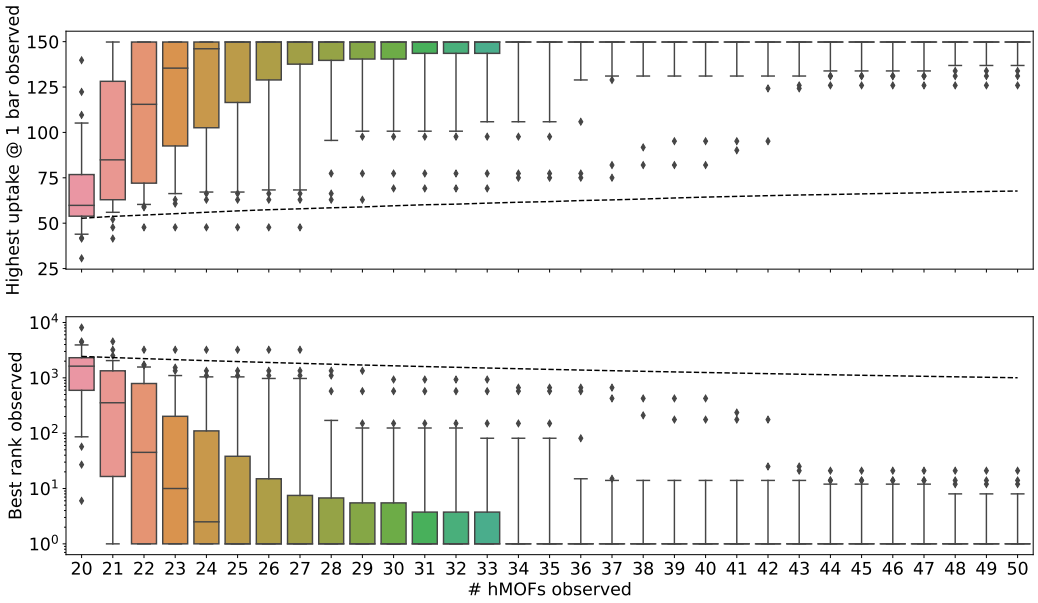

Figure 5: The LR + GP model also identifies high uptake candidates under 1 bar CH_4. The dashed line indicates the expected best uptake (top) and rank (bottom) if one were to randomly sample. (top) Starting with a random sample of 20 hMOFs, Bayesian optimization identifies hMOFs with uptake capacities of around 130 v/v or higher. (bottom) This corresponds to consistently identifying top 15 materials in the dataset.

This method can be readily applied to other problems pertinent to gas adsorption. As we rely heavily on structural parameters to predict accurate gas uptake capacities here, we expect gas uptake of other gasses that rely on dispersion forces to perform similarly well. Optimizing for high binding enthalpy at an undercoordinated metal site in a MOF is perhaps more interesting, as such predictions generally require DFT and is more computationally expensive. [35] Bayesian optimization would be able to realize greater time savings in such an application. We also note that the models presented here can be modified to predict within known physical constraints, thus improving prediction accuracy. [36] Furthermore, we anticipate exploring the acquisition function, as a pure greedy search used here is likely suboptimal. [37, 38]

5 Acknowledgements and Conflicts of Interest

This work is primarily funded by the U.S. Department of Energy, Office of Science, Office of Basic Energy Sciences, Chemical Sciences, Geosciences, & Biosciences Division under Grant No. DE-SC0019992. The authors have no conflicts of interest to declare.

References

[1] H. Li, K. Wang, Y. Sun, C. T. Lollar, J. Li, and H.-C. Zhou, "Recent advances in gas storage and separation using metal–organic frameworks," *Materials Today*, vol. 21, pp. 108–121, Mar. 2018.

[2] H. Furukawa, K. E. Cordova, M. O'Keeffe, and O. M. Yaghi, "The Chemistry and Applications of Metal-Organic Frameworks," *Science*, vol. 341, Aug. 2013.

[3] Y. J. Colón, D. A. Gómez-Gualdrón, and R. Q. Snurr, "Topologically Guided, Automated Construction of Metal–Organic Frameworks and Their Evaluation for Energy-Related Applications," *Crystal Growth & Design*, vol. 17, pp. 5801–5810, Nov. 2017.

[4] R. Anderson and D. A. Gómez-Gualdrón, "Increasing topological diversity during computational "synthesis" of porous crystals: How and why," *CrystEngComm*, vol. 21, pp. 1653–1665, Mar. 2019.

[5] C. E. Wilmer, M. Leaf, C. Y. Lee, O. K. Farha, B. G. Hauser, J. T. Hupp, and R. Q. Snurr, "Large-scale screening of hypothetical metal–organic frameworks," *Nature Chemistry*, vol. 4, pp. 83–89, Feb. 2012.

[6] P. G. Boyd, A. Chidambaram, E. García-Díez, C. P. Ireland, T. D. Daff, R. Bounds, A. Gładysiak, P. Schouwink, S. M. Moosavi, M. M. Maroto-Valer, J. A. Reimer, J. A. R. Navarro, T. K. Woo, S. Garcia, K. C. Stylianou, and B. Smit, "Data-driven design of metal–organic frameworks for wet flue gas CO 2 capture," *Nature*, vol. 576, pp. 253–256, Dec. 2019.

[7] N. S. Bobbitt, J. Chen, and R. Q. Snurr, "High-Throughput Screening of Metal–Organic Frameworks for Hydrogen Storage at Cryogenic Temperature," *The Journal of Physical Chemistry C*, vol. 120, pp. 27328–27341, Dec. 2016.

[8] S. Li, Y. G. Chung, and R. Q. Snurr, "High-Throughput Screening of Metal–Organic Frameworks for CO2 Capture in the Presence of Water," *Langmuir*, vol. 32, pp. 10368–10376, Oct. 2016.

[9] B. J. Sikora, C. E. Wilmer, M. L. Greenfield, and R. Q. Snurr, "Thermodynamic analysis of Xe/Kr selectivity in over 137 000 hypothetical metal–organic frameworks," *Chemical Science*, vol. 3, pp. 2217–2223, June 2012.

[10] Y. G. Chung, D. A. Gómez-Gualdrón, P. Li, K. T. Leperi, P. Deria, H. Zhang, N. A. Vermeulen, J. F. Stoddart, F. You, J. T. Hupp, O. K. Farha, and R. Q. Snurr, "In silico discovery of metal-organic frameworks for precombustion CO2 capture using a genetic algorithm," *Science Advances*, vol. 2, p. e1600909, Oct. 2016.

[11] I. Y. Kanal, J. A. Keith, and G. R. Hutchison, "A Sobering Assessment of Small-Molecule Force Field Methods for Low Energy Conformer Predictions," *arXiv:1705.04308 [physics]*, Aug. 2017.

[12] F. Guistino, *Materials Modelling Using Density Functional Theory: Properties and Predictions*. Oxford University Press, 2014.

[13] A. S. Rosen, M. R. Mian, T. Islamoglu, H. Chen, O. K. Farha, J. M. Notestein, and R. Q. Snurr, "Tuning the Redox Activity of Metal–Organic Frameworks for Enhanced, Selective O2 Binding: Design Rules and Ambient Temperature O2 Chemisorption in a Cobalt–Triazolate Framework," *Journal of the American Chemical Society*, vol. 142, pp. 4317–4328, Mar. 2020.

[14] H. C. Herbol, W. Hu, P. Frazier, P. Clancy, and M. Poloczek, "Efficient search of compositional space for hybrid organic–inorganic perovskites via Bayesian optimization," *npj Computational Materials*, vol. 4, pp. 1–7, Sept. 2018.

[15] M. Yamawaki, M. Ohnishi, S. Ju, and J. Shiomi, "Multifunctional structural design of graphene thermoelectrics by Bayesian optimization," *Science Advances*, vol. 4, p. eaar4192, June 2018.

[16] C. Altintas, G. Avci, H. Daglar, A. N. V. Azar, I. Erucar, S. Velioglu, and S. Keskin, "An extensive comparative analysis of two MOF databases: High-throughput screening of computation-ready MOFs for CH 4 and H 2 adsorption," *Journal of Materials Chemistry A*, vol. 7, no. 16, pp. 9593–9608, 2019.

[17] G. Avci, S. Velioglu, and S. Keskin, "High-Throughput Screening of MOF Adsorbents and Membranes for H2 Purification and CO2 Capture," *ACS Applied Materials & Interfaces*, vol. 10, pp. 33693–33706, Oct. 2018.

[18] P. Canepa, C. A. Arter, E. M. Conwill, D. H. Johnson, B. A. Shoemaker, K. Z. Soliman, and T. Thonhauser, "High-throughput screening of small-molecule adsorption in MOF," *Journal of Materials Chemistry A*, vol. 1, pp. 13597–13604, Oct. 2013.

[19] Y. G. Chung, J. Camp, M. Haranczyk, B. J. Sikora, W. Bury, V. Krungleviciute, T. Yildirim, O. K. Farha, D. S. Sholl, and R. Q. Snurr, "Computation-Ready, Experimental Metal–Organic Frameworks: A Tool To Enable High-Throughput Screening of Nanoporous Crystals," *Chemistry of Materials*, vol. 26, pp. 6185–6192, Nov. 2014.

[20] D. Goldberg, *Genetic Algorithms in Search, Optimization and Machine Learning*. Addison-Wesley Professional, thirteenth ed., 1988.

[21] R. Calandra, A. Seyfarth, J. Peters, and M. P. Deisenroth, "Bayesian optimization for learning gaits under uncertainty," *Annals of Mathematics and Artificial Intelligence*, vol. 76, pp. 5–23, Feb. 2016.

[22] F. Hutter, H. H. Hoos, and K. Leyton-Brown, "Sequential Model-Based Optimization for General Algorithm Configuration," in *Learning and Intelligent Optimization* (C. A. C. Coello, ed.), Lecture Notes in Computer Science, (Berlin, Heidelberg), pp. 507–523, Springer, 2011.

[23] B. Shahriari, K. Swersky, Z. Wang, R. P. Adams, and N. de Freitas, "Taking the Human Out of the Loop: A Review of Bayesian Optimization," *Proceedings of the IEEE*, vol. 104, pp. 148–175, Jan. 2016.

[24] B. Solnik, D. Golovin, G. Kochanski, J. E. Karro, S. Moitra, and D. Sculley, "Bayesian Optimization for a Better Dessert," in *Proceedings of the 2017 NIPS Workshop on Bayesian Optimization*, (December 9, 2017, Long Beach, USA), 2017.

[25] P. I. Frazier, "A Tutorial on Bayesian Optimization," *arXiv:1807.02811 [cs, math, stat]*, July 2018.

[26] J. R. Gardner, G. Pleiss, D. Bindel, K. Q. Weinberger, and A. G. Wilson, "GPyTorch: Blackbox Matrix-Matrix Gaussian Process Inference with GPU Acceleration," *arXiv:1809.11165 [cs, stat]*, Jan. 2019.

[27] Y. G. Chung, D. A. Gómez-Gualdrón, P. Li, K. T. Leperi, P. Deria, H. Zhang, N. A. Vermeulen, J. F. Stoddart, F. You, J. T. Hupp, O. K. Farha, and R. Q. Snurr, "In silico discovery of metal-organic frameworks for precombustion CO2 capture using a genetic algorithm," *Science Advances*, vol. 2, p. e1600909, Oct. 2016.

[28] E. C. Garrido-Merchán and D. Hernández-Lobato, "Dealing with Categorical and Integer-valued Variables in Bayesian Optimization with Gaussian Processes," *Neurocomputing*, vol. 380, pp. 20–35, Mar. 2020.

[29] D. Rogers and M. Hahn, "Extended-Connectivity Fingerprints," *Journal of Chemical Information and Modeling*, vol. 50, pp. 742–754, May 2010.

[30] T. F. Willems, C. H. Rycroft, M. Kazi, J. C. Meza, and M. Haranczyk, "Algorithms and tools for high-throughput geometry-based analysis of crystalline porous materials," *Microporous and Mesoporous Materials*, vol. 149, pp. 134–141, Feb. 2012.

[31] D. Dubbeldam, S. Calero, D. E. Ellis, and R. Q. Snurr, "RASPA: Molecular simulation software for adsorption and diffusion in flexible nanoporous materials," *Molecular Simulation*, vol. 42, pp. 81–101, Jan. 2016.

[32] E. Bichoutskaia, M. Suyetin, M. Bound, Y. Yan, and M. Schröder, "Methane Adsorption in Metal–Organic Frameworks Containing Nanographene Linkers: A Computational Study," *The Journal of Physical Chemistry C*, vol. 118, pp. 15573–15580, July 2014.

[33] N. D. Hutson and R. T. Yang, "Theoretical basis for the Dubinin-Radushkevitch (D-R) adsorption isotherm equation," *Adsorption*, vol. 3, pp. 189–195, Sept. 1997.

[34] M. Zhang, W. Zhou, T. Pham, K. A. Forrest, W. Liu, Y. He, H. Wu, T. Yildirim, B. Chen, B. Space, Y. Pan, M. J. Zaworotko, and J. Bai, "Fine Tuning of MOF-505 Analogues To Reduce Low-Pressure Methane Uptake and Enhance Methane Working Capacity," *Angewandte Chemie International Edition*, vol. 56, no. 38, pp. 11426–11430, 2017.

[35] C. Campbell, J. R. B. Gomes, M. Fischer, and M. Jorge, "New Model for Predicting Adsorption of Polar Molecules in Metal–Organic Frameworks with Unsaturated Metal Sites," *The Journal of Physical Chemistry Letters*, vol. 9, pp. 3544–3553, June 2018.

[36] L. Swiler, M. Gulian, A. Frankel, C. Safta, and J. Jakeman, "A Survey of Constrained Gaussian Process Regression: Approaches and Implementation Challenges," *Journal of Machine Learning for Modeling and Computing*, vol. 1, no. 2, pp. 119–156, 2020.

[37] T. Desautels, A. Krause, and J. W. Burdick, "Parallelizing Exploration-Exploitation Tradeoffs in Gaussian Process Bandit Optimization," *Journal of Machine Learning Research*, vol. 15, no. 119, pp. 4053–4103, 2014.

[38] F. Archetti and A. Candelieri, "The Acquisition Function," in *Bayesian Optimization and Data Science* (F. Archetti and A. Candelieri, eds.), SpringerBriefs in Optimization, pp. 57–72, Cham: Springer International Publishing, 2019.

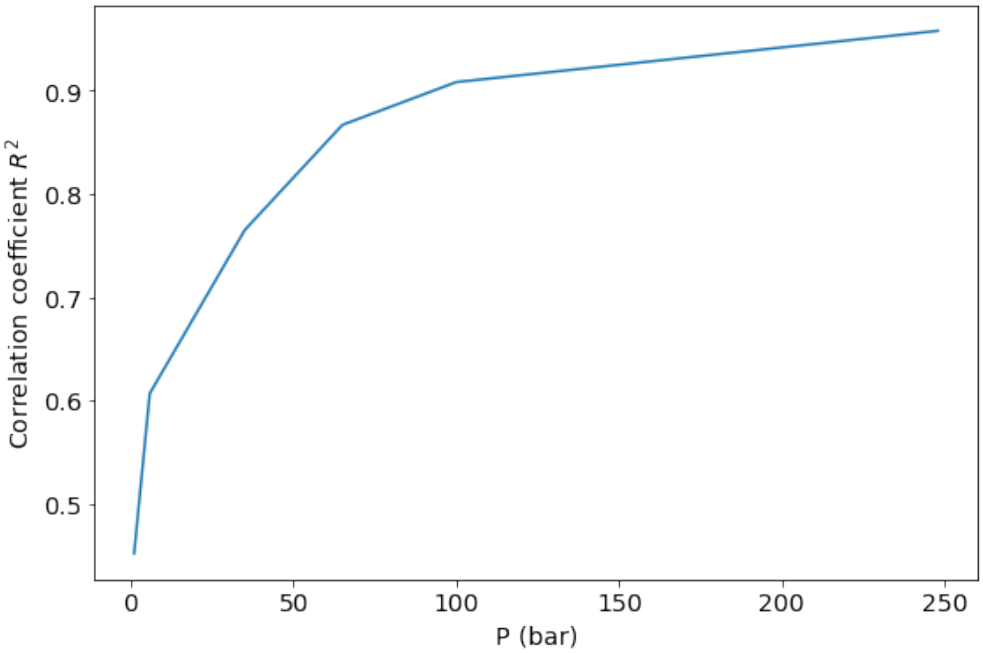

Figure 6: Correlation coefficient of linear models trained on structural features of all hMOFs where data was available. The linear model has much better predictions for higher pressure data.

A Linear Regression of Structural Features against Uptake Capacity

The simplest possible surrogate model is a linear model on structural features; this serves as a baseline for comparison against more complex models and as a basis for the LR+GP model described in this study. Below, we plot the correlation coefficients (R^2) as a function of increasing pressure CH_4 using a linear model. As seen here, increasing pressure allows for a more predictive linear model and suggests the key parameters to optimize a MOF for high pressure CH_4 uptake is purely structural. In contrast, low pressure CH_4 uptake is not predicted well by a linear model with structural features, so this suggests such a model could benefit from additional nonlinearity or more features.