Predicting the Surface Area of Metal-Organic Frameworks

Jay Siri

Abstract

Metal-organic frameworks (MOFs) are a class of nanoporous material that have applications in gas storage, catalysis, drug delivery, and many other fields. The estimation of a MOF's porosity is typically calculated as its specific surface area, either using Monte Carlo simulations or by Brunauer–Emmett–Teller (BET) approximations. In this report, we train machine learning models and compare their performance to implemented BET estimations on the CoreMOF 2019 dataset using the MOFxDB open source project. We find that machine learning methods are able to outperform BET estimation, thereby allowing for the possibility of more accurate MOF surface area predictions.

Introduction

In 2025, the Nobel Prize in Chemistry was awarded to Susumu Kitagawa, Richard Robson and Omar Yaghi "for the development of metal–organic frameworks" [10]. Metal-organic frameworks, also abbreviated as MOFs, are nanoporous, crystalline materials composed of metal ions linked together by organic components in repeating units. Due to their high porosity and relative ease of synthesis, MOFs have been explored as an exciting avenue for applications in gas adsorption (for instance, for gas storage and carbon capture), drug delivery, and catalysis [10]. Therefore, being able to estimate the surface area of MOFs and subsequently, design specific MOFs with high surface areas, is a desirable goal [6] [5].

Background

I. Literature Review

The typical method of estimating the surface area of nanoporous materials is done using Brunauer–Emmett–Teller (BET) theory, where N_2 is systematically adsorbed onto the surface of the material. The resulting isotherm can then be analyzed to approximate the material's surface area [8]. With respect to MOF surface area, BET is typically seen as the canonical calculation; Bae *et al.*, for example, apply this method to a variety of MOFs and confirm the method's validity, and Terrones *et al.* benchmark BET and other surface area estimation methods, including computational methods based on Monte Carlo simulation, to a similar range [1][12].

However, other studies have suggested that BET surface area estimation fails to accurately predict surface area by significant margins, and that machine learning (ML) may provide more accurate predictions [4]. In one such study, Datar *et al.* find that BET surface predictions overestimate the true surface area of large-surface area MOFs ($> 3500 m^2/g$) by over $> 20\%$ 54% of the time, where a Lasso linear regression model they fit only overpredicts 2% of the time and is much more accurate in general. [4]

Further, advents in the rapidly growing field of machine learning have tried, to varying degrees of success, to incorporate domain knowledge, often via enforcing some prior, to improve modeling. A survey by Park *et al.* discusses the different methods in machine learning that have been applied to studying MOFs, including feature engineering, graph neural networks, transformers, and machine-learned potentials to speed up molecular simulations of MOFs [9]. One related advancment in machine learning, but yet unexplored area with respect to MOFs, is physics-informed machine learning, of which one avenue includes incorporating the equations governing physical phenomena into the model. [11][7]

II. Dataset

In this project, we analyze and explore the efficacy of the aforementioned methods (further described in Methods) on a MOF dataset. The data in this project is queried from MOFX-DB, a large database of over 160,000 MOFs and their adsorption data [2]. Each MOF entry in the database has data such as the atoms in the MOF, isotherm data for various gases including N_2, Xe, etc., the crystal structure of the MOF, and textural properties like the pore-limiting diameter, void fraction, and surface area. The database schema of MOFX-DB and the attributes available are shown in Figure 1.

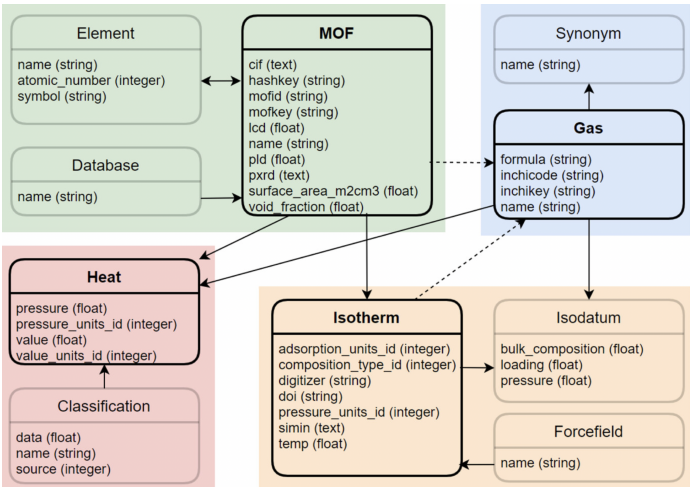


Figure 1: The MOFX-DB schema, from "MOFX-DB: An Online Database of Computational Adsorption Data for Nanoporous Materials".

Due to assumptions of the BET surface area approximation (described in Methods), it is worth noting that the only a subset of the database may be used for predictions involving BET; these MOFs come from the CoreMOF dataset, which has over 12,000 MOFs with N_2 isotherms that were simulated using RASPA and validated with experimental data [3]. The surface areas of the MOFs in this dataset were determined using the Zeo++ software [2]. In the end, we utilized 1000 MOFs from the CoreMOF 2019 dataset. The distribution of ground-truth surface areas is available in Figure 2.

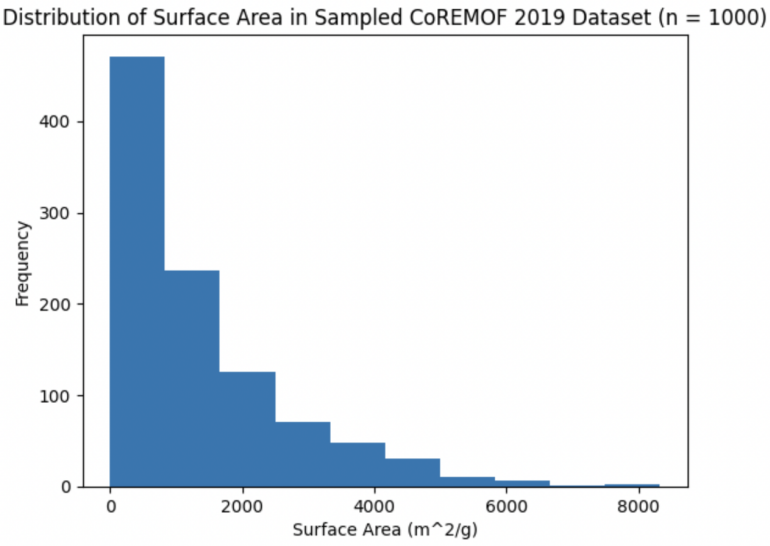


Figure 2: Histogram of 1,000 MOF surface areas from a subset of the CoreMOF 2019 dataset.

Methods

In this project, we explore and compare methods of predicting the surface area of metal-organic frameworks. First, we use rudimentary BET theory to approximate the surface area from the N_2 isotherm provided in the MOFX-DB dataset. Second, we train basic machine

learning models on the MOFX-DB dataset to predict each MOF's surface area from all the features provided in the dataset. Third, we try to combine both theory and machine learning by defining a custom loss function that incorporates BET constraints during training to guide the model's predictions.

I. BET Theory

The Brunauer–Emmett–Teller (BET) theory stipulates that the gas monolayer adsorption isotherm can be described by the equation:

$$\frac{1}{X[(\frac{P_0}{P}-1)]} = \frac{1}{X_m C} + \frac{C-1}{X_m C}(\frac{P_0}{P}) \tag{1}$$

where $\frac{P_0}{P}$ is the relative pressure, X_m is the monolayer capacity (the volume of gas adsorbed at STP), X is the weight of nitrogen adsorbed, and C is a BET constant [8]. With a N_2 adsorption isotherm, BET theory tells us that we can find a least-squares regression best-fit line through the isotherm, find the line's slope and intercept, and calculate:

$$X_m = \frac{1}{slope + intercept} \tag{2}$$

Then, we may calculate the total surface area S as:

$$S = \frac{X_m N_A A_m}{M_v} \tag{3}$$

where N_A is Avogadro's number $6.022 \times 10^{23}\ mols^{-1}$, A_m is the cross-sectional area of N_2 $1.62 \times 10^{-19}\ m^2$, and M_v is the molar volume at STP $22,414\ cm^3/mol$ [8].

In our case, since MOFX-DB provides the loading (e.g., cm^3/g) rather than volume (e.g., cm^3) of gas adsorbed, S is our resultant specific surface area. However, in order to use BET, we typically need the N_2 adsorption isotherm at $77K$ and also require data in the lower relative pressure range [8]. These assumptions limit our dataset to the CoreMOF datapoints of the MOFX-DB dataset. After filtering MOFX-DB to retrieve just appropriate data points, we calculate each MOF's monolayer capacity X_m by fitting a least-squares regression line to the isotherm data in the $0.025 - 0.35$ relative pressure range using `stats.linregress` and then calculating the surface area.

II. Machine Learning Models

Instead of using just the N_2 isotherm data, it may be advantageous to use more of the data available to infer the surface area. To do so, we train machine learning regression models, using numerical data available in MOFX-DB, especially all available gas isotherm data, as input features. The any information directly relating to the surface area was left out as the prediction label. The following models were trained:

- Linear Regression: implemented via `sklearn.linear_model.LinearRegression`, this model performs ordinary least-squares linear regression with a squared-error loss.

- Least Absolute Shrinkage and Selection Operator Regression (LASSO): implemented via `sklearn.linear_model.Lasso`, this model performs least-squares linear regression

with a regularization term and parameter ($\alpha = 0.1$) to penalize model coefficients, thereby selecting for only more important features.

- eXtreme Gradient Boosting (XGBoost): implemented via `xgboost.XGBRegressor`, this model creates an ensemble of many decision trees and then sums each tree's result to give a final prediction. We choose a squared error loss for this model.

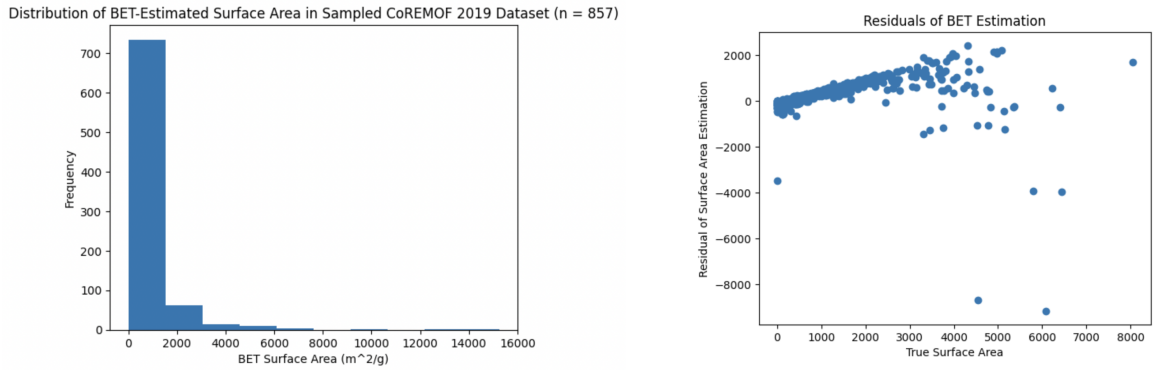
- Multi-Layer Perceptron (MLP): implemented via `sklearn.neural_network.MLPRegressor`, this is a simple neural network of 3 hidden layers, each with 15 nodes, with a squared error loss function.

In order to evaluate the models, we perform validation on a test set, leaving out 20% of the data as validation data and training on the other 80%.

Results and Discussion

For each method (BET and each of the four ML methods), we ran inference on the test dataset, then visualized the residuals for each data point in the test dataset. The distribution of surface areas was also visualized. Lastly, we calculated four error metrics on the test data: R-squared, mean squared error (MSE), mean absolute error (MAE), and mean absolute percentage error (MAPE).

In Figure 3 below, we see that BET is able to estimate the surface area with appreciable accuracy, but that it tends to overestimate a few data points to high surface areas. From the residual plot, we can also see that there is a clear positive bias since the residuals are not centered around zero, though there some very negative residuals where the estimation was much lower than the ground truth surface area.

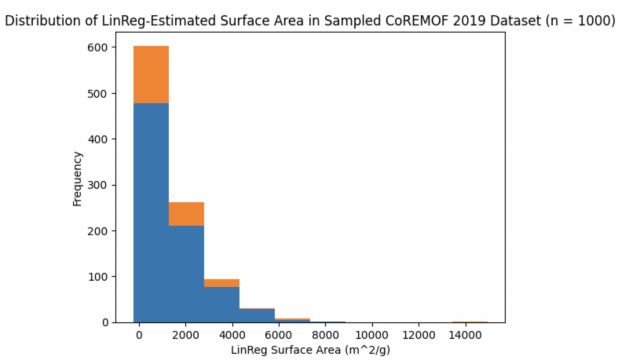


(a) BET Surface Areas

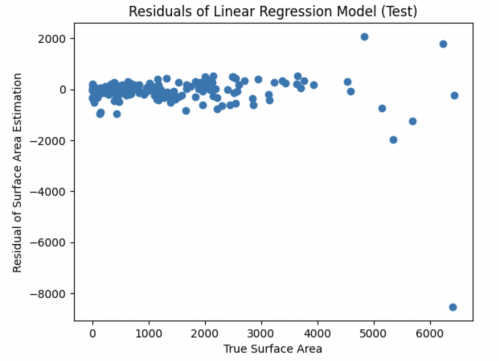
(b) BET Residuals

Figure 3: Results of BET estimation

For the machine learning models, we visualized the same results, but inference was performed on training data and visualized in the surface area distribution histograms as well, for completeness (in each histogram in Figures 4-7, orange is test data predictions, and blue is train data predictions). We see that the residuals for all the models are centered around zero, and that higher surface area predictions seem to be more difficult for these models. For both BET and each ML method, the distribution of estimated surface areas approximately matches the distribution of the ground truth surface area data (Figure 2).

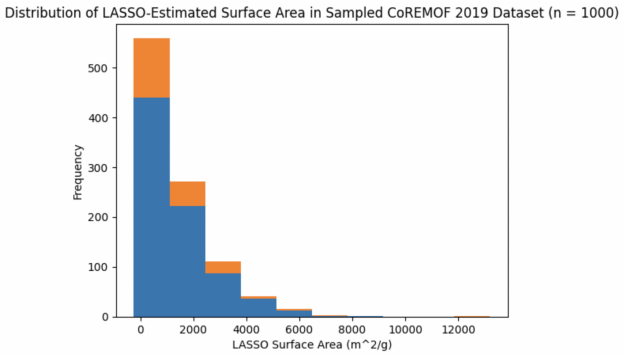


(a) Linear Regression Surface Areas

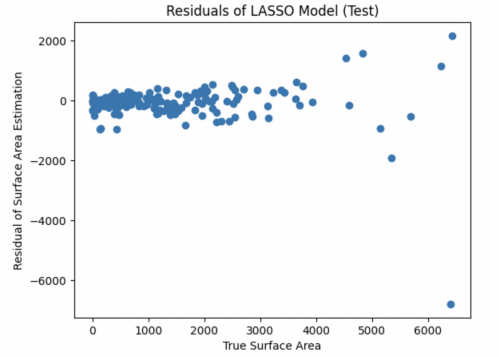


(b) Linear Regression Residuals

Figure 4: Results of Linear Regression

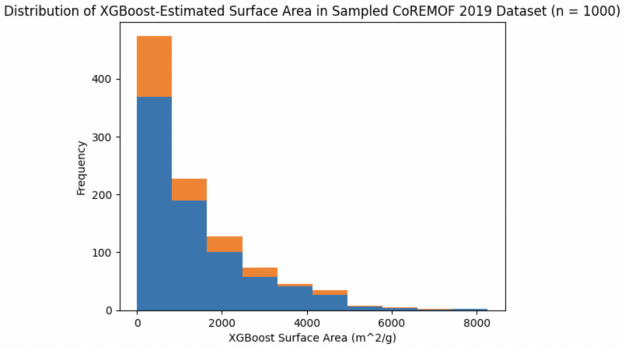


(a) LASSO Surface Areas

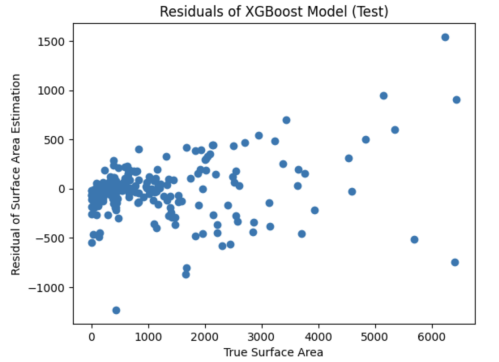


(b) LASSO Residuals

Figure 5: Results of LASSO

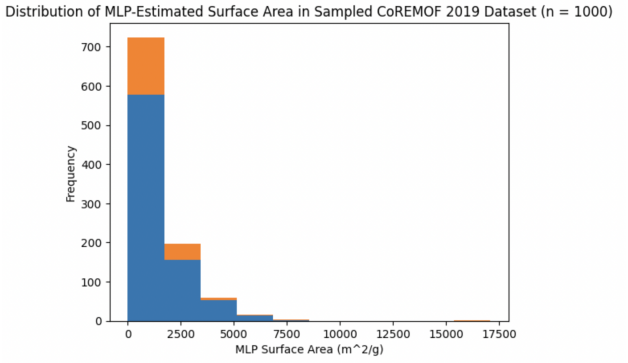


(a) Gradient Boosting Surface Areas



(b) Gradient Boosting Residuals

Figure 6: Results of Gradient Boosting



(a) Neural Network Surface Areas

(b) Neural Network Residuals

Figure 7: Results of Neural Network (Multi-Layer Perceptron)

From Table 1 below, we can see that gradient boosting (XGBoost) outperforms all other methods by most metrics, and that the model seems to be inherently more regularized and resistant to extreme, outlier predictions. Linear models are able to outperform BET as well. The deep learning method does not perform well, which may be due to the small size of the training dataset.

Method	**# of Test Data Points**	**R-squared**	**MSE**	**MAE**	**MAPE**
BET	857	0.616	542,267	410	9.32
Linear Regression	200	0.707	507,426	281	**3.05**
LASSO	200	0.779	382,579	278	3.06
Gradient Boosting	200	**0.949**	**88,857**	**199**	4.96
Neural Network	200	0.574	737,047	310	6.54

Table 1: Metrics on Surface Area Estimations

Conclusion

We show that machine learning models, especially gradient boosting (XGBoost), are able to outperform BET theory estimation on a subset of the CoreMOF 2019 dataset. Machine learning methods, however, require large amounts of labeled training data, whereas BET works out-of-the-box, from principles.

One future area of investigation could be to try to take advantage of both the sheer quantity of data and features available in MOFX-DB (as in traditional machine learning) as well as the constraints enforced by theory (as in BET theory) by using physics-informed machine learning. For example, this may be implemented by defining a custom loss function that accounts for BET estimations, then using that to train machine learning models. It is also worth noting that the models in this report were largely un-optimized, so further work could tune these models, improve architectures, expand training data, etc. Lastly, the ultimate goal for this area of research would be to estimate the surface areas without the isotherm data at all; rather, learning from other feature data.

Nevertheless, this report shows that it is possible to improve upon results from theory with machine learning methods. By improving on these surface areas estimations, it may be

possible to provide more accurate estimates of the true character of MOF structures, thereby accelerating research and development, application, and ultimately, adoption.

Code Availability

The complete code for all calculations and figures in this report can be found at: `https://github.com/jaypsiri/10.585_Final_Project`.

References

[1] Youn-Sang Bae, A. Özgür Yazaydın, and Randall Q. Snurr. "Evaluation of the BET Method for Determining Surface Areas of MOFs and Zeolites that Contain Ultra-Micropores". In: *Langmuir* 26.8 (Apr. 2010). Publisher: American Chemical Society, pp. 5475–5483. ISSN: 0743-7463. DOI: 10.1021/la100449z. URL: https://doi.org/10.1021/la100449z.

[2] N. Scott Bobbitt et al. "MOFX-DB: An Online Database of Computational Adsorption Data for Nanoporous Materials". In: *Journal of Chemical & Engineering Data* 68.2 (Feb. 2023). Publisher: American Chemical Society, pp. 483–498. ISSN: 0021-9568. DOI: 10.1021/acs.jced.2c00583. URL: https://doi.org/10.1021/acs.jced.2c00583.

[3] Yongchul G. Chung et al. "Advances, Updates, and Analytics for the Computation-Ready, Experimental Metal–Organic Framework Database: CoRE MOF 2019". In: *Journal of Chemical & Engineering Data* 64.12 (Dec. 2019). Publisher: American Chemical Society, pp. 5985–5998. ISSN: 0021-9568. DOI: 10.1021/acs.jced.9b00835. URL: https://doi.org/10.1021/acs.jced.9b00835.

[4] Archit Datar, Yongchul G. Chung, and Li-Chiang Lin. "Beyond the BET Analysis: The Surface Area Prediction of Nanoporous Materials Using a Machine Learning Method". In: *The Journal of Physical Chemistry Letters* 11.14 (July 2020). Publisher: American Chemical Society, pp. 5412–5417. DOI: 10.1021/acs.jpclett.0c01518. URL: https://doi.org/10.1021/acs.jpclett.0c01518.

[5] Omar K. Farha et al. "Metal–Organic Framework Materials with Ultrahigh Surface Areas: Is the Sky the Limit?" In: *Journal of the American Chemical Society* 134.36 (Sept. 2012). Publisher: American Chemical Society, pp. 15016–15021. ISSN: 0002-7863. DOI: 10.1021/ja3055639. URL: https://doi.org/10.1021/ja3055639.

[6] Emmett D. Goodman, Chengshuang Zhou, and Matteo Cargnello. "Design of Organic/Inorganic Hybrid Catalysts for Energy and Environmental Applications". In: *ACS Central Science* 6.11 (Nov. 2020). Publisher: American Chemical Society, pp. 1916–1937. ISSN: 2374-7943. DOI: 10.1021/acscentsci.0c01046. URL: https://doi.org/10.1021/acscentsci.0c01046.

[7] George Em Karniadakis et al. "Physics-informed machine learning". In: *Nature Reviews Physics* 3.6 (June 2021), pp. 422–440. ISSN: 2522-5820. DOI: 10.1038/s42254-021-00314-5. URL: https://doi.org/10.1038/s42254-021-00314-5.

[8] Majid Naderi. "Chapter Fourteen - Surface Area: Brunauer–Emmett–Teller (BET)". In: *Progress in Filtration and Separation*. Ed. by Steve Tarleton. Oxford: Academic Press, Jan. 2015, pp. 585–608. ISBN: 978-0-12-384746-1. DOI: 10.1016/B978-0-12-384746-1.00014-8. URL: https://www.sciencedirect.com/science/article/pii/B9780123847461000148.

[9] Junkil Park et al. "From Data to Discovery: Recent Trends of Machine Learning in Metal–Organic Frameworks". In: *JACS Au* 4.10 (Oct. 2024). Publisher: American Chemical Society, pp. 3727–3743. DOI: 10.1021/jacsau.4c00618. URL: https://doi.org/10.1021/jacsau.4c00618.

[10] *Press release*. Publisher: NobelPrize.org. Oct. 2025. URL: https://www.nobelprize.org/prizes/chemistry/2025/press-release/.

[11] Sam. Raymond. *Physics-Informed Machine Learning Using the Laws of Nature to Improve Generalized Deep Learning Models Introduction -Fusing Data and Simulation.* Publisher: MathWorks. URL: https://www.mathworks.com/content/dam/mathworks/mathworks-dot-com/images/responsive/supporting/events/matlab-expo-2021/proceedings/matlab-expo-2021-physics-informed-machine-learning-using-the-laws-of-nature-to-improve-generalized-deep-learning-models-edt.pdf.

[12] Gianmarco Terrones et al. "SESAMI APP: An Accessible Interface for Surface Area Calculation of Materials from Adsorption Isotherms". In: *Journal of Open Source Software* 8 (June 2023), p. 5429. DOI: 10.21105/joss.05429.