Tree-based Quantile Active Learning for automated discovery of MOFs

Ashna Jose[1*], **Emilie Devijver**[2], **Noël Jakse**[1], **Valérie Monbet**[3], **Roberta Poloni**[1]
[1]Univ. Grenoble Alpes, CNRS, Grenoble INP, SIMaP, F-38000 Grenoble, France
[2]Univ. Grenoble Alpes, CNRS, Grenoble INP, LiG, F-38000 Grenoble, France
[3]IRMAR, CNRS, UMR-CNRS 6625, Rennes, 35000, France
[*]ashna.jose@grenoble-inp.fr

Abstract

Metal-organic frameworks (MOFs), formed through coordination bonds between metal ions and organic ligands, are promising materials for efficient gas adsorption, due to their ultrahigh porosity, chemical tunability and large surface area. Because over a hundred thousand hypothetical MOFs have been reported to date, brute force discovery of the best performer MOF for a specific application is not feasible. Recently, predicting material properties using machine learning algorithms has played a crucial role in scanning large databases, but this often requires large labeled training sets, which is not always available. To address this, active learning, where the training set is constructed iteratively by querying only informative labels, is necessary. Moreover, in most cases, a very specific range of the property of interest is desirable. We employ a novel regression tree-based quantile active learning algorithm that uses partitions of a regression tree to select new samples to be added to the training set. It thereby limits the sample size while maximizing the prediction quality over a quantile of interest. Tests on benchmark MOF data sets demonstrate that focusing on a specific quantile is effective in learning regression models to predict electronic band gaps and CO_2 adsorption in the regions of interest, from a very limited labeled data set.

1 Introduction

Metal-organic frameworks (MOFs) [1, 2], formed through coordination bonds between metal ions and organic ligands, are promising materials for efficient gas capture and separation [3, 4], due to their ultrahigh porosity, chemical tunability and large surface area [5, 6]. They have also been shown to be potential candidates for water harvesting [7], catalysis [8] and sensing [9], thus evoking an interest in their electronic properties [10, 11, 12]. Since they are built with metal nodes and organic linkers, the myriads of possible combinations of these lead to innumerable MOFs which makes discovery of novel MOFs with a certain property of interest challenging. As experimentally synthesizing such large numbers of MOFs is not viable, computational techniques like density functional theory (DFT) [13] and molecular simulations [14] have been used to screen them with a relatively low cost.

In addition to high throughput screening approaches, machine learning (ML) has become a very important tool to predict properties of MOFs [15, 16, 17, 18, 19, 20, 21]. By training highly accurate ML models, one can further reduce the cost of performing molecular simulations for each new structure. The work of Randall Q. Snurr et. al. [22], for example, predicts PBE band gaps for over 14000 MOFs using a graph neural network (CGCNN [23]). As the availability of a large labeled data set is not always guaranteed, it prevents the use of deep neural networks in the low data regime. Therefore building (small) optimal data sets to calibrate other efficient ML models is necessary in such situations. Active learning (AL) algorithms are designed for such cases. AL aims

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

at constructing the most informative, diverse and representative training set iteratively by using an acquisition function[24, 25] to add new samples to the training set, as opposed to random sampling. This avoids labeling redundant samples, thus reducing the labeling cost. Many AL algorithms are model-free i.e. selection of new samples is done based on diversity [26] and/or good representation [27] of the input space. Model-free AL approaches combined with ML algorithms are beneficial for prediction tasks when compared to random sampling. But, the resulting sample often does not represent the target space well, since the procedure does not use any information about the labels. Adding knowledge of the targets usually assists in understanding its conditional distribution with the features, resulting in the selection of better samples for prediction purposes. Model-based AL schemes like Query by Committee (QBC [28]) accomplish this by defining an acquisition criterion based on the knowledge of the samples already labeled, and an initial model trained on a few labeled samples. Unfortunately, most model-based AL schemes exhibit high computational complexity. Another drawback of model-based AL methods is that the obtained sample is model-specific and it may not be optimal for other ML algorithms.

Furthermore, for many applications, a specific range of the target property is of interest. For example, to discover conducting MOFs, low values of band gaps are desirable. Similarly, high adsorption values are desirable for gas capture. Since available data sets are heterogeneous, focusing on a quantile of interest could be beneficial, and lead to faster discovery of materials with desirable properties at a lower cost. On the other hand, focusing only on the region of interest would render the model not generalizable, and may lead to over-fitting, so it is important to train the model on heterogeneous samples, while focusing more on the quantile of interest. Although there has been substantial work in the field of active learning in the past years, the task of making accurate predictions on a specific range of values has not been explored to the best of our knowledge.

In this paper, we aim to be data efficient in predicting electronic band gaps and adsorption properties of MOFs by adapting a recently proposed model-based AL method [29], namely Regression Tree-based Active Learning (RT-AL). We propose to extend the method to focus on a range of values of the property of interest using Quantile RT-AL (QRT-AL), and show that this decreases the labeling cost tremendously. We also succeed to demonstrate that our approach works for different quantiles of interest, low quantile for band gap predictions and high quantile for predicting adsorption properties.

The organization of the paper is as follows. Section 2 describes the method. Section 3 introduces the MOF databases and settings, followed by results and discussion. The last section concludes the paper.

2 Quantile Regression Tree-based Active Learning (QRT-AL)

Following RT-AL [29], an initial sample I_{init} of size n_{init} is constructed randomly from an unlabeled dataset and is labeled. A standard regression tree is then trained with K leaves, and is used to predict the labels for every unlabeled sample remaining. Thereafter, the leaves of the tree are used to add more samples to the training set. Conditionally to the first labeled set, the number of samples to be labeled from each leaf k, n_k^*, are distributed into the different leaves as:

$$n_k^* = n_{\text{act}} \frac{\sqrt{\pi_k \hat{\sigma}_k^2 \gamma_k}}{\sum_{\ell=1}^{K} \sqrt{\pi_\ell \hat{\sigma}_\ell^2 \gamma_k}},$$

where n_{act} are the total number of samples to be selected by QRT-AL, $\hat{\sigma}_k^2$ denotes the variance computed on the true labels in leaf k, π_k is the proportion of unlabeled samples in leaf k, and γ_k specifies the quantile interval of interest: for each leaf $1 \leq k \leq K$,

$$\gamma_k = \frac{\sum_{q=1}^{Q} w^q n_k^q}{\sum_{q=1}^{Q} n_k^q},$$

where n_k^q are the number of unlabeled samples in leaf k in quantile interval q, and w^q are weights defined depending on the quantile of interest. A good sample should be focused around this quantile, but it should also see samples everywhere, in order to have a global view of the data set. Thus, to generalize the machine learning model, the range of the response is divided into Q quantile intervals, instead of completely focusing on the quantile of interest. The new samples added to the training

set are diverse and representative of both the input and the target space and are more focused on the target values of interest, thus in principle learning the region of interest better.

After computing the number of samples to be selected from each region, they are selected using random sampling. Once the new samples are labeled, the regression tree can be retrained, leading to a more optimised model. This routine can then be repeated by adding few samples at each step, followed by retraining the regression tree, till the desired size of the training set, or a targeted accuracy of the model is obtained. The pseudo code of the algorithm is given in the supplementary material.

3 Results

We demonstrate the performance of our method to predict CO_2 adsorption for MOFs in the hypothetical MOF (hMOF) [30] database, and to predict band gaps for MOFs in the Quantum MOF (QMOF) [22, 31] database. These publicly available data sets consist of atomic structures of MOFs along with the respective target properties. These were selected as they contain two very different properties of MOFs: adsorption and band gap.

The version of the QMOF database used [22] consists of 14482 MOFs, with optimised structures and electronic band gaps (in eV) computed at the PBE-D3(BJ) using DFT. Stoichiometric descriptors (ST-120) from Meredig and Agrawal et al. [32] were used as the representations, with 103 attributes describing the elemental fractions from H–Lr and 17 statistical attributes of elemental properties. As band gap is more sensitive to electronic properties, for the low data regime, simple compositional descriptors are suitable. The ST-120 features for all the MOFs were computed using Matminer [33]. For band gaps, we are interested in the low quantile for electrical conductivity applications, so we divide the range of target values into 3 quantiles: [0,0.2), [0.2,0.5) and [0.5,1] with weights 0.70, 0.25 and 0.05 respectively, giving higher weight to the quantile of interest and reducing it as we move away from the region of interest. Note that there can be different ways of choosing the quantiles and weights (using different values of weights and finer distribution of the quantiles for instance), and we present here one of the many possible cases as a proof of concept.

The hMOF database consists of 137,652 hypothetical MOFs, with data of CO_2 adsorption in mol/g available at 0.05 and 2.5 bar pressures obtained using grand canonical Monte Carlo (GCMC) simulations [30]. We curate a set of 16 features: element-fractions of the 11 species present in the database and 5 structure-related properties, namely Pore Limiting Diameter (PLD), Largest Cavity Diameter (LCD), void fraction, gravimetric and volumetric surface area, as these features are known to directly impact the adsorption properties of transition metal complexes. For CO_2 adsorption, the high quantile intervals are of interest for efficient gas capture, so the quantiles chosen are [0,0.6), [0.6,0.8) and [0.8,1] with weights 0.05, 0.25 and 0.70 respectively.

The full data sets were split in the ratio 8:2 in a stratified manner, with 80% being the pool of MOFs from which the training set is chosen, and the remaining 20% were in the test set, used to determine the performance of the trained models. The first 20 samples (100 for hMOF) were selected using Random Sampling (RS) and the initial training set was constructed, followed by training the first regression tree using scikit-learn [34]. The minimum samples in a leaf parameter was set to 5, as suggested in [29], keeping it high enough to get meaningful variance between labeled samples, but sufficiently low for the tree to give accurate predictions. This is followed by iterative additions of MOFs (batch size indicated in the plots) to the training set till approximately 10% of the total available training pool is labeled. At each iteration, a Random Forest (RF)[1] of 50 regression trees is trained using the training set at the given iteration and its performance is measured by making predictions on the held out test set and computing the quantile MAE (QMAE), for the quantile of interest Q, defined by:

$$\text{QMAE} = \frac{1}{\sharp\{q(y_i) \in Q\}} \sum_{i:q(y_i)\in Q} (y_i - \hat{y}_i),$$

[1]Other tree-based methods like Gradient Boosting Regression Tree (GBRT) and XGBoost[35], and other ML models[29] can also be trained. Since tree-based models are interpretable and require almost no hyperparameter tuning, they are used here. Moreover, tree-based methods have been shown to train models with accuracy close to that of many deep learning models like graph neural networks[36], with far less computational complexity and more explainability, thus making RFs an apt choice.

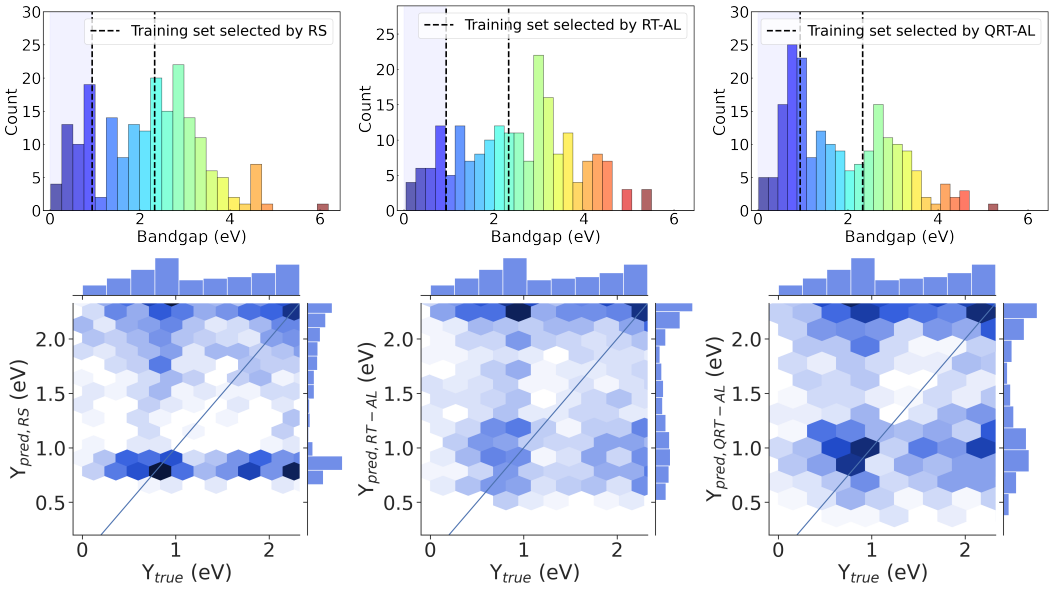

Figure 1: Label distribution of the training sets (with 200 MOFs) selected by RS, RT-AL and QRT-AL (left to right, respectively) from the training pool of the QMOF database. The vertical dashed lines depict the quantile intervals chosen and the quantile interval of interest has been shaded. Respective parity plots are shown below each case for the first two quantiles.

where y_i are true labels of the test set, and $\hat{y}_i$ is the prediction for a test sample using the RF, which is the classical MAE but computed only on test points that are in the considered quantile. Along with QRT-AL, we also train RFs on MOFs selected using random sampling and RT-AL to determine the degree of improvement of our algorithm over them. Note that QRT-AL can be used both in batch and sequential mode (see Appendix for training curves for the sequential case). The sequential case however is extremely computationally expensive, and does not lead to a reduction in QMAE when compared to the batch-mode, therefore batch-mode AL is used here.

The top row of Figure 1 shows the distribution of the labels (band gap) of 200 MOFs selected by RS, RT-AL and QRT-AL (left to right, respectively) from the QMOF dataset. We see that RT-AL samples evenly from the different regions of the label space, while RS does not, and misses some regions. Our method QRT-AL succeeds to selected higher number of samples from the low quantile (quantile of interest for band gap, shown as the shaded region), and progressively selects fewer samples for higher values of band gap. Note that even though QRT-AL selects more sample from the low quantile, it also selects samples from all other quantiles, thus ensuring that the training set is still representative enough of the complete target space. The advantage of focusing on a quantile can be seen from the parity plots in the bottom row of Figure 1, shown for the first two quantiles. For low values of band gap, the higher density of samples around the diagonal for QRT-AL indicates that QRT-AL successfully identifies more MOFs with low band gap, when compared to RT-AL and random sampling, with only 200 samples in the training set. On the other hand, random sampling predicts those values of band gaps well which correspond to the peaks of the label distribution of the true labels, while RT-AL predicts all values of band gaps with roughly the same accuracy.

Figure 2 shows QMAE plots for band gap and CO_2 adsorption prediction, along with histograms of the respective target quantities. The QMAE were averaged over 100 runs, with different train-test splits to compute variance caused by different initializations, and the standard deviation is shown as shaded regions in the plots. For band gap prediction, QRT-AL outperforms both RT-AL and RS as expected. Interestingly, the QMAE of RT-AL is higher than that of RS. This can be understood from the distribution of the band gaps. RT-AL samples well from all regions of the target space, as shown in [29]. However, as for this problem low values of band gap are of interest, sampling well from all regions is not desirable. RS samples better than RT-AL since it does not take the diversity/representativity of the input features or the labels into account. RS thereby selects fewer MOFs with high values of band gap, as there are few of them, and more MOFs with low band gap

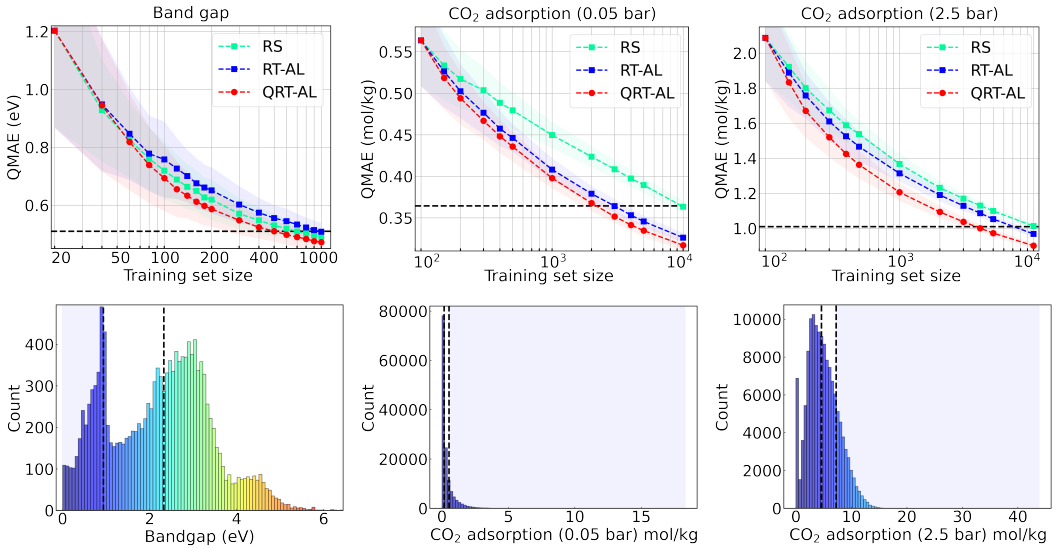

Figure 2: Quantile MAE (averaged over 100 runs) computed on the test set predicting band gaps for MOFs in the QMOF database, and CO_2 adsorption at 0.05 and 2.5 bar pressures for MOFs in the hMOF database, while sampling the training set with RS, RT-AL and QRT-AL. Distributions of the respective target properties has been shown as histograms below each case. The vertical dashed lines depict the quantile intervals chosen and the quantile interval of interest has been shaded.

values, giving better QMAE values in the low quantile region. On the contrary, for CO_2 adsorption, RT-AL selects better samples compared to RS. This happens because here the interest is in the high quantile, and from the distribution of CO_2 adsorption at both 0.05 and 2.5 bar pressures, we see that there are very few MOFs in the high quantile region. Thus, RS ends up sampling more from the peaks of the distribution, that does not correspond to the target values of interest. RT-AL, on the other hand, ensures sampling from all regions of the target, thereby also sampling from the tails of the distribution. As QRT-AL is designed to focus more on the region of interest (the tail of the full distribution in this case), it samples better than both RS and RT-AL.

Another interesting detail observed upon comparing the QMAE plots of CO_2 adsorption for 0.05 vs 2.5 bar pressures is that for 0.05 bar, QRT-AL and RT-AL both improve over RS by a large margin, while for 2.5 bar, the improvement of RT-AL over RS is minor. The performance of RS at 10000 samples is achieved by RT-AL in only 3000 samples for 0.05 bar pressure, while for 2.5 bar, 7000 samples are needed to achieve the performance of 10000 samples randomly drawn. Even though this reduction is still very significant, this difference occurs due to the highly peaked distribution of adsorption at 0.05 bar, as compared to the slightly more evenly distributed adsorption at 2.5 bar which makes it easier for RS to sample from different regions of the target space. As QRT-AL focuses specifically on the quantile of interest, it samples better than RS and RT-AL in both cases. This is why even though RS and RT-AL give similar performance when the target is more evenly distributed, QRT-AL still stands out.

4 Conclusion and Discussion

In this paper, we show that our method, QRT-AL, is a promising active learning algorithm when specific ranges of the target variables are of interest. QRT-AL significantly reduces the labeling cost, for two very different properties of metal organic frameworks, on data sets of very different sizes and selects the most informative samples for both high and low quantile cases. To the best of our knowledge, this is the first method to take into account such an objective in active learning. We are confident that this first step in quantile active learning has much greater potential beyond this MOF data set benchmark and will prove to be an interesting topic of research in the future.

Acknowledgments and Disclosure of Funding

We acknowledge the CINES, IDRIS and TGCC under project No. INP2227/72914/gen7211, as well as CIMENT/GRICAD for computational resources. This work has been partially supported by MIAI@Grenoble Alpes (ANR-19-P3IA-0003).

References

[1] Hong-Cai Zhou, Jeffrey R Long, and Omar M Yaghi. Introduction to metal-organic frameworks. *Chem. Rev.*, 112(2):673–674, February 2012.

[2] Cheng Wang, Demin Liu, and Wenbin Lin. Metal–organic frameworks as a tunable platform for designing functional molecular materials. *Journal of the American Chemical Society*, 135(36):13222–13234, 2013. PMID: 23944646.

[3] Meili Ding, Robinson W. Flaig, Hai-Long Jiang, and Omar M. Yaghi. Carbon capture and conversion using metal–organic frameworks and mof-based materials. *Chem. Soc. Rev.*, 48:2783–2828, 2019.

[4] Hailian Li, Mohamed Eddaoudi, M O'Keeffe, and O M Yaghi. Design and synthesis of an exceptionally stable and highly porous metal-organic framework. *Nature*, 402(6759):276–279, November 1999.

[5] Hao Li, Kecheng Wang, Yujia Sun, Christina T. Lollar, Jialuo Li, and Hong-Cai Zhou. Recent advances in gas storage and separation using metal–organic frameworks. *Materials Today*, 21(2):108–121, 2018.

[6] Jian-Rong Li, Ryan J. Kuppler, and Hong-Cai Zhou. Selective gas adsorption and separation in metal–organic frameworks. *Chem. Soc. Rev.*, 38:1477–1504, 2009.

[7] Husam A Almassad, Rada I Abaza, Lama Siwwan, Bassem Al-Maythalony, and Kyle E Cordova. Environmentally adaptive MOF-based device enables continuous self-optimizing atmospheric water harvesting. *Nat. Commun.*, 13(1):4873, August 2022.

[8] JeongYong Lee, Omar K. Farha, John Roberts, Karl A. Scheidt, SonBinh T. Nguyen, and Joseph T. Hupp. Metal–organic framework materials as catalysts. *Chem. Soc. Rev.*, 38:1450–1459, 2009.

[9] Arturo Gamonal, Chen Sun, A Lorenzo Mariano, Estefania Fernandez-Bartolome, Elena Guerrero-SanVicente, Bess Vlaisavljevich, Javier Castells-Gil, Carlos Marti-Gastaldo, Roberta Poloni, Reinhold Wannemacher, Juan Cabanillas-Gonzalez, and Jose Sanchez Costa. Divergent adsorption-dependent luminescence of amino-functionalized lanthanide metal-organic frameworks for highly sensitive NO2 sensors. *J. Phys. Chem. Lett.*, 11(9):3362–3368, May 2020.

[10] Lilia S. Xie, Grigorii Skorupskii, and Mircea Dincă. Electrically conductive metal–organic frameworks. *Chemical Reviews*, 120(16):8536–8580, 2020. PMID: 32275412.

[11] Eric M Johnson, Stefan Ilic, and Amanda J Morris. Design strategies for enhanced conductivity in metal-organic frameworks. *ACS Cent. Sci.*, 7(3):445–453, March 2021.

[12] Huabin Zhang, Jianwei Nai, Le Yu, and Xiong Wen (David) Lou. Metal-organic-framework-based materials as platforms for renewable energy and environmental applications. *Joule*, 1(1):77–107, 2017.

[13] Andrew S Rosen, Justin M Notestein, and Randall Q Snurr. Identifying promising metal-organic frameworks for heterogeneous catalysis via high-throughput periodic density functional theory. *J. Comput. Chem.*, 40(12):1305–1318, May 2019.

[14] Emmanuel Ren, Philippe Guilbaud, and François-Xavier Coudert. High-throughput computational screening of nanoporous materials in targeted applications. *Digital Discovery*, 1:355–374, 2022.

[15] Jake Burner, Ludwig Schwiedrzik, Mykhaylo Krykunov, Jun Luo, Peter G. Boyd, and Tom K. Woo. High-performing deep learning regression models for predicting low-pressure co2 adsorption properties of metal–organic frameworks. *The Journal of Physical Chemistry C*, 124(51):27996–28005, 2020.

[16] Sangwon Lee, Baekjun Kim, Hyun Cho, Hooseung Lee, Sarah Yunmi Lee, Eun Seon Cho, and Jihan Kim. Computational screening of trillions of metal–organic frameworks for high-performance methane storage. *ACS Applied Materials & Interfaces*, 13(20):23647–23654, 2021. PMID: 33988362.

[17] Kamal Choudhary, Taner Yildirim, Daniel W. Siderius, Assaf A. Gilad, Kusne, Austin McDannald, and Diana L. Ortiz-Montalvo. Graph neural network predictions of metal organic framework co 2 adsorption properties. 2022.

[18] Aditya Nandy, Chenru Duan, and Heather J. Kulik. Using machine learning and data mining to leverage community knowledge for the engineering of stable metal–organic frameworks. *Journal of the American Chemical Society*, 143(42):17535–17547, 2021. PMID: 34643374.

[19] Seyed Mohamad Moosavi, Balázs Álmos Novotny, Daniele Ongari, Elias Moubarak, Mehrdad Asgari, Özge Kadioglu, Charithea Charalambous, Andres Ortega-Guerrero, Amir H Farmahini, Lev Sarkisov, Susana Garcia, Frank Noé, and Berend Smit. A data-science approach to predict the heat capacity of nanoporous materials. *Nat. Mater.*, 21(12):1419–1425, December 2022.

[20] Zhonglin Cao, Rishikesh Magar, Yuyang Wang, and Amir Barati Farimani. Moformer: Self-supervised transformer model for metal–organic framework property prediction. *Journal of the American Chemical Society*, 145(5):2958–2967, 2023. PMID: 36706365.

[21] Hakan Demir, Hilal Daglar, Hasan Can Gulbalkan, Gokhan Onder Aksu, and Seda Keskin. Recent advances in computational modeling of mofs: From molecular simulations to machine learning. *Coordination Chemistry Reviews*, 484:215112, 2023.

[22] Andrew S. Rosen, Shaelyn M. Iyer, Debmalya Ray, Zhenpeng Yao, Alán Aspuru-Guzik, Laura Gagliardi, Justin M. Notestein, and Randall Q. Snurr. Machine learning the quantum-chemical properties of metal–organic frameworks for accelerated materials discovery. *Matter*, 4(5):1578–1597, 2021.

[23] Tian Xie and Jeffrey C. Grossman. Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties. *Phys. Rev. Lett.*, 120:145301, Apr 2018.

[24] James T. Wilson, Riccardo Moriconi, Frank Hutter, and Marc Peter Deisenroth. The reparameterization trick for acquisition functions. *ArXiv*, abs/1712.00424, 2017.

[25] Frank Hutter, Holger H. Hoos, and Kevin Leyton-Brown. Sequential model-based optimization for general algorithm configuration. In Carlos A. Coello Coello, editor, *Learning and Intelligent Optimization*, pages 507–523, Berlin, Heidelberg, 2011. Springer Berlin Heidelberg.

[26] Dongrui Wu, Chin-Teng Lin, and Jian Huang. Active learning for regression using greedy sampling. *Information Sciences*, 474:90–105, 2019.

[27] Ziang Liu, Xue Jiang, Hanbin Luo, Weili Fang, Jiajing Liu, and Dongrui Wu. Pool-based unsupervised active learning for regression using iterative representativeness-diversity maximization. *Pattern Recognition Letters*, 142:11–19, 2021.

[28] Robert Burbidge, Jem J. Rowland, and Ross D. King. Active learning for regression based on query by committee. In Hujun Yin, Peter Tino, Emilio Corchado, Will Byrne, and Xin Yao, editors, *Intelligent Data Engineering and Automated Learning - IDEAL 2007*, pages 209–218, Berlin, Heidelberg, 2007. Springer Berlin Heidelberg.

[29] Ashna Jose, João Paulo Almeida de Mendonça, Emilie Devijver, Noël Jakse, Valérie Monbet, and Roberta Poloni. Regression tree-based active learning. *Data Min. Knowl. Discov.*, August 2023.

[30] Christopher E Wilmer, Michael Leaf, Chang Yeon Lee, Omar K Farha, Brad G Hauser, Joseph T Hupp, and Randall Q Snurr. Large-scale screening of hypothetical metal-organic frameworks. *Nat. Chem.*, 4(2):83–89, November 2011.

[31] Andrew S Rosen, Victor Fung, Patrick Huck, Cody T O'Donnell, Matthew K Horton, Donald G Truhlar, Kristin A Persson, Justin M Notestein, and Randall Q Snurr. High-throughput predictions of metal–organic framework electronic properties: theoretical challenges, graph neural networks, and data exploration. *Npj Comput. Mater.*, 8(1), May 2022.

[32] B. Meredig, A. Agrawal, S. Kirklin, J. E. Saal, J. W. Doak, A. Thompson, K. Zhang, A. Choudhary, and C. Wolverton. Combinatorial screening for new materials in unconstrained composition space with machine learning. *Phys. Rev. B*, 89:094104, Mar 2014.

[33] Logan Ward, Alexander Dunn, Alireza Faghaninia, Nils E.R. Zimmermann, Saurabh Bajaj, Qi Wang, Joseph Montoya, Jiming Chen, Kyle Bystrom, Maxwell Dylla, Kyle Chard, Mark Asta, Kristin A. Persson, G. Jeffrey Snyder, Ian Foster, and Anubhav Jain. Matminer: An open source toolkit for materials data mining. *Computational Materials Science*, 152:60–69, September 2018.

[34] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[35] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 785–794, New York, NY, USA, 2016. Association for Computing Machinery.

[36] Dejun Jiang, Zhenxing Wu, Chang-Yu Hsieh, Guangyong Chen, Ben Liao, Zhe Wang, Chao Shen, Dongsheng Cao, Jian Wu, and Tingjun Hou. Could graph neural networks learn better molecular representation for drug discovery? a comparison study of descriptor-based and graph-based models. *J. Cheminform.*, 13(1):12, February 2021.