Adaptive Representation of MOFs in Bayesian Optimization

Mahyar Rajabi Kochi[#]
Chemical Engineering & Applied Chemistry
University of Toronto
Toronto, Canada

Negareh Mahboubi[#]
Chemical & Materials Engineering
University of Alberta
Alberta, Canada

Aseem Partap Singh Gill
Chemical Engineering & Applied Chemistry
University of Toronto
Toronto, Canada

Seyed Mohamad Moosavi[*]
Chemical Engineering & Applied Chemistry
University of Toronto
Toronto, Canada

[#]equal contribution
[*]mohamad.moosavi@utoronto.ca

Abstract

Bayesian optimization (BO) is increasingly used in molecular optimization and to guide self-driving laboratories for automated materials discovery. A crucial aspect of BO is how molecules and materials are represented as feature vectors, where both the completeness and compactness of these representations can influence the efficiency of the optimization process. Traditionally, a fixed representation is chosen by expert chemists or applying data-driven feature selection methods on available labeled datasets. However, when dealing with novel optimization tasks, prior knowledge or large datasets are often unavailable, and relying on these even can introduce bias into the search process. In this work, we demonstrate a Feature Adaptive Bayesian Optimization (FABO) framework, which integrates feature selection in Bayesian optimization process to dynamically adapt material representations throughout the optimization cycles. We demonstrate the effectiveness of this adaptive approach across several molecular optimization tasks, including the discovery of high-performing metal-organic frameworks (MOFs) in three distinct tasks, each involving unique property distributions and requiring a distinct representation. Our results show that the adaptive nature of the representation leads to outperforming random search baseline.

1 Introduction

Recent advancements in machine learning (ML) and artificial intelligence (AI) are revolutionizing molecular and materials discovery by enabling self-driving labs (SDLs) that integrate ML with lab automation and robotics [1]. At the core of SDLs lies Bayesian optimization (BO), which autonomously balances exploration and exploitation to guide experimental workflows [2]. Effective material representation is essential for BO, as high-dimensional representations can hinder performance due to

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

the curse of dimensionality [3, 4]. While methods like kernel tuning and generative embeddings show promise, they often face challenges in compressing material data for advanced systems [5, 6] . Metal-organic frameworks (MOFs), with highly tunable chemistry and vast diversity, illustrate the need for adaptable representations to accelerate discovery for diverse applications [7, 8].

2 Feature Adaptive Bayesian Optimization

The Feature Adaptive Bayesian Optimization (FABO) workflow aims to efficiently identify optimal materials from large databases while minimizing expensive experiments or simulations. Each closed-loop optimization cycle includes data labeling, updating material representations, refining the surrogate model, and selecting the next experiment using an acquisition function, as can be seen in Figure 1. FABO, similar to any other principled Bayesian optimization, relies on a surrogate model for uncertainty-aware objective function predictions and an acquisition function to balance exploitation and exploration. This study uses a Gaussian Process Regressor (GPR) for its robust uncertainty quantification and combines Expected Improvement (EI) and Upper Confidence Bound (UCB) acquisition functions.

FABO dynamically updates material representations at each cycle, refining features to enhance optimization efficiency without requiring extensive labeled data upfront. Two feature selection methods, Minimum Redundancy Maximum Relevance (mRMR) and Spearman ranking, are incorporated in FABO. mRMR balances feature relevance to the target variable and redundancy among features using mutual information metrics [9], while Spearman ranking evaluates features based on their monotonic relationship with the target using rank correlation coefficients [10]. Both methods are computationally efficient and suitable for iterative optimization compared to embedded techniques like LASSO or tree-based methods, which require hyperparameter tuning [11]. FABO selects between 5 and 20 features during each BO run, enabling adaptable and efficient exploration of the material search space.

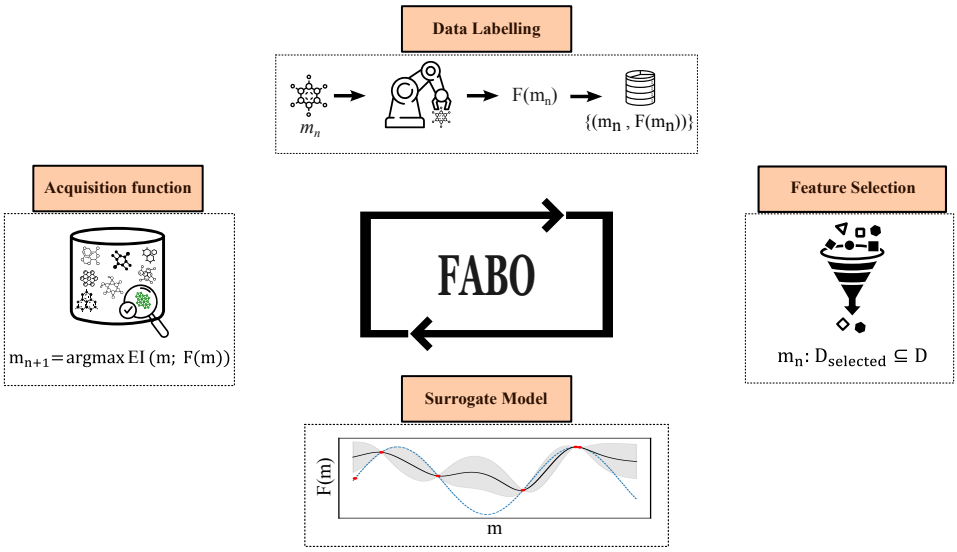

Figure 1: **Feature Adaptive Bayesian Optimization (FABO) framework.** FABO operates in an iterative feedback loop: (1) label the candidate material (m_n) computationally or experimentally ($F(m_n)$) and add it to the labeled dataset, (2) perform feature selection based on labeled data to determine the most informative representation, (3) update the surrogate model using the selected feature set (D_{selected}), and (4) apply the acquisition function to select the next experiment (m_{n+1}) for data labeling

3 Case study

This study applies FABO to discover MOFs with target properties from large databases, leveraging their complex chemistry-geometry relationships. Two datasets are used: QMOF, with 8,437 MOFs

in which band gap is the target property and acquired from high-throughput DFT, and CoRE-2019, with 9,525 MOFs in which CO_2 adsorptions at low and high pressures are properties of interest. Initial feature pools include chemical descriptors like Revised Autocorrelation Calculations (RACs) [12, 13, 14] and stoichiometric sets (Stoichiometric-45 [15] and Stoichiometric-120 [16]), alongside geometric features like pore sizes calculated using Zeo++ [17]. FABO is benchmarked against BO campaign in which features are selected randomly. The experimental budget in each single BO campaign is 250 and to mitigate the influence of initial data points on optimization outcomes, 20 BO campaigns are conducted, each with 10 randomly selected initial datasets.

We use three metrics to evaluate the quality of the acquired MOFs during the BO campaign: the best rank, the best value of the objective function, and the number of acquired materials among the top 100 materials in the dataset [18]. The search efficiency curves in Figure 2 demonstrate the high performance of FABO across all three metrics and three objectives compared to the baseline. The results in Figure 2 shows the baseline often falls short in fully capturing the complexity of structure-property relationship. In specific, for the more complex properties, like CO_2 uptake at low pressure and band gap, which involve complex chemistry, random feature selection could not identify the best set.

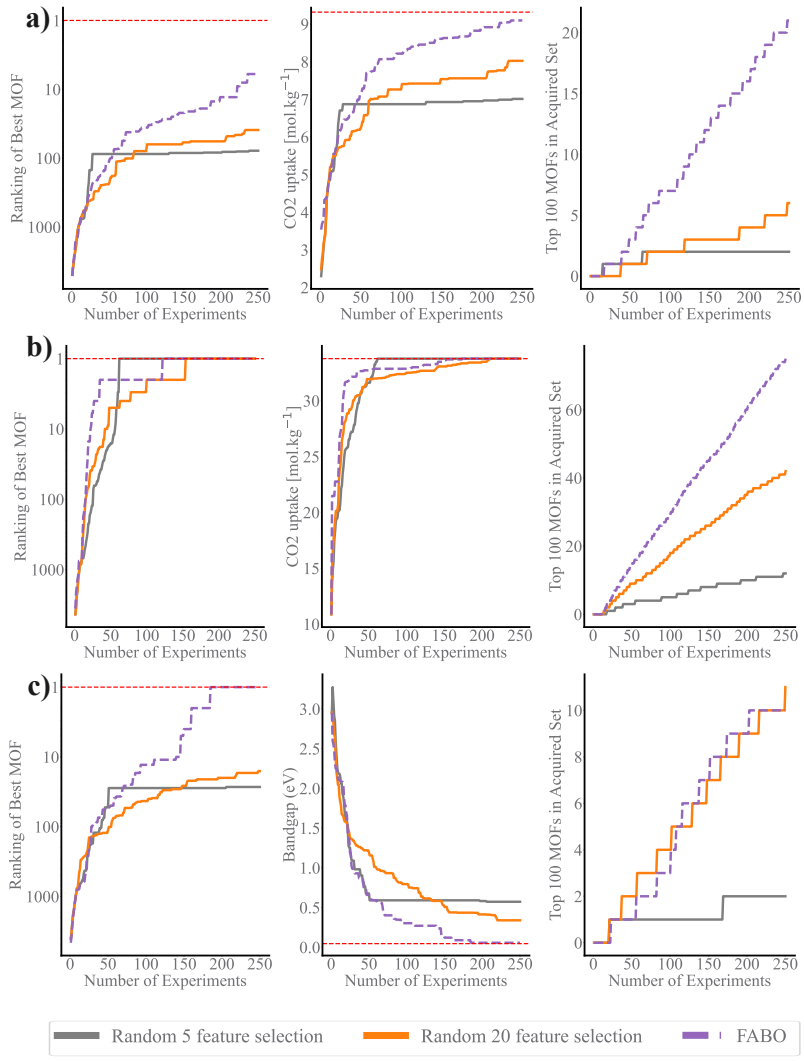

Figure 2: **Search efficiency curves for FABO, illustrating performance against BO campaigns with random feature selection.** (a) CO_2 uptake at low pressure, (b) CO_2 uptake at high pressure, and (c) band gap. The quality of the acquired set of MOFs is shown in three panels: (left) the highest rank relative to the entire dataset; (middle) the optimum value of the objective function; and (right) the number of top 100 MOFs (based on the property of interest) included in acquired MOF set.

4 Conclusion

In this work, we introduced the Feature Adaptive Bayesian Optimization (FABO) framework, which integrates feature selection into Bayesian optimization to dynamically refine material representations throughout the optimization process. Our open-source implementation of FABO is available for researchers to easily apply to their own domain-specific optimization problems. By starting from a complete feature set, FABO's integrated feature selection within BO ensures that the most relevant features are dynamically chosen to optimize the search space efficiently.

Acknowledgement

The authors gratefully acknowledge financial support from Natural Sciences and Engineering Research Council of Canada, the University of Toronto's Acceleration Consortium through the Canada First Research Excellence Fund under Grant number CFREF-2022-00042, and National Research Council of Canada under the Materials for Clean Fuels Challenge Program. The authors thank Sterling Baird and Benjamin Sanchez-Lengeling for their valuable discussions and insights.

Author contributions

Conceptualization: S.M.M., M.R.K., N.M., A.P.S.G.; Data curation: M.R.K., N.M., A.P.S.G.; Formal analysis: M.R.K., N.M., A.P.S.G.; Funding acquisition: S.M.M.; Investigation: M.R.K., N.M., A.P.S.G.; Methodology: M.R.K., N.M., A.P.S.G., S.M.M.; Project administration: S.M.M.; Software: M.R.K., N.M., A.P.S.G.; Supervision: S.M.M.; Visualization: M.R.K., N.M., A.P.S.G.; Writing: M.R.K., N.M., & S.M.M.

Competing interest

The authors declare no competing interests.

Code availability

The code base for FABO as well as the codes to reproduce results of this study are available from `https://github.com/AI4ChemS/FABO`.

References

[1] Seyed Mohamad Moosavi, Kevin Maik Jablonka, and Berend Smit. The role of machine learning in the understanding and design of materials. *Journal of the American Chemical Society*, 142(48):20273–20287, 2020.

[2] Eric Taw and Jeffrey B Neaton. Accelerated discovery of ch4 uptake capacity metal–organic frameworks using bayesian optimization. *Advanced Theory and Simulations*, 5(3):2100515, 2022.

[3] Alexander Pomberger, AA Pedrina McCarthy, Ahmad Khan, Simon Sung, CJ Taylor, MJ Gaunt, Lucy Colwell, David Walz, and AA Lapkin. The effect of chemical representation on active machine learning towards closed-loop optimization. *Reaction Chemistry & Engineering*, 7(6):1368–1379, 2022.

[4] Sterling G Baird, Jason R Hall, and Taylor D Sparks. Compactness matters: Improving bayesian optimization efficiency of materials formulations through invariant search spaces. *Computational Materials Science*, 224:112134, 2023.

[5] David Eriksson and Martin Jankowiak. High-dimensional bayesian optimization with sparse axis-aligned subspaces. In *Uncertainty in Artificial Intelligence*, pages 493–503. PMLR, 2021.

[6] Ryan-Rhys Griffiths and José Miguel Hernández-Lobato. Constrained bayesian optimization for automatic chemical design using variational autoencoders. *Chemical science*, 11(2):577–586, 2020.

[7] Seyed Mohamad Moosavi, Henglu Xu, Linjiang Chen, Andrew I Cooper, and Berend Smit. Geometric landscapes for material discovery within energy–structure–function maps. *Chemical Science*, 2020.

[8] Hiroyasu Furukawa, Kyle E Cordova, Michael O'Keeffe, and Omar M Yaghi. The chemistry and applications of metal-organic frameworks. *Science*, 341(6149):1230444, 2013.

[9] Hanchuan Peng, Fuhui Long, and Chris Ding. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on pattern analysis and machine intelligence*, 27(8):1226–1238, 2005.

[10] Jiefang Jiang, Xianyong Zhang, and Zhong Yuan. Feature selection for classification with spearman's rank correlation coefficient-based self-information in divergence-based fuzzy rough sets. *Expert Systems with Applications*, 249:123633, 2024.

[11] HongFang Zhou, JiaWei Zhang, YueQing Zhou, XiaoJie Guo, and YiMing Ma. A feature selection algorithm of decision tree based on feature weight. *Expert Systems with Applications*, 164:113842, 2021.

[12] Seyed Mohamad Moosavi, Aditya Nandy, Kevin Maik Jablonka, Daniele Ongari, Jon Paul Janet, Peter G Boyd, Yongjin Lee, Berend Smit, and Heather Kulik. Understanding the diversity of the metal-organic framework ecosystem. *Nature Communications*, 11:4068, 2020.

[13] Jon Paul Janet and Heather J Kulik. Resolving transition metal chemical space: Feature selection for machine learning and structure–property relationships. *The Journal of Physical Chemistry A*, 121(46):8939–8954, 2017.

[14] Efthymios I Ioannidis, Terry ZH Gani, and Heather J Kulik. molsimplify: A toolkit for automating discovery in inorganic chemistry. *Journal of computational chemistry*, 37(22):2106–2117, 2016.

[15] Yuping He, Ekin D Cubuk, Mark D Allendorf, and Evan J Reed. Metallic metal–organic frameworks predicted by the combination of machine learning methods and ab initio calculations. *The journal of physical chemistry letters*, 9(16):4562–4569, 2018.

[16] Bryce Meredig, Ankit Agrawal, Scott Kirklin, James E Saal, Jeff W Doak, Alan Thompson, Kunpeng Zhang, Alok Choudhary, and Christopher Wolverton. Combinatorial screening for new materials in unconstrained composition space with machine learning. *Physical Review B*, 89(9):094104, 2014.

[17] Thomas F Willems, Chris H Rycroft, Michaeel Kazi, Juan C Meza, and Maciej Haranczyk. Algorithms and tools for high-throughput geometry-based analysis of crystalline porous materials. *Microporous and Mesoporous Materials*, 149(1):134–141, 2012.

[18] Qiaohao Liang, Aldair E Gongora, Zekun Ren, Armi Tiihonen, Zhe Liu, Shijing Sun, James R Deneault, Daniil Bash, Flore Mekki-Berrada, Saif A Khan, et al. Benchmarking the performance of bayesian optimization across multiple experimental materials science domains. *npj Computational Materials*, 7(1):188, 2021.