Article information

Article title

Materials Science Optimization Benchmark Dataset for High-dimensional, Multi-objective, Multi-fidelity Optimization of CrabNet Hyperparameters

Authors

Sterling G. Baird[1]*, Jeet N. Parikh[2], Taylor D. Sparks[1]

Affiliations

1. Materials Science & Engineering, University of Utah, 122 S. Central Campus Drive, #304 Salt Lake City, Utah 84112
2. Northwood High School, 4515 Portola Pkwy, Irvine, CA 92620

Corresponding author's email address and Twitter handle

sterling.baird@utah.edu
@SterlingBaird1

Keywords

adaptive design, Bayesian optimization, formulation optimization, PseudoCrab

Abstract

Benchmarks are crucial for driving progress in scientific disciplines. To be effective, benchmarks should closely mimic real-world tasks while being computationally efficient, allowing for accessibility and repeatability. Developing surrogate models that can be indistinguishable from the ground truth observation within the explored dataset bounds dramatically reduces the computational burden of running benchmarks without sacrificing quality, but this requires a large amount of initial data. In the fields of materials science and chemistry, relevant optimization tasks can be challenging due to their complexity, which includes hierarchical, noisy, multi-fidelity, multi-objective, high-dimensional, and non-linearly correlated variables. Additionally, they may include mixed numerical and categorical variables that are subject to linear and non-linear constraints. Simulating or experimentally verifying such tasks can be difficult, which is why benchmarks are essential. This study aimed to overcome these challenges by generating 173219 quasi-random hyperparameter combinations across 23 hyperparameters and using them to train CrabNet on the Matbench experimental band gap dataset (Computational runtime: 387 RTX-2080-Ti GPU days). The results were stored in a free-tier shared MongoDB Atlas dataset, creating a regression dataset that maps hyperparameter combinations to metrics such as MAE, RMSE, computational runtime, and model size for the CrabNet model trained on the Matbench experimental band gap benchmark task. To simulate the actual simulations, heteroskedastic noise was incorporated into the regression dataset, and bad hyperparameter combinations were excluded. Percentile ranks were computed within each group of identical

parameter sets to capture heteroskedastic noise, rather than assuming Gaussian noise as is done in traditional approaches. This approach can be applied to other benchmark datasets, bridging the gap between optimization benchmarks with low computational overhead and realistically complex, real-world optimization scenarios.

Specifications table

Subject	Computational materials science
Specific subject area	Composition-based experimental band gap prediction
Type of data	Table Figure
How the data were acquired	The data was obtained by executing CrabNet v2.0.8 (available at https://github.com/sparks-baird/CrabNet) for each of the five folds of the Matbench experimental band gap task (https://matbench.materialsproject.org/Leaderboards%20Per-Task/matbench_v0.1_matbench_expt_gap/). Python code in https://github.com/sparks-baird/matsci-opt-benchmarks/blob/7c4346624895a7826ada07ff5e44c2f49eb42b9d/scripts/crabnet_hyperparameter/crabnet_hyperparameter_submitit.py was used for orchestration, with the University of Utah's Center for High-performance Computing (CHPC) resources. Jobs were sent to the SLURM scheduler using Submitit (https://github.com/facebookincubator/submitit), and results were logged in JSON format using the MongoDB Data API. The matsci-opt-benchmarks code used for this study can be found at https://github.com/sparks-baird/matsci-opt-benchmarks/tree/v0.2.1 (https://dx.doi.org/10.5281/zenodo.7694289).
Data format	Analyzed Filtered Raw
Description of data collection	The Ax Platform was used to perform a quasi-random Sobol sampling of 65536 parameter combinations, varying 23 hyperparameters within a constrained search space, with 5 repeats resulting in a total of 327680 training runs. Out of these, 173219 completed successfully, consuming 387 RTX-2080-Ti GPU days or 4614.29 CUDA core years, resulting in 41550 unique sets. To rank repeat simulations, the "dense" method with pct=True was used in pandas.core.groupby.GroupBy.rank.

Data source location	Free-tier Shared Cluster MongoDB Atlas Database
Data accessibility	Repository name: Zenodo Data identification number: 7694268 Direct URL to data: https://doi.org/10.5281/zenodo.7694268

Value of the data

- The dataset is valuable for benchmarking adaptive design methods applied to high-dimensional, constrained, multi-fidelity optimization tasks.
- Practitioners of optimization in the physical sciences can leverage this dataset to simulate real-world materials optimization tasks, such as alloy discovery, and achieve improved performance.
- The dataset is also useful in gaining insights into the hyperparameter optimization strategies used for developing compositionally restricted material property prediction models.

Objective

Industry-relevant optimization tasks in the physical sciences are often “hierarchical, noisy, multi-fidelity[1,2], multi-objective[3,4], high-dimensional[5,6], and non-linearly correlated while exhibiting mixed numerical and categorical variables subject to linear[7] and non-linear constraints.”[8,9] Existing benchmark datasets[10–15], while very useful, typically are single-objective, single-fidelity, low-dimensional, and ignore or simplify the influence of noise. The inclusion of heteroskedastic noise in a surrogate model enables us to establish a "Turing test" scenario where the surrogate model is virtually identical to the actual simulation. This approach bridges the gap between low-cost surrogate function evaluations using benchmark datasets and costly, real-world objective function evaluations by considering a multi-objective, multi-fidelity, and high-dimensional task while accounting for heteroskedastic noise.

Data description

The regression dataset contains hyperparameter sets (including repeats) spanning twenty-three hyperparameter sets and their corresponding MAE, RMSE, computational runtimes, and model size for training CrabNet.

For histogram data for the number of successful repeats see Figure 1.

For histograms of the mean absolute error, root-mean-square error, runtime, and model size, see Figure 2, Figure 3, Figure 4, and Figure 5, respectively.

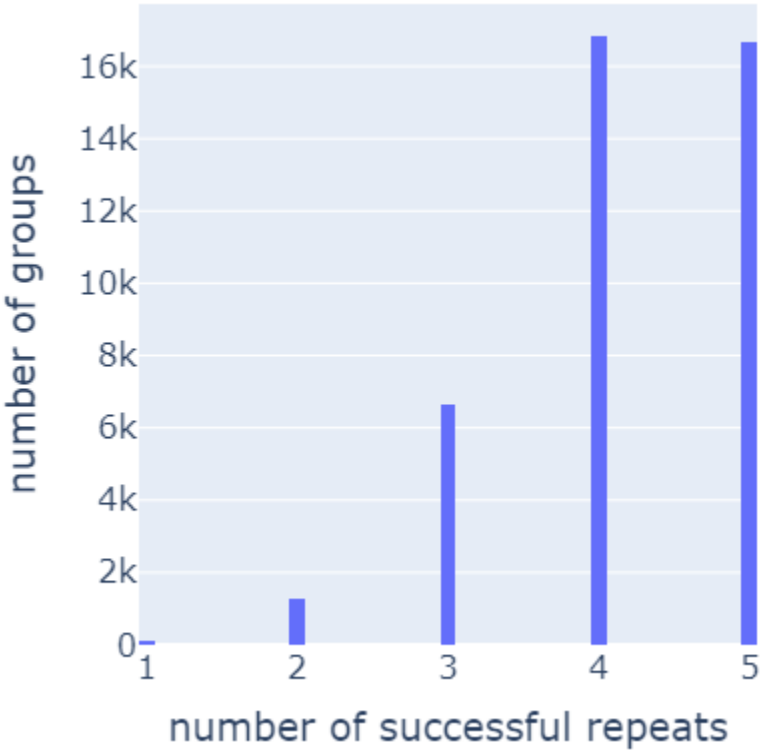

Figure 1. Histogram of number of parameter groups vs. number of successful repeats within a given group. The lowest number of repeats for a parameter set is 1, with approximately 2.6 repeats on average.

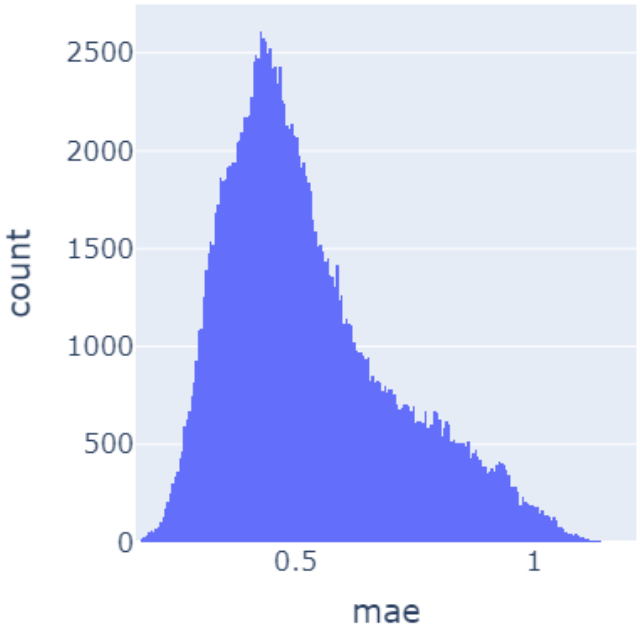

Figure 2. Histogram of number of training runs vs. mean absolute error using CrabNet on the Matbench experimental band gap task.

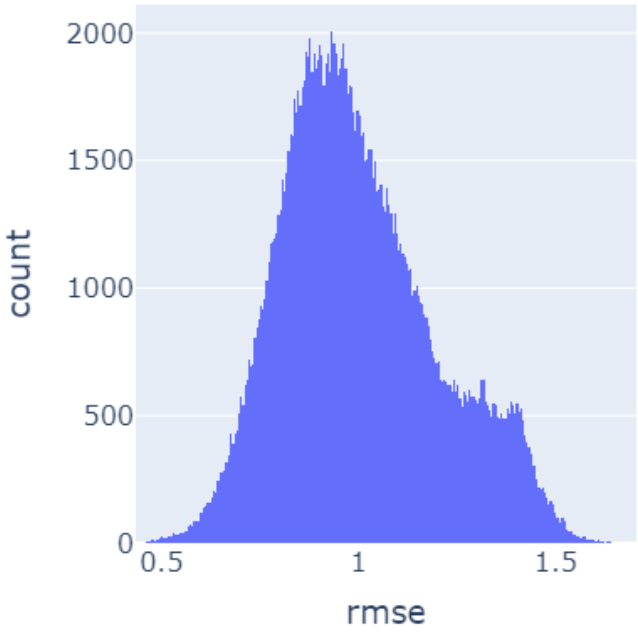

Figure 3. Histogram of number of training runs vs. root-mean-square-error using CrabNet on the Matbench experimental band gap task.

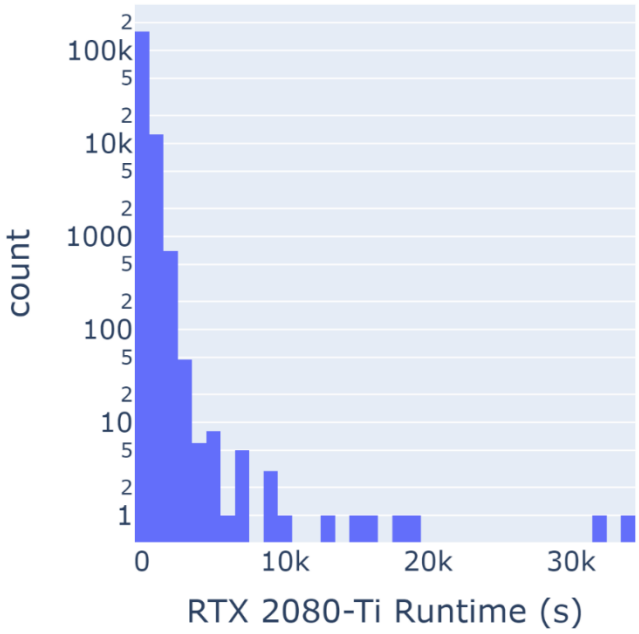

Figure 4. Histogram of number of training runs vs. GPU runtime on an RTX 2080-Ti using CrabNet on the Matbench experimental band gap task. The y-axis is log-scaled.

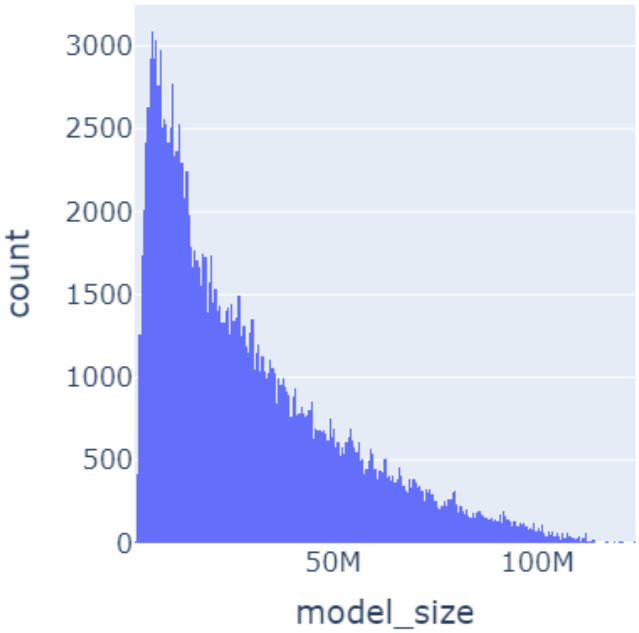

Figure 5. Histogram of number of training runs vs. model size using CrabNet on the Matbench experimental band gap task.

Experimental design, materials and methods

A vast number of CrabNet models, totaling to 173219, were trained with different hyperparameter combinations using the Ax platform's quasi-random Sobol sampling function to generate unique parameter combinations. While there can be other uses, this dataset is primarily intended as a multi-objective, multi-fidelity, high-dimensional benchmark dataset for formulation-based

optimization scenarios by scaling each of the numerical parameters to the range of 0 to 1 and applying a contrived constraint that the sum of all parameters must equal one. To realistically capture noise in the benchmark dataset, simulations were repeated for each quasi-random parameter combination. To improve efficiency and reduce latency, hyperparameter sets (including repeats) were shuffled and divided into batches, then sent to a high-performance computing environment for asynchronous evaluation. Despite some results not being completed due to either timeout or preemption, this trade-off was deemed reasonable for the efficiency and completion gains.

Results were logged to a free-tier MongoDB Atlas database and then aggregated and prepared as machine-learning-ready datasets via Python in Jupyter notebooks. For implementation details, see https://github.com/sparks-baird/matsci-opt-benchmarks/tree/main/scripts/crabnet_hyperparameter and https://github.com/sparks-baird/matsci-opt-benchmarks/tree/main/notebooks/crabnet_hyperparameter.

Ethics statements

There are no statements to declare.

CRediT author statement

Sterling G. Baird: Project administration, Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Data Curation, Writing - Original Draft, Writing - Review & Editing, Visualization. **Jeet N. Parikh**: Methodology, Software, Formal Analysis, Data Curation, Writing - Original Draft, Writing - Review & Editing, **Taylor D. Sparks**: Supervision, Funding acquisition

Acknowledgments

Funding: This work was supported by the National Science Foundation Division of Materials Research [Grant number DMR-1651668].

We thank Trupti Mohanty for work and discussion related to the future use-case of this dataset as a pseudo-materials benchmark. We acknowledge the University of Utah's Center for High Performance Computing (CHPC) for providing computational resources. The code and manuscript associated with a benchmark dataset for particle packing simulations[8] served as a starting point for the code and manuscript for this work, for which there is shared language and document structure. We acknowledge OpenAI for providing free usage of their research tool, ChatGPT, which was used during the review and editing process.

Declaration of interests

x The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

☐ The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

References

(1) Ghoreishi, S. F.; Molkeri, A.; Arróyave, R.; Allaire, D.; Srivastava, A. Efficient Use of Multiple Information Sources in Material Design. *Acta Materialia* **2019**, *180*, 260–271. https://doi.org/10.1016/j.actamat.2019.09.009.

(2) Kandasamy, K.; Vysyaraju, K. R.; Neiswanger, W.; Paria, B.; Collins, C. R.; Schneider, J.; Poczos, B.; Xing, E. P. Tuning Hyperparameters without Grad Students: Scalable and Robust Bayesian Optimisation with Dragonfly. *arXiv:1903.06694 [cs, stat]* **2020**.

(3) Hanaoka, K. Comparison of Conceptually Different Multi-Objective Bayesian Optimization Methods for Material Design Problems. *Materials Today Communications* **2022**, 103440. https://doi.org/10.1016/j.mtcomm.2022.103440.

(4) Häse, F.; Roch, L. M.; Aspuru-Guzik, A. Chimera: Enabling Hierarchy Based Multi-Objective Optimization for Self-Driving Laboratories. *Chem. Sci.* **2018**, *9* (39), 7642–7655. https://doi.org/10.1039/C8SC02239A.

(5) Baird, S. G.; Liu, M.; Sparks, T. D. High-Dimensional Bayesian Optimization of 23 Hyperparameters over 100 Iterations for an Attention-Based Network to Predict Materials Property: A Case Study on CrabNet Using Ax Platform and SAASBO. *Computational Materials Science* **2022**, *211*, 111505. https://doi.org/10.1016/j.commatsci.2022.111505.

(6) Eriksson, D.; Jankowiak, M. High-Dimensional Bayesian Optimization with Sparse Axis-Aligned Subspaces. *arXiv:2103.00349 [cs, stat]* **2021**.

(7) Baird, S.; Hall, J. R.; Sparks, T. D. The Most Compact Search Space Is Not Always the Most Efficient: A Case Study on Maximizing Solid Rocket Fuel Packing Fraction via Constrained Bayesian Optimization. ChemRxiv September 6, 2022. https://doi.org/10.26434/chemrxiv-2022-nz2w8-v2.

(8) Baird, S. G.; Sparks, T. D. Materials Science Optimization Benchmark Dataset for Multi-Fidelity Hard-Sphere Packing Simulations. ChemRxiv January 9, 2023. https://doi.org/10.26434/chemrxiv-2023-fjjk7.

(9) Baird, S.; Hall, J. R.; Sparks, T. D. *Compactness Matters: Improving Bayesian Optimization Efficiency of Materials Formulations through Invariant Search Spaces*; preprint; Chemistry, 2023. https://doi.org/10.26434/chemrxiv-2022-nz2w8-v3.

(10) Dunn, A.; Wang, Q.; Ganose, A.; Dopp, D.; Jain, A. Benchmarking Materials Property Prediction Methods: The Matbench Test Set and Automatminer Reference Algorithm. *npj Comput Mater* **2020**, *6* (1), 138. https://doi.org/10.1038/s41524-020-00406-3.

(11) De Breuck, P.-P.; Evans, M. L.; Rignanese, G.-M. Robust Model Benchmarking and Bias-Imbalance in Data-Driven Materials Science: A Case Study on MODNet. *J. Phys.: Condens. Matter* **2021**, *33* (40), 404002. https://doi.org/10.1088/1361-648X/ac1280.

(12) Wang, A.; Liang, H.; McDannald, A.; Takeuchi, I.; Kusne, A. G. Benchmarking Active Learning Strategies for Materials Optimization and Discovery. arXiv April 12, 2022. http://arxiv.org/abs/2204.05838 (accessed 2022-07-04).

(13) Liang, Q.; Gongora, A. E.; Ren, Z.; Tiihonen, A.; Liu, Z.; Sun, S.; Deneault, J. R.; Bash, D.; Mekki-Berrada, F.; Khan, S. A.; Hippalgaonkar, K.; Maruyama, B.; Brown, K. A.; Fisher III, J.; Buonassisi, T. Benchmarking the Performance of Bayesian Optimization across Multiple Experimental Materials Science Domains. *npj Comput Mater* **2021**, *7* (1), 188. https://doi.org/10.1038/s41524-021-00656-9.

(14) Henderson, A. N.; Kauwe, S. K.; Sparks, T. D. Benchmark Datasets Incorporating Diverse Tasks, Sample Sizes, Material Systems, and Data Heterogeneity for Materials Informatics. *Data in Brief* **2021**, *37*, 107262. https://doi.org/10.1016/j.dib.2021.107262.
(15) Häse, F.; Aldeghi, M.; Hickman, R. J.; Roch, L. M.; Christensen, M.; Liles, E.; Hein, J. E.; Aspuru-Guzik, A. Olympus: A Benchmarking Framework for Noisy Optimization and Experiment Planning. *Mach. Learn.: Sci. Technol.* **2021**, *2* (3), 035021. https://doi.org/10.1088/2632-2153/abedc8.