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 an essential driver of progress in scientific disciplines. Ideal benchmarks mimic real-world tasks as closely as possible, where insufficient difficulty or applicability can stunt growth in the field. Benchmarks should also have sufficiently low computational overhead to promote accessibility and repeatability. The goal is then to win a "Turing test" of sorts by creating a surrogate model that is indistinguishable from the ground truth observation (at least within the dataset bounds that were explored), necessitating a large amount of data. In materials science and chemistry, industry-relevant optimization tasks are often hierarchical, noisy, multi-fidelity, multi-objective, high-dimensional, and non-linearly correlated while exhibiting mixed numerical and categorical variables subject to linear and non-linear constraints. To complicate matters, unexpected, failed simulation or experimental regions may be present in the search space. In this study, 173219 quasi-random hyperparameter combinations were generated across 23 hyperparameters and used to train CrabNet on the Matbench experimental band gap dataset. The results were logged to a free-tier shared MongoDB Atlas dataset. This study resulted in a regression dataset mapping hyperparameter combinations (including repeats) to MAE, RMSE, computational runtime, and model size for CrabNet model trained on the Matbench experimental band gap benchmark task[1]. This dataset is used to create a surrogate model as close as possible to running the actual simulations by incorporating heteroskedastic noise. Failure cases for bad hyperparameter combinations were excluded via careful construction of the hyperparameter search space, and so were not considered as was

done in prior work. For the regression dataset, percentile ranks were computed within each of the groups of identical parameter sets to enable capturing heteroskedastic noise. This contrasts with a more traditional approach that imposes a-priori assumptions such as Gaussian noise, e.g., by providing a mean and standard deviation. A similar approach can be applied to other benchmark datasets to bridge 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 Raw
How the data were acquired	Data was acquired by running CrabNet v2.0.8 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/ with orchestration conducted using Python in https://github.com/sparks-baird/matsci-opt-benchmarks/blob/7c4346624895a7826ada07ff5e44c2f49eb42b9d/scripts/crabnet_hyperparameter/crabnet_hyperparameter_submitit.py. The Python code was run using the University of Utah's Center for High-performance Computing (CHPC) resources. Submitit https://github.com/facebookincubator/submitit was used to send jobs to the SLURM scheduler and the MongoDB Data API was used to log results in JSON format. For a snapshot of the matsci-opt-benchmarks code used, see 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	Twenty-three hyperparameters were varied in a quasi-random Sobol sampling of 65536 parameter combinations using a constrained search space via the Ax Platform, with 5 repeats (total: 327680 training runs). Of these, 173219 ran to completion (387 RTX-2080-Ti GPU days or 4614.29 CUDA core years) with 41550 unique sets.

	Repeat simulations were grouped and ranked by percentile using the “dense” method with pct=True 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 data is useful for adaptive design benchmarking of a high-dimensional, constrained, multi-fidelity task
- Optimization practitioners in the physical sciences can benefit from the data by using it to mimic real materials optimization tasks such as alloy discovery
- The data can be used to understand hyperparameter optimization efforts for compositionally restricted material property prediction models

Objective

In the fields of materials science and chemistry, industry-relevant optimization tasks are often hierarchical, noisy, multi-fidelity[3,4], multi-objective[5,6], high-dimensional[7,8], and non-linearly correlated while exhibiting mixed numerical and categorical variables subject to linear[9] and non-linear constraints. Existing benchmark datasets[1,10–14], while very useful, typically are single-objective, single-fidelity, low-dimensional, and ignore or simplify the influence of noise. By incorporating heteroskedastic noise, we create a “Turing test” of sorts with a surrogate model that is indistinguishable from the ground truth simulation. Using a simultaneously multi-objective, multi-fidelity, and high-dimensional task while considering heteroskedastic noise helps to bridge the gap between cheap-to-evaluate surrogate functions based on benchmark datasets and high-cost, real-world objective function evaluations.

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.

There are six regression models (surrogate_models.pkl) trained on all data meant for production use. These six models can be used together to create the benchmark function.

There are five cross-validation sets of six regression models (cross_validation_models_0.pkl, cross_validation_models_1.pkl, cross_validation_models_2.pkl, cross_validation_models_3.pkl, cross_validation_models_4.pkl).

The model metadata (model_metadata.json) contains the raw mean absolute error scores, the raw predictions, and the true values for each of the cross-validation folds.

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.

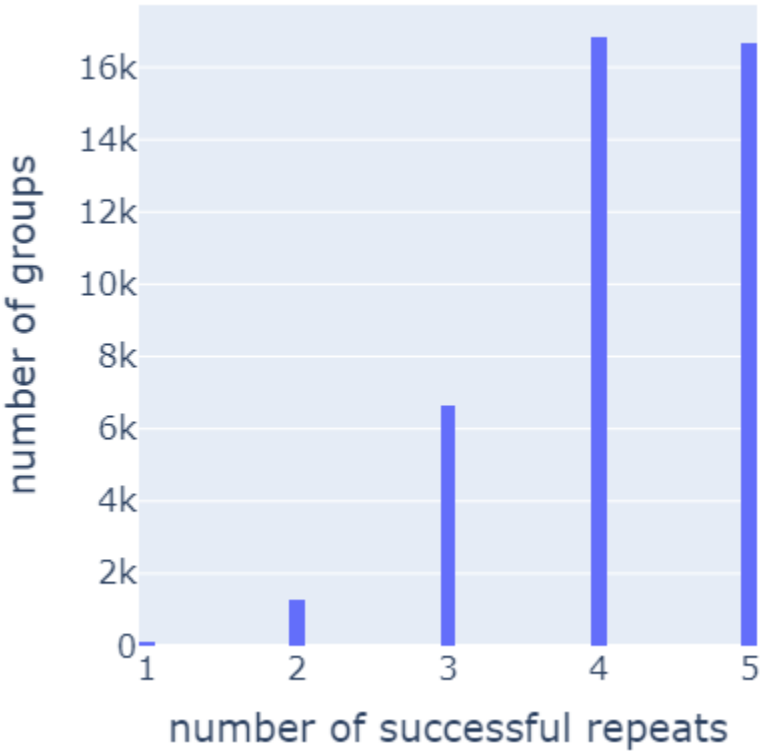

Figure 1. Histogram of number of parameter groups vs. number of successful repeats within a given group. The lowest number of repeats for a parameter set is 1, with approximately 2.6 repeats on average.

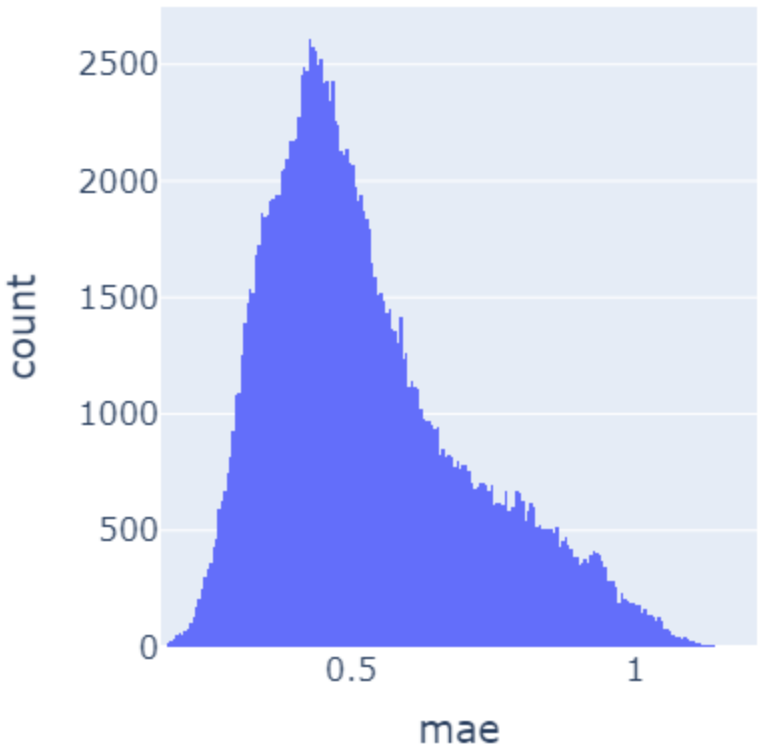

Figure 2. Histogram of number of training runs vs. mean absolute error using CrabNet on the Matbench experimental band gap task.

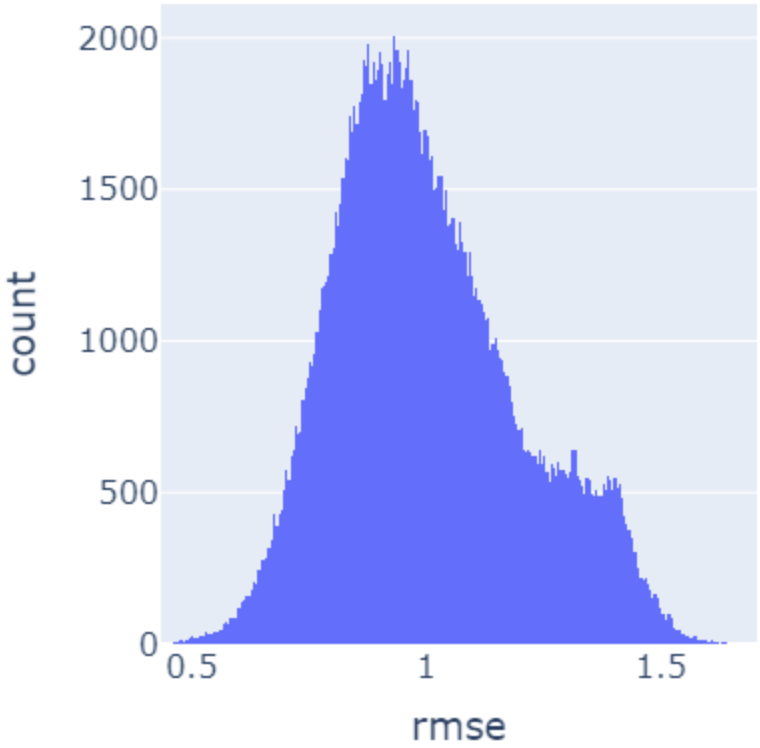

Figure 3. Histogram of number of training runs vs. root-mean-square-error using CrabNet on the Matbench experimental band gap task.

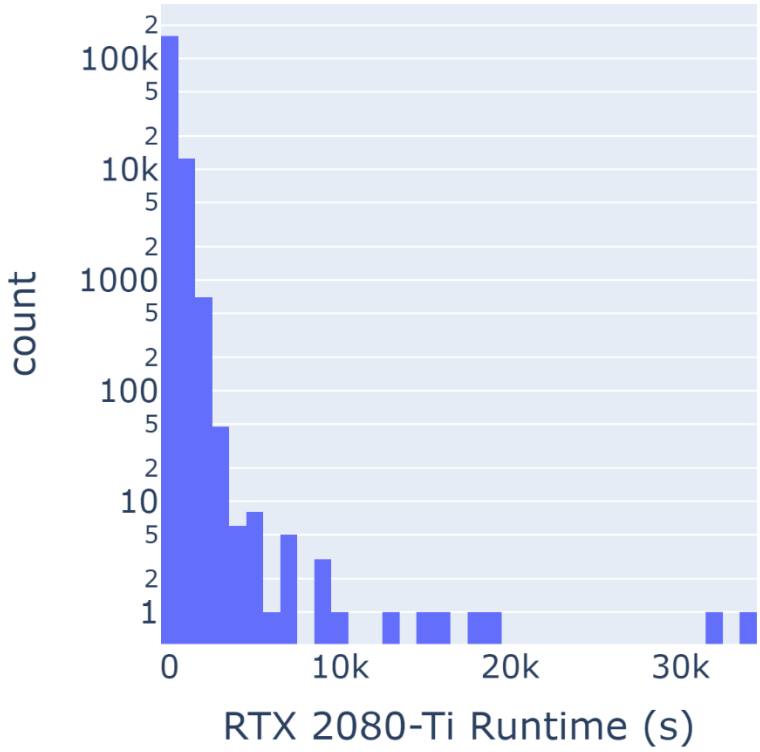

Figure 4. Histogram of number of training runs vs. GPU runtime on an RTX 2080-Ti using CrabNet on the Matbench experimental band gap task. The y-axis is log-scaled.

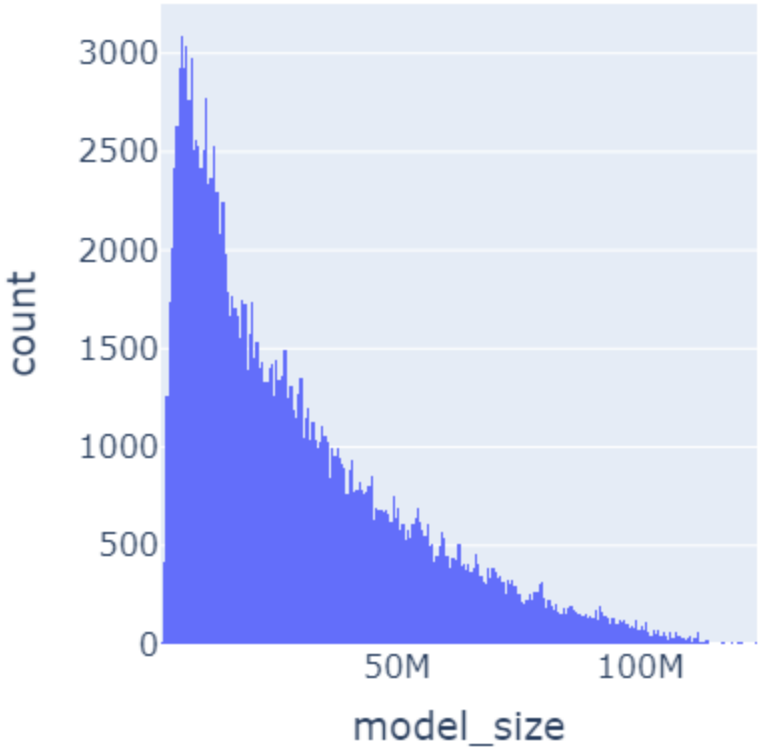

Figure 5. Histogram of number of training runs vs. model size using CrabNet on the Matbench experimental band gap task.

Experimental design, materials and methods

Hundreds of thousands (173219 in total) of CrabNet models were trained using various hyperparameter combinations. The unique parameter combinations were obtained using quasi-random Sobol sampling of the constrained feature space using the Ax platform.

Quasi-random Sobol sampling was used to generate parameter combinations to obtain a more uniform sampling of the allowable parameter space. 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 the noise for this benchmark dataset, simulations were repeated for each of the quasi-random parameter combinations. To maximize throughput and reduce latency, hyperparameter sets (including repeats) were shuffled and divided into batches and sent to a high-performance computing environment for asynchronous evaluation. Some results did not complete due to either timeout or preemption, which is seen as a reasonable trade-off for the gains in efficiency of implementation and completion.

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. Instructions for model usage will be made available at https://matsci-opt-benchmarks.readthedocs.io/.

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.

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) 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.
(2) 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.
(3) 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.

(4) 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**.
(5) 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.
(6) 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.
(7) 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.
(8) Eriksson, D.; Jankowiak, M. High-Dimensional Bayesian Optimization with Sparse Axis-Aligned Subspaces. *arXiv:2103.00349 [cs, stat]* **2021**.
(9) 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.
(10) 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.
(11) 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).
(12) 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.
(13) 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.
(14) 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.
(15) Mościński, J.; Bargieł, M.; Rycerz, Z. A.; Jacobs, P. W. M. The Force-Biased Algorithm for the Irregular Close Packing of Equal Hard Spheres. *Molecular Simulation* **1989**, *3* (4), 201–212. https://doi.org/10.1080/08927028908031373.
(16) Bezrukov, A.; Bargieł, M.; Stoyan, D. Statistical Analysis of Simulated Random Packings of Spheres. *Particle & Particle Systems Characterization* **2002**, *19* (2), 111–118. https://doi.org/10.1002/1521-4117(200205)19:2<111::AID-PPSC111>3.0.CO;2-M.
(17) Skoge, M.; Donev, A.; Stillinger, F. H.; Torquato, S. Packing Hyperspheres in High-Dimensional Euclidean Spaces. *Phys. Rev. E* **2006**, *74* (4), 041127. https://doi.org/10.1103/PhysRevE.74.041127.
(18) Lubachevsky, B. D. How to Simulate Billiards and Similar Systems. *Journal of Computational Physics* **1991**, *94* (2), 255–283. https://doi.org/10.1016/0021-9991(91)90222-7.
(19) Lubachevsky, B. D.; Stillinger, F. H. Geometric Properties of Random Disk Packings. *Journal of Statistical Physics* **1990**, *60* (5), 561–583. https://doi.org/10.1007/BF01025983.