Article information

Article title

Materials Science Optimization Benchmark Dataset for Multi-fidelity Hard-sphere Packing Simulations

Authors

Sterling G. Baird[1]*, Taylor D. Sparks[1]

Affiliations

1. Materials Science & Engineering, 122 S. Central Campus Drive, #304 Salt Lake City, Utah 84112-0056

Corresponding author's email address and Twitter handle

sterling.baird@utah.edu
@SterlingBaird1

Keywords

adaptive design, physics-based, Lubachevsky–Stillinger, force-biased algorithms, particle packing, packing generation, transfer learning, size distribution

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, 438371 random hard-sphere packing simulations representing 279 CPU days' worth of computational overhead were performed across nine input parameters with linear constraints and two discrete fidelities each with continuous fidelity parameters and results were logged to a free-tier shared MongoDB Atlas database. Two core tabular datasets resulted from this study: 1. a failure probability dataset containing unique input parameter sets and the estimated probabilities that the simulation will fail at each of the two steps, and 2. a regression dataset mapping input parameter sets (including repeats) to particle packing fractions and computational runtimes for each of the two steps. These two datasets can be used to create a surrogate model as close as possible to running the actual simulations by

incorporating simulation failure and heteroskedastic noise. 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	Physics-based geometric packing
Type of data	Table Figure
How the data were acquired	Data was acquired by running compiled C software hosted at https://github.com/VasiliBaranov/packing-generation in a two-step process orchestrated using Python in https://github.com/sparks-baird/matsci-opt-benchmarks/blob/3c0a74b1a594d5628bde232062e55804590c4e1f/src/matsci_opt_benchmarks/particle_packing/utils/packing_generation.py#L61-L181. The Python code called the compiled packing generation executable and was run using the University of Utah's Center for High-performance Computing (CHPC) resources. 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 v0.1.0 (https://zenodo.org/badge/latestdoi/577963870).
Data format	Analyzed Filtered
Description of data collection	Seven irreducible parameters plus number of particles and initial scaling factor were varied in a quasi-random Sobol sampling of 65536 parameter combinations using a constrained search space via the Ax Platform, with 15 repeats (total: 983040 simulations). Of these, 438371 ran to completion (279 CPU days) with 41228 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: 7513019 Direct URL to data: https://doi.org/10.5281/zenodo.7513019

Value of the data

- The data is useful for adaptive design benchmarking of a constrained, multi-fidelity task
- Optimization practitioners in the physical sciences can benefit from the data
- The data can be used to understand packing with different particle types in powder-bed additive manufacturing experiments and could be complemented with experimental data

Objective

In the fields of materials science and chemistry, industry-relevant optimization tasks 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. Existing benchmark datasets[8–13], while very useful, typically ignore or simplify the influence of noise and ignore the fact that certain parameter combinations will result in failure. By incorporating simulation failure and heteroskedastic noise, we create a “Turing test” of sorts with a surrogate model that is indistinguishable from the ground truth simulation. This bridges the gap between cheap-to-evaluate surrogate functions based on benchmark datasets and high-cost, real-world objective function evaluations.

Data description

The failure probability dataset contains unique input parameter sets (nine variables) and the estimated probabilities that the simulation will fail at each of the two steps (force-biased algorithm and Lubachevsky–Stillinger).

The regression dataset contains input parameter (including repeats) spanning nine variables and corresponding particle packing fractions and computational runtimes for each of the two steps (force-biased algorithm and Lubachevsky–Stillinger).

For histogram data summarizing characteristics of the two datasets, see Figure 1, Figure 2, and Figure 3.

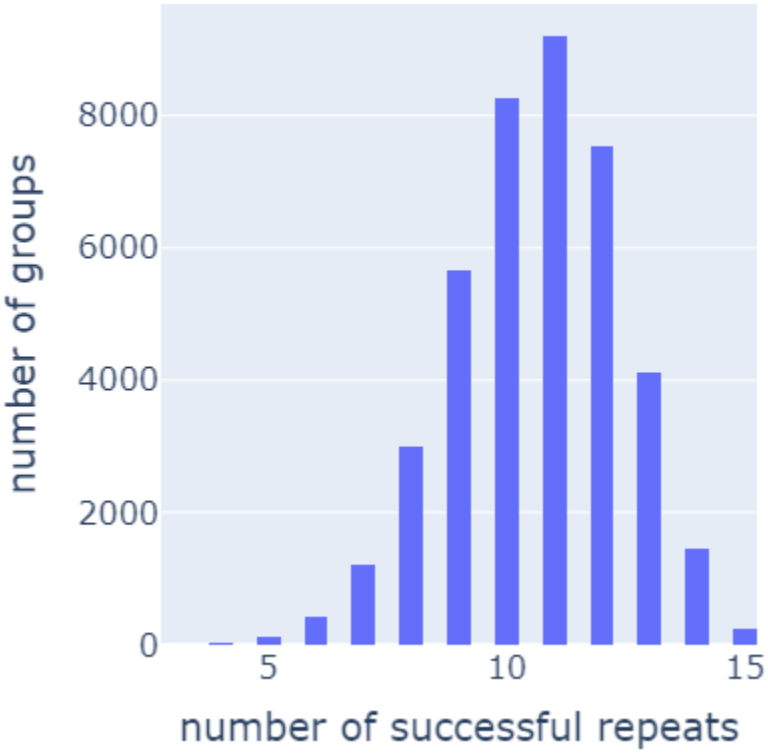

Figure 1. Histogram of number of parameter groups vs. number of successful repeats within a given group. The lowest number of repeats for a parameter set is 3, with approximately 10 repeats on average.

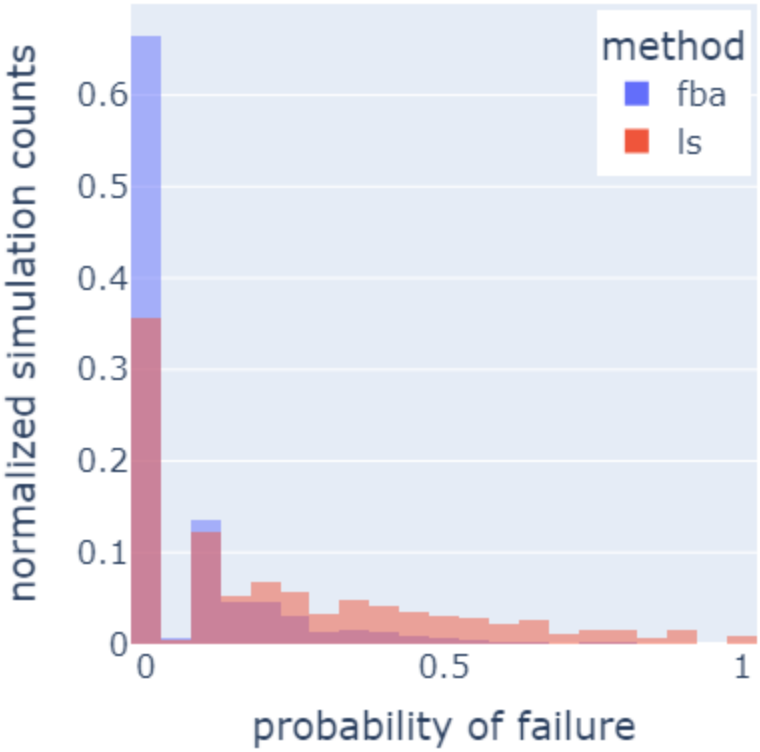

Figure 2. Histogram of normalized simulation counts vs. the probability of a simulation failing for a given parameter set. On average, the force-biased algorithm or fba (blue) is more likely to succeed than the Lubachevsky–Stillinger or ls (red) algorithm.

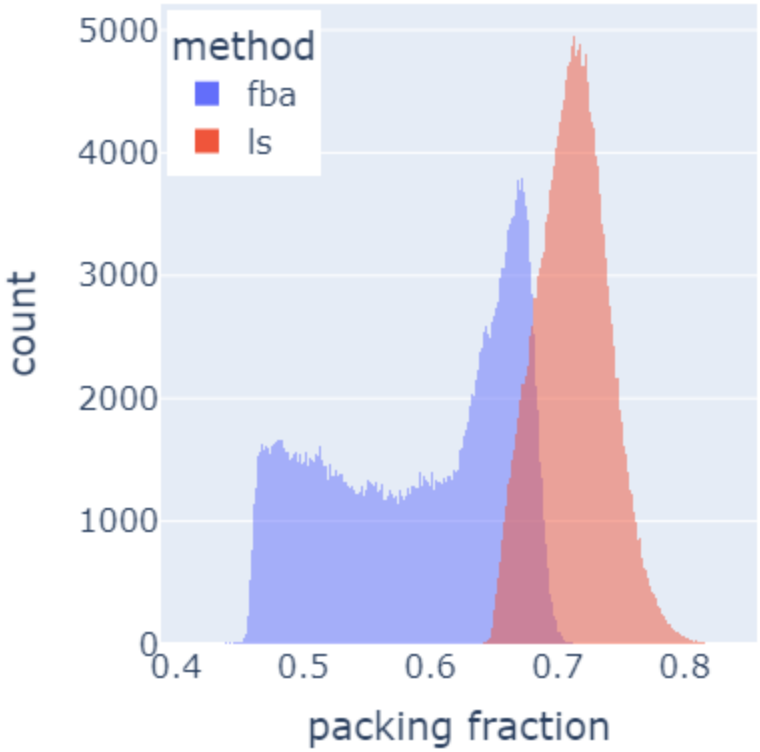

Figure 3. Histogram of number of simulations vs. packing fraction for the force-biased algorithm or fba (blue) and Lubachevsky–Stillinger or ls algorithm (red). On average, the ls algorithm tends to have higher packing fractions with a more Gaussian-like distribution than fba.

Experimental design, materials and methods

Hundreds of thousands (438371 in total) random hard-sphere packing simulations were performed using a two-step process: a force-biased algorithm[14,15] followed by the Lubachevsky–Stillinger algorithm[16–18]. Three truncated log-normal distributions were used to approximate realistic mixtures of three different particle types: i.e., if you were to separately weigh out three different powders and mix them together. Two parameters (scale and shape) describe each of the three distributions, and three additional composition parameters describe the fractional share (e.g., in terms of volume) of each of the particle types. Additionally, the number of particles (100-1000) and an initial scaling factor were allowed to vary. With a greater number of particles, denser and more realistic packs can be generated at the expense of computational cost (i.e., the fidelity parameter). The initial scaling factor affects the computational stability of the simulation; with an adequate scaling factor, the simulation is more likely to complete successfully. 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-fidelity benchmark dataset for constrained adaptive design scenarios. 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, simulation parameters (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. Most parameter combinations had at least 8 repeats. 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/particle_packing and https://github.com/sparks-baird/matsci-opt-benchmarks/tree/main/notebooks/particle_packing.

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. **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 acknowledge Vasili Baranov and Robin De Schepper for help with the packing-generation codebase. We acknowledge Jason R. Hall for discussion of particle packing simulations in the context of industry processes.

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) 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.
(9) 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.
(10) 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).
(11) 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.
(12) 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.
(13) 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.
(14) 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.
(15) 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.
(16) 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.
(17) 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.
(18) 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.