***Data in Brief* article template** *Version 13 (August 2022)*

Before you complete this template, a few important points to note:

. It is important to double check that your manuscript **meets the data article criteria** on our webpage: [What Data are Suitable for Data in Brief](https://www.journals.elsevier.com/data-in-brief/policies-and-guidelines/what-data-are-suitable-for-data-in-brief).

· The format of a data article is different from that of a traditional research article. To help you write yours, we have created a template on page 3. **We will only consider data articles submitted using this template**.

· It is mandatory to **publicly share the research data in a repository** referred to in your *Data in Brief* article (you will find information on our data sharing criteria in the template and on the journal homepage at [Public repositories to store and find data - Policies and Guidelines - Data in Brief - Journal - Elsevier](https://journals.elsevier.com/data-in-brief/policies-and-guidelines/public-repositories-to-store-and-find-data))

· Please consult the *Data in Brief* [Guide for Authors](https://www.elsevier.com/journals/data-in-brief/2352-3409/guide-for-authors) when preparing your manuscript; it highlights mandatory requirements and is packed with useful advice. We have also developed a [step-by-step video guide](https://www.journals.elsevier.com/data-in-brief/about-data-in-brief/video-how-to-submit-your-research-data-article-to-data-in-brief) to help you complete this template accurately and increase your chances of acceptance.

· If a dataset previously described in a *Data in Brief* article has been substantially expanded, you may choose to publish an “update” article. Please write to [dib-me@elsevier.com](mailto:dib-me@elsevier.com) to request a template.

**Still have questions?**

Ø  [Visit our FAQ page](https://www.journals.elsevier.com/data-in-brief/about-data-in-brief/data-in-brief-faq)

Ø Email our Scientific Editors at dib-me@elsevier.com

**Companion article instructions**

Data articles support original research. They improve research integrity and reproducibility by making the data available and discoverable and reduce unwanted duplication of research. A companion data article can be submitted directly to *Data in Brief* following the acceptance of your original research.

Please note that the article template asks for the details of your accepted manuscript so that we can link your original research article with your companion data article.

**Please use the article template to write your data article:**

· Step 1 – Write

Fill in the template provided on page 3. The template shows you exactly what reviewers are looking for in your data article.

· Step 2 – Share

Make your data publicly accessible by posting it to one of the repositories we support. You may find the list of data repositories at [FAIRsharing : Elsevier's Data Policy](https://fairsharing.org/bsg-p000100/) and [Public repositories to store and find data - Policies and Guidelines - Data in Brief - Journal - Elsevier](https://journals.elsevier.com/data-in-brief/policies-and-guidelines/public-repositories-to-store-and-find-data)

· Step 3 – Submit

Submit your article with the shared data directly to *Data in Brief* or together with your original research article.

**Data article template**

Please fill in the template below. As you complete each section, it is important to carefully read the associated instructions, which you will find in blue, italicized text.

**Please delete this line and everything above it before submitting your article. In addition, we ask you to delete the instructions, including those featured in the Article Information and Specifications table.**

- - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -

**Article information**

**Article title**

*The article title must include the word ‘data’ or ‘dataset’. Please avoid acronyms and abbreviations where possible. Where your data article supports an original research article, the title should differ from the title used in your research paper.*

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

**Authors**

*List all authors. Please mark the corresponding author with (\*).*

Sterling G. Baird1\*, Taylor D. Sparks1

**Affiliations**

*Please include the full address of each author institution.*

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

**Corresponding author’s email address and Twitter handle**

*Institutional email address preferred.*

*If you have a Twitter handle, please add it here ‘twitter: @....’*

[sterling.baird@utah.edu](mailto:sterling.baird@utah.edu)

@SterlingBaird1

**Keywords**

*Include 4-8 keywords (or phrases) to help others discover your article online. Avoid repeating words used in your title.*

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

**Abstract**

*The abstract should describe the data collection process, the analysis performed, the data, and their reuse potential. It should* ***not*** *provide conclusions or interpretive insights. Minimum length 100 words / maximum length 500 words.*

*Tip: Do not use words such as ‘study’, ‘results’ and ‘conclusions’ – a data article should only describe your data.*

Hundreds of thousands (438371 in total) random hard-sphere packing simulations were performed using a two-step process: a force-biased algorithm1,2 followed by the Lubachevsky–Stillinger algorithm3–5. Three truncated log-normal distributions were used to approximate realistic mixtures of three different particle types. 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. Two core tabular datasets resulted from this study: 1. a failure probability dataset containing unique input parameter sets and the estimated probability that the force-biased and/or Lubachevsky-Stillinger algorithms will fail, and 2. a regression dataset mapping input parameter sets (including repeats) to particle packing fractions and computational runtimes at the end of 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. The goal is to win a “Turing test” of sorts by creating a surrogate model that is indistinguishable from the ground truth simulation (at least within the dataset bounds that were explored). For the regression dataset, percentile ranks were computed within each of the groups of identical parameter sets to enable capturing heteroskedastic noise. This is in contrast 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**

*Every section of this table is mandatory. Please enter information in the right-hand column and remove all the instructions in blue, italicized text.*

|  |  |
| --- | --- |
| **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. |
| **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** | ***All data referred to in your data article must be made publicly available prior to publication.***  · *All* ***raw data*** *linked to charts, graphs or figures in your manuscript, should be shared via a data repository.*  · *If you are describing* ***secondary data*** *(data that has already been collected through primary sources and made readily available for researchers to use for their own research)****,*** *you are required to provide a list of the primary data sources used and make the full secondary dataset publicly available via a data repository.*  *Please deposit your data in a* [*data repository*](https://journals.elsevier.com/data-in-brief/policies-and-guidelines/public-repositories-to-store-and-find-data)*. View our supported data repositories on* [*FAIRsharing.org*](https://fairsharing.org/bsg-p000100/) *and our* [*short list of data repositories*](https://www.journals.elsevier.com/data-in-brief/policies-and-guidelines/public-repositories-to-store-and-find-data)*. If there is no specialized repository for your data, please use a generalized repository such as* [*Mendeley Data*](https://data.mendeley.com/)*.*  *Please indicate here where your data are hosted.*  Repository name:  Data identification number:  Direct URL to data:*e.g.,* [*https://www.data.edu.com*](https://www.data.edu.com/) *– the URL should be working at the time of submission.*  *For data that require access controls for ethical reasons (e.g., patient data), please describe how readers can request access. You must also provide a link to any Data Use Agreement (DUA) with the proprietor of the data or upload a copy as a supplementary file.*  Instructions for accessing these data:  ***Important:*** *if your data has access controls, you must provide a mechanism that allows our Editors and reviewers to access the data without compromising their anonymity. Please include these instructions with your submission. Contact the Scientific Editors (DIB-ME@elsevier.com) if you have any questions.* |
| **Related research article** | *If your data article supports an original research article, please cite the associated research article here. You should only list one article.*  ***For a published article:***  *J. van der Geer, J.A.J. Hanraads, R.A. Lupton, The art of writing a scientific article, J. Sci. Commun. 163 (2010) 51–59.* [*https://doi.org/10.1016/j.Sc.2010.00372*](https://doi.org/10.1016/j.Sc.2010.00372)  ***For an article which has been accepted and is in press:***  *J. van der Geer, J.A.J. Hanraads, R.A. Lupton, The art of writing a scientific article, J. Sci. Commun. In Press.*  *If your data article is* ***not related to a research article****, please delete this entire ‘Related research article’ row.* |

**Value of the data**

*Provide up to 6 bullet points explaining why these data are of value to the scientific community. At a minimum you must answer the first 3 bullet point questions below (then those questions should be deleted). Please keep your points brief – ideally, each one should be no longer than 400 characters. Avoid any conclusions or inferences.*

*· Why are these data useful?*

*· Who can benefit from these data?*

*· How can these data be used/reused for further insights and/or development of experiments?*

* The data is useful for constrained, multi-fidelity adaptive design benchmarking
* 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**

*Describe the reasoning and context behind the generation of this dataset. Do not make concluding, interpretive, or otherwise inferential statements about the dataset. In case your data article is related to an original research article, please briefly describe how the data article adds value to the published article (Max. 200 words).*

Existing benchmark datasets 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 goal of this section is to describe all the raw data files.*

*Individually describe the data files that appear in this article (not the article it supports), e.g., figure 1, figure 2, table 1, dataset, raw data, and data shared in the repository, etc. Ensure you refer to every file separately and provide a clear description for each one (do not just list them). Please use a table/figure/any other visual aid to familiarize readers with the data in the repository. Headings/legends should be included for tables, figures and graphs. No insights, interpretations, background, or conclusions should appear in this section.*

Chart, histogram

Description automatically generated

Chart

Description automatically generated

Chart

Description automatically generated

**Experimental design, materials and methods**

*Provide a complete description of the experimental design and methods used to acquire these data. Include any program or code files used for data filtering or analysis. It is important that this section is as comprehensive as possible. If this is a companion article, this section should contain more detail than the corresponding section of your accompanying research article, and it should be written in original text (i.e., avoid textual overlap). There is no character limit; however, no insights, interpretations, or background should be included.*

*For questionnaires, please indicate the sources from where the questionnaire items have been derived/adapted. If items are significantly adapted, please indicate the degree of adaptation (e.g. provide the original item and the adapted item and highlight the changes).*

*Tip: Do not describe your data (figures, tables, etc.) here – you should have already done that in the Data description section above.*

**Ethics statements**

There are no statements to declare.

**CRediT author statement**

*CRediT is an initiative that enables authors to share an accurate and detailed description of their diverse contributions to a published work.*

*Example of a CRediT author statement:*

***Zhang San:*** *Conceptualization, Methodology, Software* ***Priya Singh****: Data curation, Writing- Original draft preparation.* ***Wang Wu****: Visualization, Investigation.* ***Jan Jansen:*** *Supervision.* ***Ajay Kumar****: Software, Validation.:* ***Sun Qi:*** *Writing- Reviewing and Editing.*

*Please add a CRediT author statement for your data article here, using the* [*categories listed on this web*](https://www.elsevier.com/authors/journal-authors/policies-and-ethics/credit-author-statement)*page.*

**Acknowledgments**

*All contributors who do not meet the criteria for authorship should be listed in an acknowledgments section.*

*In addition, please list any funding sources in this section. List funding sources in this standard way to facilitate compliance to funder's requirements:*

*Funding: This work was supported by the National Institutes of Health [grant numbers xxxx, yyyy]; the Bill & Melinda Gates Foundation, Seattle, WA [grant number zzzz]; and the United States Institutes of Peace [grant number aaaa].*

*It is not necessary to include detailed descriptions of the program or type of grants and awards. When funding is from a block grant or other resources available to a university, college, or other research institution, submit the name of the institute or organization that provided the funding.*

*If no funding has been provided for the research, please include the following sentence:*

*This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.*

**Declaration of interests**

*Please* ***tick*** *the appropriate statement below and declare any financial interests/personal relationships which may affect your work in the box below.*

☐ 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:

*Please declare any financial interests/personal relationships which may be considered as potential competing interests here.*

**References**

*References are limited to* ***a maximum of 20 and irrelevant self-citation is not allowed****. Please cite any article you have referred to while collecting or analyzing data and preparing your manuscript.* ***If your data article supports an original research article which is published or in press, please cite the associated article here; ideally, it should be the first citation****.* ***Please also make sure to cite the corresponding dataset.***

*Please review the References section in the* [*Guide for Authors*](https://www.elsevier.com/journals/data-in-brief/2352-3409/guide-for-authors#txt68000) *for a comprehensive guide to Data in Brief reference style and then list your references below, numbering each one as you go. You can also find the referencing style* [*here*](https://endnote.com/style_download/data-in-brief/).

- - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -

***Reminder: Before you submit, please delete all the instructional text in italicized text, wherever it appears – including here. Thank you!***

***And finally, we welcome your feedback on how we can improve this template.***[*Click here*](https://forms.office.com/Pages/ResponsePage.aspx?id=P-50kiWUCUGif5-xXBBnXTeXkbO343VFrbpYVBvxdZtUM05UVjIwM0U4WlRKUldCOTNMRUQwOVRHTy4u) ***to provide anonymous feedback via a very short survey.***