





1 matbench-genmetrics: A Python library for
2 benchmarking crystal structure generative models
3 using time-based splits of Materials Project structures

4 **Sterling G. Baird** ¹, **Joseph Montoya** ², and **Taylor D. Sparks** ¹

5 ¹ Materials Science & Engineering, University of Utah, USA ² Toyota Research Institute, Los Altos, CA,
6 USA  Corresponding author

DOI: [10.xxxxxx/draft](https://doi.org/10.xxxxxx/draft)

Software

- [Review](#) 
- [Repository](#) 
- [Archive](#) 

Editor: Sophie Beck 

Reviewers:

- [@ml-evs](#)
- [@mkhorton](#)
- [@jamesrhester](#)

Submitted: 17 June 2023

Published: unpublished

License

Authors of papers retain copyright
and release the work under a
Creative Commons Attribution 4.0
International License ([CC BY 4.0](https://creativecommons.org/licenses/by/4.0/)).

7 **Summary**

8 The progress of a machine learning field is both tracked and propelled through the development
9 of robust benchmarks. While significant progress has been made to create standardized, easy-
10 to-use benchmarks for molecular discovery (e.g., ([Brown et al., 2019](#))), this remains a challenge
11 for solid-state material discovery ([Alverson et al., 2022](#); [Xie et al., 2022](#); [Zhao et al., 2022](#)). To
12 address this limitation, we propose matbench-genmetrics, an open-source Python library for
13 benchmarking generative models for crystal structures. We use four evaluation metrics inspired
14 by Guacamol ([Brown et al., 2019](#)) and Crystal Diffusion Variational AutoEncoder (CDVAE) ([Xie](#)
15 [et al., 2022](#))—validity, coverage, novelty, and uniqueness—to assess performance on Materials
16 Project data splits using timeline-based cross-validation. We believe that matbench-genmetrics
17 will provide the standardization and convenience required for rigorous benchmarking of crystal
18 structure generative models. A visual overview of the matbench-genmetrics library is provided
19 in [Figure 1](#).

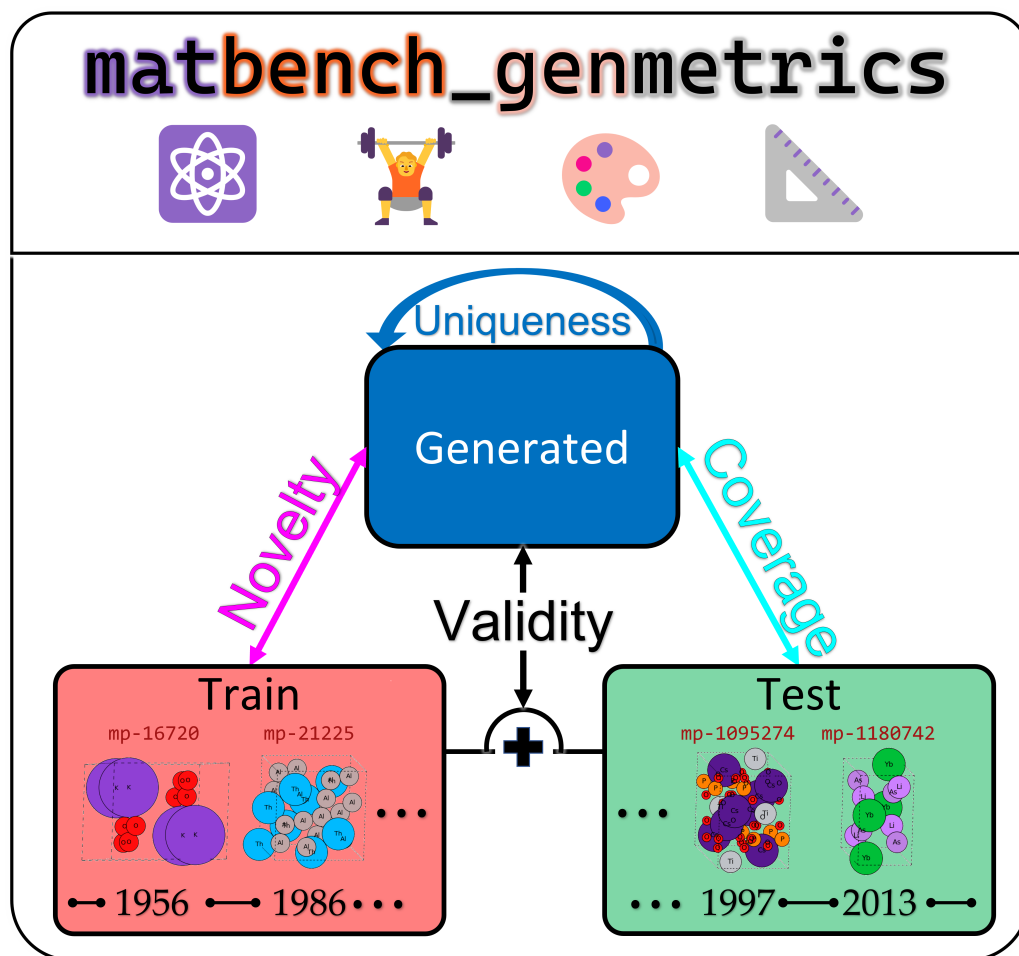


Figure 1: Summary visualization of matbench-genmetrics to evaluate crystal generative model performance using validity, coverage, novelty, and uniqueness metrics based on calendar-time splits of experimentally determined Materials Project database entries. Validity is the comparison of distribution characteristics (space group number) between the generated materials and the training and test sets. Coverage is the number of matches between the generated structures and a held-out test set. Novelty is a comparison between the generated and training structures. Finally, uniqueness is a measure of the number of repeats within the generated structures (i.e., comparing the set of generated structures to itself). For in-depth descriptions and equations for the four metrics described above, see <https://matbench-genmetrics.readthedocs.io/en/latest/readme.html> and <https://matbench-genmetrics.readthedocs.io/en/latest/metrics.html>.

Statement of need

In the field of materials informatics, where materials science intersects with machine learning, benchmarks play a crucial role in assessing model performance and enabling fair comparisons among various tools and models. Typically, these benchmarks focus on evaluating the accuracy of predictive models for materials properties, utilizing well-established metrics such as mean absolute error (MAE) and root-mean-square error (RMSE) to measure performance against actual measurements. A standard practice involves splitting the data into two parts, with one serving as training data for model development and the other as test data for assessing performance (Dunn et al., 2020).

However, benchmarking generative models, which aim to create entirely new data rather than focusing solely on predictive accuracy, presents unique challenges. While significant progress has

been made in standardizing benchmarks for tasks like image generation and molecule synthesis, the field of crystal structure generative modeling lacks this level of standardization. Molecular generative modeling benefits from widely adopted benchmark platforms such as Guacamol (Brown et al., 2019) and Moses (Polykovskiy et al., 2020), which offer easy installation, usage guidelines, and leaderboards for tracking progress. In contrast, existing evaluations in crystal structure generative modeling, as seen in CDVAE (Xie et al., 2022), FTCP (Ren et al., 2022), PGCGM (Zhao et al., 2022), CubicGAN (Zhao et al., 2021), and CrysTens (Alverson et al., 2022), lack standardization, pose challenges in terms of installation and application to new models and datasets, and lack publicly accessible leaderboards. While these evaluations are valuable within their respective scopes, there is a clear need for a dedicated benchmarking platform to promote standardization and facilitate robust comparisons.

In this work, we introduce matbench-genmetrics, a materials benchmarking platform for crystal structure generative models. We use concepts from molecular generative modeling benchmarking to create a set of evaluation metrics—validity, coverage, novelty, and uniqueness—which are broadly defined as follows:

- **Validity:** a measure of how well the generated materials match the distribution of the training dataset
- **Coverage:** the ability to successfully predict known materials which have been held out
- **Novelty:** generating structures which are close matches to examples in the training set are penalized
- **Uniqueness:** the number of repeats within the generated structures

matbench-genmetrics is comprised of two namespace packages. The first is `matbench_genmetrics.core`, which provides the following features:

- GenMatcher: A class for calculating matches between two sets of structures
- GenMetrics: A class for calculating validity, coverage, novelty, and uniqueness metrics
- MPTSMetrics: class for loading `mp_time_split` data, calculating time-series cross-validation metrics, and saving results
- Fixed benchmark classes for 10, 100, 1000, and 10000 generated structures

Additionally, we introduce the `matbench_genmetrics.mp_time_split` namespace package as a complement to `matbench_genmetrics.core`. It provides a standardized dataset and cross-validation splits for evaluating the mentioned four metrics. Time-based splits have been utilized in materials informatics model validation, such as predicting future thermoelectric materials via word embeddings (Tshitoyan et al., 2019), searching for efficient solar photoabsorption materials through multi-fidelity optimization (Palizhati et al., 2022), and predicting future materials stability trends via network models (Aykol et al., 2019). Recently, Hu et al. (Zhao et al., 2022) used what they call a rediscovery metric, referred to here as a coverage metric in line with molecular benchmarking terminology, to evaluate crystal structure generative models. While time-series splitting wasn't used, they showed that after generating millions of structures, only a small percentage of held-out structures had matches. These results highlight the difficulty (and robustness) of coverage tasks. By leveraging timeline metadata from the Materials Project database (Jain et al., 2013) and creating a standard time-series splitting of data, `matbench_genmetrics.mp_time_split` enables rigorous evaluation of future discovery performance.

The `matbench_genmetrics.mp_time_split` namespace package provides the following features:

- downloading and storing snapshots of Materials Project crystal structures via `pymatgen` (Ong et al., 2013)
- modification of `pymatgen` search criteria to fetch custom datasets
- utilities for post-processing Materials Project entries
- convenience methods to access the snapshot dataset
- predefined scikit-learn `TimeSeriesSplit` cross-validation splits (Ong et al., 2013)

In future work, metrics will serve as multi-criteria filters to prevent manipulation. Stand-alone metrics can be “hacked” by generating nonsensical structures for novelty or including training structures to inflate validity scores. To address this, multiple criteria are considered simultaneously for each generated structure, such as novelty, uniqueness, and filtering rules like non-overlapping atoms, stoichiometry, or checkCIF criteria (Spek, 2020). Additional filters based on machine learning models can be applied for properties like negative formation energy, energy above hull, ICSD classification, and coordination number. Applying machine-learning-based structural relaxation using M3GNet (Chen & Ong, 2022) (e.g., as in CrysTens (Alverson et al., 2022)) before filtering is also of interest. Contributions related to multi-criteria filtering, enhanced validity filters, and implementing a benchmark submission system and public leaderboard are welcome.

We believe that the matbench-genmetrics ecosystem is a robust and easy-to-use benchmarking platform that will help propel novel materials discovery and targeted crystal structure inverse design. We hope that practitioners of crystal structure generative modeling will adopt matbench-genmetrics, contribute improvements and ideas, and submit their results to the planned public leaderboard.

Acknowledgements

We acknowledge contributions from Kevin M. Jablonka (@kjappelbaum), Matthew K. Horton (@mkhorton), Kyle D. Miller (@kyledmiller), and Janosh Riebesell (@janosh). S.G.B. and T.D.S. acknowledge support by the National Science Foundation, USA under Grant No. DMR-1651668. We acknowledge OpenAI for the use of ChatGPT for basic proofreading and editing.

References

- Alverson, M., Baird, S., Murdock, R., & Sparks, T. (2022). *Generative adversarial networks and diffusion models in material discovery*. <https://doi.org/10.26434/chemrxiv-2022-6l4pm>
- Aykol, M., Hegde, V. I., Hung, L., Suram, S., Herring, P., Wolverton, C., & Hummelshøj, J. S. (2019). Network analysis of synthesizable materials discovery. *Nature Communications*, 10(1), 2018. <https://doi.org/10.1038/s41467-019-10030-5>
- Brown, N., Fiscato, M., Segler, M. H. S., & Vaucher, A. C. (2019). GuacaMol: Benchmarking Models for de Novo Molecular Design. *Journal of Chemical Information and Modeling*, 59(3), 1096–1108. <https://doi.org/10.1021/acs.jcim.8b00839>
- Chen, C., & Ong, S. P. (2022). A universal graph deep learning interatomic potential for the periodic table. *Nature Computational Science*, 2(11), 718–728. <https://doi.org/10.1038/s43588-022-00349-3>
- Dunn, A., Wang, Q., Ganose, A., Dopp, D., & Jain, A. (2020). Benchmarking materials property prediction methods: The Matbench test set and Automatminer reference algorithm. *Npj Computational Materials*, 6(1), 1–10. <https://doi.org/10.1038/s41524-020-00406-3>
- Jain, A., Ong, S. P., Hautier, G., Chen, W., Richards, W. D., Dacek, S., Cholia, S., Gunter, D., Skinner, D., Ceder, G., & Persson, K. A. (2013). Commentary: The Materials Project: A materials genome approach to accelerating materials innovation. *APL Materials*, 1(1), 011002. <https://doi.org/10.1063/1.4812323>
- Ong, S. P., Richards, W. D., Jain, A., Hautier, G., Kocher, M., Cholia, S., Gunter, D., Chevrier, V. L., Persson, K. A., & Ceder, G. (2013). Python Materials Genomics (pymatgen): A robust, open-source python library for materials analysis. *Computational Materials Science*, 68, 314–319. <https://doi.org/10.1016/j.commatsci.2012.10.028>

- 125 Palizhati, A., Torrisi, S. B., Aykol, M., Suram, S. K., Hummelshøj, J. S., & Montoya, J.
126 H. (2022). Agents for sequential learning using multiple-fidelity data. *Scientific Reports*,
127 12(1), 4694. <https://doi.org/10.1038/s41598-022-08413-8>
- 128 Polykovskiy, D., Zhebrak, A., Sanchez-Lengeling, B., Golovanov, S., Tatanov, O., Belyaev,
129 S., Kurbanov, R., Artamonov, A., Aladinskiy, V., Veselov, M., Kadurin, A., Johansson,
130 S., Chen, H., Nikolenko, S., Aspuru-Guzik, A., & Zhavoronkov, A. (2020). Molecular
131 sets (MOSES): A benchmarking platform for molecular generation models. *Frontiers in*
132 *Pharmacology*, 11. <https://doi.org/10.3389/fphar.2020.565644>
- 133 Ren, Z., Tian, S. I. P., Noh, J., Oviedo, F., Xing, G., Li, J., Liang, Q., Zhu, R., Aberle,
134 A. G., Sun, S., Wang, X., Liu, Y., Li, Q., Jayavelu, S., Hippalgaonkar, K., Jung, Y., &
135 Buonassisi, T. (2022). An invertible crystallographic representation for general inverse
136 design of inorganic crystals with targeted properties. *Matter*, 5(1), 314–335. <https://doi.org/10.1016/j.matt.2021.11.032>
- 137
138 Spek, A. L. (2020). *checkCIF* validation ALERTS: What they mean and how to respond.
139 *Acta Crystallographica Section E Crystallographic Communications*, 76(1), 1–11. <https://doi.org/10.1107/S2056989019016244>
- 140
141 Tshitoyan, V., Dagdelen, J., Weston, L., Dunn, A., Rong, Z., Kononova, O., Persson, K. A.,
142 Ceder, G., & Jain, A. (2019). Unsupervised word embeddings capture latent knowledge
143 from materials science literature. *Nature*, 571(7763), 95–98. <https://doi.org/10.1038/s41586-019-1335-8>
- 144
145 Xie, T., Fu, X., Ganea, O.-E., Barzilay, R., & Jaakkola, T. (2022). Crystal Diffusion Vari-
146 ational Autoencoder for Periodic Material Generation. *arXiv:2110.06197 [Cond-Mat,*
147 *Physics:physics]*. <https://arxiv.org/abs/2110.06197>
- 148 Zhao, Y., Al-Fahdi, M., Hu, M., Siriwardane, E. M., Song, Y., Nasiri, A., & Hu, J. (2021).
149 High-throughput discovery of novel cubic crystal materials using deep generative neural
150 networks. *Advanced Science*, 8(20), 2100566. <https://doi.org/10.1002/advs.202100566>
- 151 Zhao, Y., Siriwardane, E. M. D., Wu, Z., Hu, M., Fu, N., & Hu, J. (2022). *Physics Guided*
152 *Generative Adversarial Networks for Generations of Crystal Materials with Symmetry*
153 *Constraints* (No. arXiv:2203.14352). arXiv. <https://arxiv.org/abs/2203.14352>