




Foundry-ML - Software and Services to Simplify Access to Machine Learning Datasets in Materials Science


KJ Schmidt ^{1,2*}, Aristana Scourtas ^{1,2*}, Logan Ward ², Steve Wangen³, Marcus Schwarting ⁴, Isaac Darling⁴, Ethan Truelove⁴, Aadit Ambadkar¹, Ribhav Bose¹, Zoa Katok¹, Jingrui Wei⁵, Xiangguo Li⁵, Ryan Jacobs ⁵, Lane Schultz⁵, Doyeon Kim⁵, Michael Ferris ⁶, Paul M. Voyles ⁵, Dane Morgan ⁵, Ian Foster ^{1,2,4}, and Ben Blaiszik ^{1,2}

¹ Globus, University of Chicago, Chicago, IL, United States of America ² Data Science and Learning Division, Argonne National Laboratory, Lemont, IL, United States of America ³ Data Science Institute, University of Wisconsin-Madison, Madison, WI, United States of America ⁴ Department of Computer Science, University of Chicago, Chicago, IL, United States of America ⁵ Department of Materials Science and Engineering, University of Wisconsin-Madison, Madison, WI, United States of America ⁶ Department of Computer Science, University of Wisconsin-Madison, Madison, WI, United States of America * These authors contributed equally.

DOI: [10.21105/joss.05467](https://doi.org/10.21105/joss.05467)

Software

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

Editor: Fei Tao 

Reviewers:

- [@duhd1993](#)
- [@marshallmcdonnell](#)

Submitted: 21 April 2023

Published: 22 January 2024

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/)).

Summary

The application of open science and machine learning to scientific, engineering, and industry-relevant problems is a critical component of the cross-department U.S. Artificial Intelligence (AI) strategy highlighted e.g., by the AI Initiative, the recent National AI Strategy report ("[Strengthening and Democratizing the u.s. Artificial Intelligence Innovation Ecosystem - an Implementation Plan for a National Artificial Intelligence Research Resource](#)," 2023), the Year of Open Data, Materials Genome Initiative ([Pablo et al., 2019](#); [Ward & Warren, 2015](#)), and more. A key aspect of these strategies is to ensure that infrastructure exists to make datasets easily accessible for training, retraining, reproducing, and verifying model performance on chosen tasks. However, the discovery of high-quality, curated datasets adhering to the FAIR principles (findable, accessible, interoperable and reusable) remains a challenge.

To overcome these dataset access challenges, we introduce Foundry-ML, software that combines several services to provide researchers capabilities to publish and discover structured datasets for ML in science, specifically in materials science and chemistry. Foundry-ML consists of a Python client, a web app, and standardized metadata and file structures built using services including the Materials Data Facility([Blaiszik et al., 2016, 2019](#)) and Globus ([Ananthakrishnan et al., 2018](#); [Chard et al., 2015](#)). Together, these services work in conjunction with Python software tooling to dramatically simplify data access patterns, as we show below.

Statement of need

The processes by which high-quality structured science datasets are published and accessed remains decentralized, without shared standards, and scattered with some exceptions (e.g., Wu et al. (2018)). With Foundry-ML, we provide 1) a simple Python interface that allows users to access structured ML-ready materials science and chemistry datasets with just a few lines of code, 2) a prototype web-based interface for dataset search and discovery, and 3) software that enables users to publish their own ML-ready datasets in a self-service manner.

Foundry-ML focuses foremost on accessibility and reproducibility. Figure 1 shows an example of how, with just a few lines of code, researchers can access a curated collection of ML-ready datasets, the associated metadata describing the dataset contents, split details (e.g., train, test, validate), and other information (e.g., number of entries). As of Q1 2023, we have collected and made available 30 datasets in Foundry with data representations including tabular data (e.g., csv, Excel), key-value data (e.g., JSON), image sets, and hierarchical data (e.g., HDF5).

a Load Dataset Information

```
from foundry import Foundry

f = Foundry()

f = f.load("10.18126/c5z9-zej7")
```

b Learn About Dataset Contents

Data for: Ab initio control of zeolite synthesis and intergrowth with high-throughput simulations
Schwalbe-Koda, Daniel; Gómez-Bombarelli, Rafael

key	type	units	description
crystal_id	input		unique identifier associated with each point. It is the unique index of each JSON entry.
zeolite	input		IZA code of the zeolite.
SMILES	input		SMILES string of the guest adsorbed in the zeolite.
isobutylene	input		isobutylene of the guest adsorbed in the zeolite.
Volume (Angstroms)	input	Angstrom ³	Volume of the OSDA, given in Angstrom ³ .
Axis 1 (Angstrom)	input	Angstrom	first principal component of the OSDA, given in Angstrom.
Axis 2 (Angstrom)	input	Angstrom	second principal component of the OSDA, given in Angstrom.
isobutylene?	input	kJ/mol	If the pair is known in the literature, the value is equal to 1. Otherwise, it is 0.
lattice	input		lattice matrix of the crystal.
input	input		table containing the atomic number and the (x, y, z) cartesian coordinates (in Angstrom) of all atoms in the unit cell.

Split details: train, test, validate, split

Keys with info: crystal_id, zeolite, SMILES, isobutylene, Volume, Axis 1, Axis 2, isobutylene?, lattice, input

Information about the dataset: data_type, number_of_entries, last_update, domain, n_items

c Download and Use Dataset

```
res = f.load_data()
X, y = res['train']
```

Daniel Schwalbe-Koda, Rafael Gómez-Bombarelli

X.head()

crystal_id	zeolite	SMILES	isobutylene	Volume	Axis 1	Axis 2	isobutylene?	lattice	input
0	1018126	CC(C)C	CC(C)C	1018126	1018126	1018126	0	[[[0.0, 0.0, 0.0], [0.0, 0.0, 0.0], [0.0, 0.0, 0.0]]]	0.0
1	1018126	CC(C)C	CC(C)C	1018126	1018126	1018126	0	[[[0.0, 0.0, 0.0], [0.0, 0.0, 0.0], [0.0, 0.0, 0.0]]]	0.0
2	1018126	CC(C)C	CC(C)C	1018126	1018126	1018126	0	[[[0.0, 0.0, 0.0], [0.0, 0.0, 0.0], [0.0, 0.0, 0.0]]]	0.0
3	1018126	CC(C)C	CC(C)C	1018126	1018126	1018126	0	[[[0.0, 0.0, 0.0], [0.0, 0.0, 0.0], [0.0, 0.0, 0.0]]]	0.0
4	1018126	CC(C)C	CC(C)C	1018126	1018126	1018126	0	[[[0.0, 0.0, 0.0], [0.0, 0.0, 0.0], [0.0, 0.0, 0.0]]]	0.0

y.head()

	Binding (SiO2)	Binding (OSDA)	Directivity (SiO2)	Competition (SiO2)	Competition (OSDA)	Templating
0	-0.792990	-6.343916	13.610253	16.963938	135.894435	19.428617
1	-3.000776	-24.006207	11.402466	16.312994	104.841808	18.325914
2	-7.733204	-41.243755	6.670038	12.812365	109.974071	15.525417
3	-2.536347	-20.290778	11.866895	15.167858	95.902317	18.026902
4	-2.536347	-20.290778	11.866895	15.167858	120.006328	19.195592

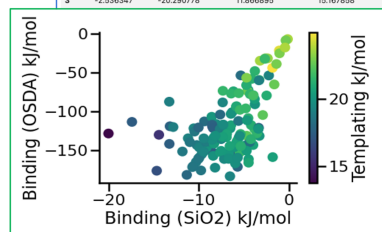


Figure 1: A Foundry-ML use case for zeolite design. (a) A user instantiates the Foundry-ML Python client and loads the descriptive metadata using the DOI. (b) Descriptive metadata includes information about the keys included in the datasets, associated units, and a short description. The metadata also include information about the dataset including the associated splits (e.g., train, test, validate), and the amount of data included. (c) A user can then load the data using the `load_data` function. This function returns a Pandas or Dask dataframe for tabular data. The zeolite dataset shown here, its metadata, and the data itself from researchers Daniel Schwalba-Koda and Rafael Gomez-Bombarelli.

Foundry-ML is built upon a solid base. We have developed Foundry-ML using the Materials Data Facility (MDF) (Blaiszik et al., 2016, 2019) and Globus services like Auth, Transfer, and Search. Foundry-ML users can upload large datasets (MDF supports multi-TB databases, with potentially millions of files), making them easy to share, use, and discover by the rest of the scientific community. All datasets are made available through the Foundry-ML software, the Foundry-ML webapp and also via Globus endpoints that support both Globus and HTTPS access.

Beyond just simplified data access, enhanced interpretability is a key feature of Foundry-ML. Foundry-ML datasets have required metadata (see Figure 1b) that are provided by the authors of each dataset. All metadata are stored in Globus Search (Chard et al., 2015) to facilitate queries. To make these metadata easily usable by Foundry-ML users, query helpers are provided via the Foundry-ML Python client to perform common actions e.g., listing all datasets, selecting datasets by DOI, and more.

In addition to the Python software interface to each dataset, we have developed a prototype web interface (Figure 2) that lists all datasets with instructions on how to access them and key features of each dataset (e.g., number of entries, inputs, targets, type of data, tags, free text description). While the examples presented here come from the domains of materials science and chemistry, Foundry-ML is designed to be domain agnostic, and since similar problems exist in other domains, we expect these approaches to generalize. Generalizing to other domains will allow the same software and services to help solve similar problems across scientific domains.

Datasets

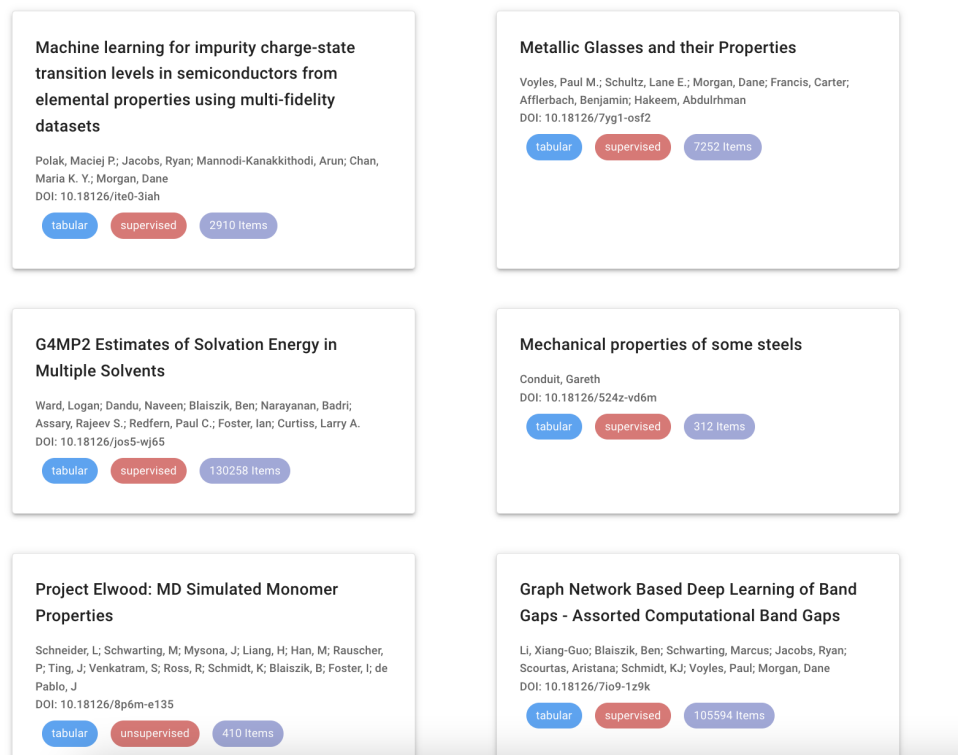


Figure 2: Foundry Website UI for browsing Datasets. This figure shows a web user interface for browsing available datasets with summary information about the datasets.

Usage

Foundry has been successfully used in educational curricula ([Stan et al., 2021](#)) and to publish datasets by research teams at the University of Chicago, Argonne National Lab, the University of Toronto ([Huang et al., 2022](#)), 3M ([Schneider et al., 2022](#)), the University of Wisconsin ([Li et al., 2021](#); [Wei et al., 2021](#)), MIT ([Schwalbe-Koda et al., 2021](#)) [Figure 2](#), and many more. In [Figure 2](#), we highlight a use case for the ML-guided design of organic structure-directing agents (OSDAs) to promote zeolite formation from the team of Gomez-Bombarelli at MIT. By using only the Foundry-ML software and the dataset DOI [Figure 1a](#), which could be cited in a paper or retrieved from the Foundry-ML web app or software, a researcher can load descriptive metadata [Figure 1b](#) to understand the dataset contents, and load the data [Figure 1c](#) for analysis, exploration, and replication. A notebook showcasing this use case is available at in the GitHub examples linked in the Documentation section below.

Future Directions

In future work, we intend to add capabilities to Foundry-ML that enable publication and connection of datasets with ML models creating a combined ecosystem of datasets and models. This work will be completed in collaboration between two National Science Foundation (NSF) projects, (#1931306) “Collaborative Research: Framework: Machine Learning Materials Innovation Infrastructure” and (#2209892) “Garden: A FAIR Framework for Publishing and Applying AI Models for Translational Research in Science, Engineering, Education, and Industry”.

We also plan to generalize the metadata and extend these capabilities to scientific datasets in domains beyond materials and chemistry.

Documentation

Detailed Foundry-ML documentation is available via GitBook at the following location [GitBook documentation](#). We have also have compiled [example notebooks](#) that show how to publish, retrieve, and use select Foundry-ML datasets.

Acknowledgements

This work was supported by the National Science Foundation under NSF Award Number: 1931306 “Collaborative Research: Framework: Machine Learning Materials Innovation Infrastructure”. **MDF** This work was performed under the following financial assistance award 70NANB19H005 from U.S. Department of Commerce, National Institute of Standards and Technology as part of the Center for Hierarchical Materials Design (CHiMaD).

References

- Ananthakrishnan, R., Blaiszik, B., Chard, K., Chard, R., McCollam, B., Pruyne, J., Rosen, S., Tuecke, S., & Foster, I. (2018). Globus platform services for data publication. In *Proceedings of the practice and experience on advanced research computing* (pp. 1–7). <https://doi.org/10.1145/3219104.3219127>
- Blaiszik, B., Chard, K., Pruyne, J., Ananthakrishnan, R., Tuecke, S., & Foster, I. (2016). The materials data facility: Data services to advance materials science research. *Jom*, 68(8), 2045–2052. <https://doi.org/10.1007/s11837-016-2001-3>
- Blaiszik, B., Ward, L., Schwarting, M., Gaff, J., Chard, R., Pike, D., Chard, K., & Foster, I. (2019). A data ecosystem to support machine learning in materials science. *MRS Communications*, 9(4), 1125–1133. <https://doi.org/10.1557/mrc.2019.118>
- Chard, K., Pruyne, J., Blaiszik, B., Ananthakrishnan, R., Tuecke, S., & Foster, I. (2015). Globus data publication as a service: Lowering barriers to reproducible science. *2015 IEEE 11th International Conference on e-Science*, 401–410. <https://doi.org/10.1109/eScience.2015.68>
- 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), 138. <https://doi.org/10.1038/s41524-020-00406-3>
- Huang, B., Lilienfeld, O. A. von, Krogel, J. T., & Benali, A. (2022). *arXiv Preprint arXiv:2210.06430*. <https://doi.org/10.1021/acs.jctc.2c01058>
- Li, X.-G., Blaiszik, B., Schwarting, M. E., Jacobs, R., Scourtas, A., Schmidt, K., Voyles, P. M., & Morgan, D. (2021). Graph network based deep learning of bandgaps. *The Journal of Chemical Physics*, 155(15), 154702. <https://doi.org/10.1063/5.0066009>
- Pablo, J. J. de, Jackson, N. E., Webb, M. A., Chen, L.-Q., Moore, J. E., Morgan, D., Jacobs, R., Pollock, T., Schlom, D. G., Toberer, E. S., & others. (2019). New frontiers for the materials genome initiative. *Npj Computational Materials*, 5(1), 41. <https://doi.org/10.1038/s41524-019-0173-4>
- Polykovskiy, D., Zhebrak, A., Sanchez-Lengeling, B., Golovanov, S., Tatanov, O., Belyaev, S., Kurbanov, R., Artamonov, A., Aladinskiy, V., Veselov, M., & others. (2020). Molecular sets (MOSES): A benchmarking platform for molecular generation models. *Frontiers in Pharmacology*, 11, 565644. <https://doi.org/10.3389/fphar.2020.565644>

- Schneider, L., Schwarting, M., Mysona, J., Liang, H., Han, M., Rauscher, P. M., Ting, J. M., Venkatram, S., Ross, R. B., Schmidt, K., & others. (2022). In silico active learning for small molecule properties. *Molecular Systems Design & Engineering*, 7(12), 1611–1621. <https://doi.org/10.1039/D2ME00137C>
- Schwalbe-Koda, D., Kwon, S., Paris, C., Bello-Jurado, E., Jensen, Z., Olivetti, E., Willhammar, T., Corma, A., Román-Leshkov, Y., Moliner, M., & others. (2021). A priori control of zeolite phase competition and intergrowth with high-throughput simulations. *Science*, 374(6565), 308–315. <https://doi.org/10.1126/science.abh3350>
- Stan, T., James, J., Pruyne, N., Schwarting, M., Yeom, J., Voorhees, P., Blaiszik, B. J., Foster, I., & Emery, J. D. (2021). *Machine learning in materials science: Image analysis using convolutional neural networks in MatCNN*. <https://nanohub.org/resources/35361>
- Strengthening and democratizing the u.s. Artificial intelligence innovation ecosystem - an implementation plan for a national artificial intelligence research resource. (2023). In *WHOSTP*. The United States Government. <https://www.ai.gov/wp-content/uploads/2023/01/NAIRR-TF-Final-Report-2023.pdf>
- Ward, C. H., & Warren, J. A. (2015). *Materials genome initiative: Materials data*. US Department of Commerce, National Institute of Standards; Technology.
- Wei, J., Blaiszik, B., Morgan, D., & Voyles, P. (2021). Benchmark tests of atom-locating CNN models with a consistent dataset. *Microscopy and Microanalysis*, 27(S1), 2518–2520. <https://doi.org/10.1017/S1431927621008989>
- Wu, Z., Ramsundar, B., Feinberg, E. N., Gomes, J., Geniesse, C., Pappu, A. S., Leswing, K., & Pande, V. (2018). MoleculeNet: A benchmark for molecular machine learning. *Chemical Science*, 9(2), 513–530. <https://doi.org/10.1039/C7SC02664A>