Noname manuscript No.
(will be inserted by the editor)

Benchmark AFLOW Data Sets for Machine Learning

**Conrad L. Clement · Steven K. Kauwe ·
Taylor D. Sparks***

Received: date / Accepted: date

Abstract Materials informatics is increasingly finding ways to exploit machine learning algorithms. Techniques such as decision trees, ensemble methods, support vector machines, and a variety of neural network architectures are used to predict likely material characteristics and property values. Supplemented with laboratory synthesis, applications of machine learning to compound discovery and characterization represent one of the most promising research directions in materials informatics. A shortcoming of this trend, in its current form, is a lack of standardized materials data sets on which to train, validate, and test model effectiveness. Applied machine learning research depends on benchmark data to make sense of its results. Fixed, predetermined data sets allow for rigorous model assessment and comparison. Machine learning publications that don't refer to benchmarks are often hard to contextualize and reproduce. In this data descriptor article, we present a collection of data sets of different material properties taken from the AFLOW database. We describe them, the procedures that generated them, and their use as potential benchmarks. We provide a compressed ZIP file containing the data sets, and a GitHub repository of associated Python code. Finally, we discuss opportunities for future work incorporating the data sets and creating similar benchmark collections.

Keywords AFLOW · benchmark data sets · machine learning · materials informatics

1 Introduction

The previous decade saw widespread interest in machine learning, affecting and sometimes transforming fields and industries. Machine learning has ex-

University of Utah
Department of Materials Science & Engineering
Corresponding author: sparks@eng.utah.edu

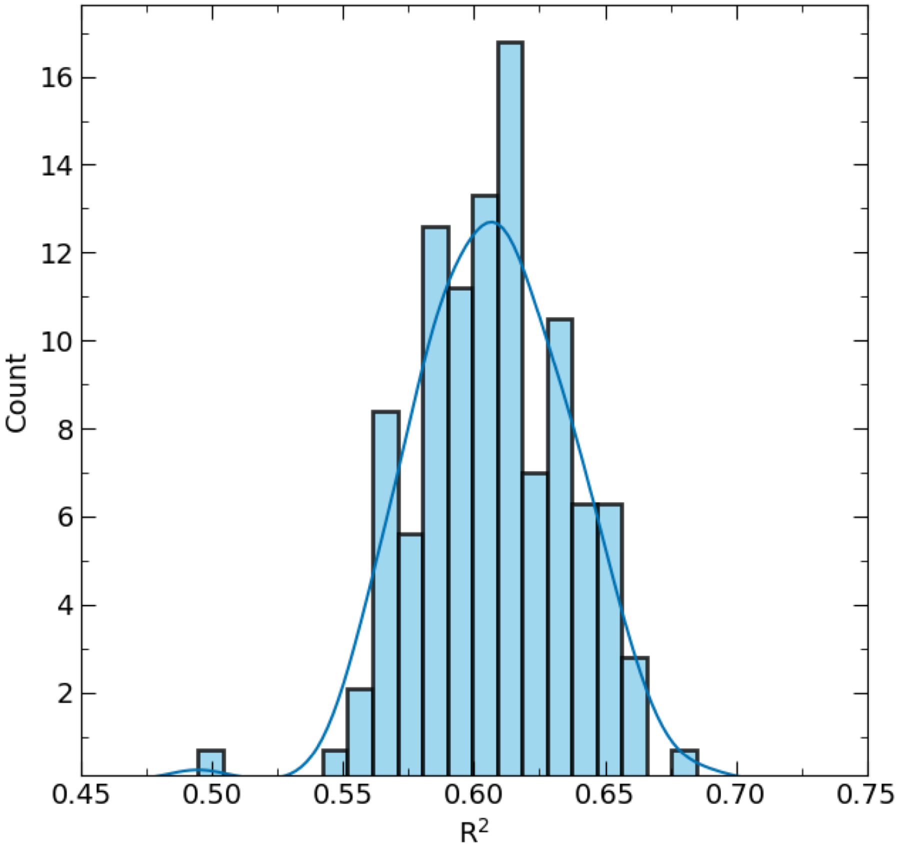

Fig. 1 A histogram of the distribution of R^2 measures for 150 ridge regressions predicting compound shear moduli. Each regression model was trained on 75% of the available composition-based data and tested on 25%, each using a unique `random_state` input to the `train_test_split` function.

isted for more than half a century as an area of research [1]. The swell of attention to it, in recent years, has been driven largely by advances in neural network algorithms, and deep neural networks in particular. However, algorithm development is only part of the story. Other key factors that have made machine learning so useful include special purpose GPU chips, the general increase in computing power associated with Moore's Law, and the unprecedented availability of training data.

Successfully applying machine learning tools to a scientific or industrial domain depends on access to large, high-quality data sets [2], together with the software infrastructure needed to process them. In the case of materials science and informatics, sources of training data include shared research databases such as AFLOW [3], the Materials Project [4], the Inorganic Crystal Structure Database (ICSD) [5], and the Open Quantum Materials Database (OQMD) [6]. A more comprehensive list of materials databases can be found in Hill et al. [7] Examples of software specifically designed for materials data analysis include the *matminer* [8] and *pymatgen* [9] Python libraries, Quantum ESPRESSO [10], and the proprietary data platform available through Citrine Informatics [11].

A good overview of the kinds of problems in materials science to which machine learning is applicable and which could benefit from having benchmarks can be found in Schmidt et al. [12]. However, the need for standard benchmarks isn't unique to materials science. The motivation and push toward establishing them exists in any field where machine learning interacts with real-world data. Benchmark development can be seen in mining biological data to address a similar set of problems as materials informatics, as in Olson et al. [13] Widely used benchmark data sets are often driven by and coevolve with advances in algorithm development. Some notable cases of this in image processing and computer vision include the MNIST data set of handwritten digits [14], widely used for training character recognition systems, as well the CIFAR-10 and CIFAR-100 data sets [15]. In the case of MNIST, character recognition represents a well-defined task amenable to different algorithms, and so is an excellent problem for benchmarking. One such candidate in materials science is the prediction of band gap values, for which there are many disparate approaches including ensemble methods [16] and mixed classification-regression models [17]. Therefore, benchmarking band gap predictions could serve as a leading example of reproducible machine learning work in materials science. Cataloging benchmark model performance won't be insightful in every case, such as with formation energy; this is because formation energy can often be calculated to "chemical accuracy" using density functional methods [18]. However, in general, we believe that benchmark data sets can contribute significantly to the use of machine learning in materials science.

It's important to note the reason that partitioned data sets themselves are provided with this article, rather than simply publishing the procedures and code needed to create them. A different instance of the training/testing data set partitions, being randomly generated with the *scikit-learn* function `model_selection.train_test_split`, will almost certainly be different than those provided. The importance of this can be seen in Figure 1, in which ridge regressions are trained using the `Oliynyk` composition-based feature vector (CBFV) [19] to predict values for the shear modulus data included with this article. The train/test split was determined randomly and independently for each instance of the regression model, resulting in a 0.18 spread of measured R^2 values. Without reference to fixed training and testing sets, a researcher using the same methods might honestly report an R^2 value anywhere along this distribution. Here it's clear that the shared data sets themselves represent the standard for comparison, and not just their relative proportions.

2 Data Description

Our data sets were created from `aflowlib.org`, using the website's provided API for structuring database queries. The list of available material properties for which one can query AFLOW is provided in the file `valid_targets.csv`. `valid_targets.csv` contains 180 distinct properties in total, though it should be noted that many of them have no physical meaning. Rather, those properties provide information about the computational environment used to calculate the physical properties.

Responses from AFLOW were then returned from the queries as JavaScript Object Notation (JSON) files, in the directory `property_files`. Each JSON file encodes numerical property information for a distinct material compound. Using a Python script, we then iterated through the JSON files according to a property of interest (e.g. shear modulus), and collected the associated values for each property into a single comma-separated value (CSV) file. These files all take the form of `property_name.csv`, and are stored in the directory simply labeled `data`.

Following this, the script `process_data.py` was called to partition the single data set for each property into three subsets, for model training, validation, and testing, in proportions of 70%, 15%, and 15% of the original property file, respectively. These smaller CSV files are grouped by property name in the `processed_data` folder, ready for use. Duplicate values were dropped prior to splitting, and properties with fewer than 1,000 data points were not split or included in `processed_data`.

Not included with this article are the Crystallographic Information Structure (CIF) files [20] which encode structural information for each compound. However, these can be retrieved with the `download_cif_files` function in `process_data.py`. Examples of work in which structural information has been exploited as a feature for training machine learning models include [21] and [22].

2.1 Supporting Data and Citations

The collected data sets, representing the essential contribution of this work, can be accessed at `https://doi.org/10.6084/m9.figshare.11954742` together in a single supplementary ZIP file. The code used to create them is available at `https://github.com/oconradh/benchmark_aflow`. When using them as machine learning benchmarks, please cite this article in addition to the AFLOW database from which the data sets were originally accessed. Please also take care to respect the non-commercial license with which AFLOW provides its data.

2.2 Data Utility

The goal of this work is to supply and contextualize standardized training, validation, and testing data sets for the 180 properties available via the AFLOW API, for use in materials informatics research. Additionally, we provide the exact procedure by which the data was retrieved and transformed, in the form of a publicly available GitHub repository. In doing this, we aim to support further work in materials informatics by providing baseline standards for machine learning model evaluation against these data sets. By using them as fixed reference points for model validation and performance, one can guarantee meaningful comparison across classes of machine learning algorithms, parameterizations of a particular model, and model featurization schemes.

Proper use of the data sets as benchmarks will involve training one's machine learning model on the 70% training set, tuning hyperparameters on the 15% validation set, and assessing overall model performance on the 15% test set. Importantly, the model must not be exposed to the test set at any point prior to final performance evaluation [23]. It's possible to encounter a situation in which model validation is of little or no importance, and the validation data better used in testing. In this case, it would reasonable to combine the provided validation and test sets into one larger, 30%, test set. This would need to be discussed, however, and could no longer be compared against the original three set scheme.

We make no claims to the empirical or theoretical accuracy of the data, having not produced it ourselves. Importantly, no data cleaning or other transformations were applied. Successful use in machine learning may require first filtering out missing or physically impossible values, or transforming a data set's distribution. Examples here might include the removal of compounds with an electronic energy band gap (`Egap`) of zero, or extreme atomic energy (`energy_atom`) outliers. It's beyond our scope to discuss each kind of property being provided; more information for interpreting them and their derivations can be found at the AFLOW website, `www.aflow.org`.

2.3 Future Work

A natural continuation of this project would be to extend it to more properties and additional sources of data. Promising and useful sources here include (but are certainly not exhausted by) those mentioned above: the Materials Project, the ICSD, and potentially new databases derived from Quantum ESPRESSO calculations.

Additionally, in the spirit of establishing informatics benchmarks, one could publish and maintain a comprehensive survey of learning algorithm performance for given data sets and featurizations. In this sense the metric being standardized would be the optimal, fully-validated implementation of a machine learning algorithm with respect to a particular materials data set. Scores could then be determined by conventional measures of performance, such as coefficient of determination (R^2), root-mean-square-error (RMSE), etc. and represent the best observed performance for an entire class of algorithms.

Finally, another project would be to establish qualitative standards for materials informatics data sets based on existing domain knowledge. These would be shared heuristics and constraints that particular materials data sets (and especially benchmark sets) ought to satisfy in order to be both physically meaningful and useful in a machine learning context.

While these additional directions could help to establish more benchmarks, the standard data sets themselves are only as good as the work that uses them. Any research project that systematically incorporates our benchmarks as a means of evaluating machine learning methods would build on the work presented here, and further illustrate the importance of benchmarking.

Acknowledgements The authors gratefully acknowledge support from the NSF CAREER Award DMR 1651668. The authors thank the creators of AFLOW for the creation of the database and for making its contents available for this article. In addition, the authors express their gratitude to the open-source software community, for developing the excellent tools used in this research, including Python and the *pandas*, *numpy*, *matplotlib*, and *sklearn* Python libraries, among others.

The authors declare that they have no conflict of interest.

References

1. David Donoho. 50 years of data science. *Journal of Computational and Graphical Statistics*, 26(4):745–766, 2017.
2. Ram Seshadri and Taylor D Sparks. Perspective: interactive material property databases through aggregation of literature data. *APL Materials*, 4(5):053206, 2016.
3. Stefano Curtarolo, Wahyu Setyawan, Gus LW Hart, Michal Jahnatek, Roman V Chepulskii, Richard H Taylor, Shidong Wang, Junkai Xue, Kesong Yang, Ohad Levy, et al. AFLOW: an automatic framework for high-throughput materials discovery. *Computational Materials Science*, 58:218–226, 2012.
4. Anubhav Jain, Shyue Ping Ong, Geoffroy Hautier, Wei Chen, William Davidson Richards, Stephen Dacek, Shreyas Cholia, Dan Gunter, David Skinner, Gerbrand Ceder, et al. Commentary: The materials project: A materials genome approach to accelerating materials innovation. *Apl Materials*, 1(1):011002, 2013.
5. Mariette Hellenbrandt. The inorganic crystal structure database (ICSD)—present and future. *Crystallography Reviews*, 10(1):17–22, 2004.
6. James E Saal, Scott Kirklin, Muratahan Aykol, Bryce Meredig, and Christopher Wolverton. Materials design and discovery with high-throughput density functional theory: the open quantum materials database (OQMD). *Jom*, 65(11):1501–1509, 2013.
7. Joanne Hill, Gregory Mulholland, Kristin Persson, Ram Seshadri, Chris Wolverton, and Bryce Meredig. Materials science with large-scale data and informatics: unlocking new opportunities. *Mrs Bulletin*, 41(5):399–409, 2016.
8. Logan Ward, Alexander Dunn, Alireza Faghaninia, Nils ER Zimmermann, Saurabh Bajaj, Qi Wang, Joseph Montoya, Jiming Chen, Kyle Bystrom, Maxwell Dylla, et al. Matminer: An open source toolkit for materials data mining. *Computational Materials Science*, 152:60–69, 2018.
9. Shyue Ping Ong, William Davidson Richards, Anubhav Jain, Geoffroy Hautier, Michael Kocher, Shreyas Cholia, Dan Gunter, Vincent L Chevrier, Kristin A Persson, and Gerbrand Ceder. Python materials genomics (pymatgen): A robust, open-source python library for materials analysis. *Computational Materials Science*, 68:314–319, 2013.

10. Paolo Giannozzi, Stefano Baroni, Nicola Bonini, Matteo Calandra, Roberto Car, Carlo Cavazzoni, Davide Ceresoli, Guido L Chiarotti, Matteo Cococcioni, Ismaila Dabo, et al. Quantum ESPRESSO: a modular and open-source software project for quantum simulations of materials. *Journal of physics: Condensed matter*, 21(39):395502, 2009.
11. Citrination. *www.citrination.com*.
12. Jonathan Schmidt, Mário RG Marques, Silvana Botti, and Miguel AL Marques. Recent advances and applications of machine learning in solid-state materials science. *npj Computational Materials*, 5(1):1–36, 2019.
13. Randal S Olson, William La Cava, Patryk Orzechowski, Ryan J Urbanowicz, and Jason H Moore. PMLB: a large benchmark suite for machine learning evaluation and comparison. *BioData mining*, 10(1):36, 2017.
14. Li Deng. The MNIST database of handwritten digit images for machine learning research [best of the web]. *IEEE Signal Processing Magazine*, 29(6):141–142, 2012.
15. Alex Krizhevsky, Vinod Nair, and Geoffrey Hinton. CIFAR-10 and CIFAR-100 datasets. *www.cs.toronto.edu/kriz/cifar.html*.
16. Steven K Kauwe, T Welker, and Taylor D Sparks. Extracting knowledge from dft: experimental band gap predictions through ensemble learning. *MRS Commun*, 2018.
17. Ya Zhuo, Aria Mansouri Tehrani, and Jakoah Brgoch. Predicting the band gaps of inorganic solids by machine learning. *The journal of physical chemistry letters*, 9(7):1668–1673, 2018.
18. Yubo Zhang, Daniil A Kitchaev, Julia Yang, Tina Chen, Stephen T Dacek, Rafael A Sarmiento-Pérez, Maguel AL Marques, Haowei Peng, Gerbrand Ceder, John P Perdew, et al. Efficient first-principles prediction of solid stability: Towards chemical accuracy. *Npj Computational Materials*, 4(1):1–6, 2018.
19. Ryan Murdock, Steven Kauwe, Anthony Wang, and Taylor Sparks. Is Domain Knowledge Necessary for Machine Learning Materials Properties? 2 2020.
20. Sydney R Hall, Frank H Allen, and I David Brown. The crystallographic information file (CIF): a new standard archive file for crystallography. *Acta Crystallographica Section A: Foundations of Crystallography*, 47(6):655–685, 1991.
21. Tian Xie and Jeffrey C Grossman. Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties. *Physical review letters*, 120(14):145301, 2018.
22. Kristof T Schütt, Huziel E Sauceda, P-J Kindermans, Alexandre Tkatchenko, and K-R Müller. SchNet–a deep learning architecture for molecules and materials. *The Journal of Chemical Physics*, 148(24):241722, 2018.
23. Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media, 2009.