Investigating Data Reusability in Density Functional Theory Studies

Rob Fleur[1], Addy Ireland[1], Xintong Zhao[1], Scott McClellan[1], Eric Paltoo[1], Tianyu Su[2], Channyung Lee[2], Yuan An[1], Xiaohua Hu[1], Elif Ertekin[2], and Jane Greenberg[1]
[1]*Metadata Research Center*, *Drexel University*, Philadelphia, USA
[2]*Mechanical Science and Engineering*, *University of Illinois at Urbana-Champaign*, Urbana, USA

I. Overview

Over the last decade, there has been a significant increase in supporting reproducible computational research (RCR) [1]. The global adoption of the FAIR principles [2] stands as a key indicator of this trend. Specifically, federal and global research funding agencies have increasingly mandated scientific data and related products, such as code and algorithms, be made Findable, Accessible, Interoperable, and Reusable (FAIR) [2].

The Materials Genome Initiative (MGI) has motivated the need for sharing data to support data-driven materials design and served as a bridge for advancing FAIR in materials science across the field of materials science. Density Functional Theory (DFT) presents an example of material science researchers seeking to advance FAIR principles and supporting RCR, and is the focus of this research.

Scholarly big data serves as the source for studying this topic and is the focus of the research presented in this paper. The work that follows provides background information on DFT and FAIR, presents results of our analysis leveraging scholarly big data to assess data sharing, and discusses next steps.

II. Background - Density Functional Theory and the FAIR Principles

DFT is a method used in several areas of materials science. DFT is a powerful quantum mechanical computational method to investigate the electronic structure of systems of atomic nuclei and electrons [5], [6]. It has become a standard technique in many branches of chemistry, materials science, and physics. Formally, DFT is based on the Hohenberg-Kohn theorems [3], which help reformulate the many-body Schrodinger equation into a system of equations to be solved for the ground state electron density and total energy. Although many formulations of DFT exist, the Kohn-Sham framework is most commonly used [4]. Using DFT, it is possible to estimate physical properties such as lattice constants, binding energies, stability, and transition states. Approximate descriptions of optical, electronic, vibrational, and other features are also possible. Over the last 30 years, massive numbers of DFT calculations have enabled property prediction at scale, some properties for which experimental measurements are challenging. Today DFT underlies almost all databases of properties of materials computed from first-principles. It is often used to guide the search and discovery of new materials with target properties that have never been explored before. While quite a few experimental studies now routinely include DFT calculations to explain findings in published works, there is also a large number of dedicated DFT studies as well.

DFT plays a pivotal role in computational materials science, enabling quantum mechanical descriptions of the electronic properties of solids, molecules, and other diverse classes of materials. Its success and widespread application arise largely because DFT simulations provide reasonable accuracy at small computational cost. Nowadays, a variety of software/codes are available, making DFT simulations more accessible to a wide range of scientists and engineers for predicting properties and facilitating materials design and optimization. Although detailed aspects of implementation vary amongst different codes, most share a similar underlying structure, facilitating comparison of user-selected simulation parameters and corresponding outputs. User choices such as the exchange-correlation functional, basis set, convergence criteria, and other numerical parameters can all introduce variability in DFT results.

As the workhorse method of the first-principles modeling community, DFT is increasingly automated, e.g., to build large databases or as a component of a multi-scale modeling framework. Consequently, reproducibility of DFT results underlies scientific credibility for much of the first-principles modeling community. Therefore assessment of DFT user community practices with respect to FAIR principles is imperative for DFT practitioners in adopting the FAIR principles. This is achieved through a variety of mechanisms, including reporting of simulation parameters in scholarly publications, providing simulation parameters and/or input/output files as supplemental data, and by publishing workflows, simulation inputs, and simulation outputs in online materials data repositories [7], [8].

III. Method

We conducted an empirical study involving three main steps: (1) corpus collection and processing, (2) target articles filtering, and (3) supplemental data analysis.

Corpus Collection and Processing: We collected 172,000 research articles under materials science category along with their supplemental materials from the American Chemical

This work is funded by NSF OAC # 2118201

Society (ACS). After processing the collected data into usable forms, all research articles are in XML format; supplemental materials are stored in various formats such as CIF, PDF, dataset, and so forth.

Target Article Filtering: We used a basic dictionary related to DFT research, which was generated by domain scientists, to identify relevant articles. The dictionary contained the following keywords: density functional theory (DFT), exchange correlation, Quantum Espresso, VASP, plane wave basis set, effective core potential (ECP), and pseudopotential. Articles that contained at least one keyword were retained as our target articles set. We obtained 10,034 articles after keyword matching.

Supplemental Data Analysis: We assessed the supplemental data based on the FAIR principles, which includes article metadata, type of materials, and in-text data citations. Supplemental materials were also assorted according to their presence in either the ACS, Figshare repository, and other external repositories.

IV. Initial Findings

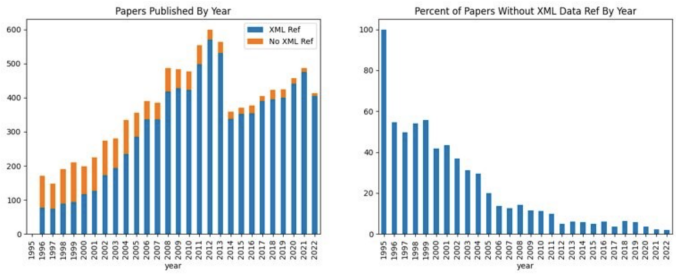

Fig. 1. Proportion of Articles Publishing their Data Vs. No Supplemental Data

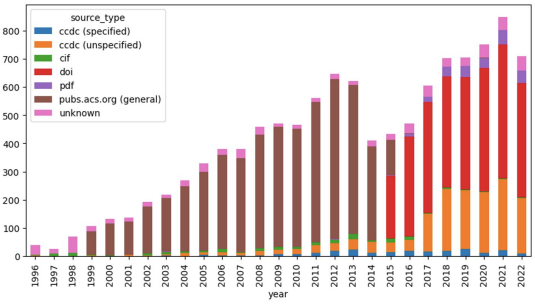

Fig. 2. Distribution of Supplemental Data Type by Year

The proportion of articles publishing their data is shown in figure 1. The proportion of articles that do not publish their data steadily decreases over time and there was 5% of articles without supplemental data in 2022. A detailed data type distribution is demonstrated in figure 2.

However, there are challenges to re-use published data: Within our filtered DFT corpus, 81% of articles contain at least one supplemental document. PDFs may be great for easily putting together a single document containing descriptions, charts, and tables, but when it comes time to interoperate on that data is where the perks of a PDF start degrading.

V. Conclusion and Future Work

The research presented here confirms that DFT researchers are taking important steps to publish data underlying their research. The trends shown in the figures confirm researchers publishing their data. Furthermore, figure 2 shows the broadening diversity of supplemental data sources as well as shifts in how data is referenced.

Our future work involves pursue two key goals. First, we will systematically classify papers focusing on DFT based on an expanded set of subject matter specific terms and to evaluate the extent of DFT simulation reporting. Focusing on understanding an expanded set of subject matter specific terms which should increase accuracy of the model. The initial dictionary of terms and phrases provided by the the collaborator allowed us to discern broad contours regarding DFT-related articles. Our second key goal is to analyze the availability of DFT-related files or additional supplementary files from the publications. This is important because the papers which discuss DFT fall into several categories. The first type of papers make passing mention of DFT calculations which often occur in the discussion section and are considered to have low relevance because they do not offer in depth engagement of the process. The second type of paper, which is often of greater relevance, employs DFT terminology in the methods section of a paper. A third type of paper broadly develops or refines DFT as a method without application to a specific experiment. Finally, future work may lead to the construction of a repository for DFT data including input parameters and outputs, facilitating data citation and ensuring the reproducibility of DFT simulations.

References

[1] J. Leipzig, D. Nüst, C. T. Hoyt, S. Soiland-Reyes, K. Ram, and J. Greenberg, "The role of metadata in reproducible computational research." arXiv, 2020. doi: 10.48550/ARXIV.2006.08589.

[2] M. D. Wilkinson et al., "The FAIR Guiding Principles for scientific data management and stewardship," 2016, doi: 10.1038/sdata.2016.18.

[3] P. Hohenberg and W. Kohn, "Inhomogeneous Electron Gas," Physical Review, vol. 136, no. 3B. American Physical Society (APS), pp. B864–B871, Nov. 09, 1964. doi: 10.1103/physrev.136.b864.

[4] W. Kohn and L. J. Sham, "Self-Consistent Equations Including Exchange and Correlation Effects," Physical Review, vol. 140, no. 4A. American Physical Society (APS), pp. A1133–A1138, Nov. 15, 1965. doi: 10.1103/physrev.140.a1133.

[5] R. O. Jones, "Density functional theory: Its origins, rise to prominence, and future," Reviews of Modern Physics, vol. 87, no. 3. American Physical Society (APS), pp. 897–923, Aug. 25, 2015. doi: 10.1103/revmodphys.87.897.

[6] K. Burke, "Perspective on density functional theory," The Journal of Chemical Physics, vol. 136, no. 15. AIP Publishing, Apr. 17, 2012. doi: 10.1063/1.4704546.

[7] C. Draxl and M. Scheffler, "The NOMAD laboratory: from data sharing to artificial intelligence," JPhys materials, vol. 2, no. 3, pp. 36001-, 2019, doi: 10.1088/2515-7639/ab13bb.

[8] B. Blaiszik, K. Chard, J. Pruyne, R. Ananthakrishnan, S. Tuecke, and I. Foster, "The Materials Data Facility: Data Services to Advance Materials Science Research," JOM (1989), vol. 68, no. 8, pp. 2045–2052, 2016, doi: 10.1007/s11837-016-2001-3.