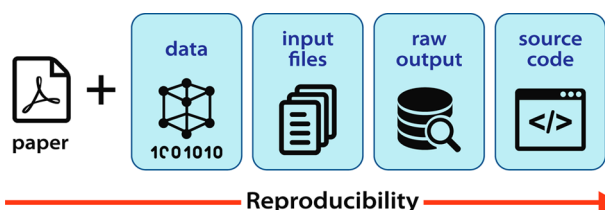


Reproducible Research in Computational Chemistry of Materials



The reproducibility of experimental findings is a crucial underlying tenet of the scientific method, and the awareness of its central place in the advancement of knowledge has been long recognized, tracing back to Ibn al-Haytham (أبو علي، الحسن بن الحسن بن الهيثم) in the 11th century and later to the scientists of the Renaissance. While the way we perform and report research has evolved immensely, reproducibility remains as vital today as it was then—and maybe more, in a world where 2.5 million scientific articles are published worldwide every year. A quantitative analysis of the causes of retraction of research articles, published in 2012,¹ concludes that 22% of retractions cite a “known artifact” in the data or “unexplained irreproducibility” as the cause of the retraction, without any research misconduct or ethical violation. This statistic highlights the need to improve the current standards for the reproducibility of published works, by creating and promoting policies for reproducible science. This parallels the recent advances by journals, professional associations, academic institutions, and funding agencies in the domain of publication ethics (see, for example, the guidelines developed by the Committee on Publication Ethics² and nowadays adopted by many scientific journals).

In this landscape the swift rise of computational science, which has led to exciting developments and breakthroughs in all fields of academic research, requires us to reinterpret the traditional understanding of reproducibility to apply it to computational results from *in silico* experiments.³ Computer calculations being a finite series of operations on a fully deterministic machine, they are in theory fully replicable and can be rerun to yield bit-for-bit identical results. However, in practice this does not translate at all into full reproducibility of computations, despite the modest cost of online hosting data, owing to the complexity of the modern hardware and software stacks. I have, like many colleagues, been frustrated a few times in the past by difficulties in reproducing results from the published literature—or by obtaining structures or data related to published findings, for comparison with my own work. This is the case even when the authors would be fully willing to share their data at the time of publication: over time the data might have been lost, or the person working on it gone, and there is no real incentive to go dig up files from 5 or 10 years back—especially in cases where the group does not work in this area anymore.

Browsing through a recent issue of a flagship materials chemistry journal gives a good idea of how important computational methods have become in our field and what

progress we still have to make toward reporting these studies in a way that promotes reproducibility. Out of 41 full research articles in an issue, I counted 7 that were exclusively or predominantly theoretical in nature (17%), with 13 more that included some computational results in figures or tables (thus a total of 49%). For most of the papers, there was little to provide any help to a researcher willing to reproduce the calculations: the molecular and crystal structures discussed were not provided except as snapshots in figures; input files for calculations were not provided. In a few cases, where computational work was one of several characterization methods, the level of description of the work performed (whether in the main text or as supporting information) actually appeared too low to allow reproducing the calculations. Three papers provided full structural models, three others reported extensive data sets of computed properties, and one gave the source code used for analysis. However, in both cases, these were available only in PDF format over multiple pages.

Here, I summarize the different aspects of research reproducibility in computational chemistry and materials science and delineate the current practices of the community (both researchers and publishers) in this respect.

■ OPENING THE DATA

The first necessary step toward research reproducibility and the collective advancement of knowledge is that scientists make available the results of their experiments and calculations. Numerical quantities measured are presented in the text, in tables, or in graphs. Moreover, all structures—molecular or crystallographic—discussed should be made available, not only in graphical form (for example, as 3D representation) but as structure files. Most journals nowadays include this requirement in their instructions to authors, although it is unfortunately not always enforced. It is also good practice for the structures to be given as machine-readable files in standard formats (XYZ, PDB, or CIF files) rather than to list coordinates in a PDF document as Supporting Information. This lowers the barrier for others to visualize the data or reuse it in future works, as well as improves the discoverability and indexing by future data-mining projects.

Some computational studies can produce a large amount of data (chemical structures as well as computed properties), for example, in works of high-throughput screening or large scale calculations on existing structure databases. This requires thorough curation of the data to produce a usable database, including all properties produced in a standard, machine readable format (such as XML or JSON) accompanied by the documentation of the format itself. Care should also be taken to make sure that the database published contains all relevant metadata, for example, the calculation methods and method parameters employed (if they vary between structures) or the conditions of the calculation (elapsed time, platform on which it was run, etc.). When possible, this data can then be cross-

Published: April 11, 2017

linked with the earlier database(s) from which it was built. An excellent example of this practice is the recent inclusion of computed elastic properties for inorganic crystalline compounds⁴ from the Materials Project;⁵ the computed data set was first published online (DOI: 10.5061/dryad.h505v) and linked to in the paper and then integrated into the Materials Project web-based explorer. It original contained computed data for 1,181 materials, while that number has since then gone up to 4,375.

Finally, let us note that the choices made for hosting the data online impact its long-term availability. While hosting data on one's group Web site may seem practical, it does not guarantee availability in the long term. The use of larger-scale institutional repositories should be preferred—or not-for-profit repositories with reasonable guarantees of long-term storage and independent archiving. Getting a stable URL, that will not change over time, is also important: some online services, such as Zenodo,⁶ offer Digital Object Identifiers (DOI) for data stored in public GitHub repositories.⁷

Recent years have seen a large improvement in data availability, coming not only from individual author practices but also from changes in several journals' policies, which now request statements declaring the accessibility of the data and its location.⁸ Statements that "data is available upon request" may rapidly become a thing of the past—only to be used in special circumstances where the data cannot be freely released.

■ OPENING THE INPUT AND OUTPUT FILES

In addition to the publication of curated data that is presented and discussed in the paper, full reproducibility of computational studies requires to make available to the research community at large the input files that were used to start the computations. A complete set of input files—including all parameters and initial configuration of the system—contains all information about a given computation and includes many "technical" parameters that may not always be included in an article's methodology section. This is useful to both the reader and the reviewer, who is tasked with judging whether the methods applied are sound. Moreover, it provides information directly in machine-readable form, lowering the bar for replication studies and avoiding potential human errors—avoiding, for example, rekeying an entire force field input file from article tables.

In addition to input files, in a context where studies routinely use large amounts of high-performance computing (HPC) resources and storage is relatively cheap, the next logical step in improving reproducibility is to store and publish the output files (or raw data) from the computations—in all or part. This allows readers to study the details of the computations ran, even if they do not have access or do not wish to spend large CPU time rerunning the calculation. It also enables other researchers to use data from the raw output that may not have been exploited by the original authors but could be of interest for other purposes—atomic charges for analysis or force field derivation, vibrational eigenvectors, etc. This usefulness is, however, balanced by the fact that raw calculation outputs can take up large amounts of disk space: neither publishers nor authors would be willing pay the cost of long-term hosting for unreasonably large files, and the output files thus need to be curated (trimmed or selected) before being published. We also note in passing that some restrictive software licenses do not allow you to publish output files—requiring you, for example, to redact execution times or computational performance data.

A novel project has been born, whose goal is to build a distributed database of computational chemistry results. ioChem-BD⁹ has a two-pronged focus of converting data from several codes into a common data standard, as well as allowing data management, search, and manipulation—all the way to publishable supporting information files. It allows storage of the data, a graphic user interface to manipulate it, and APIs to connect it to other databases.


■ OPENING THE SOFTWARE

Finally, we turn our focus now to the core of the computational chemistry research: the software. Indeed, even with the publication of computation input files, full reproducibility requires the use of the same software. Even the same method, implemented in two different codes, can lead to different results. This means that the use of commercial software or unpublished in-house codes can restrict the reproducibility of a computational study by the research community. I consider this an encouragement to promote and use software that is either open source or at least freely available to the academic community. Furthermore, in the case where one uses commercial (or not publicly available) software, there is an even stronger incentive to publish online the raw outputs of the computations, thus mitigating the issue.

The remarks above are true not only of large computational software packages, such as quantum chemistry packages or molecular simulation suites, but also of the smaller pieces of software created during the course of the research: analysis codes, postprocessing tools, visualization scripts, etc. Those are sometimes overlooked, yet often crucial results depend on their correctness—and the details of analysis methods are not always fully documented in the articles themselves.

Finally, let us note that even for software packages whose source is freely available, differences in the exact results obtained can vary depending on software version but also on the nature of the hardware used, on the other software used on the computer (operating system, compiler, mathematical library, etc.), and on runtime parameters such as number of CPU cores or HPC nodes used. Thus, reproducibility of research and publication of data is particularly important in computational sciences,¹⁰ even when exact replicability of the simulations is not technically possible—because the calculations used hardware that has become obsolete, for example.

In conclusion, the ease and low cost with which we can nowadays host reasonable amounts of data online is an encouragement to publish more data accompanying scientific articles, going beyond the traditional "supporting information" files to include input files and output files of computations as well as source code. This will improve reproducibility of existing work and strengthen confidence of the wider public in the role of computational chemistry tools in materials science.

François-Xavier Coudert*

Chimie ParisTech, PSL Research University, CNRS, Institut de Recherche de Chimie Paris, 75005 Paris, France

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: fx.coudert@chimie-paristech.fr. Web: <http://coudert.name>. Twitter: @fxcoudert.

ORCID

François-Xavier Coudert: 0000-0001-5318-3910

Notes

Views expressed in this editorial are those of the author and not necessarily the views of the ACS.

■ REFERENCES

- (1) Grieneisen, M. L.; Zhang, M. A Comprehensive Survey of Retracted Articles from the Scholarly Literature. *PLoS One* **2012**, *7*, e44118.
- (2) COPE guidelines, available online at <http://publicationethics.org/resources/guidelines> (accessed on March 15, 2017).
- (3) Stodden, V.; McNutt, M.; Bailey, D. H.; Deelman, E.; Gil, Y.; Hanson, B.; Heroux, M. A.; Ioannidis, J. P. A.; Taufer, M. Enhancing reproducibility for computational methods. *Science* **2016**, *354*, 1240–1241.
- (4) de Jong, M.; Chen, W.; Angsten, T.; Jain, A.; Notestine, R.; Gamst, A.; Sluiter, M.; Krishna Ande, C.; van der Zwaag, S.; Plata, J. J.; Toher, C.; Curtarolo, S.; Ceder, G.; Persson, K. A.; Asta, M. Charting the complete elastic properties of inorganic crystalline compounds. *Sci. Data* **2015**, *2*, 150009.
- (5) 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. The Materials Project: A materials genome approach to accelerating materials innovation. *APL Mater.* **2013**, *1*, 011002.
- (6) <https://zenodo.org> (accessed on March 15, 2017).
- (7) <https://github.com> (accessed on March 15, 2017).
- (8) Dealing with data. *Nat. Mater.* **2017**, *16*, 1.
- (9) <http://www.iochem-bd.org> (accessed on March 15, 2017).
- (10) Lejaeghere, K.; Bihlmayer, G.; Björkman, T.; Blaha, P.; Blügel, S.; Blum, V.; Caliste, D.; Castelli, I. E.; Clark, S. J.; Dal Corso, A.; et al. Reproducibility in density functional theory calculations of solids. *Science* **2016**, *351*, aad3000.