THE JOURNAL OF PHYSICAL CHEMISTRY Letters

Viewpoint

pubs.acs.org/JPCL

Open Source and Open Data Should Be Standard Practices

More that 30 years before the term "Open Source" became popular, the computational chemistry community had invented a way for chemists to exchange the source code for their programs with other research groups. The Quantum Chemistry Program Exchange (QCPE) was one important factor (among many) that led to the explosion of computational chemistry.[1] The advantages of exchanging source code were clear to the participants in the QCPE. The exchange of code was efficient, avoiding the need for students to reinvent the wheel for each new project. The QCPE acted as a permanent repository for orphaned code and provided support for new users of these programs. One of the biggest advantages of the QCPE was that it acted as a form of publication and recognition for a type of intellectual work that is not captured well by papers and citations in journals. Throughout the 1980s and 1990s, it became common for papers in ACS journals to list specific QCPE program numbers in references.

The free availability of scientific *data* has also had substantial benefits. For example, the protein data bank (PDB), the ubiquitous repository for protein structures, has enabled reuse of the primary structural data but has also opened up new avenues of research using statistical and meta-analysis of the structures. This kind of research was perhaps not expected by the original depositors of the data, but it has provided enormously valuable insights. Again, the PDB accession code acts as a way of assigning recognition and credit for open data even if the work that led to the structural data has not been published in a journal article.

The practical advantages of sharing code and data are important, but there are now strong *scientific* reasons for making open source and open data the accepted norm. The chief reason is the growing sense that science has reached a reproducibility crisis,[2,3] where a number of scientists are admitting that their own organizations have had difficulty reproducing the results of prior publications. The reproducibility crisis has been blamed on many factors: an overemphasis on novelty as a requirement for publication, poor statistical analyses, the loss of lab expertise through graduation of students and postdocs, changing versions of code and data files, and the inadequacy of methodology sections at describing all of the steps necessary to carry out the work.

Reproducibility. One of the foundations of science is that independent scientists should be able to subject theories and models to similar tests in different locations, on different equipment, at different times and get similar answers. The reason that scientific papers provide comprehensive details in methodology sections is to allow skeptical researchers to verify experimental results for themselves.

Because so much of modern science now relies on numerical experiments and computer simulations, we must also pay careful attention to reproducibility in modeling and simulation. Numerical experiments on simple models and small data sets can be reproducible both *in principle* and *in practice* with little effort on the part of a skeptical researcher. However, as numerical experiments become more complex and the data sets become larger, calculations that are reproducible in principle are no longer reproducible in practice without access to the code, data, and the meta-data that describes how the data is organized.

Much of modern computational chemistry has now passed this complexity threshold. Asking a skeptical researcher to reproduce large electronic structure calculations or complex molecular mechanics simulations places an impossible burden on these researchers unless the work provides public access to the code and data that generated the results. **Access to the code and data should therefore be an *expectation for publication and review* in the chemical literature.**

One of the recommendations of the reproducible research standard[4] is that code components of research should be released under an open source license,[5] whereas data (which is covered by a different set of copyright laws) should be released under a CC0 license.[6] These recommendations are sensible and should be adopted by our community.

Advantages of Open Source. The practical advantages of open source are still just as important as they were for the QCPE. Open Source directly lowers costs to research grants and funding agencies, and there are significant additional savings due to the reuse of software components. Being able to access the source code that generated the results in a paper allows scientists who are learning the topic to "see under the hood" and to recreate these calculations without having to reinvent every piece of a complex code. This is an enormous efficiency savings.

Having source code publicly available also brings unexpected and welcome collaborative opportunities. Many groups that have released open source scientific codes report similar stories of researchers across the globe submitting bug reports, bug fixes, code enhancements, or documentation to a project simply because it was available for them to use and modify. Unexpected collaborations can sometimes yield real scientific output.

Because Open Source has become common in the tech sector, the tools for making code available come with useful features that scientists have yet to adopt widely. For example, a code-sharing service like GitHub[7] provides automatic revision control and assigns a revision number for each modification to the code. This can enhance reproducibility by providing exact versioning information in a publication. GitHub[7] and figshare[8] can also assign digital object identifiers (DOIs), which may help assign recognition and credit for source code contributions.[9]

Making scientific code publicly available provides early career researchers who are seeking employment with an online portfolio of their work. Code that is written with the knowledge that it will be publicly released often ends up cleaner (and less buggy) than the quick script that will never see the light of day. Researchers that end up leaving a field will leave behind a repository of knowledge for those that follow in their footsteps. Publicly accessible open source therefore provides a continuous repository of knowledge that closed code does not.

Received: February 9, 2015
Accepted: March 12, 2015
Published: April 2, 2015

ACS Publications © 2015 American Chemical Society

DOI: 10.1021/acs.jpclett.5b00285
J. Phys. Chem. Lett. 2015, 6, 1168−1169

Downloaded via 187.56.87.210 on November 13, 2021 at 14:23:44 (UTC).
See https://pubs.acs.org/sharingguidelines for options on how to legitimately share published articles.

Challenges. The biggest challenges to widespread adoption of Open Source and Open Data in science are social, and not technical, issues. Publications and citations are still the primary currency for *recognizing effort and attributing credit* in science. These bibliometric measures made sense in the 17th and 18th centuries when scientific publishing was starting up, but they are relatively coarse-grained measures that overlook the intellectual work of creating software and curating data. Software development and maintenance and data collection and curation have not been afforded the same level of visibility in scholarly publishing as experimental scientific techniques. Forward-thinking scholarly publishers should be helping to fill this gap in the reputational system that motivates scientists. One positive step would be for journals to request and publish information on *all of the software* (including specific revisions) that was used to produce figures and process the data for published papers. This would help the creators of this software to track and understand how their software is being used and to understand what kind of impact is has made on the scientific community. Third party services like Impactstory[10] are making headway on this problem, but the journals have a role to play in making attribution and recognition more representative of actual impact.

The other challenge to open source in science is a cluster of issues related to the *sustainability* of software development efforts. Scientific software is often developed by domain scientists with little background in common software engineering practices. Software development and maintenance tasks are inherited from previous group members by researchers with widely varying skill levels and with little training on good coding practices. Because they are not trained in software development, scientists often create code that is impossible to maintain. This is an area where open source can help. The main repositories for open source software (GitHub[7] and SourceForge[11]) enforce sensible coding practices like revision control, and connections to the open source world will help scientists learn good software engineering techniques.

There are real costs associated with maintaining scientific software. The open source community has struggled with the idea of how to generate revenue to support software development and maintenance. There are a number of business models that could keep the scientific review of software open while allowing for revenue to maintain the software. These models include: selling support, splitting the software into an Open Source computational engine and a paid interface, selling consulting, training, or computing services, and contracting software development services to add additional features. The key is finding a model that maintains the ability of skeptical researchers to critically review the important parts of the code.

Outlook. Although the rise of the Internet meant the end for the QCPE as a code distribution service, it has made it even easier for scientists to make their code and data publicly available. Today, the reproducibility crisis has made it urgent that we insist on open source and open data as standard practices in science. However, we have still not recovered the most important aspect of the QCPE effort: the alternative form of publication and recognition that can be used as a form of scientific currency.

J. Daniel Gezelter*

Department of Chemistry and Biochemistry, University of Notre Dame, Notre Dame, Indiana 46556, United States

AUTHOR INFORMATION

Corresponding Author

*E-mail: gezelter@nd.edu.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

The author thanks Steven Corcelli, T. Daniel Crawford, Jan H. Jensen, and John Parkhill for helpful comments on an earlier version of this Viewpoint.

REFERENCES

(1) Boyd, D. B. Quantum Chemistry Program Exchange, Facilitator of Theoretical and Computational Chemistry in Pre-Internet History. In *Pioneers of Quantum Chemistry*; Strom, E. T., Wilson, A. K., Eds.; American Chemical Society: Washington, DC, 2013; Vol. *1122*, Chapter 8, pp 221−273.

(2) Ioannidis, J. P. A. Why Most Published Research Findings Are False. *PLoS Med.* **2005**, *2*, e124.

(3) Prinz, F.; Schlange, T.; Asadullah, K. Believe it or not: how much can we rely on published data on potential drug targets? *Nat. Rev. Drug Discovery* **2011**, *10*, 712−712.

(4) Stodden, V. The Legal Framework for Reproducible Scientific Research: Licensing and Copyright. *Comput. Sci. Eng.* **2009**, *11*, 35−40.

(5) Open Source Initiative. http://opensource.org/licenses (accessed Feb 2015).

(6) Creative Commons. https://creativecommons.org/choose/zero/ (accessed Feb 2015).

(7) Github. https://github.com (accessed Feb 2015).

(8) Figshare. http://figshare.com (accessed Feb 2015).

(9) Mozilla Science Lab. http://mozillascience.github.io/code-research-object/ (accessed Feb 2015).

(10) Impactstory. https://impactstory.org (accessed Feb 2015).

(11) Sourceforge. http://sourceforge.net (accessed Feb 2015).

DOI: 10.1021/acs.jpclett.5b00285
J. Phys. Chem. Lett. 2015, 6, 1168−1169