THE JOURNAL OF PHYSICAL CHEMISTRY C
Viewpoint
Cite This: J. Phys. Chem. C 2017, 121, 28212−28213
pubs.acs.org/JPCC

Is the Supporting Information the Venue for Reproducibility and Transparency?

Making research data, software, and data processing tools readily available to the public could significantly enhance the impact of scientific publications. Openness and transparency could address reproducibility concerns and accelerate scientific progress.[1,2] While archives and repositories would be preferable for securing data in the long term, the Supporting Information (SI) already allocates significant space to provide auxiliary files, links, and essential information needed to make scientific findings immediately reproducible as well as data processing protocols and numerical procedures executable. Furthermore, the SI document could ensure that data that might not fit in the tight confines of a journal article lives online even when laboratories move on to other projects, close, and lose track of their data.

The change for openness and transparency might be difficult to embrace, depending on the specific nature of the project. Nevertheless, the trend to make raw research data and open-source software packages available to the public has been building steady momentum. Openness in research data and software sharing is already making transformative contributions to the communication of research findings, critical for ensuring reproducibility as well as training of the next generation of scientists.

Significant progress on transparency has already been made in various fields, such as molecular and protein crystallography,[3,4] that will likely continue inspiring the broader scientific community. Databases and repositories in the public domain have been successfully established and proved transformational for a wide range of studies of small molecules[3,5] and proteins,[4] including extensive theoretical work and detailed analysis that builds upon data provided by expert crystallographers.[4] Much of this work is expedited by using machine-readable formats such as PDB[4] and CIF,[3] providing essential data that assists the task of peer review. At the same time, comparative studies of reported model structures as well as homology models are routinely performed, enabling studies that would never be possible if crystallography data were confined only to the research groups who reported the data.[3]

Similarly, sophisticated data search engines, such as Google Scholar and PubMed, together with the tools of data science have transformed the way the scientific community operates, allowing for instant accessibility of publications from anywhere and at any time.[6] In the theoretical/computational field, resources such as GitHub allow for massive dissemination of open-source software and codes under a stable URL.[7] In addition to commercial software, open-source software allows the global community to build upon codes developed by other researchers when they opt not to invest time and resources to redevelop tools that have already been published.[8] Such resources ensure reproducibility of reported results and secure codes that might otherwise run the risk of being lost in obscurity when developers no longer support them.[8]

An outstanding question is whether the physical chemistry community has sufficiently embraced the trend for openness or whether there are opportunities for further developments. Some enthusiasts of transparency suggest that the SI should provide enough data, instructions, and information for direct reproducibility of the reported results by the reviewers. Others suggest that immediate reproducibility should be a requirement for publication. While that suggestion is under debate, simple steps could already be taken to make publications more impactful and reported findings readily reproducible. For example, some researchers suggest that the universal, but difficult to process, PDF might not always be the most appropriate form for the SI. In particular, the PDF format is not ideal for data mining and facile analysis in the context of other data.[2,9] This aspect could be most relevant to methodology papers that intend to introduce approaches that others would adopt after reproducing the results reported in the publication.[1]

On the experimental front, the SI could include accessibility to the raw data as well as the files necessary for data processing (e.g., Igor Pro, Excel, Prism, etc.). For example, NMR studies could benefit from access to raw data provided through the SI. While the free induction decay (FID) contains all of the research data of an NMR experiment, it is routinely made unavailable in favor of Fourier-transformed plots and tables, which themselves are prone to unspecified data processing methods and human error.[10] The transformation of the data may result in the loss of valuable information such as line widths, which comment on the dynamical nature of the system, as well as information such as the field strengths essential to reproduce the experiments.[10] Including the FID as part of the Supporting Information (and making it part of a repository system) could allow for faster correction of mischaracterized molecules and identification of potential impurities that, while minute, may still be important (e.g., could affect biological activity).[10] While FID data is typically found in proprietary formats, free software packages could convert the data into a standard format and enable readers to regenerate the plots in the paper for themselves.[10]

The field of computational chemistry is especially conducive to making full use of the SI. Traditionally, publications have provided information for “reproducibility” in the form of keywords and parameters in sentence form as well as figures of molecular structures that might not be always sufficient for reproducibility. More recently, researchers have started to provide XYZ coordinates in the SI. However, some researchers suggest that input files should be provided because they are often essential for immediate reproducibility. In addition to coordinates, reproducibility of electronic structure calculations might require essential information from input files, such as the convergence criteria and implemented algorithms, which are often decisive in reproducibility even when using the same XYZ coordinates. Molecular dynamics (MD) simulations also require many “details” of the simulation conditions, beyond

Published: December 28, 2017

DOI: 10.1021/acs.jpcc.7b11665

the usual information provided in the manuscripts such as protonation states of amino acids, force-field parameters, and the constant parameters for simulations. Given the empirical nature of the force-field parameters and the great variety of molecules, solids, and proteins for which MD simulation is used, researchers often have to tweak the available force-field parameters. Thus the availability of the tweaked force-field parameters could be warranted in the SI. In a similar vein, specific parameters of the simulation (using rigid bonds or not, frequency of full electrostatic computation, etc.) may be crucial for the reviewing process and also for the training of budding scientists. Whereas the output data of an MD simulation is quite cumbersome, the SI might serve as the proper place to include the input files clearly depicting the simulation parameters. In particular, the inclusion of input files could assist experimentalists and those who do not routinely perform calculations by providing all of the parameters and keywords in the required format for execution. Such practice would aid reproducibility and benefit scientists that could use models and examples from previous studies for their own inputs. In addition, some researchers wonder whether access to the software employed for the reported calculations should be granted to the reviewers.[7,9] Even without access to the software, input files would certainly be more useful than the usual tables reported in the papers, which are themselves prone to error in reading or writing.[7] While open-source software would alleviate the concern of reproducibility, uploading output files (perhaps deleting performance data if required by specific software makers) could allow for more extensive analysis and facile processing.[7] Similarly, the SI could be the place to store the programs and postprocessing tools such as scripts that might be too small to be worth including in a repository. Tutorials and examples on how to run the software and calculations would not only make it easier to reproduce the reported findings[9] but also help new researchers learn the ropes of software packages, including new members of the research group reporting the publication.

While academic Web sites are convenient for hosting materials, there are several outstanding questions that remain to be addressed. For example: Are more stable repositories such as URL and even a DOI necessary to ensure security and robustness?[7] What else could the physical chemistry community do to promote a healthy culture of data sharing, openness, and transparency? Should the physical chemistry community and federal agencies promote and reward increased visibility of raw research data and tools that are ultimately taxpayer funded?[2] Would concerns about increased workload from dealing with correspondence and preparing the data be addressed by supplying tutorials, using specialized repositories, or embargoing data for several years?[2] Would concerns about ownership of code be mitigated by selecting the appropriate license agreement, which could allow free academic applications while charging for commercial use?[8] Are there more practical ways to ensure security and accessibility of research data?[7] Many of these questions merit discussion by the physical chemistry community as well as by the broader scientific forum.

Benjamin Rudshteyn
Atanu Acharya
Victor S. Batista*

Department of Chemistry and Energy Sciences Institute,
Yale University, New Haven, Connecticut 06520, United
States

AUTHOR INFORMATION

Corresponding Author

*E-mail: victor.batista@yale.edu.

ORCID

Benjamin Rudshteyn: 0000-0002-9511-6780
Atanu Acharya: 0000-0002-6960-7789
Victor S. Batista: 0000-0002-3262-1237

Notes

The authors declare no competing financial interest.
This Viewpoint is jointly published in *The Journal of Physical Chemistry A, B,* and *C*.

ACKNOWLEDGMENTS

V.S.B. acknowledges financial (computing time support) from the National Science Foundation (XSEDE), Air Force Office of Scientific Research (Copper), Department of Energy (NERSC and EMSL), the National Institutes of Health, Yale (YCRC), and the TomKat Charitable Trust. B.R. acknowledges support from the National Science Foundation Graduate Research Fellowship under Grant DGE-1122492. We thank Mr. Ray Kelly for helpful discussions and for editing the manuscript.

REFERENCES

(1) Chawla, D. S. Taking on Chemistry's Reproducibility Problem *Chemistry World* **2017**. https://www.chemistryworld.com/news/taking-on-chemistrys-reproducibility-problem/3006991.article.

(2) Gewin, V. Data Sharing: An Open Mind on Open Data. *Nature* **2016**, *529*, 117−119.

(3) Bruno, I. J.; Groom, C. R. A Crystallographic Perspective on Sharing Data and Knowledge. *J. Comput.-Aided Mol. Des.* **2014**, *28*, 1015−1022.

(4) Berman, H. M. The Protein Data Bank: A Historical Perspective. *Acta Crystallogr., Sect. A: Found. Crystallogr.* **2008**, *64*, 88−95.

(5) Groom, C. R.; Bruno, I. J.; Lightfoot, M. P.; Ward, S. C. The Cambridge Structural Database. *Acta Crystallogr., Sect. B: Struct. Sci., Cryst. Eng. Mater.* **2016**, *72*, 171−179.

(6) Falagas, M. E.; Pitsouni, E. I.; Malietzis, G. A.; Pappas, G. Comparison of PubMed, Scopus, Web of Science, and Google Scholar: Strengths and Weaknesses. *FASEB J.* **2008**, *22*, 338−342.

(7) Coudert, F.-X. Reproducible Research in Computational Chemistry of Materials. *Chem. Mater.* **2017**, *29*, 2615−2617.

(8) Pirhadi, S.; Sunseri, J.; Koes, D. R. Open Source Molecular Modeling. *J. Mol. Graphics Modell.* **2016**, *69*, 127−143.

(9) Walters, W. P. Modeling, Informatics, and the Quest for Reproducibility. *J. Chem. Inf. Model.* **2013**, *53*, 1529−1530.

(10) Bisson, J.; Simmler, C.; Chen, S.-N.; Friesen, J. B.; Lankin, D. C.; McAlpine, J. B.; Pauli, G. F. Dissemination of Original NMR Data Enhances Reproducibility and Integrity in Chemical Research. *Nat. Prod. Rep.* **2016**, *33*, 1028−1033.

DOI: 10.1021/acs.jpcc.7b11665