Journal of
Chemical Information
and Modeling

Letter

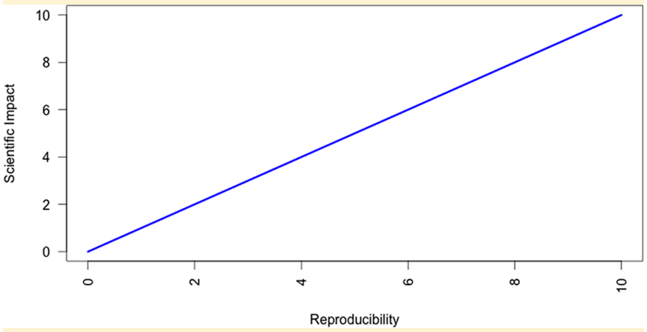
pubs.acs.org/jcim

Terms of Use

Modeling, Informatics, and the Quest for Reproducibility

W. Patrick Walters*

Vertex Pharmaceuticals, Inc., 130 Waverly St., Cambridge, Massachusetts 02139, United States



ABSTRACT: There is no doubt that papers published in the Journal of Chemical Information and Modeling, and related journals, provide valuable scientific information. However, it is often difficult to reproduce the work described in molecular modeling and cheminformatics papers. In many cases the software described in the paper is not readily available, in other cases the supporting information is not provided in an accessible format. To date, the major journals in the fields of molecular modeling and cheminformatics have not established guidelines for reproducible research. This letter provides an overview of the reproducibility challenges facing our field and suggests some guidelines for improving the reproducibility of published work.

One of the fundamental requirements for scientific progress is the ability to reproduce and build upon the work of others. In most scientific disciplines, reproducibility is a prerequisite for publication. In the field of organic synthesis, publications are expected to include comprehensive experimental sections that detail procedures and allow others to repeat the work. A synthesis paper must also include experimentally determined spectra to not only verify the identity of the compounds synthesized but also to allow others repeating the work to determine whether they have isolated the correct product. Reproducibility is also important in a number of computational disciplines such as physics, statistics, and bioinformatics, where the inclusion of source code for new methods as well as example data is the expected norm.

Unfortunately this is not the case for molecular modeling and cheminformatics. Papers in our discipline rarely, if ever, include source code. In many cases, even program executables are not available for others to evaluate. This has led to the unfortunate evolution of a somewhat dysfunctional "trust me; it works" culture where very little is shared. My intent is not to question the integrity of those who publish in this and related journals but to request some changes that will allow us to move forward as a community. I'm not sure what originally led those working in our field to not release source code. Some may have been motivated by a desire to secure intellectual property that would lead to future financial gains. Others, who are not professional programmers, may not consider their "research code" worthy of publication. Academic groups may feel that proprietary code provides an advantage when applying for grants. This argument could, of course, be removed if funding agencies required grant recipients to release their source code, as suggested in a recent editorial in *Science*.[1]

It is possible that the wide array of computer hardware platforms available 20 years ago would have made supporting a particular code base more difficult. However, over the last 10 years, the world seems to have settled on a small number of hardware platforms, dramatically reducing any sort of support burden. In fact, modern virtual machine technologies have made it almost trivial to install and run software developed in a different computing environment.

Other factors may have also affected the situation. Technology transfer offices at some universities have become more aggressive in pushing groups to monetize their research. Actually, cost is a small component of the difficulty created by tech transfer offices. Many industrial groups will not consider licensing academic code due to the time-consuming and often tedious process of negotiating a licensing agreement. Industrial groups are typically no better than academics when it comes to releasing code. Companies may believe that software provides a competitive advantage and prevent employees from publishing source code or data. However, this view does seem to be at odds with publishing the work in the first place. It seems that a proper publication would reveal as much as distributing the source code.

There has been a recent move by some academic groups to alleviate this situation by creating versions of new computational methods that are accessible through Web interfaces. While this is a step in the right direction, it does not address the core issue. The Web interface allows the methods to be tested and compared, but it is still a "black box" that does not allow others to build on the work. Another drawback to the Web interface model is that scientists in industry, who may be able to provide valuable feedback on the methodology, will not be able to upload proprietary compounds through the Web.

To make matters worse, the data used to generate or validate a new method, when included, are typically not in an easily machine-readable format. Computational papers often only include a common or IUPAC name for the compounds used for validation or testing of methodology. While it is possible to use Internet resources such as Wikipedia to generate a structure table from names, the work required to obtain a machine-readable structure representation often borders on heroic.

Published: June 12, 2013

When chemical structures are provided, they are typically presented as structure drawings that cannot be readily translated into a machine-readable format. When groups do go to the trouble of redrawing dozens of structures, it is almost inevitable that structure errors will be introduced. Years ago, one could have argued that too many file formats existed and that it was difficult to agree on a common format for distribution. However, over the last 10 years, the community seems to have agreed on SMILES and MDL Mol or SD files as a standard means of distribution. Both of these formats can be easily processed and translated using freely available software such as OpenBabel,[2] CDK,[3] and RDKit.[4] At the current time, there do not seem to be any technical impediments to the distribution of structures and data in electronic form.

Responsible reviewers can ensure reproducibility by insisting that others be able to easily test and extend work described in a manuscript submitted for publication. However, we would make more progress if the *Journal of Chemical Information and Modeling* (JCIM) and related journals adopted the following guidelines.

(1) Wherever possible, source code should be provided for new computational methods. The source code can be a reference implementation of a method or algorithm and does not need to include a graphical interface. If it is not possible to release the source code for a new method, authors should provide a sufficient justification. Reviewers and editors will then consider this explanation. Any paper that does not comply with the reproducibility guidelines will include this explanation when published. In cases where it is not possible to release code due to intellectual property or other limitations, an executable version of the new method should be readily accessible. Commercial products should provide time limited licenses to facilitate evaluation and comparison of published methods.

(2) Any chemical structures and data mentioned in the paper should be made available in a commonly used (SDF or SMILES) format. Distribution of data in pdf format is not sufficient.

(3) Any publications that employ commercial or open-source software should include scripts or parameter files as well as data files that will enable others to easily reproduce the work.

(4) A clear easy to follow description of any new method should be a key criterion during the review process. Wherever possible, a paper should contain a simple worked example that demonstrates the application of the method. Parameter values and intermediate results for example compounds should be included as part of the supporting material.

(5) Reviewers should put particular emphasis on the reproducibility of the method described in a manuscript. Each reviewer should evaluate the description of the method, as well as the presence of associated code, data, or executables, to ensure that the results can be independently reproduced.

As suggested by one of the anonymous reviewers, it would be useful to add the following questions to the review form. (1) Is the method reproducible based on the authors' description? (2) Is sufficient public data included to reproduce the work? (3) Have the authors provided links to downloadable binaries? (4) Have the authors provided links to downloadable source code? Rather than rejecting a paper based on a single negative response to one of these questions, the editors could use a "weight of evidence" approach in determining whether a paper would be accepted.

This is not the first call for reproducibility in our field. As pointed out in a blog post by Apodaca,[5] concerns over reproducibility of computational methods were published almost 20 years ago in this journal.[6] Recent papers by Landrum,[7] Ince,[8] and Neylon[9] have raised similar concerns. While the editors of JCIM[10] and the Journal of Medicinal Chemistry[11] have published guidelines on the use of proprietary data, their efforts only address a portion of the problem. In order to reproduce computational methods, both data and code need to be readily available in an immediately usable form.

The community is beginning to realize the need for reproducibility. A recent conference on free energy methods in drug discovery[12] featured numerous discussions on how code and data sets could be shared between groups. Collaboration is a key to discovery, and journals like JCIM need to do more to promote reproducibility of published work. The guidelines above, while not perfect, provide a starting point for a discussion of how we can improve and advance our field.

AUTHOR INFORMATION

Corresponding Author

*E-mail: pat_walters@vrtx.com Phone: (617) 341-6242.

Notes

The views expressed here are those of the author and not necessarily those of the author's employer, Vertex Pharmaceuticals, Inc.

The authors declare no competing financial interest.

REFERENCES

(1) Morin, A.; Urban, J.; Adams, P.; Foster, I.; Sali, A.; Baker, D.; Sliz, P. Shining light into black boxes. *Science* **2012**, *336*, 159−160.

(2) Open Babel: The Open Source Chemistry Toolbox. http://openbabel.org/wiki/Main_Page (accessed May 27, 2013).

(3) The Chemistry Development Kit. http://sourceforge.net/projects/cdk/ (accessed May 27, 2013).

(4) RDKit: Cheminformatics and Machine Learning Software. http://www.rdkit.org/ (accessed May 27, 2013).

(5) Depth First. http://depth-first.com/articles/2006/08/23/readily-available-without-infringements-or-restrictions (accessed May 27, 2013).

(6) Figueras, J. Letter to the editor. Comment on editorial on software distribution in science. *J. Chem. Inf. Comput. Sci.* **1984**, *24*, 276−276.

(7) Landrum, G. A.; Stiefl, N. Is that a scientific publication or an advertisement? Reproducibility, source code and data in the computational chemistry literature. *Future Med. Chem.* **2012**, *4*, 1885−1887.

(8) Ince, D. C.; Hatton, L.; Graham-Cumming, J. The case for open computer programs. *Nature* **2012**, *482*, 485−488.

(9) Neylon, C.; Aerts, J.; Brown, C. T.; Coles, S. J.; Hatton, L.; Lemire, D.; Millman, K. J.; Murray-Rust, P.; Perez, F.; Saunders, N. Changing computational research. The challenges ahead. *Source Code Biol. Med.* **2012**, *7*, 1−2.

(10) Jorgensen, W. L. QSAR/QSPR and proprietary data. *J. Chem. Inf. Model.* **2006**, *46*, 937−937.

(11) Stahl, M.; Bajorath, J. Computational medicinal chemistry. *J. Med. Chem.* **2011**, *54*, 1.

(12) 2012 Workshop on Free Energy Methods in Drug Design. http://www.alchemistry.org/wiki/index.php?title=2012_Workshop_on_Free_Energy_Methods_in_Drug_Design (accessed May 27, 2013).