

Towards “Reproducibility-as-a-Service”

Tom Crick¹, Samin Ishtiaq² and Benjamin A. Hall³

¹Department of Computing & Information Systems, Cardiff Metropolitan University, UK

²Microsoft Research Cambridge, UK

³MRC Cancer Unit, University of Cambridge, UK

¹tcrick@cardiffmet.ac.uk

²samin.ishtiaq@microsoft.com

³bh418@mrc-cu.cam.ac.uk

Abstract

The reproduction and replication of novel results has become a major issue for a number of scientific disciplines. In computer science and related computational disciplines such as systems biology, the issues closely revolve around the ability to implement novel algorithms and models. Taking an approach from the literature and applying it to a new codebase frequently requires local knowledge missing from the published manuscripts and project websites. Alongside this issue, benchmarking, and the development of fair — and publicly available — benchmark sets present another barrier.

In this paper, we outline several suggestions to address these issues, driven by specific examples from a range of scientific domains. Finally, based on these suggestions, we propose a new open automated platform for scientific software development which effectively abstracts specific dependencies from the individual researcher and their workstation, allowing easy sharing and reproduction of results. This new cyberinfrastructure for computational science offers the potential to incentivise a culture change and drive the adoption of new techniques to improve the efficiency of scientific exploration.

1 Introduction

Marc Andreessen (co-author of Mosaic, the first widely used Web browser) boldly stated in 2011 that “*software is eating the world*” [1]. It is true: we clearly live in a computational world, with our everyday communications, entertainment, shopping, banking, transportation, national security, etc, all heavily dependent on (or replaced by) software.

This is particularly true for science and engineering. A 2012 report by the Royal Society stated that computational techniques have “*moved on from assisting scientists in doing science, to transforming both how science is done and what science is done*” [2]. Many of the examples discussed in this paper take advantage of a fundamental advantage of computer science and more broadly, computational science: the unique ability to share the raw outputs of their work as software and datafiles. New experiments, simulations, models, benchmarks, even proofs cannot be done without software. And this software does not consist of simple hack-together, use-once, throw-away scripts; scientific software repositories contain thousands, perhaps millions, of lines of code and they increasingly need to be actively supported and maintained. More importantly, with reproducibility being a fundamental tenet of science, they need to be re-useable.

However, if we closely analyse the scientific literature related to software tools it often does not appear to be adhering to these rules [3]. How many of them are reproducible? How many explain their experimental methodologies, in particular the basis for their benchmarking? In particular, can we (re)build the code [4]? We, the authors, are perhaps as guilty as anyone in the past, where we have published papers [5, 6] with benchmarks and promises of code to be released in the near future.

There are numerous reasons why the wider scientific community is in this state. We are currently undergoing significant changes to academic dissemination and publication, especially the open access movement, with new models being proposed [7–9]. Journals such as Nature, PLoS Computational Biology and Bioinformatics explicitly require that source code and data is made available online under some form of open source license. While these initiative are great, they are often optional, seem piecemeal, and do little to enable verification or validation of scientific results at a later stage. Even within the same field, there are different ideas of what defines reproducibility.

Nevertheless, the reproduction and replication of reported scientific results has become a widely discussed topic within the scientific community [10–12]. Whilst the retraction of several studies has drawn the focus of many commentators, automated systems, which allow easy reproduction of results, offer the potential to improve the efficiency of scientific exploration and drive the adoption of new techniques. But just publishing (linked) scientific data is not enough to ensure the required reusability [13]. There exists a wider socio-cultural problem that pervades the scientific community, with estimates that as much as 50% of published studies, even those in top-tier academic journals, cannot be repeated with the same conclusions by an industrial lab [14]. There are numerous non-technical impediments to making software maintainable and re-useable. The pressure to “make the discovery” and publish quickly disincentivises careful software curation. Releasing code prematurely is often seen to give your competitors an advantage, but we should be shining light into these “black boxes” [11]. In essence: better software, better research [15].

Nevertheless, there is promising existing work in this area [16–19], as well as a number of manifestos for reproducible research and community initiatives, such as the Recomputation Manifesto [20]¹ and cTuning [21]², along with curated recommendations on where to publish research software³.

However, things can, should and need to be much better. Building upon previous work [22], we present a call to action, along with a set of recommendations which we hope will lead to better, more sustainable, more re-useable software, to move towards an imagined future practice and usage of scientific software development. The basis for many of these recommendations is predicated on the fundamental scientific tenets of openness and sharing.

2 We Need to Talk About Reproducibility

2.1 *Can I Implement Your Algorithm?*

Reproducibility (replication, repeatability, et al.) is a fundamental tenet of good science. Yet many descriptions of algorithms are too high-level, too obscure, too poorly-defined to allow an easy re-implementation by a third party. A step in the algorithm might say: “*We pick an element from the frontier set*” but which element do you pick? Will the first one do? Why will any element suffice? Sometimes the author would like to give more implementation detail but is constrained by the paper page limit. Sometimes the authors’ description in-lines other algorithms or data structures that perhaps only that author is familiar with.

Recommendation I: We recommend here that a paper must describe the algorithm in such a way that it is implementable by any reader of that algorithm. This is subjective, of course. Therefore, we also recommend that relevant scientific conferences have a special track for papers that re-implement past papers’ algorithms, techniques or tools, as well as incentives to support sharing of computational artefacts; reproducibility should not just be discussed at conference and workshops convened explicitly for that purpose (for example, a number of high-profile computer science conferences now explicitly acknowledge the importance of reproducibility, promoting community-driven reviewing and validation⁴; we have recently proposed a new reproducibility model for the computer-aided formal analysis and verification research community [23]).

2.2 *Set The Code Free*

There can be no better proof that your algorithm works, than if you provide the source code of an implementation. Software development is hard, but sharing and re-using code is relatively easy.

Many years ago, Richard Stallman (founder of the GNU Project and Free Software Foundation) postulated that all code would be free [24] and we would make our money by consulting on the code. As it turns out, this is now the case for a significant part of the computing industry. There are, of course, hard commercial pressures for keeping code closed-source. Even in the scientific domain, scientists and their collaborators may wish to hold onto their code as a competitive advantage, especially if there exists larger competitors who could use the available code to “reverse scoop” the inventors, charging into a promising new research area opened by the inventors.

¹<http://www.recomputation.org/>

²<http://ctuning.org/>

³<http://www.software.ac.uk/resources/guides/which-journals-should-i-publish-my-software>

⁴<http://ctuning.org/cm/wiki/index.php?title=Reproducibility>

Closed source is one thing. Licenses that deny the user from viewing, modifying, or sharing the source are another thing. There are, however, even licences on widely adopted tools like GAUSSIAN [25] that prohibit even analysing software performance and behaviour. For example, a wide variety of licenses exist for molecular dynamics software, with different degrees of openness (GROMACS uses the GNU Lesser General Public License (LGPL) [26], CHARMM and Desmond are Academic/Commercial software licences [27, 28], Amber and NAMD are custom open-like licences). Z3 is an example from the verification area: the code itself is not open source, but the MSR-LA license that allows the source code to be read, copied, forked for academic use, provides researchers in the field much more than before [29].

Recommendation II: There is little doubt that, if science wants to be open and free, then the code that underlies it too needs to be open and free. Code that is available for browsing, modifying, and forking facilitates testing and comparison, and promotes competition. We recommend that code be published under an appropriate open source license [30]; while we defer legal discussion of the specifics of any particular licences, BSD and Apache are good, flexible ones.

Ultimately: set the code free. Put it on a public space such as GitHub, where it is easy to share and fork. You should embrace the spirit of the (somewhat tongue-in-cheek) CRAPL academic-strength open source license⁵ and publish your code – it is good enough [10].

2.3 *Be A Better Person*

If you have the appropriate skills and the experience, you can always create better software. We have seen the emergence of successful initiatives, such as the Software Sustainability Institute⁶, Software Carpentry⁷ and the UK Community of Research Software Engineers⁸, in cultivating world-class research through software, developing software skills and raising the profile of research software engineers.

Many scientists will not have had any formal, or even informal, training in scientific software development. Even basic training in software engineering concepts like version control, unit testing, build tools, etc, can help improve the quality of the software written enormously [31]. Interestingly, many of these concepts are taught to computer science undergraduates, but it could be argued that they are taught at the wrong time of their careers, without the experience of complex, long-running projects.

Recommendation III: Software development skills should be regarded as fundamental literacies for scientists and engineers: we recommend that basic programming and computational skills are taught as core at undergraduate and postgraduate level.

2.4 *The Lingua Franca of Computational Science*

There is no other scientific or technical field where its participants can just make up a non-principled artefact like a programming language so easily. In a way, it shows how much of a “commons” computer science has become, that anyone can create a new programming language, API, framework or compiler. This clearly has its advantages and disadvantages.

High-level languages are generally more readable than their competitors. The “density” of a program is often seen to be a good thing, but it is not always the case that a shorter Haskell program (for example) is easier to maintain than a longer Python/C++ one. Nevertheless, what is important is the readability of the code itself. A good example here is from the world of automatic theorem proving: the SSReflect language is much more readable than the original, standard Coq language [32]. SSReflect uses mathematicians’ vernacular for script commands, allows reproducibility of automatic proof-checking because parameters are named rather than numbered. Even though these proof scripts are really only ever going to be run by a machine, they seek to maintain the basic mathematical idea that a proof should be readable by another mathematician.

High-level programming languages impose constraints like types: that you can never add a number and a string is the most basic example, but ML’s functors provide principled ways of plugging in components with

⁵<http://matt.might.net/articles/crapl/>

⁶<http://www.software.ac.uk/>

⁷<http://software-carpentry.org/>

⁸<http://www.rse.ac.uk>

their implementations completely hidden. Aggressive type checking avoids a subset of bugs which can arise due to incorrectly written functions e.g. well publicised problems with a NASA Mars orbiter⁹. A further example is a pressure coupling bug¹⁰ in GROMACS [26], which arose due to the inappropriate swapping of a pressure term with a stress tensor. A further extension of types, a concept called units of measure that is implemented in languages such as F#, can deal with these kinds of bugs at compile time. Similarly, problems found using in-house software for crystallography led to the retraction of five papers [33], due to a bug which inverted the phases.

Recommendation IV: The use of a principled, high-level programming language in which to write your software helps hugely with the maintainability, robustness and openness of the software produced.

2.5 *Test It To See*

Some models may be chaotic and influenced by floating-point errors (e.g. molecular dynamics), further frustrating testing. For example: Sidekick is an automated tool for building molecular models and performing simulations [34]. Each system is simulated from an different initial random seed, and under most circumstances this is the only difference expected between replicas. However, on a mixed cluster with both AMD and Intel microprocessors on the nodes, the difference in architecture was found to alter the number of water molecules added to each system by one. This meant that the same simulation performed on different architectures would diverge. Similarly, in a different simulation engine, different neighbour searching strategies gave divergent simulations due to the differing order in which forces were summed.

A further example is the handling of pseudo-random number generation in Avida [35], an open source scientific software platform for conducting and analysing experiments with self-replicating and evolving computer programs. In order to produce consistent random number generation across platforms, it may be necessary to code bespoke random number generators within the system, which is not ideal for sharing and reproducibility.

Recommendation V: Despite these challenges to testing, unshared code is ultimately untestable. Testing new complex scientific software is difficult – until the software is complete, unit tests may not be available. You should thus aim to link to/from publicly-shared code: shared code is inherently more test-able.

2.6 *Lineage (or: “Standing On The Shoulders Of Giants”)*

Research software is not just software – it is the instantiation of novel algorithms and data structures (or at least novel applications of data structures). Thus, lineage is important:

Recommendation VI: Code should always include links to papers publishing key algorithms and the code should include explicit relationships to other projects on the repository (i.e. *Project B* was branched from *Project A*). This ensure that both the researchers and software developers working upstream of the current project are properly credited, encouraging future sharing and development. Remember, the people who did the research are not necessarily the same people as the developers and maintainers of the software, so it is important to reward both appropriately with citations (a good way of doing this is the use of CITATION files¹¹).

2.7 *YMMV*

The tweet in Figure 1 is sad but worryingly true, highlighting the perils of reproducible research¹². Often, the tool that the paper describes does not exist for download. Or runs only on one particular bespoke platform. Or might run for the author, for a while, but will ‘bit-rot’ so quickly that even the author cannot compile it in a couple of month’s time.

Recommendation VII: Providing the source code of the tool helps, of course. But you must also provide details of precisely *how* you built and wrote the software. For example:

⁹<http://www.cnn.com/TECH/space/9909/30/mars.metric.02/>

¹⁰<http://redmine.gromacs.org/issues/14>

¹¹<http://blog.rtwilson.com/encouraging-citation-of-software-introducing-citation-files/>

¹²<http://www.phdcomics.com/comics.php?f=1689>



Figure 1: #overlyhonestmethods on Twitter

[source: <https://twitter.com/ianholmes/status/288689712636493824>]

- you should provide the compiler and build toolchain;
- you should provide build tools (e.g. Makefiles/Ant/etc) and comprehensive build instructions;
- you should list or link to all non-standard packages and libraries that you use;
- you should note the specifics of the hardware and OS used.

This may appear to be significant extra overhead for researchers, but GitHub APIs, continuous integration servers, virtual machines and cloud environments can make it easier; see Section 3 for more on this.

2.8 Data Representations and Formats

We often do not, and should not, care how things are stored on disk, what their precise representations are. But a common, constrained, standard representation is good for passing tests or models around between different tools. A properly described representation, like the SMT-LIB format¹³ for Satisfiability Modulo Theory (SMT) solvers, where both the syntax and semantics are well understood, hugely aids developing tools, techniques and benchmarks.

Another example, from biology, is that of the standard representation of qualitative networks and Boolean networks [36, 37]. These networks can be expressed in SMV format, but this would mean that standard qualitative/Boolean network behaviours have to be hard-coded for each variable, introducing the possibility for errors. In the BioModelAnalyzer tool [38], the XML contains *only* the modifiable parameters limiting the possibility for error.

Recommendation VIII: Avoid creating new representations when common formats already exist. Use existing extensible internationally standardised representations and formats to facilitate sharing and re-use.

2.9 World Records

The benchmarks the tool describes are fashioned only for this instance of this time. They might claim to be from the Windows device driver set, but the reality is that they are stripped down versions of the originals. Stripped down so much as to be useless to anyone but the author vs. the referee. It is worse than that really: enough benchmarks are included to beat other tools. The comparisons are never fair (neither are

¹³<http://smt-lib.org>

other peoples’ comparisons against your tool). If every paper has to be novel, then every benchmark, too, will be novel; there is no monotonic, historical truth in new, synthetically-crafted benchmarks. It is as if, in order to beat Usain Bolt’s 100m world record time, you make him wear boots on a muddy icy track, weighing him down with 50kg of excess weight. Given this set up, you could surely hope to beat his 9.63s time on a shorter length track.

Recommendation IX: Benchmarks should be public. They should allow anyone to contribute, implying that the tests are in a standard format. Further, these benchmarks must be heavily curated. Every test/assertion should be justified. Papers should be penalised if they do not use these public benchmarks. While there are some domains in which it may not be immediately possible to share full benchmarks sets, this should be the exception (with justification) rather than the norm.

A good example of some of these points is the RCSB Protein Data Bank¹⁴ and Systems Biology Markup Language [39]. The software ones we know of, the SMT Competition¹⁵, SV-COMP¹⁶ and Termination Problems Data Base¹⁷ are on that journey. Such repositories would allow the tests to be taken and easily analysed by any competitor tool.

2.10 Welcome to Web 2.0

Virtual machines (VMs) in the cloud also make the testing of scaling properties more simple. If you have a tool that you claim is more efficient, you could put together a cluster of slow nodes in the cloud to demonstrate how well the software scales for parallel calculations. Cloud computing is cheap, and getting cheaper. Algorithms that used to require massive HPC resources can now be run cheaply by bidding on the VM spot market. The Web is a great leveller: use and share workflows and web services [40,41].

Recommendation X: The Web and the cloud really do open up a whole new way of working. Even small, seemingly trivial features like putting up a web interface to your tool and its tests will allow users who are not able to install necessary dependencies to explore the running of the tool [42]. Ultimately, this can lead to making an “executable paper” appear on the Internet. The interactive *Try F#*¹⁸ and *Z3* tutorials¹⁹ are a great start that begin to expose what can be done in this area.

3 A Model for Reproducible Research Software

A service for reproducibility is intended to play three important roles. It should:

1. Demonstrate that a piece of code can be compiled, run and behaves as described, without manual intervention from the developer.
2. Store and link specific artefacts with their linked publications or other publicly-accessible datasets.
3. Allow new benchmarks to be added, by users other than the developer, to widen the testing and identify potential bugs.

The whole premise of our previous paper [43] is that *algorithms* (implementations) and *models* (benchmarks) are inextricably linked. Algorithms are designed for certain types of models; models, though created to mimic some physical reality, also serve to stress the current known algorithms. An integrated autonomous open cloud-based service can make this link explicit.

By developing a cloud-based, centralised service, which performs automated code compilation, testing and benchmarking (with associated auditing), we will link together published implementations of algorithms and input models. This will allow the prototype to link together software and data repositories, toolchains, workflows and outputs, providing a seamless automated infrastructure for the verification and validation of

¹⁴<http://www.pdb.org>

¹⁵<http://smtcomp.sourceforge.net/2014/>

¹⁶<http://sv-comp.sosy-lab.org/2015/>

¹⁷<http://termination-portal.org/wiki/TPDB>

¹⁸<http://www.tryfsharp.org/Learn>

¹⁹<http://rise4fun.com/Z3/tutorial/guide>

scientific models and in particular, performance benchmarks. The program of work will lead the cultural shift in both the short and long-term to move to a world in which computational reproducibility helps researchers achieve their goals, rather than being perceived as an overhead.

A system as described here has several up-front benefits: it links research papers more closely to their outputs, making external validation easier and allows interested users to explore unaddressed sets of models. Critically, it helps researchers across computational science to be more productive, rather than reproducibility being an overhead on their day-to-day work. In the same way that tools such as GitHub make collaborating easier while simultaneously allowing effortless sharing, we envisage our system being similarly usable for sharing and testing algorithms, software, models and benchmarks online.

Suppose you have come up with a better algorithm to deal with some standard problem. You write up the paper on the algorithm, and you also push an implementation of your algorithm to the our cloud environment’s section on this standard problem. The effect of pushing your implementation is to register your program as a possible competitor in this standard problem competition. There are several dozen widely-agreed tests on this problem already on our cloud environment’s database. Maybe, after some negotiation due to your novel approach to this standard problem, you add some of your own tests to the database too.

Pushing your code activates the environment’s continuous integration system. The cloud pulls in all the dependencies your code needs, on the platforms you specify, and runs all the benchmarks. This happens every time you push. It also happens every time one of your dependencies (a library, a firmware upgrade for your platform, a new API) changes too. This system (presented in Figure 2) would integrate with publicly available source code repositories, automates the build, testing and benchmarking of algorithms and benchmarks. It would allow testing models against competing algorithms, and the addition of new models to the test suite (either manually or from existing online repositories).

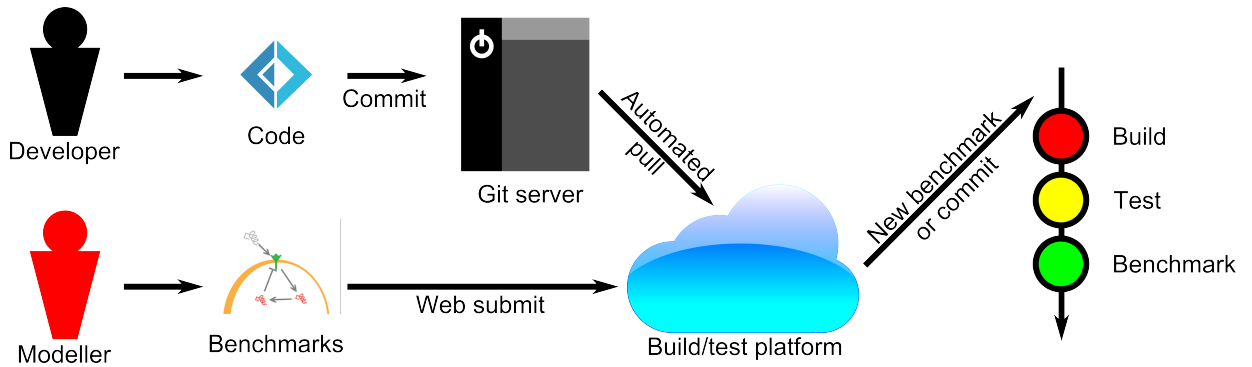


Figure 2: Proposed reproducibility service workflow

If we are truly serious about addressing the systemic socio-technical issues in scientific disciplines that are underpinned by leveraging software and computational techniques, then the proposal above would bring together almost all of the points we have discussed in this paper to provide an open research infrastructure for all. There are already several web services that already aim to do many of these things (for example, a repository for disseminating the computational models associated with publications in the social and life sciences [44]), so a service that can integrate most if not all of these features is possible. Such a service would then allow algorithms and models to evolve together, and be reproducible from the outset. Something more complete, and stamped with the authority of the major domain conferences/journals/professional societies, would mean that your code would never ‘bit-rot’, and no one would have problems reproducing the implementation of your published algorithm.

4 Next Steps

Following the proposal of such a system, the question becomes: *how do we encourage widespread uptake, or even standardisation?* Such a service would appear to be non-trivial, given the large numbers of tools and workflows that could potentially require to be supported by the service. Furthermore, after such a service has been implemented, how do we ensure it is *useful* and *usable* for researchers.

The benefits to the wider computational research community from a cultural change to favour reproducibility are clear and as such we should aim through software cyberinfrastructure and sharable, community curated research workflows to mitigate these costs. Furthermore, we can reasonably expect the distinct

needs of specific research communities to evolve over time, and initial implementations of the platform may require refinement in response to user feedback (supporting the critical cultural change by improving the efficiency of researchers). As such, if the wider research community is to move to requiring reproducibility, it seems most reasonable that this is staggered over a number of years to allow for both of these elements to develop, until eventually all researchers are required to use the service.

The key question for different research communities then becomes: *how to initialise this change?* Such a requirement creates a set of new costs to researchers, both in terms of time spent ensuring that their tools work on the centralised system (in addition to their local implementation), but also potentially in terms of equipment (in terms of running the system). Such costs may be easier to bear for some groups compared to others, especially those with large research groups who can more easily distribute the tasks, and it is important that the service does not present a barrier to early career researchers and those with efficient budgets (this type of cost analysis is not unique to reproducibility efforts – it has been estimated that a shift to becoming exclusively open access for a journal may lead to a ten-fold increase in computer science publication costs [45]).

Nevertheless, this proposed new cyberinfrastructure could have a profound impact on the way that computational science is performed, repositioning the role of models, algorithms and benchmarks and accelerating the research cycle, perhaps truly enabling a “fourth paradigm” of data intensive scientific discovery [46]. But ultimately, an open discussion and understanding of what reproducibility means for the wider computational science research community is important: we all need to explicitly state that this is worthwhile and address it, or don’t bother doing it at all.

References

- [1] Marc Andreessen. Why Software Is Eating The World. *The Wall Street Journal*, August 2011. Available online: <http://online.wsj.com/news/articles/SB10001424053111903480904576512250915629460>.
- [2] Royal Society. Science as an open enterprise. 2012. Available from: <https://royalsociety.org/policy/projects/science-public-enterprise/report/>.
- [3] Editorial. Devil in the details. *Nature*, 470(7334):305–306, 2011.
- [4] Christian Collbery, Todd Proebsting, Gina Moraila, Akash Shankaran, Zuoming Shi, and Alex M. Warren. Measuring Reproducibility in Computer Systems Research. Technical report, Department of Computer Science, University of Arizona, 2014.
- [5] Tom Crick, Marina De Vos, Martin Brain, and John Fitch. Generating Optimal Code using Answer Set Programming. In *Proceedings of 10th International Conference on Logic Programming and Non-monotonic Reasoning (LPNMR’09)*, volume 5753 of *Lecture Notes in Computer Science*, pages 554–559. Springer, 2009.
- [6] Josh Berdine, Byron Cook, and Samin Ishtiaq. SLayer: Memory Safety for Systems-Level Code. In *Proceedings of the 23rd International Conference on Computer Aided Verification (CAV 2011)*, volume 6806 of *Lecture Notes in Computer Science*, pages 178–183. Springer, 2011.
- [7] David De Roure. Replacing the Paper: The Twelve Rs of the e-Research Record. <http://www.scilogsg.com/eresearch/replacing-the-paper-the-twelve-rs-of-the-e-research-record/>, November 2011.
- [8] Victoria Stodden, Peixuan Guo, and Zhaokun Ma. Toward Reproducible Computational Research: An Empirical Analysis of Data and Code Policy Adoption by Journals. *PLoS ONE*, 8(6), 2013.
- [9] Grigori Fursin and Christophe Dubach. Community-Driven Reviewing and Validation of Publications. In *Proceedings of the 1st ACM SIGPLAN Workshop on Reproducible Research Methodologies and New Publication Models in Computer Engineering (TRUST’14)*, pages 1–4. ACM Press, 2014.
- [10] Nick Barnes. Publish your computer code: it is good enough. *Nature*, 467(753), 2010.
- [11] A. Morin, J. Urban, P. D. Adams, I. Foster, A. Sali, D. Baker, and P. Sliz. Shining Light into Black Boxes. *Science*, 336(6078):159–160, 2012.

- [12] Lucas N. Joppa, Greg McInerny, Richard Harper, Lara Salido, Kenji Takeda, Kenton O’Hara, David Gavaghan, and Stephen Emmott. Troubling Trends in Scientific Software Use. *Science*, 340(6134):814–815, 2013.
- [13] Sean Bechhofer, Iain Buchan, David De Roure, Paolo Missier, John Ainsworth, Jiten Bhagata, Philip Couch, Don Cruickshank, Mark Delderfield, Ian Dunlop, Matthew Gamble, Danus Michaelides, Stuart Owen, David Newman, Shoaib Sufi, and Carole Goble. Why linked data is not enough for scientists. *Future Generation Computer Systems*, 29(2):599–611, 2013.
- [14] Lev Osherovich. Hedging against academic risk. *Science-Business eXchange*, 4(15), 2011.
- [15] Carole Goble. Better Software, Better Research. *IEEE Internet Computing*, 18(5):4–8, 2014.
- [16] S.E. Sim, S. Easterbrook, and R.C. Holt. Using benchmarking to advance research: a challenge to software engineering. In *Proceedings of the 25th International Conference on Software Engineering (ICSE 2003)*, pages 74–83. IEEE Press, 2003.
- [17] Fernando Seabra Chirigati, Matthias Troyer, Dennis Shasha, and Juliana Freire. A Computational Reproducibility Benchmark. *IEEE Data Engineering Bulletin*, 36(4):54–59, 2013.
- [18] Victoria Stodden and Sheila Miguez. Best Practices for Computational Science: Software Infrastructure and Environments for Reproducible and Extensible Research. *Journal of Open Research Software*, 2(1):1–6, 2014.
- [19] Victoria Stodden, Sheila Miguez, and Jennifer Seiler. ResearchCompendia.org: Cyberinfrastructure for Reproducibility and Collaboration in Computational Science. *Computing in Science & Engineering*, 17(12), 2015.
- [20] Ian P. Gent. The Recomputation Manifesto. Available from: <http://arxiv.org/abs/1304.3674>, April 2013.
- [21] Grigori Fursin, Renato Miceli, Anton Lokhmotov, Michael Gerndt, Marc Baboulin, Allen D. Malony, Zbigniew Chamski, Diego Novillo, and Davide Del Vento. Collective mind: Towards practical and collaborative auto-tuning. *Scientific Programming*, 22(4):309–329, 2014.
- [22] Tom Crick, Benjamin A. Hall, and Samin Ishtiaq. “Can I Implement Your Algorithm?”: A Model for Reproducible Research Software. In *Proceedings of 2nd International Workshop on Sustainable Software for Science: Practice and Experiences (WSSSPE2)*, 2014.
- [23] Tom Crick, Benjamin A. Hall, and Samin Ishtiaq. Dear CAV, We Need to Talk About Reproducibility. Available from: <http://arxiv.org/abs/1502.02448>, February 2015.
- [24] Richard M. Stallman. *Free Software Free Society: Selected Essays of Richard M. Stallman*. Free Software Foundation, 2010.
- [25] Jim Giles. Software company bans competitive users. *Nature*, 429(6989), 2004.
- [26] Berk Hess, Carsten Kutzner, David van der Spoel, and Erik Lindahl. GROMACS 4: Algorithms for Highly Efficient, Load-Balanced, and Scalable Molecular Simulation. *Journal of Chemical Theory and Computation*, 4(3):435–447, 2008.
- [27] B. R. Brooks, C. L. Brooks, A. D. Mackerell, L. Nilsson, R. J. Petrella, B. Roux, Y. Won, G. Archontis, C. Bartels, S. Boresch, A. Caffisch, L. Caves, Q. Cui, A. R. Dinner, M. Feig, S. Fischer, J. Gao, M. Hodoscek, W. Im, K. Kuczera, T. Lazaridis, J. Ma, V. Ovchinnikov, E. Paci, R. W. Pastor, C. B. Post, J. Z. Pu, M. Schaefer, B. Tidor, R. M. Venable, H. L. Woodcock, X. Wu, W. Yang, D. M. York, and M. Karplus. CHARMM: The biomolecular simulation program. *Journal of Computational Chemistry*, 30(10):1545–1614, 2009.
- [28] Kevin J. Bowers, Edmond Chow, Huafeng Xu, Ron O. Dror, Michael P. Eastwood, Brent A. Gregersen, John L. Klepeis, Istvan Kolossvary, Mark A. Moraes, Federico D. Sacerdoti, John K. Salmon, Yibing Shan, and David E. Shaw. Scalable algorithms for molecular dynamics simulations on commodity clusters. In *Proceedings of the 2006 ACM/IEEE Conference on Supercomputing*. IEEE Press, 2006.

- [29] Leonardo de Moura. Releasing the Z3 source code. 2012. Available online: <http://leodemoura.github.io/blog/2012/10/02/open-z3.html>.
- [30] Open Source Licenses. <http://opensource.org/licenses>.
- [31] Greg Wilson. Software carpentry: Getting scientists to write better code by making them more productive. *Computing in Science & Engineering*, 8(6), 2006.
- [32] Georges Gonthier, Beta Ziliani, Aleksandar Nanevski, and Derek Dreyer. How to make ad hoc proof automation less ad hoc. *Journal of Functional Programming*, 23(4):357–401, 2013.
- [33] Greg Miller. A Scientist’s Nightmare: Software Problem Leads to Five Retractions. *Science*, 314(5807):1856–1857, 2006.
- [34] Benjamin A. Hall, Khairul Bariyyah Abd Halim, Amanda Buyan, Beatrice Emmanouil, and Mark S. P. Sansom. Sidekick for membrane simulations: Automated ensemble molecular dynamics simulations of transmembrane helices. *Journal of Chemical Theory and Computation*, 10(5):2165–2175, 2014.
- [35] Charles Ofria and Claus O. Wilke. Avida: A Software Platform for Research in Computational Evolutionary Biology. *Artificial Life*, 10(2):191–229, 2004.
- [36] S. A. Kauffman. Metabolic stability and epigenesis in randomly constructed genetic nets. *Journal of Theoretical Biology*, 22(3):437–67, 1969.
- [37] M A Schaub, T A Henzinger, and J Fisher. Qualitative networks: a symbolic approach to analyze biological signaling networks. *BMC Systems Biology*, 1:4, 2007.
- [38] David Benque, Sam Bourton, Caitlin Cockerton, Byron Cook, Jasmin Fisher, Samin Ishtiaq, Nir Piterman, Alex Taylor, and Moshe Y Vardi. BMA: visual tool for modeling and analyzing biological networks. In *Proceedings of the 24th International Conference on Computer Aided Verification (CAV 2012)*, volume 7358 of *Lecture Notes in Computer Science*, pages 686–692. Springer, 2012.
- [39] C. Chaouiya, D. Berenguier, S. M. Keating, A. Naldi, M. P. van Iersel, N. Rodriguez, A. Drager, F. Buchel, T. Cokelaer, B. Kowal, B. Wicks, E. Goncalves, J. Drier, M. Page, P. T. Monteiro, A. von Kamp, I. Xenarios, H. de Jong, M. Hucka, S. Klamt, D. Thieffry, N. Le Novère, J. Saez-Rodriguez, and T. Helikar. SBML qualitative models: a model representation format and infrastructure to foster interactions between qualitative modelling formalisms and tools. *BMC Systems Biology*, 7, 2013.
- [40] Tom Crick, Peter Dunning, Hyunsun Kim, and Julian Padget. Engineering Design Optimization using Services and Workflows. *Philosophical Transactions of the Royal Society A*, 367(1898):2741–2751, 2009.
- [41] S. Olabarriaga, G. Pierantoni, G. Taffoni, E. Sciacca, M. Jaghoori, V. Korkhov, G. Castelli, C. Vuerli, U. Becciani, E. Carley, and B. Bentley. Scientific Workflow Management – For Whom? In *Proceedings of 10th IEEE International Conference on e-Science (e-Science 2014)*, pages 298–305. IEEE Press, 2014.
- [42] Benjamin A. Hall, Ethan Jackson, Alex Hajnal, and Jasmin Fisher. Logic programming to predict cell fate patterns and retrodict genotypes in organogenesis. *Journal of The Royal Society Interface*, 11(98), 2014.
- [43] Tom Crick, Benjamin A. Hall, Samin Ishtiaq, and Kenji Takeda. “Share and Enjoy”: Publishing Useful (and Usable) Scientific Models. In *Proceedings of 1st International Workshop on Recomputability*, 2014.
- [44] Nathan D. Rollins, C. Michael Barton, Sean Bergin, Marco A. Janssen, and Allen Lee. A Computational Model Library for publishing model documentation and code. *Environmental Modelling & Software*, 61:59–64, 2014.
- [45] Moshe Y. Vardi. Openism, IPism, Fundamentalism, and Pragmatism. *Communications of the ACM*, 57(8), 2014.
- [46] Tony Hey, Stewart Tansley, and Kristin Tolle, editors. *The Fourth Paradigm: Data-Intensive Scientific Discovery*. Microsoft Research, 2009.