

“Share and Enjoy”: Publishing Useful and Usable Scientific Models

Tom Crick
Department of Computing
Cardiff Metropolitan University
Cardiff, UK
Email: tcrick@cardiffmet.ac.uk

Benjamin A. Hall and Samin Ishtiaq and Kenji Takeda
Microsoft Research
Cambridge, UK
Email: {benhall,samin.ishtiaq,kenji.takeda}@microsoft.com

Abstract—The reproduction and replication of reported scientific results is a hot topic within the academic community. The retraction of numerous studies from a wide range of disciplines, from climate science to bioscience, has drawn the focus of many commentators, but there exists a wider socio-cultural problem that pervades the scientific community. Sharing data and models often requires extra effort, and this is currently seen as a significant overhead that may not be worth the time investment.

Automated systems, which allow easy reproduction of results, offer the potential to incentivise a culture change and drive the adoption of new techniques to improve the efficiency of scientific exploration. In this paper, we discuss the value of improved access and sharing of the two key types of results arising from work done in the computational sciences: models and algorithms. We propose the development of an integrated cloud-based system underpinning computational science, linking together software and data repositories, toolchains, workflows and outputs, providing a seamless automated infrastructure for the verification and validation of scientific models and in particular, performance benchmarks.

Keywords—*Reproducibility, Benchmarks, Models, Cloud Services, e-Infrastructure, Computational Science, Open Science*

I. INTRODUCTION

Two key types of results arise from work done in the computational sciences: *models* and *algorithms*. Models represent an abstraction of reality, and their behaviour is expected to be reliably reproduced even if different algorithms are used. This validation of a model’s behaviour can be impacted by a number of factors relating to the specific techniques used, but similar approaches are expected to give broadly the same results. In contrast, when new algorithms are proposed to replace or supplement existing algorithms, they are expected to verifiably replicate the results of other algorithms.

However, neither class of result exists in isolation: a new algorithm is dependent on a set of models (or benchmarks) to demonstrate its new capabilities. Equally, model development can both necessitate the development of new algorithms and highlight the differences between alternative approaches. Whilst algorithms and their implementations have been highlighted as a potential barrier to reproducibility [1], in this paper we discuss the value of improved access and sharing of models in reducing mistakes and in generating new scientific insights.

We describe efforts to reproduce computational models and algorithms, specifically the multitude of issues relating to benchmarking of models and algorithms. We conclude with thoughts on where efforts should be focused 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.

The reproduction and replication of reported scientific results is a widely discussed topic within the scientific community [2]–[4]. 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. Nevertheless, this is 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 [5]. Furthermore, just publishing (linked) scientific data is not enough to ensure the required reusability [6].

Specific examples of the benefits of reproducible workflows to researchers are now appearing in the scientific literature. For example, new tools for membrane protein simulation [7], [8] demonstrate how complex workflows can be automated, preventing errors and differences arising from manual execution, whilst making it faster to perform new analyses. More complex tools, such as Copernicus, aim to automate more generic molecular dynamics workflows [9]. Alongside this, recent work in executable biology [10] showed how a new class of models, representing a defined property of biological networks, defeated an existing algorithm for proving stability. As such, it was the broader application of a new algorithm to additional models (or benchmarks) which highlighted an unexplored but important phenomena the algorithm could not address.

In each case, these tools 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. However, despite this advantage, and alongside ongoing – and significant – changes to the traditional models of academic dissemination and publication [11]–[13], there remain cultural and technical barriers to both the sharing and reimplementation of algorithms [1]. These include widely discussed topics, such as choices of programming language and software licence, as

The authors would like to acknowledge the use of the Sirius Cybernetics Corporation’s motto for the title of this paper.

well as unreported technical details of the implementations and ensuring that developers get credit for their work¹. One fundamental barrier to sharing remains – the overhead in time and effort required work with enough information to make models, algorithms and data reusable by a third-party can be significant.

However, even when this is considered, the testing of the benchmarks themselves may be non-trivial. Benchmarks may be tailored to the specific problems being addressed by the algorithm, and may not be expected to cover all edge cases. Implementation details, such as pseudo-random number generation, floating-point rounding behaviour and order-of-execution, may affect benchmark results. Furthermore, for high performance computing applications, repeating benchmark results may not be possible by other groups without identical hardware resources. Here we extend a previous proposal [1] by specifically discussing the problems posed by models, considering the issues surrounding sharing and analysing benchmark sets.

II. THE NATURE OF MODELS

A. Abstraction levels

A model describes reality at some level of abstraction. The more detailed it is, the more ‘faithful’ it often purports to be, but also then the more special-purpose (and so useless to many others). It is an important aspect of the modelling task as to what level (abstract vs. concrete) to model at. But often it is implicit, embodied but not embedded, in the model.

An example of this comes from the treatment of floating-point conversions in qualitative networks in systems biology [14]. Each variable in a network has an algebraic target function which describes how the variable should change at each step. The variables themselves are integers, and the target function may return a float, which must be converted to an integer for the update. This can be done in the target function itself, but if the function returns a float, the specific implementation dictates if this is a rounding, floor or ceiling function. This implicit (to the model) but specific (from the implementation) may change the results of the modelling.

Another case of showing how the implementation of qualitative networks may change the model is the treatment of variable ranges within the model. Whilst the formalism allows the variables in a model to have differing ranges of integers, the mechanism of conversion is not specified in the formalism. As such, this is another area where implementations can dictate the precise behaviour of the model, and thus needs to be explicitly annotated on the model.

A final example is the handling of pseudo-random number generation in Avida [15], 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.

B. Benchmark repositories and annotation

A model, or a test, needs to be annotated. A benchmark is a set of models that have been put together for some explicit purpose. Perhaps the directory structure of the benchmarks indicates this purpose, perhaps assertions in the models indicates this purpose; in short, the benchmarks needs to be curated. If the benchmark is public (allowing anyone to contribute), then the curation is even more necessary to make the models reusable.

Once there is a set of tests, there is the issue of how independent the tests are from each other. The concept of “composability” is a fundamental one in computer science. Say we have two functions $f : A \rightarrow B$ and $g : B \rightarrow C$. Their composition $f;g : A \rightarrow C$ in some category with structure means that, if f has property P and g has property Q , then $f;g$ has property $P \oplus Q$, where \oplus is some combinatorial operator in the domain of discourse. What this abstract characterisation means is that a program can be tested by testing its parts, whole system testing can be done by unit testing.

But this is only the case if the system can be decomposed, and we know that in many important areas such as machine learning and computational science, the models are not often decomposable. We have not been explicit about it – people normally are not, especially for models – but we have been discussing algorithms and models assuming that they are truly in-divisible objects. We have not required them to be composable or decomposable. What can this mean practically? If we have an algorithm A that claims to run on model M with result R , then there is no reason to assume that a slight modification of A will also have result R . Or that A running on a $M * N$, for some operator $*$, has a suitably extended R result.

We thus need to be very careful when algorithms are running on models automatically and asynchronously (or due to events beyond our control), on a global scale, with an effect such as performance results that matter to third-parties. Both algorithms and models will need careful curating. Some good examples of such benchmarks are the UCI Machine Learning Repository², 2009 Netflix Prize benchmarks³, SMT Competition⁴, SV-COMP⁵, Answer Set Programming Competition⁶, and the Termination Problem Database⁷ are on that journey. Such repositories would allow the tests to be taken and easily analysed by any competitor tool. There has been some work towards developing this connected infrastructure: for example, knowledge management systems to preserve and share complete auto-tuning and machine learning setups for optimisation, collecting all related artefacts and their software and hardware dependencies besides just performance data [16].

III. WORKFLOW OF META-MODELS

A. Protocols as scripts

Studying the behaviours of complex models is non-trivial. Whilst concise methods sections of papers may give a rep-

²<http://archive.ics.uci.edu/ml/>

³<http://www.netflixprize.com/>

⁴<http://smtcomp.sourceforge.net/2014/>

⁵<http://sv-comp.sosy-lab.org/2015/>

⁶<https://www.mat.unical.it/aspcomp2014/>

⁷<http://termination-portal.org/wiki/TPDB>

¹e.g. Software Sustainability Institute (<http://www.software.ac.uk>) and the UK Community of Research Software Engineers (<http://www.rse.ac.uk>)

representative minimal working protocol (or workflow), missing details may present barriers to its reproduction. This is exacerbated by the inclusion (or, depending on the case, omission) of manual transformation steps which may subtly change the model. Data format conversions may be non-trivial and performed manually. These may involve ad hoc scripts, which might not be part of any of the explicitly shared codebase. This can be supported by open protocols, stored in electronic lab notebooks during the process of model building. However even in this case, assumed knowledge may prevent simple replication by a third party.

One common approach to tackling complex protocols (or workflows) in computational sciences is to automate the process by scripting the laborious elements, such as in the Taverna Workflow Management System [17] for a range of disciplines from heliophysics [18] to multi-disciplinary design optimisation in engineering [19].

An specific example for simulating molecular dynamics is Sidekick [20]: in its early steps, it builds an initial model of a α -helical peptide, performs an energy minimisation of the peptide *in vacuo*, solvates the peptide, adds counter-ions, and runs a second energy minimisation. This is all done without any user actions, and the subsequent replicates are performed with different random seeds to collect accurate statistics. Even in this ideal case however, variations may arise between replicates. In testing on a hybrid AMD/Intel cluster, one of the authors found that the solvation step added a variable number of water molecules. In a molecular dynamics simulation, this is sufficient to cause two simulations with otherwise identical starting states to diverge over time. As such, the inherent properties of the model and simulation should be taken into account in the overall protocol design, and noted in any attempts to reproduce the behaviour. Building, curating and sharing of scientific workflows can provide consistently reusability, but this can require additional effort on the scientist. Workflows are often most effective where scientific processes need to be repeatable many times, therefore amortising the upfront cost of creating the workflow and its components. This approach may not be appropriate for more exploratory science, where the researcher tends to use a more interactive process with their data and models. A successful example of community-building in this space is the myExperiment project [21], which aims to make it easy to find, use and share scientific workflows and other research objects.

In several disciplines, electronic lab notebooks have become the norm. These tools, combined with open repositories such as FigShare⁸ and ZappyLab⁹, facilitate the sharing of protocols. This may be needed for legal compliance (e.g. drug trials), but has been successfully used in large research consortia, for example the use of Accelrys Notebook (formerly Contur ELN)¹⁰ by the structural genomics consortium in Oxford. Similarly, ZappyLab aims to build a free, standardised protocol repository for the life sciences. Within computational sciences, efforts to mine these repositories could offer the potential to convert manual protocols and work flows into prototype scripts, to aid reproducibility.

B. Performance and scalability

A key question in reproducing research in a computational context is whether performance is a key issue. For models of the physical world, such as computational fluid dynamics and molecular dynamics, it is the resulting physics that is typically important to the end user, rather than how fast it took to solve the computational problem. In algorithms research, performance can be the key research result, and therefore reproducing this is important. Another example is in high performance computing where scalability of code (e.g. Gromacs¹¹, NAMD¹², Desmond¹³). Here the aim is to make simulations run more efficiently over large numbers of cores/nodes. On-demand cloud resources such as Microsoft Azure and Amazon Elastic Compute Cloud offer a potentially attractive (and cost-effective) route to reproducing computational experiments.

An important question is which performance metric to use. Wall clock time is commonly used, but this does not allow for long-term performance reproducibility as any such benchmarking is a snapshot in time. This is true whether the underlying hardware the software is running on is physical or virtual hardware. Some “op count” is a more interesting measure. In many cases, the cost of hardware and system artefacts are important but often overlooked, such as for solvers in Answer Set Programming [22]. Also, other structural properties of the models the algorithms are running on, are more interesting. In the field of system biology, whether an algorithm can prove properties like termination, stability, interesting start conditions, etc, are useful measurements of whether one algorithm is better than another. Recent initiatives such as the Recomputation Manifesto [23], explicitly overlooks performance metrics, instead focusing on ensuring future reproducibility¹⁴, with runtime performance regarded as a secondary issue.

IV. FUTURE OUTLOOK

The whole premise of this paper is that *algorithms* (implementations) and *models* (benchmarks) are inextricably linked and should go together. 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 cloud-based service can make this link explicit.

In the software development world, no one would (should) commit to a project without first running the smoke tests. You could be clever and run the tests via the version control system’s pre-commit hook. That way you would never forget to run the tests. All of this can be done, at scale, on the cloud now. Services such as Jenkins¹⁵, Visual Studio Online¹⁶, etc, schedule the tests to run as soon as you commit. We envisage moving to a world in which benchmarks become available online, in the same vein as open access of publications and

⁸<http://www.figshare.com/>

⁹<http://www.zappylab.com/>

¹⁰<http://accelrys.com/products/eln/contur/>

¹¹<http://www.gromacs.org/>

¹²<http://www.ks.uiuc.edu/Research/namd/>

¹³http://www.deshawresearch.com/resources_desmond.html

¹⁴<http://www.recomputation.org/>

¹⁵<http://jenkins-ci.org/>

¹⁶<http://www.visualstudio.com/en-us/products/what-is-visual-studio-online-vs.aspx>

research data. It seems a small step to hook these Continuous Integration (CI) systems up to the algorithm implementations that are written to run on these benchmarks.

Suppose you have come up with a better algorithm to deal with some of these benchmarks. You write up the paper on the algorithm but, more importantly, you also register the implementation of your algorithm at this open service, as a possible algorithm to run on this benchmark set. The benchmarks live in distributed git (or similar) repositories. Some of the servers that house these repositories are CI servers. Now, when you push a commit to your algorithm, or someone else pushes a commit to theirs, or when someone else adds a new benchmarks, the service's CI system is triggered. It is also activated with the addition of a new library, firmware upgrades, API changes, etc. All registered algorithms are run on all registered models, and the results are published. The CI servers act as an authoritative source, analogous to the Linux Kernel Archives¹⁷, of results for these algorithms running on these benchmarks.

There are already several web services that nearly do all of this things. We feel that a service that can integrate some of these features is possible that would then allow algorithms and models to evolve together, and always be reproducible.

A system as described here has several up-front benefits: it link papers more closely to their outputs, making external validation easier and allows interested users to explore unaddressed sets of models. Critically, it helps researchers to be more productive, rather than 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 hope that we can design and build a system that is similarly usable for sharing and testing benchmarks online.

In summary, this proposed new e-infrastructure could have a profound impact on the way that computational science is performed, repositioning the role of models, algorithms and benchmarks and accelerate the research cycle (perhaps truly enabling a "fourth paradigm" of data intensive scientific discovery [24]). Furthermore, it would effect a cultural change by reducing overheads and improving the efficiency of researchers.

REFERENCES

- [1] T. Crick, B. A. Hall *et al.*, "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, (submitted, under review).
- [2] N. Barnes, "Publish your computer code: it is good enough," *Nature*, vol. 467, no. 753, 2010.
- [3] A. Morin, J. Urban *et al.*, "Shining Light into Black Boxes," *Science*, vol. 336, no. 6078, pp. 159–160, 2012.
- [4] L. N. Joppa, G. McInerney *et al.*, "Troubling Trends in Scientific Software Use," *Science*, vol. 340, no. 6134, pp. 814–815, 2013.
- [5] L. Osherovich, "Hedging against academic risk," *Science-Business eXchange*, vol. 4, no. 15, 2011.
- [6] S. Bechhofer, I. Buchan *et al.*, "Why linked data is not enough for scientists," *Future Generation Computer Systems*, vol. 29, no. 2, pp. 599–611, 2013.
- [7] P. J. Stansfeld and M. S. Sansom, "From coarse grained to atomistic: A serial multiscale approach to membrane protein simulations," *Journal of Chemical Theory and Computation*, vol. 7, no. 4, pp. 1157–1166, 2011.
- [8] B. A. Hall, E. Jackson *et al.*, "Logic programming to predict cell fate patterns and retrodict genotypes in organogenesis," *Journal of The Royal Society Interface*, vol. 11, no. 98, 2014.
- [9] S. Pronk, P. Larsson *et al.*, "Copernicus: a new paradigm for parallel adaptive molecular dynamics," in *Proceedings of 2011 International Conference for High Performance Computing, Networking, Storage and Analysis (SC'11)*. ACM Press, 2011.
- [10] B. Cook, J. Fisher *et al.*, "Finding instability in biological models," in *Proceedings of the 26th International Conference on Computer Aided Verification (CAV 2014)*, 2014.
- [11] D. De Roure, "Replacing the Paper: The Twelve Rs of the e-Research Record," <http://www.scilogs.com/eresearch/replacing-the-paper-the-twelve-rs-of-the-e-research-record/>, November 2011.
- [12] V. Stodden, P. Guo *et al.*, "Toward Reproducible Computational Research: An Empirical Analysis of Data and Code Policy Adoption by Journals," *PLoS ONE*, vol. 8, no. 6, 2013.
- [13] G. Fursin and C. 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)*. ACM Press, 2014, pp. 1–4.
- [14] M. A. Schaub, T. A. Henzinger *et al.*, "Qualitative networks: a symbolic approach to analyze biological signaling networks," *BMC Systems Biology*, vol. 1, no. 4, 2007.
- [15] C. Ofria and C. O. Wilke, "Avida: A Software Platform for Research in Computational Evolutionary Biology," *Artificial Life*, vol. 10, no. 2, pp. 191–229, 2004.
- [16] G. Fursin, R. Miceli *et al.*, "Collective mind: Towards practical and collaborative auto-tuning," *Scientific Programming*, vol. 22, no. 4, pp. 309–329, 2014.
- [17] K. Wolstencroft, R. Haines *et al.*, "The Taverna workflow suite: designing and executing workflows of Web Services on the desktop, web or in the cloud," *Nucleic Acids Research*, vol. 41, no. W1, 2013.
- [18] A. L. Blanc, J. Brooke *et al.*, "Workflows for Heliophysics," *Journal of Grid Computing*, vol. 11, no. 3, pp. 481–503, 2013.
- [19] T. Crick, P. Dunning *et al.*, "Engineering Design Optimization using Services and Workflows," *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 367, no. 1898, pp. 2741–2751, 2009.
- [20] B. A. Hall, K. B. A. Halim *et al.*, "Sidekick for membrane simulations: Automated ensemble molecular dynamics simulations of transmembrane helices," *Journal of Chemical Theory and Computation*, vol. 10, no. 5, pp. 2165–2175, 2014.
- [21] D. De Roure, C. Goble *et al.*, "The Design and Realisation of the myExperiment Virtual Research Environment for Social Sharing of Workflows," *Future Generation Computer Systems*, vol. 25, no. 5, pp. 561–567, 2009.
- [22] M. Brain and M. De Vos, "The Significance of Memory Costs in Answer Set Solver Implementation," *Journal of Logic and Computation*, vol. 19, no. 4, pp. 615–641, 2009.
- [23] I. P. Gent, "The Recomputation Manifesto," April 2013, available from: <http://arxiv.org/abs/1304.3674>.
- [24] T. Hey, S. Tansley *et al.*, Eds., *The Fourth Paradigm: Data-Intensive Scientific Discovery*. Microsoft Research, 2009.

¹⁷<https://www.kernel.org/>