https://doi.org/10.1038/s41563-023-01699-7

Open computational materials science

The materials modelling community is emerging as a champion for reproducible and reusable science. Aron Walsh discusses how FAIR databases, collaborative codes and transparent workflows are advancing this movement.

Reproducibility should be a solved problem in computational materials science. In a typical research workflow, a set of input files is transformed into a set of output files using a well-defined algorithm or model. If complete inputs are provided, and the associated programs are available, then the results are repeatable, reproducible and reusable by others. Of course, there are situations where this deterministic chain is broken. Humans make mistakes, codes contain bugs or are updated, and computer resources are insufficient; hardships that computational scientists learn to endure and overcome.

There has been a positive transition towards open science and data protocols in the materials modelling community. Calculations performed using independent implementations of density functional theory show convergence towards a single value[1]. Community benchmarks have been established for supervised machine-learning models of materials properties and stability[2], while checklists have been introduced for the reporting of data-driven statistical models[3]. At the same time, obstacles remain. The field continues to evolve with more complex algorithms and larger datasets that are less straightforward to repeat and report. One recent application of large-language models to crystal structure generation, CrystalLLM, trained a model with 25 million parameters using more than 2 million crystallographic information files. Such a feat is made possible by the open materials data infrastructure that has been established.

To support reproducible research, we must go "beyond the traditional 'supporting information' files to include input files and output files of computations as well as source code"[4]. There has been progress in this direction through structure and property databases that satisfy the findable, accessible, interoperable and reusable (FAIR)[5] principles. Services such as the Novel Materials Discovery (NOMAD) Laboratory[6] and Materials Cloud[7] have been built as community platforms to store, process and interact with materials data and models. They are free and open source, and easy to use, and a digital object identifier (DOI) can be generated for each dataset uploaded. Input and output files from many codes are converted into searchable formats that are interoperable through a common application programming interface. For more general-purpose less-structured reports, Zenodo is powered by the CERN data centre and provides a service that has benefited thousands of researchers and projects from around the world. These solutions are complemented by a growing number of curated databases and datasets for specific properties, and application areas. The pioneering Materials Project, Open Quantum Materials Database and AFLOW databases cover many known and hypothetical crystals. The Open Catalyst Project provides systematic training data for machine-learning models of catalysis, while OpenKIM is a curated repository for interatomic potentials, and MatNavi combines experimental and simulated datasets for several classes of compounds.

As the transition to open-access publishing is widening access to academic reports, the growth of open-source tools is lowering the barrier to entry for computational science. For atomic-scale modelling, the impact of toolkits such as the Atomic Simulation Environment[8] and Pymatgen[9] has been revolutionary. In the past, many scientific codes and (often cryptic, uncommented) scripts for simulation set-up and data analysis were locked within research groups, making publications difficult to reproduce or adapt, and allowing bugs to persist over long periods of time. There is no longer the need for researchers to reinvent the wheel for basic tasks, and workflows can be chronicled efficiently. Data provenance can be tracked in automated workflow systems such as AiiDA[10]. Robust version control systems, such as git, allow code changes to be monitored and tested. Effective user documentation can be populated directly from code with appropriate function annotations (docstrings), supporting the programmer's mantra "explicit is better than implicit".

Defect information file	
Compound:	LK-99
Defect species:	V_{Pb}
Charge states:	[0, −1, −2]
Supercell:	[3 3 2]
Chemical potential:	O-rich, 300 K
Correction scheme:	Kumagai–Oba

Fig. 1 | Illustration of a hypothetical 'defect information file', which is human and computer readable, while summarizing information collated from a defect calculation workflow. The stages include defect specification, calculation set-up and post-processing of the thermodynamic potentials. Inspired by the doped package.

Studies in computational materials science increasingly involve multiple calculations, covering a range of compositions, structures and properties. In such cases, the workflow becomes as important as the individual calculations, with multiple pre- and post-processing stages that can involve complex software and hardware dependencies. Versions and configurations affect reproducibility, as defaults and core functionality change, which can be addressed with an appropriate requirements or configuration file. Even something as seemingly straightforward as the calculation of a point defect in a crystal can require the combination of dozens of files and several codes. While the information collated over each processing step is perhaps deserving of its own standard information file for reporting (Fig. 1), a community consensus is required for successful adoption in each use case. At a minimum, a scripted workflow or annotated electronic notebook can provide the essential information.

Reproducibility is not simply a checkbox at the time of publication, but an ongoing

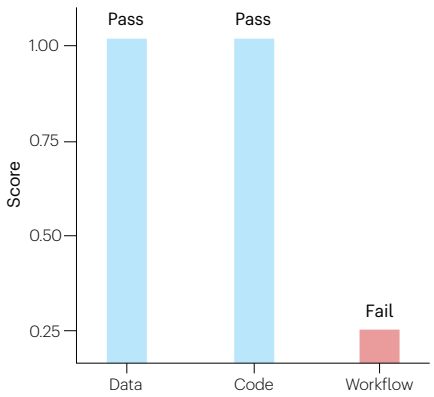

Fig. 2 | Illustration of a reproducibility review as part of the publication process. Factors may include data availability (inputs and outputs), code availability (versioned) and workflow transparency (pre- and post-processing).

process within research groups and institutes. That said, there are some primary conditions to meet the burden of reproducibility when a study is formally reported (Fig. 2). We all want to maximize the impact of our results and ensure that our chosen methodologies are robust. Beyond reproducibility, opening a project repository to the community allows others to assist with maintenance and feature development. Of course, there may be the fear of giving others access to your tailored methods, but open science motivates us to continuously raise the bar and evolve scientifically.

It must be appreciated that open research takes effort. The extra time needed to process data, descriptions in the form of metadata, annotations and uploads to repositories is far from negligible. This additional work must be appreciated in formal assessments of researchers, from PhD examinations to interview panels and promotion boards. In my experience, this is already happening, but should be formalized in assessment and selection criteria. It should also feature in the training process for the next generation of academic and industrial researchers. Open science and data concepts can be introduced in the undergraduate curriculum, emphasized in postgraduate courses and advanced by postdoctoral researchers.

There are many flavours of computational materials research that require bespoke considerations and solutions for the validation and verification of results. What we must continue to avoid are opaque reports that lead to the persistence of erroneous conclusions in the literature and wasted resources trying to duplicate them. Scientific publishers have been introducing clearer guidance and more stringent checks to support reproducibility in publications, while many national funders are also tightening their open data requirements for projects that they support. Replicable reports ultimately increase the reliability and impact of computational predictions to design new compounds, optimize properties and understand material behaviour. We are at a time when the role of computational research in materials science is of growing importance. As we run bigger and better simulations, let us also make them reproducible and repurposable.

Aron Walsh [1,2] ✉

[1]Department of Materials, Imperial College London, London, UK. [2]Department of Physics, Ewha Womans University, Seoul, South Korea. ✉e-mail: a.walsh@imperial.ac.uk

Published online: 3 January 2024

References

1. Lejaeghere, K. et al. *Science* **351**, aad3000 (2016).
2. Dunn, A., Wang, Q., Ganose, A., Dopp, D. & Jain, A. *npj Comput. Mater.* **6**, 138 (2020).
3. Artrith, N. et al. *Nat. Chem.* **13**, 505–508 (2021).
4. Coudert, F.-X. *Chem. Mater.* **29**, 2615–2617 (2017).
5. Wilkinson, M. D. et al. *Sci. Data* **3**, 160018 (2016).
6. Draxl, C. & Scheffler, M. *J. Phys. Mater.* **2**, 036001 (2019).
7. Talirz, L. et al. *Sci. Data* **7**, 299 (2020).
8. Hjorth Larsen, A. et al. *J. Phys. Condens. Matter* **29**, 273002 (2017).
9. Ong, S. P. et al. *Comput. Mater. Sci.* **68**, 314–319 (2013).
10. Huber, S. P. et al. *Sci. Data* **7**, 300 (2020).

Competing interests

The author declares no competing interests.