



Editorial

Data and code availability requirements in open science and consequences for different research environments



Artificial Intelligence in the Life Sciences (AILSCI) strongly supports code and data sharing as a central part of open science and requires reproducibility of published work. Therefore, for publication of *Original Research* in *AILSCI* including *Articles* and *Short Communications*, data and custom code must be made accessible for review and, upon acceptance of a study, provided in an open access deposition (or as machine-readable supplementary information, if the size is manageable). The custom code deposition does not necessarily need to include any script that is used in a study, but must include code that is essential for reproducing the main findings. The use of proprietary data for modeling is generally discouraged and the data deposition requirements apply if proprietary materials are used.

For models generated with open source software, sufficient details must be provided in a paper such that the models can be re-built and the calculations reproduced. Required details include parameter optimization protocols and all optimized parameter settings, which may be provided as supplementary information. Models generated with open source software can also be made available, but this is not a stringent requirement as long as sufficient details are reported.

Studies from commercial entities frequently use proprietary data and/or code, which generally conflicts with *AILSCI's* deposition requirements. However, it is also evident that it is often impossible for researchers from companies to carry out or repeat their work using relevant public data and/or code, if available at all, to meet deposition requirements for publication. Therefore, while ensuring full reproducibility of *Original Research* is a must (with a possible exception; *vide infra*) there are different avenues to publication for investigators from industry to enable the communication of interesting science and new findings from their internal research

- (1) *AILSCI* offers different article types as potential alternatives to publication of *Original Research*. Specifically, the *Perspective*, *Conceptual Analysis*, *Viewpoint*, and *Methods & Protocols* article types provide opportunities for communication of new scientific developments, concepts, or case studies without the need for full data disclosure. Hence, a research study might be presented in different ways, for example, more generally describing new approaches and applications such that reproducibility of specific results would not be essential. This might be facilitated in a *Perspective* or *Conceptual Analysis* paper. Alternatively, in a *Methods & Protocols* article, a new computational approach or predictive model might be introduced and only a brief exemplary application presented, for which only a small amount of data would need to be made available, together with the model.

- (2) Reporting new software and its applications generally requires deposition of the code, and this also applies to commercial software. However, a *Methods & Protocols* article reporting a new algorithm or methodology might contain detailed pseudo-code to enable re-coding by others, instead of granting access to the software. Of course, this would not be feasible for large software packages. However, when reporting a new approach with custom code in a *Methods & Protocols* article, reproducibility does not generally imply that an implementation must be provided. By contrast, if a new proprietary/commercial software would be benchmarked against publicly available programs, claiming superior performance, *AILSCI* would not publish the study without making the new software available to the public. In this case, the specific results must be fully reproducible. Studies reporting software with restricted access will be considered by the editors on a case-by-case basis. This also includes the use of commercial software by industrial or academic investigators if the main findings of a study depend on the use of programs with restricted access. Importantly, predictions must also be reproducible if they are experimentally confirmed.
- (3) As an exception for *Original Research*, *AILSCI* is introducing a possible editorial waiver on deposition requirements for proprietary data. This waiver option can be considered only if a study reports novel findings that further advance the field and the authors provide convincing evidence that comparable results could not be obtained on the basis of current public data. In such exceptional cases, the scientific significance of new findings would outweigh data deposition requirements. For example, this might apply to the analysis of or predictive modeling based on in-house data sets from drug discovery that have evolved over time such as pharmacological or compound profiling data. Such data are typically sparse in the public domain and often unique to a pharma company. Knowledge extraction from such data might yield unprecedented findings of general relevance. However, in such cases, a stringent requirement is the demonstrated novelty of the findings and conclusions drawn from them. For example, building standard prediction models based on proprietary data and reporting high accuracy for internal test sets would not be sufficient. Authors who would like to consider the data deposition waiver option may contact the editor-in-chief to apply for a waiver with supporting evidence. Such requests will be considered on a case-by-case basis by the editors and are subject to editorial discretion. If a waiver is granted and the manuscript accepted after review, the published version will need to contain

an acknowledgement that the editors have issued a waiver on the data deposition requirements for this study in light of the significance of the reported findings.

Finally, it is noted that *Methods & Protocols* articles are also suitable for reporting practical computational/AI solutions for relevant tasks. For example, these might include calculation workflows, scripts, and notebooks or adaptations and extension of available code for specific applications.

Declaration of Competing Interest

The author declares that there are no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

No data was used for this article.

Jürgen Bajorath*

Department of Life Science Informatics, B-IT, Lamarr Institute for Machine Learning and Artificial Intelligence, LIMES Unit Chemical Biology & Medicinal Chemistry, University of Bonn, Friedrich-Hirzebruch-Allee 5/6, D-53115 Bonn, Germany

* Corresponding author.

E-mail address: bajorath@bit.uni-bonn.de.