Keeping checks on machine learning

Community-driven initiatives are proposing standards to improve the reporting and reproducibility of machine learning in biology. We support these developments, some of which are described in this month's special issue.

The Methods section is an essential part of most primary research papers in the life sciences. Details of experimental and analytic procedures facilitate understanding and examination by readers and allow results to be replicated. Many journals, including *Nature Methods*, also ask authors to fill in checklists that cover key methodological issues, such as the steps used in data processing and statistical analysis.

In contrast to these relatively straightforward procedures, biologists are now also enjoying the age of big data and deep learning. Advanced machine learning methods have gained traction as a result of their impressive performance when tackling complex tasks. Their power stems from highly complex model architecture, vast numbers of parameters and massive training data. However, these factors also lead to severe barriers to ensuring reproducibility of the results and reusability of the methods, and make them daunting to describe in a typical Methods section. Overlooking this issue can have various negative ramifications, ranging from overconfidence in the reported performances to failure in new applications. In this sense, the reproducibility crisis that concerns experimental sciences is expanding its shadow to computational biology.

In response to this challenge, communities are teaming up to develop solutions. In this month's special issue, two independent groups of researchers have proposed recommendations and guidelines for reporting machine learning approaches used in the life sciences. The DOME recommendations devised by Walsh et al. target four major aspects when reporting supervised machine learning applications: data, optimization, model and evaluation. A machine learning reporting checklist consisting of questions to the author is designed to assist in writing and reviewing machine learning-related papers. Matschinske et al. present the AIMe standards and registry, which generates electronic reports describing biomedical artificial intelligence methodology and makes them available via a database. By querying and examining a report associated with a machine learning-related paper, readers and reviewers can evaluate its content and quality.

Despite differences in their specific requirements and implementations, the two proposals share several similarities. First, both are community-based efforts that involve researchers from multiple countries and institutes. This is vital for establishing standards and consensus for such a broad topic, which needs substantial efforts for brain-storming, discussion and refinement. Second, both standards are open to community feedback and further improvement. Governance mechanisms have also been developed to support their future revision. Third, both DOME and AIMe are formulated as minimal standards that include only the most essential requirements for sufficient and clear machine learning reporting in the life sciences. This makes them broadly applicable, whereas it is likely that more specific reporting standards will be more effective for specific types of machine learning algorithm and application areas.

Nicely exemplifying this last point is a Comment by Laine et al. on using deep learning in bioimage analysis. After elaborating on strategies to validate deep learning predictions and choose appropriate tools, the authors discuss essentials for reporting deep learning methodology that are specific to bioimaging. As this example shows, endeavors to propose and promote research area-specific guidelines for machine learning reporting will be valuable for meeting the unique needs of different communities.

Reporting alone, however, cannot address all the problems in reproducibility and reusability. Material and data availability has become another norm in scientific publication. Following a similar vein, Heil et al. propose a three-level ranking system to assess reproducibility of machine learning in biology. The bronze standard mandates availability of data, models and code, the gold standard requires the highest level of reproducibility, including full automation of the analysis, and the silver standard stands between them. Compared to experimental materials such as plasmids and cell lines, data, models and code should be easier to widely share and utilize, thus facilitating efforts to improve reproducibility and reusability in computational biology.

But still, it is not an easy job to meet these standards or achieve the gold standard, especially considering the increasing complexity of multi-step analysis now commonly performed in computational biology studies. As suggested by Heil and colleagues, a neat way to handle these multi-step analyses is to use software tools called workflow managers. Also in this issue, a Perspective by Wratten et al. provides a comprehensive overview of major workflow managers and their features and utilities. Beyond their role of improving reproducibility and reusability, workflow managers are also useful in developing and optimizing complex computational pipelines.

Even with these standards and tools at hand, it's still possible that current strategies for strengthening reproducibility and reusability might be outpaced by the fast advancement of machine learning technologies and our aspirations to understand and model live systems. As discussed in a Perspective by AlQuraishi and Sorger, advanced deep learning methods are now enabling the rise of differentiable biology, which uses differentiable models to integrate multi-modal data across multiple scales. Finally, Avsec et al. describe the development of a novel deep learning method based on transformer, a highly successful architecture in natural language processing, to achieve substantially improved DNA sequence-based prediction of gene expression. These and more upcoming advanced applications of machine learning in biology will drive the continuous evolution of our thinking and tools for improvement in reproducibility and reusability.

Biologists are not alone when tackling these challenges, with recent calls from chemistry (N. Artrith et al., *Nat. Chem.* **13**, 505–508; 2021) and medicine (*Nat. Med.* **26**, 1320–1324; 2020). Although substantial work will be needed, we believe this is necessary to maximize the potential of machine learning in biology. As a journal devoted to publishing new methods, we understand the importance of making computational results and methods reproducible and reusable (*Nat. Methods* **18**, 695; 2021). With the communities we serve, we are keen to explore and improve our journal standards in this area. ❐

Published online: 4 October 2021
https://doi.org/10.1038/s41592-021-01300-6