Moving towards reproducible machine learning

We provide some recommendations on how to report machine learning-based research in order to improve transparency and reproducibility.

The increasing availability of data in various fields has made data-driven research an important asset in computational science. Extracting valuable insights from data, however, can be a daunting task: datasets have become larger, and as data collection tools and operations become more sophisticated, data analysis has also become more complex. Machine learning algorithms have attracted a great deal of attention from the research community for dealing with these challenges; perhaps most notably, deep learning has become a standard tool for analyzing large, complex datasets, at times achieving impressive levels of accuracy that advance science to a great extent, such as what we have seen in battery research[1], structural biology[2,3], and chemistry[4].

But with great power comes great responsibility. The field of machine learning has faced a "reproducibility crisis"[5] due to the lack of transparency and reporting surrounding the steps taken to build data-driven models. From data collection and curation to model selection and training, all of these steps are essential to better understand how accurate, robust, general, and practical the reported models are, beyond the usual accuracy numbers often reported in papers. As machine learning becomes more popular and widely used in many different fields, it becomes crucial to make sure that researchers are reporting all of these details to allow proper reproducibility of the results.

Fortunately, the research community has paid attention to this issue, and different communities have made several recommendations on how to best report machine learning research in order to improve transparency and reproducibility[6–9]. Based on these recommendations and our own experience, we would like to use this Editorial to discuss some guidelines and suggestions that can be useful to the entire computational science research community when reporting their machine learning-based results.

Data reporting

An important step when constructing a model is the collection and selection of the datasets, as the quality of the model greatly depends on the quality and characteristics of the data. The data collection process needs to be properly discussed and reported, as there can be biases (intentional and/or unintentional) with regards to the selected data sources. Any identified biases and attempts to mitigate them should also be properly discussed, so that other researchers can be aware of the limitations when using the reported models. If synthetic data is used, the data generation process, including any assumptions that are considered, needs to be described in detail.

Raw datasets are in fact rarely used, since they may have several inconsistencies, errors, and outliers that can ultimately impact the quality of the model. In addition, data might need to be converted to a specific format and representation in order to be used for a specific model. Therefore, the data cleaning and data curation steps are critical to the research, and as such, these steps must also be reported in detail.

Finally, there are three specific datasets that are essential for model development: training, validation, and test datasets. The training dataset, as the name suggests, is used to train and generate the model; the model 'learns' from this dataset. The validation dataset is used to evaluate the performance of the model for different hyperparameter values, and to detect overfitting. Finally, the test dataset is used to assess the performance of the model. It is very important that the selection of these different sets is properly explained, as these can substantially impact the performance and robustness of the model.

Model reporting

There is a large amount of machine learning models from which researchers can choose. Higher model complexity can come with the cost of reduced transparency and interpretability[6], and may not always be the best choice; in addition, training times can vary substantially depending on the model. Therefore, the choice of the model and its level of complexity needs to be properly justified. For research using deep learning, it is a good idea to run and report on ablation studies to better understand the neural network architecture and whether or not some components can be removed with no loss of performance.

While the machine learning community has effectively harnessed the power of new computing architectures, such as supercomputers and graphics processing units, training a model can still be very time-consuming, especially depending on its level of complexity. Additionally, not every researcher may have access to more sophisticated hardware resources. Reporting the time that is taken for training is thus essential for informing readers on how practical this step could be within the context of their own available resources.

Another important consideration is that machine learning models can have different sources of randomness, such as random initializations, dropout, and data shuffling, to name a few. If possible, seeding the pseudorandom number generators used in the models, and reporting these choices, is a good idea for ensuring consistent results.

Availability of data, code, and models

It goes without saying that it is crucial for the code and data to be made publicly available to the community; not only the code for training, validating, and testing the model should be made available, but also for the data collection, cleaning, and curation steps. Differences in hardware architectures and software library versions may also lead to many inconsistencies, and thus, it is essential to properly report these details. Trained models should also be made available[7], since, as mentioned before, the training step may require a substantial amount of resources: having the trained models available lowers the barrier for other researchers to be able to reuse these models in their own research, and it also makes it easier to examine whether or not the models can generalize to other data.

Final remarks

It is worth noting that this is not meant to be a comprehensive list of guidelines for machine learning-based research, and there are certainly other issues not discussed here, such as privacy-related challenges[7] and ethical considerations[10], that are also very important to take into account. Instead, our goal is to start a conversation with the broader community of computational scientists about this topic and hopefully improve the overall

reporting of research results. We have already seen great initiatives from different groups[6–9], and we look forward to seeing more engagement from our research community in order to make machine learning more transparent and reproducible. ❐

Published online: 12 October 2021
https://doi.org/10.1038/s43588-021-00152-6

References

1. Severson, K. A. et al. *Nat. Energy* **4**, 383–391 (2019).
2. Jumper, J. et al. *Nature* **596**, 583–589 (2021).
3. Baek, M. et al. *Science* **373**, 871–876 (2021).
4. Hermann, J., Schätzle, Z. & Noé, F. *Nat. Chem.* **12**, 891–897 (2020).
5. Hutson, M. *Science* **359**, 725–726 (2018).
6. Artrith, N. et al. *Nat. Chem.* **13**, 505–508 (2021).
7. Heil, B. J. et al. *Nat. Methods* **18**, 1132–1135 (2021).
8. Mateen, B. A. et al. *Nat. Mach. Intell.* **2**, 554–556 (2020).
9. Norgeot, B. et al. *Nat. Med.* **26**, 1320–1324 (2020).
10. *Nat. Mach. Intell.* **3**, 367 (2021).