Setting the standards for machine learning in biology

David T. Jones [1,2]

Machine learning is a branch of artificial intelligence (AI) involving computer programs that are able to improve their own performance through experience (training). The diverse applications of new 'deep learning' approaches with neural networks are now expanding into the field of biology. But these applications to biological data require more scrutiny and caution to increase the standards of publishing and allow the AI revolution in biology to take off.

Artificial intelligence revolution

In the area of machine learning, deep (multi-layered) neural networks — encompassing algorithms that are loosely modelled on the connectivity of a human brain assembled into multiple (several to hundreds) functional layers — have recently gained particular popularity[1].

Almost every field of scientific research is seeing a flood of papers describing how deep learning can be applied to almost any problem for which there is sufficient available data. The reasons for this are manifold. First, the technology itself has become trivial to use by any reasonably competent programmer, owing to the availability of software that has made it possible for anyone to carry out deep learning experiments, which would have been difficult even for experienced computer scientists just a few years ago. Hardware advances such as the use of cheap graphics processors to massively accelerate machine learning have also played an important part. Now, it is not even necessary to buy the hardware, as it can be used, sometimes free of charge, via a number of cloud services. Finally, many more training opportunities have become available, leading to a massive expansion of the artificial intelligence (AI) workforce.

Machine learning methods, and neural networks in particular, have been applied to biology for several decades[2], but recently the field has seen a true explosion of interest[3]. The attraction of tackling problems for which solutions could benefit the health and well-being of millions of people is hard to resist. However, the traditional academic publishing processes in biomedical sciences are still not well prepared to handle this mass popularization of computer-led biological data science. The problem we face is that many of the papers that we are now seeing as a result of the AI revolution are not advancing the field, because these techniques are not appropriately used: either they do not offer any improvement in comparison with existing methods or their experimental design is flawed; often both.

Problem suitability and advance

The first thing to realize is that deep learning approaches have been successful in only a fairly narrow range of applications where certain criteria of data can be met: the data are available in large quantities; they are characterized by high dimensionality (meaning that each sample comprises very many variables); and they are highly structured (meaning that there is some kind of graphical relationship between the variables). Images are a perfect sample for deep learning approaches — they provide many thousands of variables (pixels) that can be clearly grouped into well-defined objects (for example, the pixels that represent a nose on a face). Text or audio data are also suitable.

Of course, in biology there are many data sets that also fit these requirements, such as sequence data (text) or microscopy data (images). There are also other, perhaps less obvious applications. For example, deep learning has recently greatly advanced our ability to predict protein tertiary structure from amino acid sequence[4]. This has been possible because protein folds can be considered as 2D maps of inter-atomic distances, which can be analysed in a similar way to images. However, not every biological data set is suitable for analysis with deep learning. One example is the analysis of single nucleotide polymorphisms (SNPs) in genomic data. Here, the presence or absence of known SNPs in a genome is certainly of high dimensionality, with millions of known SNPs, but the data are unstructured. This type of data is called categorical data, which can only be expressed in the form of tables with no specific row order. SNP data can still be grouped by annotating them to a specific gene or chromosome, but this is insufficient to make them suitable for deep learning, and other methods of analysis are likely to perform better.

As a result of this limitation in terms of data suitability, evaluation of a paper should never start with the assumption that just because it uses deep learning, it must be state-of-the-art. Instead, it should start with the opposite belief and the reviewers should allow the results

[1]*Department of Computer Science, University College London, London, UK.*

[2]*The Francis Crick Institute, London, UK.*

e-mail: d.t.jones@ucl.ac.uk

https://doi.org/10.1038/s41580-019-0176-5

of the paper to convince them otherwise. Accordingly, authors should demonstrate clearly that their data meet the suitability criteria mentioned above. They should also emphasize the scientific advance provided by the deep learning approach by properly comparing it against existing approaches.

Data leakage and bias

A perplexing issue associated with early applications of AI approaches to biological data, which unfortunately is still pervasive in current publications, is lack of reproducibility in terms of performance: when the algorithms are re-tested on new data, their performance is often worse than that claimed in the original publications. Why should this be?

When establishing a new deep learning method, it is standard practice to use cross-validation to estimate the performance of predictive models on unseen data — available data are divided into two sets: the 'training set' used to train the model, and the 'test set', used to evaluate the final model. A common variant of this would be '*n*-fold' cross-validation where *n* different splits of the data are made, with the final results being the average of the *n* training test runs.

Ideally, the test set should be used once only. However, it is now generally accepted that the test set can be used subsequently on different algorithms, thereby allowing selection of the method that offers best performance; this typically selects the best model for unseen data.

The core issue lies with data selection between the training set and the test set, which is necessary to avoid bias that would otherwise cause overfitting of the model to the training data (meaning that the program becomes unreliable when exposed to previously unseen data). The situation when this bias has not been prevented is commonly referred to as data leakage. Generally, it is assumed that random sampling is the best way to avoid selection bias. Nevertheless, this assumption does not generally hold true in biology, because biological data samples will frequently not be independent, and so random sampling is invalid. These dependencies often result either from phylogenetic effects (that is, arising from evolutionary relationships between genes or proteins) or population effects, which arise from present day experimental sampling biases. Both of these can be difficult to assess and manage[5–7]. For example, in protein structure prediction it is common practice to deal with evolutionary bias by simply checking if two proteins share more than 25% sequence identity and consider proteins below this threshold as evolutionarily unrelated. Whilst this approach is well-meaning, it is inherently flawed, because there are many cases of related proteins with zero sequence identity. Population biases frequently arise in clinical trials. For example, patient samples might be taken from a whole family, with some ending up in the training set and others in the test set. Beyond evolutionary and population biases, any underlying bias, resulting from, amongst others, the unequal distributions of species (for example, prokaryotes versus eukaryotes), experimental technique or even instrument can interfere with a proper cross-validation process.

Improving publishing standards

To work towards improving standards of publishing in this area, authors and reviewers of biological papers employing approaches using deep learning must take the above issues seriously. From a reviewer's perspective, it can sometimes be hard or even impossible to tease apart the technical details and establish whether, and if so, how exactly, data leakage might have occurred. To do so requires not just machine learning expertise, but also a good understanding of the experiments that generated the data, and hence a thorough understanding of the biology. Researchers with expertise in both areas are going to be rare, and so measures must be put in place to assist reviewers in properly assessing these types of manuscript. A good starting point may be to employ on-submission checklists that request reporting on the design of the computational experiments from the authors (see REF.[5] for example). This would point towards potential flaws in experimental design and would help to ensure that pertinent questions are asked of authors during the review process.

1. Min, S., Lee, B. & Yoon, S. Deep learning in bioinformatics. *Brief. Bioinf.* **18**, 851–869 (2017).
2. Stormo, G. D., Schneider, T. D., Gold, L. & Ehrenfeuch, A. Use of the perceptron algorithm to distinguish translation initiation sites in *E. coli*. *Nucleic Acids Res.* **10**, 2997–3011 (1982).
3. Ching, T., Himmelstein, D. S., Beaulieu-Jones, B. K. & Greene, C. S. Opportunities and obstacles for deep learning in biology and medicine. *J. R. Soc. Interface* **15**, 20170387 (2017).
4. AlQuraishi, M. AlphaFold at CASP13. *Bioinformatics*, btz422 (2019).
5. Walsh, I., Pollastri, G. & Tosatto, S. C. E. Correct machine learning on protein sequences: a peer-reviewing perspective. *Brief. Bioinf.* **17**, 831–840 (2016).
6. Chicco, D. Ten quick tips for machine learning in computational biology. *BioData Min.* **10**, 35 (2017).
7. Tabe-Bordbar, S., Emad, A., Zhao, S. D. & Sinha, S. A closer look at cross-validation for assessing the accuracy of gene regulatory networks and models. *Sci. Rep.* **8**, 6620 (2018).

Competing interests
The author declares no competing interests.