"For chemists, the AI revolution has yet to happen".

Henry Rzepa

Published May 25, 2023

Citation

Rzepa, H. (2023). "For chemists, the AI revolution has yet to happen". *Henry Rzepa's Blog*. https://doi.org/10.59350/rsd95-j8640

This editorial from Nature[cite]10.1038/d41586-023-01612-x[/cite] is a timely reminder of the importance of data. But also, not just any data, but "*accurate and accessible training data*". Accessible of course is one of the attributes of FAIR (Findable, Accessible, Interoperable and Re-usable). The editorial also states "*data need to be recorded in agreed and consistent formats, which they are not at present*". That is covered by the I and R of FAIR, often applied in conjunction with metadata recording the Media type that the data is held in (See DOI https://doi.org/jvk9 for examples of the use of Media types in chemical computation and chemical NMR). Again, "*The best possible training sets would also include data on negative outcomes*". This relates to the separation of the two publication processes, namely the article itself (or the story behind the data) and the data itself as a first class scientific object. Thus when we publish FAIR data in association with articles, the data archive will often contain data that is not used in the article itself (perhaps because it led to a negative outcome), but is nevertheless part of the FAIR data collection for that topic. Even if the data does not lead to journal publication, publishing it in a data repository means it will not be lost. Somebody (or AI software) may still find it useful.

Whilst the acronym AI is increasingly used and hyped up, I would argue that FAIR should accompany the use of the term AI in most cases (as indeed it is at eg.[cite]10.1002/mrc.5186[/cite]). Amongst other benefits, FAIR implies a metadata descriptor record is present, which if richly populated, would help address the "*accurate*" of "a*ccurate and accessible*" by adding context. As we show here[cite]10.1002/mrc.5186[/cite], FAIR is also "*AI-Ready*". Indeed an often used alternative expansion of the acronym is "FAIR is AI-Ready". It is indeed designed to be so if the metadata is sufficiently rich. I also remind that an IUPAC working party is working to produce recommendations to help with this aspect.[cite]10.1515/pac-2021-2009[/cite]

My final comment adds to the requirement of "*accurate and accessible training data*". I would reformulate this as "*accurate, accessible and* ***complete*** *training data*". Much data in chemical science is recorded on an instrument, or computed using modelling software. As it emerges from the instrument or the software package, it can be said to be "complete". Nothing has been thrown away at this stage. But think of eg NMR data. This is acquired as a FID, and then subjected to analysis (A Fourier Transform, after weighting, which does introduce potential artefacts into the data!). It is the latter data type that is invariably published, often in a visual (PDF) form which may lack numerical accuracy and which is machine processable only with difficulty. Or think of crystallography, where data emerges as diffraction images and is then transformed into structure factors and coordinates. Only the last form is often published (as a CIF file), but the original data is almost never so (see[cite]10.1021/acsomega.7b00482[/cite] for an example where complete crystallographic data is published). Then again, chemical computations. The full record of the computation is often produced as a "checkpoint" or "interoperability format" (see eg DOI: 10.14469/hpc/10043) which contains the computed wavefunction and which can be re-used to compute a wide variety of new properties. But most articles currently record computational data simply as a set of atom coordinates. If you are really lucky, you might get some keywords used to run the calculation. But nothing which would eg allow an AI-algorithm to easily compute a property it might need. We cannot be sure that a machine learning/AI procedure might not benefit from such complete data.

So, FAIR and AI are conjoined, they each need the other and should not be separated. And to repeat, where data is transformed before being published, please also add the **complete** dataset, not just any reduced form.

Post DOI: 10.14469/hpc/12586
