nature

For chemists, the AI revolution has yet to happen

Machine-learning systems in chemistry need accurate and accessible training data. Until they get it, they won't achieve their potential.

Many people are expressing fears that artificial intelligence (AI) has gone too far – or risks doing so. Take Geoffrey Hinton, a prominent figure in AI, who recently resigned from his position at Google, citing the desire to speak out about the technology's potential risks to society and human well-being.

But against those big-picture concerns, in many areas of science you will hear a different frustration being expressed more quietly: that AI has not yet gone far enough. One of those areas is chemistry, for which machine-learning tools promise a revolution in the way researchers seek and synthesize useful new substances. But a wholesale revolution has yet to happen – because of the lack of data available to feed hungry AI systems.

Any AI system is only as good as the data it is trained on. These systems rely on what are called neural networks, which their developers teach using training data sets that must be large, reliable and free of bias. If chemists want to harness the full potential of generative-AI tools, they need to help to establish such training data sets. More data are needed – both experimental and simulated – including historical data and otherwise obscure knowledge, such as that from unsuccessful experiments. And researchers must ensure that the resulting information is accessible. This task is still very much a work in progress.

Take, for example, AI tools that conduct retrosynthesis. These begin with a chemical structure a chemist wants to make, then work backwards to determine the best starting materials and sequence of reaction steps to make it. AI systems that implement this approach include 3N-MCTS, designed by researchers at the University of Münster in Germany and Shanghai University in China[1]. This combines a known search algorithm with three neural networks. Such tools have attracted attention, but few chemists have yet adopted them.

To make accurate chemical predictions, an AI system needs sufficient knowledge of the specific chemical structures that different reactions work with. Chemists who discover a new reaction usually publish results exploring this, but often these are not exhaustive. Unless AI systems have comprehensive knowledge, they might end up suggesting starting materials with structures that would stop reactions working or lead to incorrect products[2].

A generalist generative-AI system such as ChatGPT, developed by OpenAI in San Francisco, California, is simply data-hungry. To apply such a generative-AI system to chemistry, hundreds of thousands – or possibly even millions – of data points would be needed. A more chemistry-focused AI approach trains the system on the structures and properties of molecules. In the language of AI, molecular structures are graphs. In molecules, chemical bonds connect atoms – just as edges connect nodes in graphs.

> **"The best possible training sets would also include data on negative outcomes."**

The AlphaFold protein-structure-prediction tool[3] uses such a graph-representation approach. It is trained on a formidable data set: the information in the Protein Data Bank, which was established in 1971 to collate the growing set of experimentally determined protein structures and currently contains more than 200,000 structures. AlphaFold provides an excellent example of the power AI systems can have when furnished with sufficient high-quality data.

So how can other AI systems create or access more and better chemistry data? One possible solution is to set up systems that pull data out of published research papers and existing databases, such as an algorithm created by researchers at the University of Cambridge, UK, that converts chemical names to structures[4]. This approach has accelerated progress in the use of AI in organic chemistry.

Another potential way to speed things up is to automate laboratory systems. Existing options include robotic materials-handling systems, which can be set up to make and measure compounds to test AI model outputs[5,6]. However, at present this capability is limited, because the systems can carry out only a relatively narrow range of chemical reactions compared with a human chemist.

There is another, particularly obvious solution: AI tools need open data. How people publish their papers must evolve to make data more accessible. This is one reason why *Nature* requests that authors deposit their code and data in open repositories (see go.nature.com/3ohkfce). It is also yet another reason to focus on data accessibility, above and beyond scientific crises surrounding the replication of results and high-profile retractions. Chemists are already addressing this issue with facilities such as the Open Reaction Database (see go.nature.com/42ayugc).

But even this might not be enough to allow AI tools to reach their full potential. The best possible training sets would also include data on negative outcomes, such as reaction conditions that don't produce desired substances. And data need to be recorded in agreed and consistent formats, which they are not at present.

Chemistry applications require computer models to be better than the best human scientist. Only by taking steps to collect and share data will AI be able to meet expectations in chemistry and avoid becoming a case of hype over hope.

1. Segler, M. H. S., Preuss, M. & Waller, M. P. *Nature* **555**, 604–610 (2018).
2. Struble, T. J. *et al. J. Med. Chem.* **63**, 8667–8682 (2020).
3. Jumper, J. *et al. Nature* **596**, 583–589 (2021).
4. Lowe, D. M., Corbett, P. T., Murray-Rust, P. & Glen, R. C. *J. Chem. Inf. Model.* **51**, 739–753 (2011).
5. Coley, C. W. *et al. Science* **365**, eaax1566 (2019).
6. Angello, N. H. *et al. Science* **378**, 399–405 (2022).