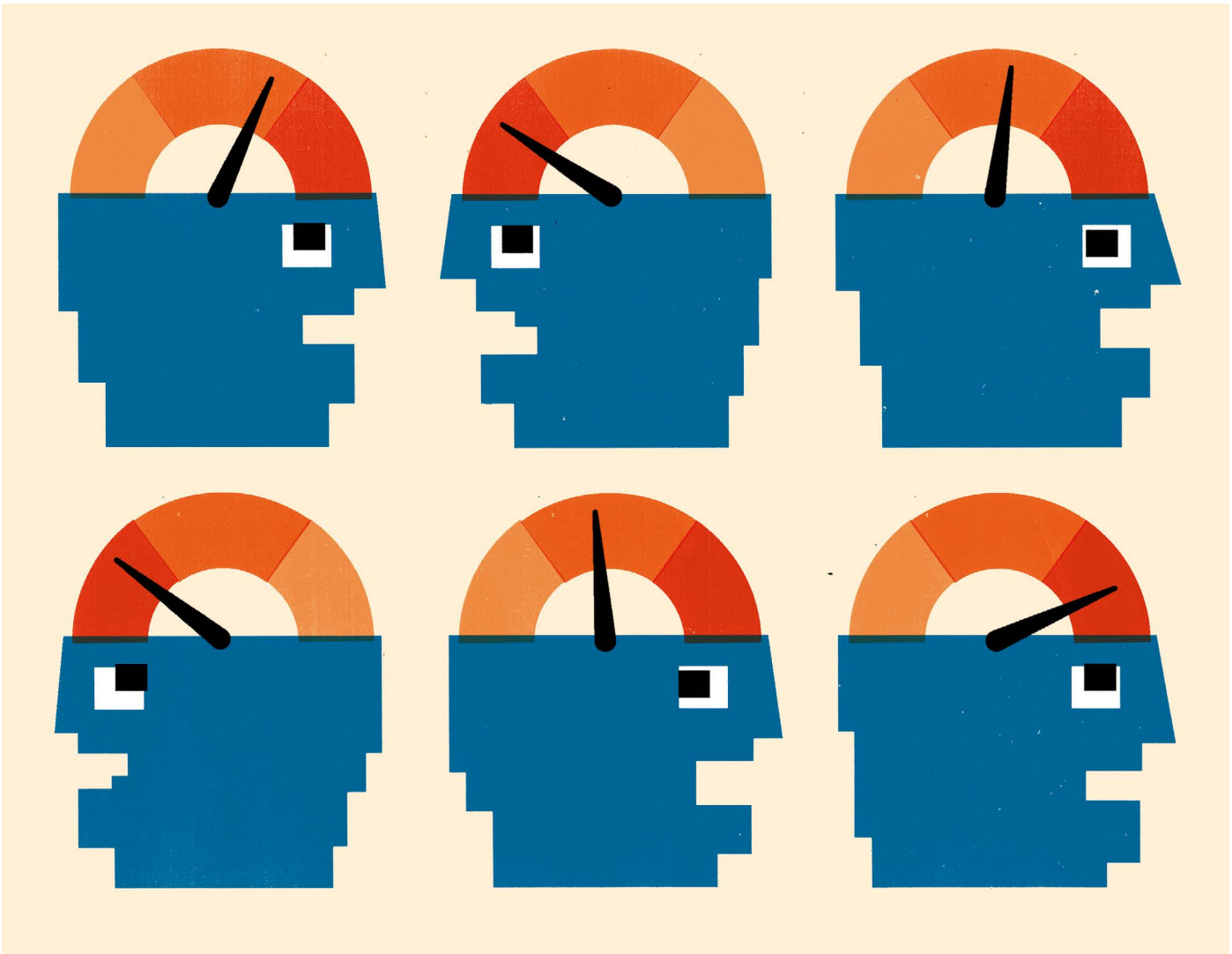

ILLUSTRATION: THE PROJECT TWINS

BUILDING BETTER BENCHMARKS

There's a problem with advancing AI for science: researchers can't agree on what constitutes progress. **By Michael Brooks**

Anshul Kundaje sums up his frustration with the use of artificial intelligence in science in three words: "bad benchmarks propagate".

Kundaje researches computational genomics at Stanford University in California. He is keen to incorporate any form of artificial intelligence (AI) that helps to accelerate progress in his field — and countless researchers have stepped up to offer tools for this purpose. But finding the ones that work best is becoming ever harder because some researchers have been making questionable claims about the AI models they have developed. These claims can take months to check. And they often turn out to be false — mainly because the benchmarks used to demonstrate and compare performance of these tools are not fit for purpose.

By then, it's often too late: Kundaje and his colleagues are left playing whack-a-mole after the flawed benchmarks have been adopted and 'improved' by enthusiastic, but naive, users. "In the meantime, everyone has been using these [benchmarks] for all kinds of wrong stuff, and then you have wrong information and wrong predictions out there," he says.

This is just one reason why a growing number of scientists worry that, until benchmarking is radically improved, AI systems designed to accelerate progress in science will have the opposite effect.

A benchmark is a test that can be used to compare the performance of different methods, just as the standard length of a metre provides a way to assess the accuracy of a ruler. "It's the standardization and definition of what we mean by progress," says Max Welling, a machine-learning researcher and co-founder of CuspAI, an AI company based in Cambridge, UK. Good benchmarks allow a user to choose the best method for a particular application, or to determine whether more conventional algorithms might give a better result. "But the first question," says Welling, "is, what do we mean by 'better'?"

It's a surprisingly deep question. Does 'better' mean faster? Cheaper? More accurate? If you're buying a car, you'll need to consider a wide range of factors, such as acceleration, boot capacity and safety, each with its own degree of importance to you. AI benchmark tools are no different – for some applications, speed might not matter as much as accuracy, for instance.

But it's even more complicated than that. If your benchmark is badly designed, the information it gives you could be misleading. If there's 'leakage', in which the benchmarking relies on data that were used to train the algorithm, the benchmark becomes more of a game of memory than a test of problem-solving. Or the test might just be irrelevant to your needs: it might be overly specific, for instance, hiding a system's inability to answer the broad swathe of questions you're interested in.

This is a problem that Kundaje and his colleagues have identified with DNA language models (DNALMs), which AI developers think could assist the discovery of interesting regulatory mechanisms in a genome. Around 1.5% of the human genome is made up of protein-coding sequences that provide templates for creating RNA (transcription) and proteins (translation). Between 5% and 20% of the genome is made up of non-coding regulatory elements that coordinate gene transcription and translation. Get the DNALMs right, and they could help to interpret and discover functional sequences, predict the consequences of altering those sequences, and redesign them to have specific, desired properties.

So far, however, DNALMs have fallen short of these goals. According to Kundaje and his colleagues, that is partly because they are not being used for the right tasks. They are being designed to compare favourably against benchmark tests, many of which evaluate usefulness not to key biological applications but rather to surrogate objectives that the models can meet[1]. The situation is not unlike schools that 'teach to the test' – you end up with students (or AI tools) that are qualified to pass a test, but do little else.

Kundaje and his colleagues at Stanford University have found these crucial shortcomings in several popular DNALM benchmarks, data sets and metrics. For example, one key task is evaluating a model's ability to rank functional genetic variants: changes in DNA sequences that can influence disease risk or molecular function in cells. Although some DNALMs are simply not evaluated on this task, others use flawed benchmark data sets that fail to account for 'linkage disequilibrium', the non-random association of genetic variants.

That makes it harder to isolate the true functional variants, a flaw that yields unrealistic estimates of these models' abilities to pinpoint such variants. It's a rookie error, Kundaje says. "This doesn't require deep domain knowledge – it's genetics 101."

Transparency and puffery

Inadequate benchmarks are creating a similar teaching-to-the-test problem in a range of scientific disciplines. But the failures don't happen only because it is challenging to create a good benchmark: it's often because there's not enough pressure to do better, according to Nick McGreivy, who completed his PhD in the application of AI in physics last year at Princeton University in New Jersey.

Most people who use AI for science seem content to allow the developers of AI tools to evaluate their usefulness using their own criteria. That's like letting pharmaceutical companies decide whether their drug should go to market, McGreivy says. "The same people who evaluate the performance of AI models also benefit from those evaluations," he says. That means that, even if research isn't deliberately fraudulent, it can be biased.

Lorena Barba, a mechanical and aerospace engineer at the George Washington University in Washington DC, has a similar perspective. Science is suffering because of "poor transparency, glossing over limitations, closet failures, overgeneralization, data negligence, gatekeeping and puffery" in attempts to put AI to work in real-world settings, as she put it in a 2023 talk at the Platform for Advanced Scientific Computing Conference in Davos, Switzerland.

"The same people who evaluate the performance of AI models also benefit from those evaluations."

Barba's own field is fluid dynamics – which involves the study of problems such as smoothing the flow of air over an aircraft's wings to improve fuel efficiency. Doing that involves solving partial differential equations (PDEs), but that isn't straightforward: most PDEs can't be solved through numerical analysis. Instead, the solutions must be approximated through a process that is similar to (expertly guided) trial and error.

The mathematical tools that accomplish this are known as standard solvers. Although they are relatively effective, they also require significant computational resources. That's why many people in fluid dynamics hope that AI – specifically machine-learning approaches – can help them to do more with fewer resources.

Machine learning is the form of AI that has seen the most progress in the past five years – mainly because of the availability of training data. Machine learning involves feeding data into an algorithm that looks for patterns or makes predictions. The parameters of the algorithm can be tweaked to optimize the usefulness of the predictions.

In theory, machine learning could deliver solutions to PDEs faster and using fewer computing resources than conventional methods. The trouble is, if you cannot trust that the benchmarks used to evaluate performance are useful or reliable, how can you trust the output of the models they validate?

McGreivy and his colleague Ammar Hakim, a computational physicist at Princeton University, have conducted an analysis of published 'improvements' to standard solvers and found that 79% of the papers they studied make problematic claims[2]. Much of that is to do with benchmarking against what they term weak baselines. This can come from unfair comparisons: machine learning for PDE could be seen as more efficient in terms of computing resources – a shorter runtime, for example – than a standard solver. But unless the solutions have similar accuracy, the comparison is meaningless. The researchers suggest that comparisons must be made at either equal accuracy or equal runtime.

Another source of weak benchmarking is comparing an AI application with non-AI numerical methods that are relatively inefficient. In 2021, for instance, data scientist Sifan Wang, who is now at Yale University in New Haven, Connecticut, and computer scientist Paris Perdikaris at the University of Pennsylvania in Philadelphia claimed that their machine-learning-based solver for a different class of differential equations yielded a 10-to-50-fold speed-up compared with a conventional numerical solver [3]. But as Chris Rackauckas, a computer scientist at the Massachusetts Institute of Technology in Cambridge, pointed out in a video, the pair weren't comparing it with state-of-the-art numerical solvers, some of which could do the job 7,000 times faster – just running on a standard laptop – than Wang and Perdikaris' approach.

"To be fair to [Perdikaris], after I had pointed this out, they did edit their paper," Rackauckas says. However, he adds, the original paper is the only version that is accessible without a paywall, and so still engenders false hope concerning AI's promise in this area.

There are many such misleading claims, McGreivy warns. The scientific literature is "not a reliable source for evaluating the success of machine learning at solving PDEs", he says. In fact, he remains unconvinced that machine learning has anything to offer in this area. "In PDE research, machine learning has been and remains a solution looking for a problem," he says.

Johannes Brandstetter, a machine-learning researcher at Johannes Kepler University in Linz, Austria, and co-founder of an AI-driven physics

NICHOLAS MCGREIVY

Nick McGreivy found that some published improvements to AI models made misleading claims.

simulation start-up company called Emmi AI, is more optimistic. He points to the Critical Assessment of Structure Prediction (CASP) competition that enabled machine learning to assist with the prediction of 3D protein structures from their amino-acid sequences[4].

CASP provides a widely accepted, independent, double-blind, standardized test of protein structure prediction. Its rigour led to the successful protein-structure predictions made by Google DeepMind's AlphaFold software, and it led David Baker, a biochemist at the University of Washington in Seattle, and his team to use deep learning — a more complex form of machine learning — to design proteins for specific applications. Demis Hassabis and John Jumper at Google DeepMind in London shared the 2024 Nobel Prize in Chemistry with Baker for this work. It might be possible, Brandstetter says, to solve the PDE benchmarking issue through something similar.

'Not completely incompetent'

Work in knot theory suggests that machine learning could also be useful in mathematics — although there are benchmark-based reasons to be cautious about AI's potential there, too, says Mengdi Wang, a theoretical computer scientist at Princeton University.

"All the benchmarks are very narrow," Mengdi Wang says. When it comes to mathematics, for example, the models memorize problem-solving patterns and therefore can look extremely competent when fed the right questions — which are the ones presented in public demonstrations of their performance.

Mengdi Wang was part of a team that showed that this success is extremely brittle[5], even for advanced benchmarks, such as the MATH Dataset, OlympiadBench and the American Invitational Mathematics Examination. The team changed some numbers in the maths problems so that they had to be solved in entirely different ways. Failing to pick up on those tweaks, "all the models' performance dropped pretty significantly", Mengdi Wang says.

And then there's the problem of benchmarking an AI that can 'do science'. In September last year, ChatGPT developer OpenAI launched OpenAI o1, a model that the company says "places among the top 500 students in the US in a qualifier for the USA Math Olympiad" and "exceeds human PhD-level accuracy on a benchmark of physics, biology, and chemistry problems". The benchmark in question is GPQA (graduate-level Google-proof Q&A benchmark), a "challenging dataset of 448 multiple-choice questions written by domain experts in biology, physics, and chemistry", as its creators put it.

But are those few hundred questions enough to create the broad knowledge and range of capabilities one would expect of a good graduate student? That certainly was not the experience of Terence Tao, a mathematician at the University of California, Los Angeles. Tao had the opportunity to experiment with OpenAI o1 last September, and commented on social media that the experience "seemed roughly on par with trying to advise a mediocre, but not completely incompetent, (static simulation of a) graduate student."

Tao was among the authors of a 2024 report[6], commissioned by former US president Joe Biden, on the prospects of using AI to supercharge scientific research. One of the report's conclusions was that without thoughtfully developed benchmark metrics and validation procedures, "AI systems can give unreliable outputs whose quality is difficult to evaluate, and which could be harmful for a scientific field and its applications."

The report recommended that those benchmarks should be developed and imposed by government agencies, such as the US National Science Foundation and the National Institute of Standards and Technology. "I think using the federal workforce to help regulate some of these things would be very helpful," says physicist Laura Greene at Florida State University in Tallahassee, another of the report's authors.

Researchers in China have made their own suggestions for developing new benchmarks, offering assessment tools, AI-ready data sets and multidimensional assessments of model attributes, such as knowledge, understanding, reasoning and values[7]. The work currently spans mathematics, physics, chemistry, life sciences and Earth and space sciences.

Barba, too, has co-authored a study that examined how to make scientific machine learning more trustworthy — including the development of reliable benchmarks[8]. She and her co-authors say the first step is always ensuring the problem in question is extremely well-defined, with a clear understanding of previous knowledge and the specific quantities of interest. With that in place, a tightly constrained process of verification, validation and credibility building could lead to an AI tool that is demonstrably useful to scientists. In the meantime, the authors ask that researchers "thoroughly document their choices and explain the significance of the problems they address".

Whether you think that will happen probably depends on whether you're an optimist or a pessimist. Kundaje is the former: "There needs to be a serious effort to create exemplar benchmarks, but there's definitely enough motivation from all parties concerned, so it will happen," he says. His team, for instance, has introduced a suite of benchmark tasks and data sets (called DART-Eval) across varying degrees of difficulty to calibrate the utility of DNALMs more rigorously.

Until good benchmarks are in place across the board, however, Mengdi Wang encourages pragmatism — and conversations with AI-savvy colleagues — among any scientists looking to use machine learning. "You need to go into it with your eyes open," she says. "You might have a domain expertise, but you really need AI expertise to understand what goes on behind the curtain."

Michael Brooks is a science writer in Lewes, UK.

1. Patel, A. *et al.* Preprint at arXiv https://doi.org/10.48550/arXiv.2412.05430 (2024).
2. McGreivy, N. & Hakim, A. *Nature Mach. Intell.* **6**, 1256–1269 (2024).
3. Wang, S. & Perdikaris, P. Preprint at arXiv https://doi.org/10.48550/arXiv.2106.05384 (2021).
4. Brandstetter, J. *Nature Mach. Intell.* **7**, 2–3 (2025).
5. Huang, K. *et al.* Preprint at arXiv https://doi.org/10.48550/arXiv.2502.06453 (2025).
6. President's Council of Advisors on Science and Technology. *Supercharging Research: Harnessing Artificial Intelligence to Meet Global Challenges* (Executive Office of the President, 2024).
7. Qin, C. *et al.* Preprint at arXiv https://doi.org/10.48550/arXiv.2503.13503 (2025).
8. Jakeman, J. D., Barba, L. A., Martins, J. R. R. A. & O'Leary-Roseberry, T. Preprint at arXiv https://doi.org/10.48550/arXiv.2502.15496 (2025).