https://doi.org/10.1038/s41587-023-02103-0

The perpetual motion machine of AI-generated data and the distraction of ChatGPT as a ‘scientist’

As a longtime researcher at the intersection of artificial intelligence (AI) and biology, for the past year I have been asked questions about the application of large language models and, more generally, AI in science. For example: “Since ChatGPT works so well, are we on the cusp of solving science with large language models?” or “Isn’t AlphaFold2 suggestive that the potential of AI in biology and science is limitless?” And inevitably: “Can we use AI itself to bridge the lack of data in the sciences in order to then train another AI?”

I do believe that AI—equivalently, machine learning—will continue to advance scientific progress at a rate not achievable without it. I don’t think major open scientific questions in general are about to go through phase transitions of progress with machine learning alone. The raw ingredients and outputs of science are not found in abundance on the internet, yet the tremendous power of machine learning lies in data—and lots of them.

A major distinguishing factor of the sciences (specifically, biology, chemistry and physics), as compared to AI fields such as natural language processing and computer vision, is the relative lack of publicly available data suitable for these domains (see Box 1). There are simply vastly fewer data existing in the sciences, and these are often siloed by academics and companies. Acquiring appropriate scientific data for AI typically requires not only highly trained humans, but also high-end facilities with expensive equipment, making for an overall costly and slow endeavor compared to humans simply going about their day by adding to the vast trove of images, text, audio and video on the World Wide Web. As one example, it has been estimated that “The replacement cost of the entire PDB [Protein Data Bank] archive is conservatively estimated at ~US$20 billion”[1]—these are the data used by DeepMind to train AlphaFold2.

BOX 1

Comparison of sizes of publicly available datasets in several domains

With respect to the amount of data available in different fields, it’s difficult to make detailed comparisons, and this summary is by no means comprehensive; but in broad strokes, we note the following. The open-access paired image-text dataset LAION (https://laion.ai/blog/laion-5b/) has nearly 6 billion paired examples[6], and the Common Crawl dataset had, as of June 2023, ~3 billion web pages comprising ~400 terabytes, with billions of new pages added each month (https://commoncrawl.org/). In contrast, in the sciences, the number of protein sequences in UniRef as of May 2023 was ~250 million, and in the decade 2012–2022 this went up by ~ 150 million (https://www.ebi.ac.uk/uniprot/TrEMBLstats). Supervised datasets in the sciences include that used by AlphaFold2[2], which was trained on ~170,000 proteins and their structures with an additional 350,000 unlabeled sequences from UniClust30 (https://deepmind.google/discover/blog/alphafold-a- solution-to-a-50-year-old-grand-challenge-in-biology/). As of 2022, there were 1,663 RNA structures[7]. ChemSpider contained 128 million chemical structures as of November 2023 (https://www.chemspider.com/). The Open Reaction Database (docs.open-reaction-database.org) currently contains 2.5 million examples of organic reaction data[8]. In a different category from experimentally collected data are data computationally generated using approximate simulations of physics, such as those using density functional theory (DFT) computations. For example, Open Catalyst 2022 contains 62,000 DFT relaxations for oxides[9]. Open Direct Air Capture 2023 contains 38 million DFT calculations on 8,800 metal–organic framework materials[10] and the Materials Project[11] contains information for 155,000 materials as of December 2023 (https://materialsproject.org).

DeepMind researchers performed a fantastic feat with the development of AlphaFold2[2], substantially moving the needle on the protein structure prediction problem. Protein structure prediction is a tremendously important challenge with actual and still-to-be-realized impact. It was also, arguably, the only challenge in biology, or possibly in all the sciences, that they could have tackled so successfully. There are several reasons for this that together made for an extremely rare success. First, the problem of protein structure prediction is easily defined quantitatively. Second, there were sufficient existing data to use in training a complex, supervised model for this narrowly defined problem—data that had been slowly and expensively collected over decades. Finally, it was possible to assess the accuracy of the results by way of held-out proteins whose structures were already known, yielding a precise, yet human-interpretable, quantitative metric that should more or less translate to many common use cases. Very few problems in the sciences are lucky enough to have all of these characteristics. In fact, the most interesting and impactful questions may not yet be formulated at all, let alone in a manner suitable for machine learning, or with existing suitable data, or even a way to readily generate suitable data for machine learning. Moreover, even for the problem of protein structure prediction, many important unsolved questions remain, primarily those of conformational dynamics and contextual effects[3], which will undoubtedly require yet more data to effectively tackle.

There is the hope that we can bridge scientific data gaps by using AI to generate synthetic data. But we simply cannot get something for nothing—fresh information must be injected into the system one way or another for there to be a win. Just as we cannot build a perpetual motion machine, we cannot generate new information in a trivial cycle of information processing. The topic of AI-generated 'synthetic' data can be somewhat nuanced and technical but can be made intuitively accessible.

One might say that AlphaFold2 used synthetic structure data predicted from the model itself to improve its own accuracy. Notably, AlphaFold2 used real, unlabeled sequence data to do so, a strategy classically known as semi-supervised learning, which has both a long history and theoretical underpinnings. Generally, such a strategy may or may not be helpful. That the specific instantiation of semi-supervised learning in AlphaFold2 used real protein sequences fed through the model to obtain AI-predicted labels does not make these data entirely AI generated or synthetic—they remain anchored on real protein sequences. Moreover, semi-supervised learning is not a magic bullet for the problem of lack of supervised data—it can only help so much. Anyone touting the use of AI-based synthetic data should be able to explain in clear terms where the new information is coming from, relative to how they plan to use it. If one cannot reasonably do so, it's almost certain that the procedure will not be useful. As my good friend and computer vision colleague Alyosha Efros recounts, "A few times each year, a physician with fewer than 2,000 MRI images contacts me about using our generative models to generate more training data. Now I know which doctors not to go to."

One path to generating useful synthetic data is to integrate human knowledge. For example, we can augment a labeled protein structure dataset by rotating protein structure labels in the original training dataset by a random amount, while keeping the underlying sequences as they were, and adding these synthesized pairs to the training data. In doing so, one is encoding the human belief, in this case corresponding to physical reality, that a protein 3D structure is, essentially, the same structure even if rotated in space. It is telling that one can entirely replace such a 'data augmentation' strategy by instead encoding this belief directly into the architecture of the neural network—a testament to the fact that a trained model has a data equivalence. Jahanian et al. put it elegantly: "Generative models can be viewed as a compressed and organized copy of a dataset"[4]. Neither data augmentation nor symmetry-encoding architecture is necessarily the better strategy. Critically, both are founded on the same information—information that is readily identifiable. In both cases, a human injects information, either by way of manipulated (rotated), augmented data to which they are declaring an invariance, or by changing the model architecture to encode that same invariance.

Might we use synthetic data from one model to help train another AI model? Indeed, we can use an auxiliary model, AI based or otherwise (e.g., biophysics based), to generate synthetic data for another AI model. For example, we might use physics-based simulations (such as molecular dynamics simulations) to generate data for one AI model that otherwise has access to only a small amount of true, experimental data. We can use one AI model to help another so long as the two models hold different information, either by way of the data they were trained on or through their inductive biases (biases induced by the model architecture, loss function, parameter initialization and/or optimization procedure used to train it). For any strategy of generating data from one AI model to be useful to train another (or the same model), there cannot be a trivial cycle in the pipeline. For example, we cannot generate data from a generative model only to directly feed these generated data back into that same model with the same learning objective—doing so is analogous to trying to build a perpetual motion machine. For a data-generating procedure to be useful, any cycle must have feedback in it to inject new information, such as filtering those generated data with an external procedure, human or machine based. One classical strategy of combining different AI models together, that of ensemble learning, can be quite powerful, and provably so when the models differ in their predictions from each other[5]. More differences mean more information. Of course, if the different information is sufficiently incorrect, this procedure might degrade rather than help the model performance. This lesson should be taken more broadly, as there truly is no free data lunch.

Will generative and other machine learning models help us to make progress in the sciences? Undoubtedly, yes. They already are. A generative model is, fundamentally, a probability density model, because it captures the probability distribution of the data, shaped by our hands in the way of inductive biases, explicitly known or not. We can use generative models to 'score' unseen data samples to see whether they 'belong' in the set of training data. We can extract learned representations from them that may themselves yield scientific insights or prove useful for downstream computational tasks. We can also generate 'new' samples from them. But we should not forget that those 'new' samples and extracted representations span only the raw information of bits that was used to train them in the first place. The hope is that the model gives us more useful and ready access to those bits of information.

As for ChatGPT and its relatives, these will undoubtedly continue to provide a new generation of incredibly useful literature synthesis tools. These literature-based tools will drive new engines of profound convenience, previously impossible and only dreamed of, such as those providing medical diagnosis and beyond—also those not yet dreamed of. But they will not themselves, anytime soon, be virtual scientists. AI can help us to understand the data we've collected, and with enough of it, to generalize from them, within reason. It can also help us to decide what to measure, initially or iteratively. In order to probe the limits of current scientific knowledge, however, we need data that we don't already have. For this, we'll just need to get back to the bench and do more experiments.

Jennifer Listgarten[1,2] ✉
[1]Electrical Engineering and Computer Science, University of California, Berkeley, Berkeley, CA, USA. [2]Center for Computational Biology, University of California, Berkeley, Berkeley, CA, USA.
✉e-mail: jennl@berkeley.edu

Published online: 25 January 2024

References

1. Burley, S. K. et al. *Nucleic Acids Res.* **51**, D488–D508 (2023).
2. Jumper, J. et al. *Nature* **596**, 583–289 (2021).
3. Terwilliger, T. C. et al. *Nat. Methods* https://www.nature.com/articles/s41592-023-02087-4 (2023).
4. Jahanian, A., Puig, X., Tian, Y. & Isola, P. Generative models as a data source for multiview representation learning. Preprint at *arXiv* https://arxiv.org/abs/2106.05258 (2022).
5. Dietterich, T. G. In *Multiple Classifier Systems (MCS 2000)*, Lecture Notes in Computer Science Vol. 1857 (Springer, 2000).
6. Schuhmann, C. et al. LAION-5B: an open large-scale dataset for training next generation image-text models. Preprint at *arXiv* https://arxiv.org/abs/2210.08402v1 (2022).
7. Deng, J. et al. *Fundam. Res.* **3**, 727–737 (2023).
8. Kearnes, S. M. et al. *J. Am. Chem. Soc.* **143**, 18820–18826 (2021).

9. Tran, R. et al. *ACS Catal.* **13**, 3066–3084 (2022).
10. Sriram, A. et al. The Open DAC 2023 dataset and challenges for sorbent discovery in direct air capture. Preprint at *arXiv* https://arxiv.org/abs/2311.00341v2 (2023).
11. Jain, A. et al. *APL Mater.* **1**, 11002 (2013).

Acknowledgements

Thanks to Tyler Bonnen, James Bowden, Jennifer Doudna, Lisa Dunlap, Alyosha Efros, Nicolo Fusi, Aaron Hertzmann, Hanlun Jiang, Aditi Krishnapriyan, Jitendra Malik, Sara Mostafavi, Hunter Nisonoff and Ben Recht for helpful comments on this piece as it was taking shape.

Competing interests

The author declares no competing interests.