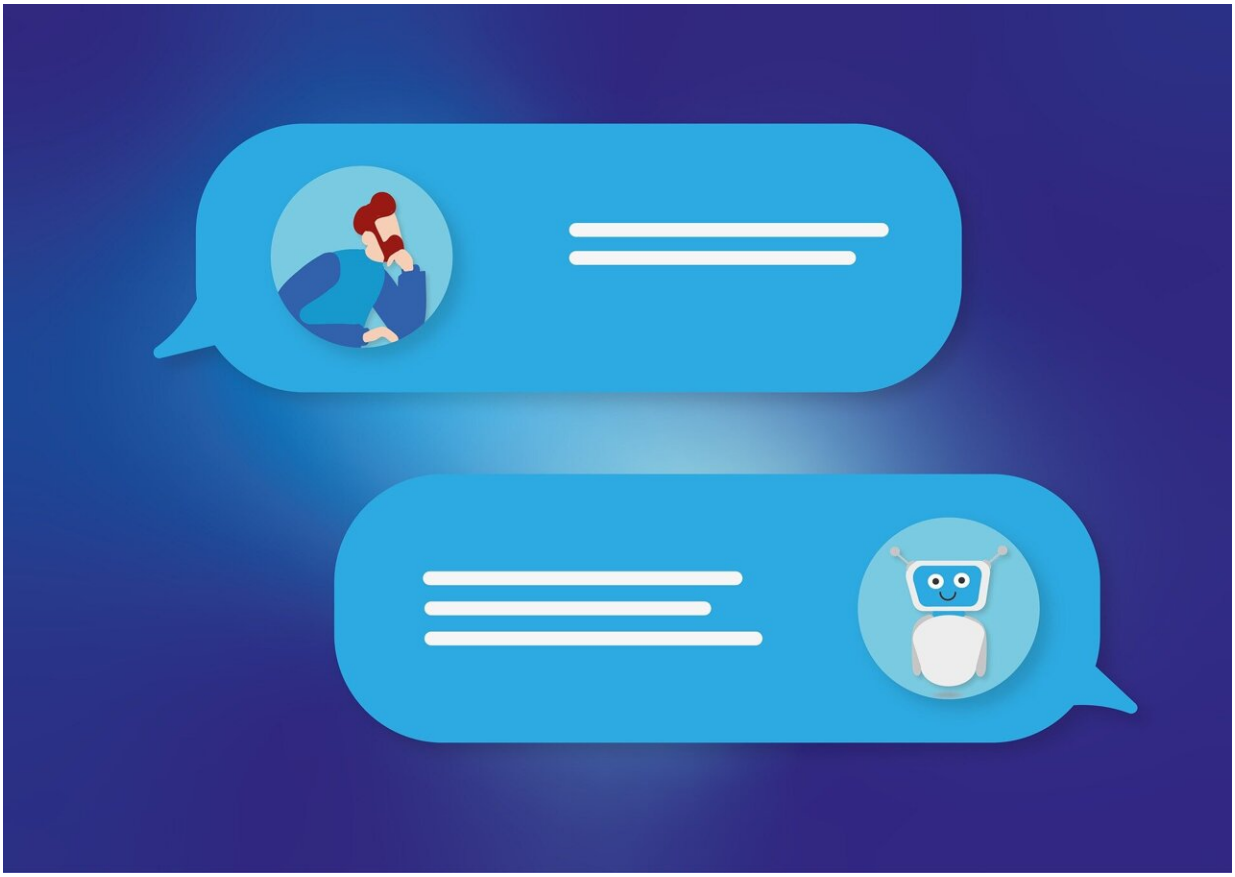
Q&A: How generative AI could help accelerate biomedical research

November 3 2023, by Corie Lok



Credit: Pixabay/CC0 Public Domain

The recent explosion of generative AI tools has prompted many discussions in virtually all fields about the benefits and risks of these

technologies. These tools, including ChatGPT, Bard and others, have been trained on huge amounts of content and can produce text and images that often look eerily like human-generated content.

At the Broad Institute of MIT and Harvard, a group of researchers, software engineers, administrators, and communicators (yes, us) has been exploring the use of these chatbots and similar tools, surveying the community and developing recommendations.

To dive deeper into this topic, we spoke with Mehrtash Babadi, an institute scientist, director of computational methods, and a machine learning and AI expert in Broad's Data Sciences Platform. He talked about how generative AI techniques can be used not just to analyze human language but also the language of genes and cells—raw biological data—to shed light on how cells and tissues work in health and disease.

He also shared his thoughts on the benefits of language-based generative models like ChatGPT, Bard, and GitHub Copilot for writing computer code, developing hypotheses, and other tasks.

"I think these systems will become increasingly useful not only for software engineers and programmers, but also for basically everyone else in every profession in the same sense that a search engine has become an indispensable part of our lives for accessing information," said Babadi, who routinely uses ChatGPT to search the internet and write emails and research summaries.

The following conversation was edited (by humans) for length and clarity.

How have you been thinking about generative AI in biology?

Generative AI is something that has been brewing for a long time in the machine learning community, going back to the fundamental tenets of Bayesian statistics. We've been using those for a long while, for modeling various aspects of biology like genomic variation, experimental artifacts, single-cell biology, and other areas.

Now with the advancement of these models, their combination with deep neural networks, vast amounts of training data and computing power, and in particular the progress of these models in generating images and natural language, they have really exploded and all of a sudden everybody is excited about them.

We are now thinking about how the same approaches that have been so successful in modeling natural language and images could be used for learning the intrinsic, innate language of biological systems like cells and tissues, and predicting their fate and response to various stimuli in silico. That's an area of active research for us, and we have made a little bit of headway, but there's a lot of work that needs to be done.

Can you explain more about how generative AI can be used to analyze biological data?

Right now, there's a lot of excitement about ChatGPT and similar conversational AI systems, and for good reasons, because these are really capable and powerful systems, and there's also a lot of emerging work in the field showing that these models also have a good grasp of biology. You can ask them questions like "what is the function of this gene?" and they will tell you because they have read textbooks and papers. So the models have learned what we know about biology.

And that's exactly the problem, because we don't know much about many aspects of biology! Our understanding of biology is still evolving

and is very biased and some of the literature is not even reproducible. The natural language models are trained on that substrate, and so they're subject to the same biases and incomplete understandings of biology that we are subject to.

So we are trying to directly learn the language of biological systems from raw biological measurements and data without any human interpretation in between.

How would a researcher use a generative model trained on raw biological data?

For example, you can envision a generative model that's been trained on biological data describing how certain tissues or cells work, and then using that model to generate data that describe new cell states or new tissues. You can even make models that you could prompt with something like "here's a cell in a tissue, generate another cell nearby" to make predictions about how different cells might work together to form a tissue, as an example.

These models could also be fine-tuned on interventional data, such as genetic or pharmacological screens, to learn to predict future screens. In a nutshell, generative models have the potential to computationalize many aspects of cell and tissue biology and perturbation screens.

What becomes very interesting now is to interface these models of cells and tissues with natural language models. So we can take natural language models and the more unbiased and comprehensive models of cells, and then fuse them together into a system that is more powerful than each of them separately. That's an active area of AI research called multimodal generative AI, where one basically combines generative models of different modalities, or interface them together, and allow

them to talk to each other.

The advantage of this is that with the models based on the innate language of biology, you avoid the bias that's inherent in the natural language models of biology. But you can use the natural language models to allow a human scientist to put in the right prompts.

What progress has been made in multimodal generative AI for biology?

We do now have multimodal generative AI of natural language and text, natural language and images, but generative AI models of biological systems are still in their infancy. We have yet to see multimodal AI systems that combine natural language with the language of biological systems.

Can generative AI be useful for hypothesis generation?

One potential example I can think of is a typical drug development project, where we want to understand the underlying mechanism for a disease and then identify a therapeutic target. Right now, this is typically done through a combination of subject matter expert insights and the design of very smart experiments that test smart hypotheses using innovative techniques to manipulate cells and whatnot.

But as we do more and more of these types of experiments, each of these experiments is a sort of lesson for a generative AI system that says "here's a cell and here's how we intervened and here's what happened." And the more of these lessons we catalog, the more we can teach a generative AI system to predict future experiments without us needing to do all of them in the lab. There is this immense opportunity to reuse all

of the experimental data that we've collected so far.

But won't some of those predictions be wrong?

Even if these generative models are sometimes wrong, they're not entirely wrong. This means that if, for example, you use them to identify a certain therapeutic target, if the systems are appropriately trained, it is highly likely that at least some of those targets actually make some sense.

That's probably one of the best applications of these systems: to take their outputs as potential hypotheses and then subject them to experimental validation. Depending on the nature of the outcome, the resulting data from the follow-up experiments will either reinforce the model's belief or correct it, ultimately making it slightly more accurate for future queries.

Let's talk about the natural language models like ChatGPT. How useful are these tools for coding and software development?

Some of us use GitHub Copilot, which is a system that helps coders and programmers write some of the more standard, boilerplate parts of code, rather than the most innovative and challenging parts. These systems are really good at helping you write parts of your code that everybody knows how to write, but you still need to do it anyway.

These systems are also really good at helping you document your code and comment on your code. So we're using these systems right now for these purposes and as smarter versions of the conventional code-completion systems.

Do you have any concerns about these language

models, like inaccuracies or potential misuse?

The challenge is that these models are well known to "hallucinate" once in a while or just very confidently lie. So you have to do your own fact-checking. As for misuse, I'm less worried about the science and engineering communities because scientists and engineers are, by training, skeptics and they tend to not take things at face value. So even if they use a generative AI system to help them solve a problem, they would test the output of these systems.

I think where I'd be worried more is how these systems could be exploited in other areas, such as generating misinformation and in other discourses where people are not as inclined to do their own fact-checking. That's where I'm worried, especially because these systems can generate content much, much faster than we can. So it's very easy to flood the space, so to speak, with lots of deliberately false, AI-generated content.

But as tools for biological research and software development, I think there's a lot of promise in helping to make some parts of research more efficient. The pace at which we're generating data, which is exponentially increasing, is far exceeding our expert capability to make sense of that data. That's where generative AI and in general, machine learning and other AI methods, could become extremely useful to help us uncover the regularities, commonalities, and differences in all this data in a way that is less biased and also more efficient and faster than we humans can do.

Provided by Broad Institute of MIT and Harvard