Artificial intelligence system predicts consequences of gene modifications

May 31 2023, by Sarah C. P. Williams

Geneformer, the new AI model developed by Theodoris and her colleagues, can be used across many areas of biology and help discover possible drug targets for disease. Credit: Gladstone Institutes

Researchers at Gladstone Institutes, the Broad Institute of MIT and Harvard, and Dana-Farber Cancer Institute have turned to artificial intelligence (AI) to help them understand how large networks of interconnected human genes control the function of cells, and how

disruptions in those networks cause disease.

Large language models, also known as foundation models, are AI systems that learn fundamental knowledge from massive amounts of general data, and then apply that knowledge to accomplish new tasks—a process called transfer learning. These systems have recently gained mainstream attention with the release of ChatGPT, a chatbot built on a model from OpenAI.

In the new work, published in the journal *Nature*, Gladstone Assistant Investigator Christina Theodoris, MD, Ph.D., developed a foundation model for understanding how genes interact. The new model, dubbed Geneformer, learns from massive amounts of data on gene interactions from a broad range of human tissues and transfers this knowledge to make predictions about how things might go wrong in disease.

Theodoris and her team used Geneformer to shed light on how heart cells go awry in heart disease. This method, however, can tackle many other cell types and diseases too.

"Geneformer has vast applications across many areas of biology, including discovering possible drug targets for disease," says Theodoris, who is also an assistant professor in the Department of Pediatrics at UC San Francisco. "This approach will greatly advance our ability to design network-correcting therapies in diseases where progress has been obstructed by limited data."

Theodoris designed Geneformer during a postdoctoral fellowship with X. Shirley Liu, Ph.D., former director of the Center for Functional Cancer Epigenetics at Dana-Farber Cancer Institute, and Patrick Ellinor, MD, Ph.D., director of the Cardiovascular Disease Initiative at the Broad Institute—both authors of the new study.

A network view

Many genes, when active, set off cascades of molecular activity that trigger other genes to dial their activity up or down. Some of those genes, in turn, impact other genes—or loop back and put the brakes on the first gene. So, when a scientist sketches out the connections between a few dozen related genes, the resulting network map often looks like a tangled spiderweb.

If mapping out just a handful of genes in this way is messy, trying to understand connections between all 20,000 genes in the human genome is a formidable challenge. But such a massive network map would offer researchers insight into how entire networks of genes change with disease, and how to reverse those changes.

"If a drug targets a gene that is peripheral within the network, it might have a small impact on how a cell functions or only manage the symptoms of a disease," says Theodoris. "But by restoring the normal levels of genes that play a central role in the network, you can treat the underlying disease process and have a much larger impact."

Artificial intelligence 'transfer learning'

Typically, to map gene networks, researchers rely on huge datasets that include many similar cells. They use a subset of AI systems, called machine learning platforms, to work out patterns within the data. For example, a machine learning algorithm could be trained on a large number of samples from patients with and without heart disease, and then learn the gene network patterns that differentiate diseased samples from healthy ones.

However, standard machine learning models in biology are trained to

only accomplish a single task. In order for the models accomplish a different task, they have to be retrained from scratch on new data. So, if researchers from the first example now wanted to identify diseased kidney, lung, or brain cells from their healthy counterparts, they'd need to start over and train a new algorithm with data from those tissues.

The issue is that for some diseases, there isn't enough existing data to train these machine learning models.

In the new study, Theodoris, Ellinor, and their colleagues tackled this problem by leveraging a machine learning technique called "transfer learning" to train Geneformer as a foundational model whose core knowledge can be transferred to new tasks.

First, they "pre-trained" Geneformer to have a fundamental understanding of how genes interact by feeding it data about the activity level of genes in about 30 million cells from a broad range of human tissues.

To demonstrate that the transfer learning approach was working, the scientists then fine-tuned Geneformer to make predictions about the connections between genes, or whether reducing the levels of certain genes would cause disease. Geneformer was able to make these predictions with much higher accuracy than alternative approaches because of the fundamental knowledge it gained during the pretraining process.

In addition, Geneformer was able to make accurate predictions even when only shown a very small number of examples of relevant data.

"This means Geneformer could be applied to make predictions in diseases where research progress has been slow because we don't have access to sufficiently large datasets, such as rare diseases and those

affecting tissues that are difficult to sample in the clinic," says Theodoris.

Lessons for heart disease

Theodoris's team next set out to use transfer learning to advance discoveries in heart disease. They first asked Geneformer to predict which genes would have a detrimental effect on the development of cardiomyocytes, the muscle cells in the heart.

Among the top genes identified by the model, many had already been associated with heart disease.

"The fact that the model predicted genes that we already knew were really important for heart disease gave us additional confidence that it was able to make accurate predictions," says Theodoris.

However, other potentially important genes identified by Geneformer had not been previously associated with heart disease, such as the gene TEAD4. When the researchers removed TEAD4 from cardiomyocytes in the lab, the cells were no longer able to beat as robustly as healthy cells.

Therefore, Geneformer had used transfer learning to make a new conclusion: Even though it had not been fed any information on cells lacking TEAD4, it correctly predicted the important role that TEAD4 plays in cardiomyocyte function.

Finally, the group asked Geneformer to predict which genes should be targeted to make diseased cardiomyocytes resemble healthy cells at a gene network level. When the researchers tested two of the proposed targets in cells affected by cardiomyopathy (a disease of the heart muscle), they indeed found that removing the predicted genes using

CRISPR gene editing technology restored the beating ability of diseased cardiomyocytes.

"In the course of learning what a normal gene network looks like and what a diseased gene network look like, Geneformer was able to figure out what features can be targeted to switch between the healthy and diseased states," says Theodoris. "The transfer learning approach allowed us to overcome the challenge of limited patient data to efficiently identify possible proteins to target with drugs in diseased cells."

"A benefit of using Geneformer was the ability to predict which genes could help to switch cells between healthy and disease states," says Ellinor. "We were able to validate these predictions in cardiomyocytes in our laboratory at the Broad Institute."

The researchers are planning to expand the number and types of cells that Geneformer has analyzed in order to keep boosting its ability to analyze gene networks. They've also made the model open-source so that other scientists can use it.

"With standard approaches, you have to retrain a model from scratch for every new application," says Theodoris. "The really exciting thing about our approach is that Geneformer's fundamental knowledge about gene networks can now be transferred to answer many biological questions, and we're looking forward to seeing what other people do with it."

More information: Patrick Ellinor, Transfer learning enables predictions in network biology, *Nature* (2023). DOI: 10.1038/s41586-023-06139-9.
www.nature.com/articles/s41586-023-06139-9

Provided by Gladstone Institutes