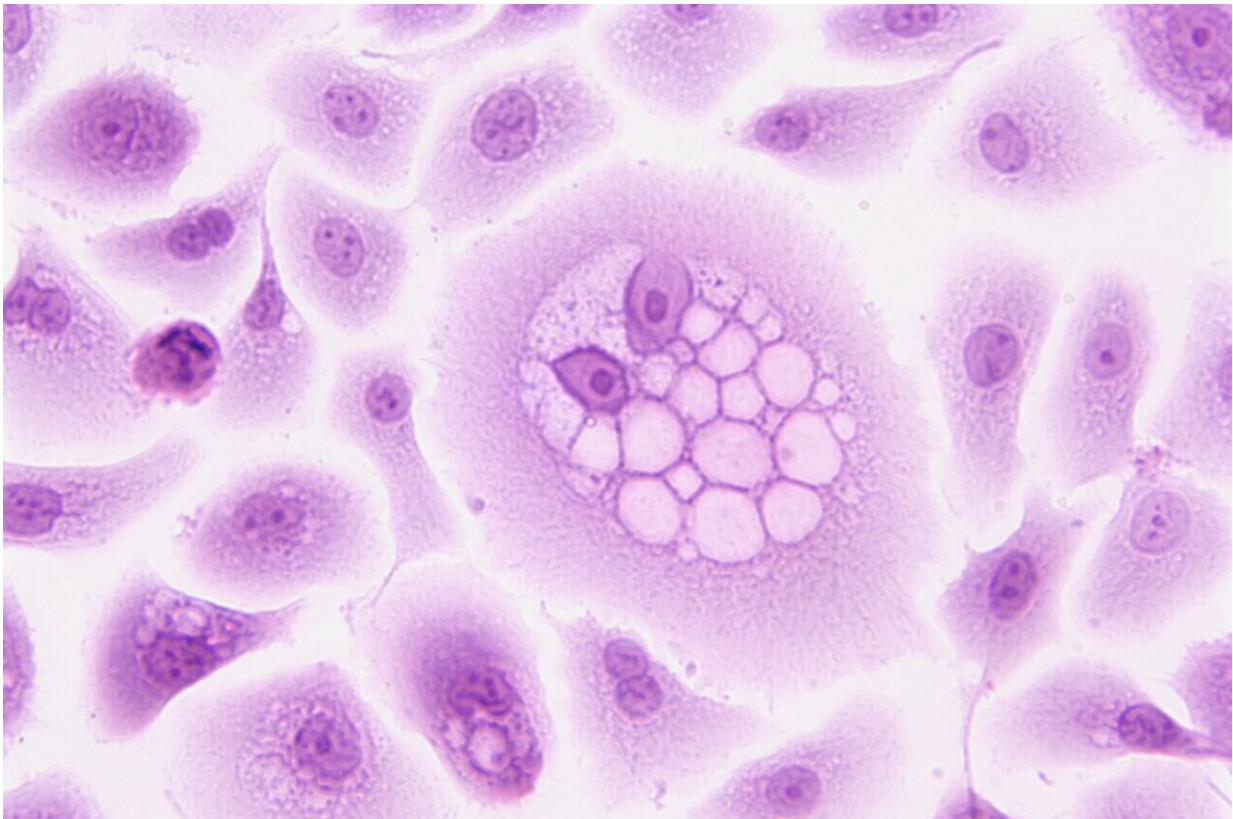# Computational biologists develop AI that predicts inner workings of cells

January 8 2025

Credit: Unsplash/CC0 Public Domain

Using a new artificial intelligence method, researchers at Columbia University Vagelos College of Physicians and Surgeons can accurately predict the activity of genes within any human cell, essentially revealing

the cell's inner mechanisms. The system, [described](#) in *Nature*, could transform the way scientists work to understand everything from cancer to genetic diseases.

"Predictive generalizable computational models allow to uncover biological processes in a fast and accurate way. These methods can effectively conduct large-scale computational experiments, boosting and guiding traditional experimental approaches," says Raul Rabadan, professor of systems biology and senior author of the new paper.

Traditional research methods in biology are good at revealing how cells perform their jobs or react to disturbances. But they cannot make predictions about how cells work or how cells will react to change, like a cancer-causing mutation.

"Having the ability to accurately predict a cell's activities would transform our understanding of fundamental [biological processes](#)," Rabadan says. "It would turn biology from a science that describes seemingly random processes into one that can predict the underlying systems that govern cell behavior."

In recent years, the accumulation of massive amounts of data from cells and more powerful AI models are starting to transform biology into a more predictive science. The 2024 Nobel Prize in Chemistry was awarded to researchers for their groundbreaking work in using AI to predict protein structures. But the use of AI methods to predict the activities of genes and proteins inside cells has proven more difficult.

New AI method predicts gene expression in any cell

In the new study, Rabadan and his colleagues tried to use AI to predict which genes are active within specific cells. Such information about gene expression can tell researchers the identity of the cell and how the

cell performs its functions.

"Previous models have been trained on data in particular cell types, usually [cancer cell lines](#) or something else that has little resemblance to normal cells," Rabadan says. Xi Fu, a graduate student in Rabadan's lab, decided to take a different approach, training a machine learning model on gene expression data from millions of cells obtained from normal human tissues. The inputs consisted of genome sequences and data showing which parts of the genome are accessible and expressed.

The overall approach resembles the way ChatGPT and other popular "foundation" models work. These systems use a set of training data to identify underlying rules, the grammar of language, and then apply those inferred rules to new situations.

"Here it's exactly the same thing: we learn the grammar in many different cellular states, and then we go into a particular condition—it can be a diseased [cell type] or it can be a normal cell type—and we can try to see how well we predict patterns from this information," says Rabadan.

Fu and Rabadan soon enlisted a team of collaborators, including co-first authors Alejandro Buendia, now a Stanford Ph.D. student formerly in the Rabadan lab, and Shentong Mo of Carnegie Mellon, to train and test the new model.

After training on data from more than 1.3 million human cells, the system became accurate enough to predict [gene expression](#) in cell types it had never seen, yielding results that agreed closely with experimental data.

New AI methods reveal drivers of a pediatric cancer

Next, the investigators showed the power of their AI system when they asked it to uncover still-hidden biology of diseased cells, in this case, an inherited form of pediatric leukemia.

"These kids inherit a gene that is mutated, and it was unclear exactly what it is these mutations are doing," says Rabadan, who also co-directs the cancer genomics and epigenomics research program at Columbia's Herbert Irving Comprehensive Cancer Center.

With AI, the researchers predicted that the mutations disrupt the interaction between two different transcription factors that determine the fate of leukemic cells. Laboratory experiments confirmed AI's prediction. Understanding the effect of these mutations uncovers specific mechanisms that drive this disease.

AI could reveal 'dark matter' in genome

The new computational methods should also allow researchers to start exploring the role of genome's "dark matter"—a term borrowed from cosmology that refers to the vast majority of the genome, which does not encode known genes—in cancer and other diseases.

"The vast majority of mutations found in cancer patients are in so-called dark regions of the genome. These mutations do not affect the function of a protein and have remained mostly unexplored," says Rabadan. "The idea is that using these models, we can look at mutations and illuminate that part of the genome."

Rabadan is working with researchers at Columbia and other universities, exploring different cancers, from brain to blood cancers, learning the grammar of regulation in normal cells, and how cells change in the process of cancer development.

The work also opens new avenues for understanding many diseases beyond cancer and potentially identifying targets for new treatments. By presenting novel mutations to the computer model, researchers can now gain deep insights and predictions about exactly how those mutations affect a cell.

Coming on the heels of other recent advances in artificial intelligence for biology, Rabadan sees the work as part of a major trend: "It's really a new era in biology that is extremely exciting; transforming biology into a predictive science."

More information: Raul Rabadan et al, A foundation model of transcription across human cell types, *Nature* (2025). DOI: 10.1038/s41586-024-08391-z. www.nature.com/articles/s41586-024-08391-z

Provided by Columbia University Irving Medical Center