# Columbia University develops AI model to guide gene therapies



Researchers at Columbia University have developed a groundbreaking artificial intelligence model named the General Expression Transformer (GET), designed to predict how the genes within a cell influence its behaviour. This advancement offers significant potential to enhance understanding of cancer and various genetic diseases, with implications for the creation of cell-specific gene therapies.

Using methodologies akin to those employed in the development of popular language processing programs like ChatGPT, GET has been trained to decipher the complex rules that govern gene expression. Gene expression is a fundamental biological process that dictates how genes are activated—not only determining the types of proteins produced by a cell, but also controlling the quantities of these proteins. Such proteins are critical to diverse bodily functions, from movement to cognitive processes.

The newly developed model is in its nascent stages but could potentially parallel the role of AlphaFold2, an AI that accurately predicts protein structures and was recently awarded the Nobel Prize in chemistry for its contributions to science. As noted by Raul Rabadan, director of the Program for Mathematical Genomics at Columbia and a key author of the research published in the journal Nature, “Biology is being transformed into something that is a predictive science. We’re seeing a revolution in biology.”

Mark Gerstein, a professor at Yale School of Medicine, highlighted the systematic efforts over the last 15 to 20 years by scientists to make predictions regarding gene regulation, derived from extensive datasets. These datasets encapsulate various human cell types, measuring gene expression and the interaction of transcription factors—proteins pivotal for gene regulation.

In an innovative departure from prior research, the training of GET utilised data from normal human tissue cells rather than abnormal cells associated with cancers. Xi Fu, a graduate student in Rabadan’s lab, led this effort by analysing data from over 1.3 million cells across 213 different types within the human body. The findings revealed that GET could predict outcomes for a specific cell type, like astrocytes from the central nervous system, even when this type was not part of its training dataset.

Mike Pazin from the National Human Genome Research Institute praised the model's ability to learn and apply knowledge across cell types: “In some ways, it’s like if I handed somebody a bunch of books in English, and said: ‘Okay, now this is in Russian. What does it say?’ … I would be like, ‘Wow, is that possible?’”

Jian Ma, a professor of computational biology at Carnegie Mellon University, explained the importance of this work in tackling one of biology’s core challenges: understanding how identical genomes can result in a vast array of behaviours across different cell types. He remarked that while a human body’s cells all share the same DNA, they express distinct gene sets based on their type—about 20,000 genes that can be selectively turned on or off.

The implications of mastering this ‘regulatory grammar’ are significant for human health, according to Yang E. Li from Washington University School of Medicine. “We want to learn the grammar and prioritise the key players in different cell types,” he stated, highlighting how disruptions in this grammar can lead to various diseases.

Researchers are optimistic that GET’s predictive capabilities could aid in designing targeted gene therapies, potentially correcting genetic mutations without affecting other cell types. Rabadan noted that the model could facilitate the creation of gene therapies that express genes solely in affected cell types, potentially revolutionising treatment strategies for diseases.

Additionally, the ability to accurately predict gene regulation may assist scientists in narrowing down the overwhelming number of experiments required to decipher the effects of genetic mutations, especially in complex diseases like cancer, which may exhibit thousands of mutations. Rabadan pointed out the difficulty faced by scientists in determining which mutations are significant, suggesting a more streamlined approach to research may emerge thanks to advancements in AI models like GET.

Source: [Noah Wire Services](https://www.noahwire.com)

## Bibliography

1. <https://www.biorxiv.org/content/10.1101/2023.09.24.559168v1.full> - This article introduces the General Expression Transformer (GET) model, its training on chromatin accessibility data and sequence information, and its ability to predict gene expression in unseen cell types with experimental accuracy.
2. <https://www.biorxiv.org/content/10.1101/2023.09.24.559168v1.full> - It details the model's performance in predicting regulatory activity, inferring regulatory elements and regulators, and identifying physical interactions between transcription factors across 213 human fetal and adult cell types.
3. <https://www.biorxiv.org/content/10.1101/2023.09.24.559168v1.full> - The article highlights GET's ability to predict gene expression in specific cell types, such as astrocytes, even when these types were not part of the training dataset.
4. <https://www.biorxiv.org/content/10.1101/2023.09.24.559168v1.full> - It explains how GET learns transcriptional regulatory syntax from chromatin accessibility data and successfully predicts gene expression in both seen and unseen cell types.
5. <https://www.biorxiv.org/content/10.1101/2023.09.24.559168v1.full> - The article discusses the model's versatility in being transferred to different sequencing platforms and measurement techniques, and its potential as a prescreening tool for cell type-specific regulatory elements.
6. <https://www.biorxiv.org/content/10.1101/2023.09.24.559168v1.full> - It mentions the model's outperformance of previous state-of-the-art models in identifying cis-regulatory elements and novel and known upstream regulators of fetal hemoglobin.
7. <https://academic.oup.com/bioinformatics/article/40/5/btae189/7649318?login=false> - Although not directly related to GET, this article on TransGEM illustrates the broader context of using Transformer-based models for gene expression and molecular biology, supporting the innovative use of AI in biological research.
8. <https://academic.oup.com/bib/article/25/1/bbad438/7459582?login=false> - This article on GPformer, another Transformer-based model, highlights the general trend of using deep learning and Transformer architectures in genomic and gene expression predictions, reinforcing the methodology used in GET.
9. <https://www.biorxiv.org/content/10.1101/2023.09.24.559168v1.full> - The article emphasizes the importance of understanding the 'regulatory grammar' of gene expression, which is crucial for human health and the development of targeted gene therapies.
10. <https://www.biorxiv.org/content/10.1101/2023.09.24.559168v1.full> - It discusses the potential of GET in aiding the design of targeted gene therapies and streamlining research on genetic mutations, particularly in complex diseases like cancer.
11. <https://www.biorxiv.org/content/10.1101/2023.09.24.559168v1.full> - The article underscores the revolutionary impact of AI models like GET on biology, transforming it into a predictive science, as highlighted by Raul Rabadan and other researchers.
12. <https://www.washingtonpost.com/science/2025/01/09/ai-predicts-gene-activity/> - Please view link - unable to able to access data