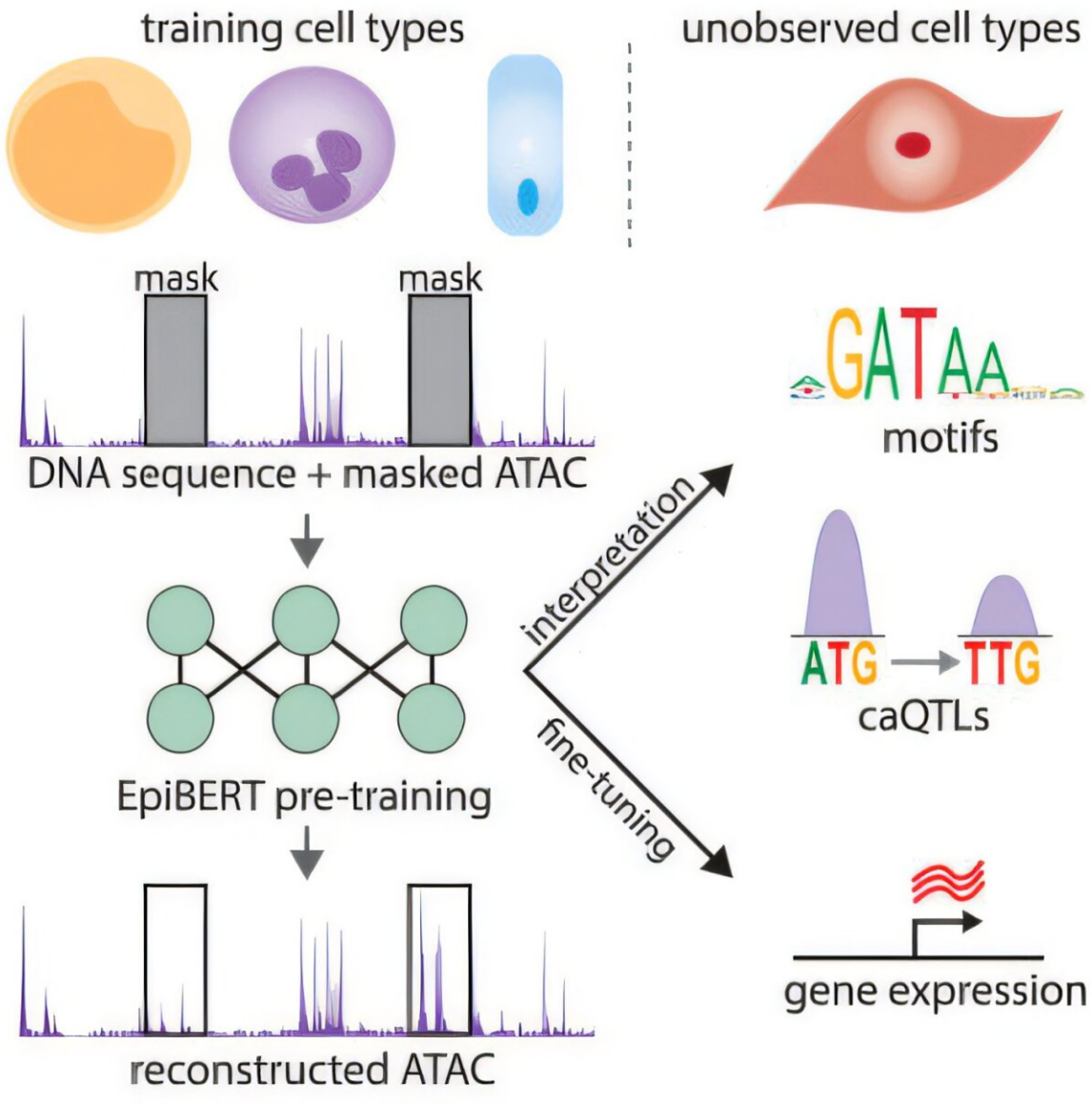
AI model learns generalized 'language' of regulatory genomics, predicts cellular stories

January 29 2025



Credit: *Cell Genomics* (2025). DOI: 10.1016/j.xgen.2025.100762

A team of investigators from Dana-Farber Cancer Institute, The Broad Institute of MIT and Harvard, Google, and Columbia University have created an artificial intelligence model that can predict which genes are expressed in any type of human cell. The model, called EpiBERT, was inspired by BERT, a deep learning model designed to understand and generate human-like language.

The work appears in *Cell Genomics*.

Every cell in the body has the same genome sequence, so the difference between two types of cells is not the genes in the genome, but which genes are turned on, when, and how many. Approximately 20% of the genome codes for regulatory elements determine which genes are turned on, but very little is known about where those codes are in the genome, what their instructions look like, or how mutations affect function in a cell.

EpiBERT was trained on data from hundreds of human cell types in multiple phases. It was fed the genomic sequence, which is 3 billion base pairs long, along with maps of chromatin accessibility that inform which of these sequences are unwound from the chromosome and read by the cell.

The model was first trained to learn the relationship between DNA sequence and chromatin accessibility across large chunks of the genome in a specific cell type. It then used these learned relationships to predict which genes were active in the corresponding cell type. It accurately identified regulatory elements—parts of the genome recognized by transcription factors—and their influence on gene expression across

many cell types, building a "grammar" that is generalizable and predictable.

This grammar-building process can be likened to the way a large language model, such as ChatGPT, learns to build meaningful sentences and paragraphs from many examples of text. The EpiBERT model can process accessibility and predict functional bases as well as RNA expression for a never-before-seen cell type.

EpiBERT will shed light on how genes are regulated in cells, and potentially, how the regulatory systems of those cells can be mutated in ways that lead to diseases such as cancer.

More information: Nauman Javed et al, A multi-modal transformer for cell type agnostic regulatory predictions, *Cell Genomics* (2025). DOI: 10.1016/j.xgen.2025.100762

Provided by Dana-Farber Cancer Institute