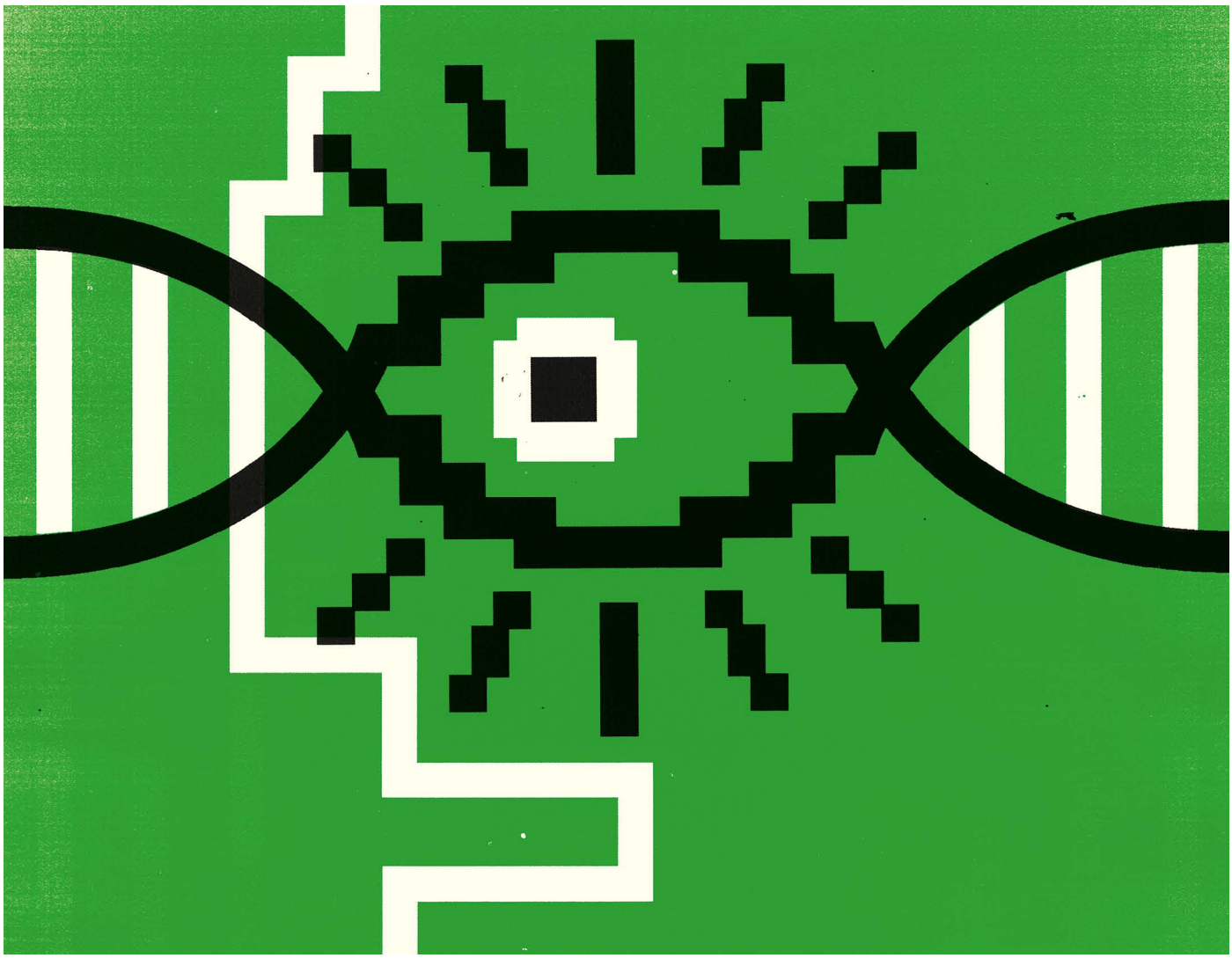

ILLUSTRATION: THE PROJECT TWINS

# DECODING THE GENOME

Scientists are seeking to decipher the role of non-coding DNA in the human genome, helped by a suite of artificial-intelligence tools. **By Jeffrey M. Perkel**

In 1862, Victor Hugo reportedly wrote to his publisher to ask how his newly published novel *Les Misérables* was selling, with a single character query: "?" The response: "!"

This story of one of the world's most concise correspondences is apocryphal. But some genome-focused artificial intelligence (AI) systems can, like the French writer's publisher, respond meaningfully to equally short prompts.

Instead of the detailed queries required to use the chatbot ChatGPT effectively, Evo, an AI model trained on some 300 billion nucleotide bases, including 80,000 microbial whole-genome sequences, will – prompted with '#' – dream up a new sequence of mobile DNA. It does so on the basis of other such biological systems that the model has been exposed to (see go.nature.com/3jvp922). Given a prompt such as '030', an AI tool called regLM can spit out 200-base sequences that are predicted to exhibit regulatory activity in any of three human cell lines (go.nature.com/4jpttm8).

Evo and regLM are part of a fast-growing suite of tools that aim to internalize, decode, interpret and build on the grammar of the genome – especially the vast portion that does not code for proteins. Think AlphaFold, but for regulatory DNA, which are sequences that control gene expression.

When Google DeepMind released AlphaFold in 2020, the company claimed it had solved a decades-old 'grand challenge' in biology – predicting a protein's 3D shape from its sequence alone. But the non-coding fraction of the genome could prove to be an even grander challenge.

A given sequence of amino acids will generally fold into the same shape, whatever the cellular context. That predictability is not true of the genome, in which short, functional sequence motifs – gene promoters and enhancers, transcription start and stop sites and so on – can be scattered across long stretches of seemingly purposeless DNA. These motifs might overlap, interact over long distances, bind to competing protein factors or respond to signals that are only present in specific cells or at certain times in development. They are also tightly wrapped within chromatin, a complex of DNA and protein, which might be more or less accessible to external proteins depending on what the cell is doing.

"How proteins are encoded in the genome, the code of how genes are expressed, when and where, how much – is one of the most fascinating problems in biology," says Stein Aerts, a computational biologist at the VIB Center for AI & Computational Biology and the Catholic University of Leuven (KU Leuven) in Belgium. But with training, AI tools can detect subtle differences between sequences and predict what they do and how they behave, identifying crucial motifs and even estimating the impact of altering them. From there, AI models can attempt to predict the physiological impact of genetic variants and even guide the design of new sequences with specified functions.

These tools are not perfect, and researchers cannot even agree on how best to assess their performance. But that makes the field exciting. "It's so clear that it's a solvable problem," says Julia Zeitlinger, a developmental and computational biologist at the Stowers Institute for Medical Research in Kansas City, Missouri, who developed an AI model called BPNet and uses it to decode the mechanistic sequence rules of gene regulation, "but it's not clear how".

PAUL JOSEPH

**Carl de Boer is a biomedical engineer at the University of British Columbia in Canada.**

## Of puppies and puffins

DeepSEA, one of the first genomic AI tools, was published[1] ten years ago this month by computational biologists Jian Zhou and Olga Troyanskaya at Princeton University in New Jersey.

DeepSEA is a convolutional neural network (CNN) – the same kind of deep-learning architecture used to teach computers to classify images as, say, a cat or a dog. Zhou and Troyanskaya trained a model on epigenetics data, including transcription-factor binding, chromatin accessibility and histone modifications, from a public research project called the Encyclopedia of DNA Elements (ENCODE). The model learnt to predict the presence of such features in 1,000-base segments of DNA it had never encountered.

DeepSEA's training enabled it to tease apart the biological consequence and severity of sequence variants associated with human disease. For instance, one breast-cancer-associated sequence variant called rs4784227 seems to strengthen the binding of a DNA-binding protein called FOXA1, whereas a variant associated with the blood condition α-thalassemia creates a possible binding site for GATA1, a transcription factor involved in blood-cell development.

Since then, the field has exploded. David Kelley, a principal investigator at the biotechnology company Calico Life Sciences in South San Francisco, California, has created or co-created multiple AI models, many with canine-inspired names. These include Akita[2] (for predicting 3D genome folding), Basset[3] and Basenji[4] (for regulatory-sequence prediction) and Borzoi[5], which predicts gene expression across the length of a gene.

These models raised a litter of variants: Basset begat Malinois, and Borzoi begat Scooby. Other researchers have built their own (non-canine) models including Puffin, ChromBPNet and more.

Not all are CNNs. Enformer – a model that predicts both gene expression and epigenetic data over long distances – and Borzoi, for instance, "use both convolution blocks and transformer blocks", says Kelley, whose laboratory developed both models. "The convolution blocks are great for capturing the local sequence patterns, and then the transformer blocks help look around a larger region to consider the local patterns in a broader context before predicting the data." But whatever the architecture, they come in two basic forms, says Anshul Kundaje, who researches computational genomics at Stanford University in California. Supervised and sequence-to-function models are trained on functional genomic data – gene expression or chromatin accessibility, for instance – and learn to predict the function of DNA sequences they have never encountered. Often working at or near single-nucleotide resolution, these models can identify key motifs, such as functionally important protein-binding sites, and predict the significance of altering them. DeepSEA is one; Kundaje's ChromBPNet, which predicts regions of chromatin accessibility, another.

**"There's just so many different biochemical mechanisms that could happen on DNA."**

The other class is unsupervised or self-supervised 'genomic language models' (gLMs). Like ChatGPT, they are trained on vast quantities of text – in this case, genomic sequence data – and are tasked with either predicting the next base (or 'token') in a sequence or filling in missing bases on the basis of surrounding context. These models "are not trying to predict the activity of a sequence, they're trying to predict the composition of a sequence", says Avantika Lal, a machine-learning scientist at biotechnology firm Genentech in South San Francisco.

With machine-learning scientist Gökçen Eraslan and their colleagues at Genentech, Lal co-developed regLM, a language model that they trained by labelling regulatory sequences with succinct markers of activity[6] – for instance, '04<sequence>' to indicate strong expression in one cell line and low activity in another. The model is therefore not strictly unsupervised, says Eraslan – he calls it a 'function-to-sequence' model. But those same labels can then be used to prompt regLM to create new sequences with predicted behaviours.

Evo 2, announced in February[7], was trained on 9.3 trillion DNA base pairs – "a representative snapshot of genomes spanning all observed evolution", as the resulting bioRxiv preprint paper puts it. It could then identify intron–exon boundaries, predict the impacts of mutations and generate 'realistic' gene and genomic sequences, among other things.

## Models made simple

Genomic AI models can also be distinguished by the type of regulatory interactions they predict, Kundaje says. Sequence-to-function models mostly identify important DNA motifs (which, because their function depends on their proximity to the regulated gene, are said to act in *cis*) without regard to the biology that occurs there.

*Trans* models, by contrast, aim to identify which genes regulate which other genes, for instance, to tease apart networks of gene regulation. (They are called *trans* because the

factors that mediate this regulation act at a distance.) But this, says Kundaje, "is still very fraught and very problematic" because *trans* models — which are generally trained on data such as RNA expression — must infer causal relationships without data that can reveal causality. There's no guarantee that two genes are directly linked just because their expression rises and falls in tandem. Even if they are, it's not necessarily obvious in which direction the relationship works: does A regulate B or vice versa? If these models are then asked to predict the impact of a perturbation — for example, what happens if a given gene is knocked out — the models often fail.

Models can include both *cis* and *trans* elements, says Sushmita Roy, a computational biologist at the University of Wisconsin–Madison, for instance by building regulatory networks on the basis of chromatin accessibility data and weighting those predictions by gene expression. But perhaps the first model to truly bridge the divide, Kundaje says, is Scooby — a single-cell version of Borzoi (go.nature.com/3upffnp). By leveraging both chromatin accessibility and transcriptional data from the same cells, Scooby predicts genome features and cell state simultaneously. "It is one of the first *cis–trans* models," he says.

Sequence-to-function models can also probe other aspects of gene regulation. In 2024, teams led by Zhou (who is now at the University of Texas Southwestern Medical Center in Dallas), Kundaje and Charles Danko, a computational biologist at Cornell University in Ithaca, New York, independently described sequence-to-function models capable of predicting sites of transcription initiation[8–10].

Zhou used his team's model, Puffin, to identify the common features and placement of key regulatory elements around sites of transcription initiation, including binding sites for the transcription factors YY1, SP1, CREB and Initiator. Danko's team trained its AI model on matched genome sequences and transcription initiation data from 58 individuals, creating a suite of models that were, he says, "for the first time aware of how differences between individuals in their genome sequence influence the pattern" of transcription initiation.

Collectively, says Zhou, these studies begin to tease apart the motifs that regulate the positioning and strength of transcription initiation, including that of the transcription factor TFIID. TFIID is an essential protein complex that binds to the promoter element known as a TATA box — despite the fact that most eukaryotic promoters don't seem to contain a TATA box. "One mechanistic interpretation is that TFIID is binding the best available of the 'bad options' when it picks a site" in a TATA-less promoter, Danko explains.

Most genomic models make these predictions from relatively small inputs — anywhere from a few hundred to a few thousand bases. But gene regulation can occur over much longer tracts of genome space, and some models are able to make predictions at or near those scales. Borzoi, for instance, accepts 524 kilobases of input DNA, and Evo 2 and Google DeepMind's newly announced AlphaGenome can work with a megabase.

These models can transform those sequences into vast collections of estimated data. Given an input sequence of 196,608 bases of human DNA, for instance, Enformer outputs 2,131 predictions of transcription factor binding, 1,860 of histone modifications, 684 of chromatin accessibility and 638 of gene expression, at 128-base resolution (go.nature.com/4mbe42h).

## A finite genome

Yet despite these models' extensive 'receptive fields', they can still miss things, says Jacob Schreiber, a computational biologist at the Research Institute of Molecular Pathology in Vienna, because enhancers might exert effects that are biologically meaningful but invisible to the AI tool. "We have not cracked long-range regulation," he says.

Another challenge is that, as vast as it is, the human genome is finite — there are only about 20,000–25,000 genes, for instance, and only a fraction of those are regulated in a cell-type-specific manner. That means that for all those billions of bases, there are relatively few examples of regulatory strategies from which a model can learn.

"There's just so many different biochemical mechanisms that could happen on DNA that there are probably a very large number of them that only occur once or even zero times in our genome sequence," says biomedical engineer Carl de Boer at the University of British Columbia in Vancouver, Canada.

One approach to broadening an AI model's knowledge base is to feed it more than just reference genomes. Some model builders, for instance, train their tools on data from multiple individuals or from across the phylogenetic tree to give the models a sense of genetic diversity.

Another approach, advanced by de Boer and Jussi Taipale, a systems biologist at the University of Cambridge, UK, is to look beyond natural genomes to fully artificial DNAs[11].

As a postdoc at the Broad Institute of MIT and Harvard in Cambridge, Massachusetts, de Boer and his colleagues tested some 100 million random sequences, each of which were 80 nucleotides in length — "about a human genome's worth" — for their ability to drive expression of a fluorescent protein in yeast (*Saccharomyces cerevisiae*)[12]. (The yeast genome is made up of about 12 million bases, compared with roughly 3 billion in the human genome.) This approach, de Boer says, "is actually much better" for understanding the grammar of the genome than using genomic DNA, "because all of the signals you see in the random DNA are causal". If you see fluorescence, the sequence is active. The genome, by contrast, is a product of evolution, meaning elements might be positioned owing to selective pressures as well as function.

According to de Boer, the yeast exercise yielded two key insights. First, it reinforced that "there are probably widespread biophysical interactions happening in regulatory regions". Functional motifs were not randomly arranged in active sequences; they were positioned in specific configurations — for instance, to conform to the helical spacing of the DNA double helix.

The second insight involved the importance of low-affinity transcription-factor–DNA interactions. Even weak interactions, the team found, could exert a large influence on gene regulation, just as relatively weak chemical interactions can hold two proteins together.

TERRA BLEVINS

**Computational biologist Jacob Schreiber co-developed an algorithm called Ledidi.**

Beyond studies of genomic grammar and gene regulation, researchers can use AI models for genetic fine-mapping — identifying which human genetic variants that have been identified in genome-wide association studies (GWAS) are causal for a certain phenotype. As many as 95% of sequence differences identified in GWAS are found in non-coding DNA[13]. Researchers can also use genomic AI tools to probe mutations *in silico*, to better understand the impact of genetic variation.

And then there's sequence design. From an engineering standpoint, successfully designing a sequence from scratch demonstrates that researchers (or their AI models) have learnt something fundamental about the genome, says Zhou. "We can use this as a way to verify our understanding," he says. More practically, it can also be used to create bespoke sequences that can limit gene expression to a specific time and place, for instance for gene-therapy applications, or to design sequences that respond to specific stimuli. "I think that the applications for gene therapy and cell therapy are very clear," says Lal.

Several papers have demonstrated this approach. In 2024, Aerts and transcription biologist Alexander Stark at the Research Institute of Molecular Biology in Vienna independently reported using sequence-to-function models as 'oracles' to select and evolve sequences that would have desired behaviour in fruit flies (*Drosophila melanogaster*)[14,15], and in Aerts's case, human cells as well. Geneticist Ryan Tewhey at the Jackson Laboratory in Bar Harbor, Maine, and his team used reporter assay data to train a derivative of the Basset model, which they then used to design sequences that were active in blood, liver and neuronal cells, as well as in zebrafish and mice[16].

These studies do tend to use cell types that are highly divergent, Kundaje notes. Practical applications, for instance in gene therapy, will probably require targeting specific cell types at particular points in development, which is a harder nut to crack.

Still, the resulting elements can reveal remarkable subtleties. Aerts's team has observed overlapping regulatory codes[14], for instance — and found that it is possible to alter a piece of regulatory DNA so that only one code functions. They also discovered 'near enhancers' that could be converted into regulatory DNA with a handful of mutations — an observation that underscores how a single genetic change can inadvertently activate previously silent genes. "The creation of a new enhancer is not so difficult," he says. And the team showed that it could design sequences to target different cells from the same starting sequence.

That's not to say that the AI model itself is designing DNA. Rather, these strategies tend to take a starting sequence, use AI to select the best performers, modify each base *in silico*, and repeat. To optimize the process, Schreiber, with Stark and computational biologist William Noble at the University of Washington in Seattle, developed an algorithm called Ledidi[17]. Rather than computationally testing every possible mutation, Schreiber explains, Ledidi seeks the minimum set of edits required to impart a desired behaviour.

According to Schreiber, the software can use multiple oracles to optimize several activities at once. As a result, it is possible to use Ledidi to design extremely subtle changes, such as decreasing chromatin accessibility in a specific region without affecting the binding of a specific protein. It can also create a suite of solutions, called an affinity catalogue to help researchers to better investigate transcription biology.

Evo 2 and regLM, by contrast, are generative: given a prompt, they spit out a new sequence from scratch. In one study[18], for instance, chemical engineer Brian Hie at Stanford University and his colleagues used one version of Evo to generate new toxin–antitoxin protein pairs, which some bacteria use as a defence against viruses.

**"There are probably widespread biophysical interactions happening in regulatory regions."**

Aerts tested a generative model in his study[14], too, finding it effective but less interesting for deciphering the *cis*-regulatory code. Using an iterative design process, he explains, it's possible to probe sequences after each round of changes to gain insights into the biology of regulatory DNA. Take Schreiber's experience with an affinity catalogue that he built for the transcription factor GATA2, for instance[17]. As he studied the different solutions the model came up with, he found that some relatively weak sequences had more GATA2 motifs, whereas stronger ones did not. "The model had learnt a really sophisticated *cis*-regulatory code," he says. "It was playing with the affinity of these motifs, their spacing relative to each other, and the presence of co-factors."

## Up for debate

Researchers agree that sequence-to-function models generally work as advertised. But what they and other AI models can and should be used for remains up for debate.

In a pair of studies[19,20] published in late 2023, computational biologists Nilah Ioannidis at the University of California, Berkeley, and Sara Mostafavi at the University of Washington in Seattle and their teams independently demonstrated that genomic models struggle with a key task: explaining why the variation in gene expression differs from person to person — why one person expresses a given gene more than another, given an individual's unique constellation of gene variants. "They actually do very poorly at this task," says Ioannidis.

To complicate matters, existing benchmarks of AI performance don't necessarily address the questions that researchers want to ask, Ioannidis notes, especially regarding interpersonal genetic variation. Kundaje's team has created its own benchmark tool, called DART-Eval, to help make the evaluation of models more systematic and comprehensive. So has Ioannidis, with her tool called GUANinE.

Kundaje is even less impressed with unsupervised models. Although they perform well with coding sequences and small genomes, when it comes to mammalian regulatory DNA, he says, "I would consider them catastrophic failures."

Unsupervised models are largely unaware of the many layers of epigenetic regulation, such as transcription-factor binding and histone modification, that make genomes work. They have no idea what happens when sequences are perturbed. And of course, not every base in the genome is meaningful. Still, genomics-AI advocates see a promising future. Casey Greene, a bioinformatician at the University of Colorado Anschutz Medical Campus in Aurora, foresees a day when AI tools will be able to design an exquisitely fine-tuned piece of DNA just as natural-language instructions can generate code today, a process sometimes called vibe coding. "I want to vibe code the genome," he says.

Whether he'll be able to do that with a single-character prompt remains to be seen.

**Jeffrey M. Perkel** is Technology Editor at *Nature*.


1. Zhou, J. & Troyanskaya, O. G. *Nature Methods* **12**, 931–934 (2015).
2. Fudenberg, G., Kelley, D. R. & Pollard, K. S. *Nature Meth.* **17**, 1111–1117 (2020).
3. Kelley, D. R., Snoek, J. & Rinn, J. L. *Genome Res.* **26**, 990–999 (2016).
4. Kelley, D. R. *et al. Genome Res.* **28**, 739–750 (2018).
5. Linder, J., Srivastava, D., Yuan, H., Agarwal, V. & Kelley, D. R. *Nature Genet.* **57**, 949–961 (2025).
6. Lal, A., Garfield, D., Biancalani, T. & Eraslan, G. *Genome Res.* **34**, 1411–1420 (2024).
7. Brixi, G. *et al.* Preprint at bioRxiv https://doi.org/10.1101/2025.02.18.638918 (2025).
8. Dudnyk, K., Cai, D., Shi, C., Xu, J. & Zhou, J. *Science* **384**, eadj0116 (2024).
9. He, A. Y. & Danko, C. G. Preprint at bioRxiv https://doi.org/10.1101/2024.03.13.583868 (2024).
10. Cochran, K. *et al.* Preprint at bioRxiv https://doi.org/10.1101/2024.05.28.596138 (2024).
11. de Boer, C. G. & Taipale, J. *Nature* **625**, 41–50 (2024).
12. de Boer, C. G. *et al. Nature Biotechnol.* **38**, 56–65 (2020).
13. Schipper, M. & Posthuma, D. *Hum. Mol. Genet.* **31**, R73–R83 (2022).
14. Taskiran, I. I. *et al. Nature* **626**, 212–220 (2024).
15. de Almeida, B. P. *et al. Nature* **626**, 207–211 (2024).
16. Gosai, S. J. *et al. Nature* **634**, 1211–1220 (2024).
17. Schreiber, J. *et al.* Preprint at bioRxiv https://doi.org/10.1101/2025.04.22.650035 (2025).
18. Merchant, A. T., King, S. H., Nguyen, E. & Hie, B. L. Preprint at bioRxiv https://doi.org/10.1101/2024.12.17.628962 (2024).
19. Huang, C. *et al. Nature Genet.* **55**, 2056–2059 (2023).
20. Sasse, A. *et al. Nature Genet.* **55**, 2060–2064 (2023).