Large genome model: Open source AI trained on trillions of bases

Late in 2025, we covered the development of an AI system called Evo that was trained on massive numbers of bacterial genomes. So many that, when prompted with sequences from a cluster of related genes, it could correctly identify the next one or suggest a completely novel protein.

That system worked because bacteria tend to cluster related genes together—something that's not true in organisms with complex cells, which tend to have equally complex genome structures. Given that, our coverage noted, "It's not clear that this approach will work with more complex genomes."

Apparently, the team behind Evo viewed that as a challenge, because today it is describing Evo 2, an open source AI that has been trained on genomes from all three domains of life (bacteria, archaea, and eukaryotes). After training on trillions of base pairs of DNA, Evo 2 developed internal representations of key features in even complex genomes like ours, including things like regulatory DNA and splice sites, which can be challenging for humans to spot.

Genome features

Bacterial genomes are organized along relatively straightforward principles. Any genes that encode proteins or RNAs are contiguous, with no interruptions in the coding sequence. Genes that perform related functions, like metabolizing a sugar or producing an amino acid, tend to be

clustered together, allowing them to be controlled by a single, compact regulatory system. It’s all straightforward and efficient.

Eukaryotes are not like that. The coding sections of genes are interrupted by introns, which don’t encode for anything. They’re regulated by a sequence that can be scattered across hundreds of thousands of base pairs. The sequences that define the edges of introns or the binding sites of regulatory proteins are all weakly defined—while they have a few bases that are absolutely required, there are a lot of bases that just have an above-average tendency to have a specific base (something like “45 percent of the time it’s a T”). Surrounding all of this in most eukaryotic genomes is a huge amount of DNA that has been termed junk: inactive viruses, terminally damaged genes, and so on.

https://arstechnica.com/science/2026/03/large-genome-model-open-source-ai-trained-on-trillions-of-bases/