New AI model reads and generates genetic code across all domains of life

Bill Shaw
11 March 2026

Scientists have developed an AI model capable of reading, analyzing and generating genetic code across all known domains of life—a development with vast implications for understanding human disease, designing new treatments and advancing biological knowledge on a scale previously impossible.

The model, called Evo 2, was published in the journal *Nature* on March 4 by a team of researchers at the Arc Institute, a nonprofit biomedical research organization based in Palo Alto, California. Unlike commonly used AI models such as ChatGPT and Anthropic's Claude, which are built from text written in human languages, Evo 2 was trained entirely on DNA sequences—approximately 9 trillion base pairs drawn from bacteria, plants, animals and every other domain of life.

Patrick Hsu, co-founder and core investigator at the Arc Institute and co-senior author on the paper, told phys.org:

> Our development of Evo 1 and Evo 2 represents a key moment in the emerging field of generative biology, as the models have enabled machines to read, write, and think in the language of nucleotides.

The potential applications of such a model are revolutionary. A tool that can predict which genetic variations cause disease, generate plausible new DNA sequences, and identify the functional properties of genes across all of biology could dramatically accelerate the development of new medicines, gene therapies, and diagnostic tools. It could transform the understanding and treatment of cancer, genetic disorders, autoimmune diseases and infectious diseases. Under conditions of rational, scientifically planned social organization, such capabilities could be made available to all of humanity.

Under capitalism, however, the benefits of such breakthroughs are inevitably channeled toward profit. The pharmaceutical giants and biotech firms already developing applications on the basis of open biological AI models will patent the downstream treatments and price them to maximize shareholder returns—not to improve public health. The working class, which produces the social wealth that makes such research possible, will be largely denied access to the life-saving treatments that emerge from it.

Building the model

To build Evo 2, the scientists compiled DNA sequences from nearly 10 public genome databases into a single massive dataset called OpenGenome2. At 5.5 terabytes—far exceeding the storage capacity of a typical laptop or workstation—the dataset reflects the enormous scale of the undertaking. These sequences were contributed by hundreds of scientists across the globe and made freely available for public use—a testament to the collaborative, non-proprietary character of scientific labor that strains against the imperatives of capitalist competition.

There are two major versions of the model: Evo 2 7B, with 7 billion parameters trained on 2.3 trillion base pairs, and Evo 2 40B, with 40 billion parameters trained on the full dataset. The larger model is more powerful but requires substantially greater computational resources.

The creation of Evo 2 was made possible by StripedHyena 2, a new computational architecture that enabled training on 30 times more data than Evo 2's predecessor, Evo 1, while processing sequences of up to 1 million nucleotides at once—far longer than any previous biological AI model.

After building the model, the scientists assessed its ability to perform a range of tasks: predicting the effects of genetic mutations, identifying disease-causing variations in human DNA, detecting functional properties of different regions of the genome and generating entirely new DNA sequences.

Evo 2 successfully predicted that mutations in critical areas of DNA would be highly damaging—a well-known biological fact, but one that the model was never explicitly programmed with. This ability emerged entirely from patterns in the raw sequence data.

The model also accurately predicted whether human genetic variants—a term scientists now prefer to "mutation," since not all variations cause disease—would lead to illness. For insertions and deletions in DNA sequences, Evo 2 outperformed all existing tools. For simpler, single-letter changes in the genetic code, it performed comparably to the best tools that had not been trained on labeled examples, though it fell short of specialized models trained on curated datasets.

The distinction is important: Evo 2 is an "unsupervised" model, meaning it learned solely from raw DNA sequences without being told what to look for. Models that are trained on data that has been labeled by scientists—so-called "supervised" models—have a built-in advantage for specific tasks. That Evo 2 can match or exceed such models on many tasks, despite learning from raw data alone, is a significant achievement.

Evo 2 also accurately identified a range of features within genomes. In bacteria, it correctly identified which genetic elements were capable of moving from one location to another in the genome. In humans, it accurately identified the boundaries between introns and exons—the segments of a gene that are cut out or retained when DNA is transcribed into the messenger RNA (mRNA) that serves as the template for building proteins. Not all such boundaries are known in the human genome, so an automated tool like Evo 2 has the potential to greatly advance biological knowledge in a short period of time.

Its ability to recognize these features emerged spontaneously from patterns in the sequence data—evidence that the model has independently developed something akin to an internal understanding of how DNA encodes RNA and proteins.

Generating new genetic code

Because Evo 2 is also a generative model, it can produce new DNA sequences using a shorter sequence as a starting prompt—analogous to how ChatGPT generates text in response to a written prompt.

The scientists tested this capability by providing Evo 2 with the first portion of a gene and asking it to complete the rest. In tests across six diverse species, the model generated between 70 and nearly 100 percent of the remaining gene accurately.

In a more ambitious test, they used Evo 2 to generate entire DNA sequences encoding complex cellular structures called mitochondria—the organelles responsible for producing energy in cells. In humans, the genes encoding mitochondrial components are scattered across all 23 chromosomes as well as in the mitochondria's own DNA. Using minimal prompting, Evo 2 generated the same types and numbers of genes as encode actual mitochondria, with high similarity to the real sequences.

The scientists also used Evo 2 to generate DNA sequences with high levels of "chromatin accessibility"—a property that determines whether a segment of DNA is physically accessible to the cellular machinery that activates genes. Working in concert with two other specialized models, Evo 2 was able to produce novel sequences with the desired properties, while simpler approaches failed.

It is important to note that, while these results are highly significant, the DNA sequences Evo 2 generates must still be tested in the real world. The authors acknowledge that their evaluation methods do not guarantee that generated genomes will be functional or capable of being replicated during cell division.

Open science and the profit system

The scientists have made all versions of Evo 2 and the OpenGenome2 dataset freely available on the HuggingFace model repository, consistent with the open-source ethos that pervades the best of modern scientific research.

Hsu noted:

> Evo 2 has a generalist understanding of the tree of life that's useful for a multitude of tasks, from predicting disease-causing mutations to designing potential code for artificial life. We're excited to see what the research community builds on top of these foundation models.

The collaborative character of the work that produced Evo 2 is striking. The DNA sequences at its foundation were contributed freely by scientists around the world, compiled from public databases spanning all domains of life. The AI architecture that made it possible was publicly available. And the finished model and its curated dataset were released back to the research community.

Yet this collaborative labor did not take place outside the profit system. Evo 2's largest model was trained on 2,048 NVIDIA H100 GPUs using NVIDIA's DGX Cloud platform on Amazon Web Services—resources provided through a formal partnership between the Arc Institute and NVIDIA, whose employees are among the paper's co-authors.

The Arc Institute itself was founded with $650 million from Silicon Valley billionaires, including Patrick Collison, the CEO of the $65 billion payments company Stripe, who is both a co-founder of the institute and a co-author on the Evo 2 paper. Greg Brockman, co-founder and president of OpenAI, contributed to the project's underlying architecture during a sabbatical. Both Collison and Brockman have ties to the Trump administration and the Israeli government, the chief perpetrators of the ongoing Gaza genocide and the imperialist war against Iran.

The contradiction is clear: the most advanced biological AI model in existence was produced through collaborative, non-proprietary scientific labor—yet it was incubated within corporate and philanthropic structures that are themselves products of the capitalist accumulation of wealth. The pharmaceutical and biotech companies that will utilize Evo 2 for commercial applications face no obligation to make the resulting treatments affordable or universally accessible and will not do so.

Tools like Evo 2 have the potential to revolutionize medicine—accelerating the discovery of treatments for cancer, genetic diseases and conditions that currently have no cure. They could extend healthy life expectancy globally, transform diagnostics and make personalized genomic medicine a reality for billions of people. But under capitalism, such advances are destined to enrich a privileged few. Already, the wealthiest layers of society have access to concierge medicine and bespoke healthcare services that the vast majority of the population cannot afford. AI-driven breakthroughs in genomic medicine will deepen this chasm unless the working class intervenes to reorganize society on a socialist basis.

Unlocking the full revolutionary potential of AI—in medicine, science, education, and every other domain—requires wresting control of these technologies from the financial oligarchy and placing them under the democratic control of the working class. The International Committee of the Fourth International (ICFI) has demonstrated in practice how AI can be placed in the service of the working class, launching Socialism AI in December 2025—the world's first revolutionary Marxist AI chatbot, built on the WSWS's archive of over 125,000 articles and the foundational works of Marx, Engels, Lenin and Trotsky, and designed to advance the political education and organization of workers and youth internationally.

The development of Evo 2 is a powerful demonstration that the most significant scientific advances emerge from collaboration, openness, and the free exchange of knowledge—principles that are fundamentally incompatible with the capitalist drive for private profit. The liberation of science and technology for the benefit of all of humanity requires the socialist reorganization of society by the international working class.