Artificial Intelligence & Environment

https://doi.org/10.66178/aie-0026-0010

News

Open Access

Evo 2 writes the book of life across all domains

Guang-Guo Ying*

Citation: Ying G-G. Evo 2 writes the book of life across all domains. *AI Environ.* 2026, 1(2): 55-56.

Received: 16 April 2026

Accepted: 22 May 2026

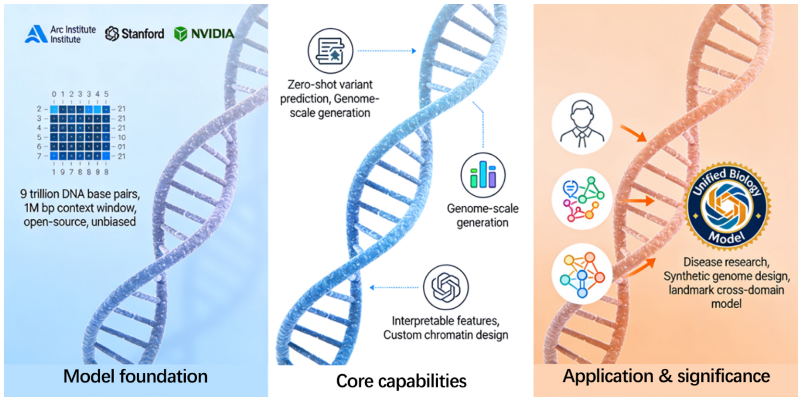
Published online: 8 June 2026

Abstract

Evo 2, a novel AI foundation model for biology developed by the Arc Institute, Stanford University, and NVIDIA, deciphers DNA grammar across all life domains with a record 1 million base pair context window. Trained on 9 trillion curated DNA base pairs, it achieves state-of-the-art zero-shot prediction of genetic variant effects, generates realistic genome-scale sequences, and features interpretable biological representations with cross-species transferability. Evo 2 also enables the design of DNA sequences with predefined chromatin accessibility patterns, validated in mouse and human cells. Fully open source with robust safety and equity design, it builds on prokaryote-focused Evo 1 and unlocks applications in human disease research and synthetic genome design. Though functional validation of generated sequences and computational efficiency need improvement, Evo 2 is a landmark advance toward a unified, cross-species model for reading, interpreting and writing genomic blueprints.

Keywords: Biology foundation model, DNA, Genome

Graphical abstract



A new artificial intelligence model that deciphers the grammar of DNA across the tree of life has demonstrated an unprecedented ability to predict the effects of genetic mutations and generate full genome-scale sequences. Named Evo 2, this model marks a major advance in the application of foundation models to biology, providing a single generalist tool that spans the central dogma from DNA to RNA to protein across both prokaryotes and eukaryotes[1].

Developed by a large multidisciplinary team led by researchers at the Arc Institute, Stanford University, and NVIDIA, Evo 2 was trained on a curated atlas of nearly 9 trillion DNA base pairs, sourced from over 8.8 trillion nucleotides of bacterial, archaeal, eukaryotic, and bacteriophage genomes. Described in *Nature*, the model is available in two versions: a 7-billion-parameter variant and a 40-billion-parameter variant. It is built on the novel StripedHyena 2 architecture, which enables a context window of up to 1 million base pairs—the longest of any DNA language model to date. This extended context allows Evo 2 to capture long-range genomic relationships, ranging from individual nucleotides to entire topologically associating domains[1].

One of the most remarkable features of Evo 2 is its capacity for zero-shot prediction of variant effects: it can assess the likelihood that a mutation is deleterious without any task-specific fine-tuning.

* Correspondence: Guang-Guo Ying (guangguo.ying@m.scnu.edu.cn)

Full list of author information is available at the end of the article.

https://doi.org/10.66178/aie-0026-0010

When tested against ClinVar, a database of clinically annotated human genetic variants, Evo 2 delivered state-of-the-art performance among unsupervised methods for noncoding variants, as well as insertions and deletions across both coding and noncoding genomic regions. It also exhibited strong performance for splice variants, ranking first among zero-shot models on SpliceVarDB. For *BRCA1*—a gene closely linked to cancer—Evo 2's likelihood scores distinguished loss-of-function variants from benign ones with high accuracy. Additionally, its internal embeddings were used to train a lightweight supervised classifier that achieved an area under the precision-recall curve (AUPRC) of 0.88, substantially outperforming zero-shot baselines[1].

Beyond predictive capabilities, Evo 2 is a powerful generative model. Given a short sequence prompt, it can complete entire genes; with a larger prompt, it can generate sequences that mimic whole genomes. For example, when prompted with a segment of the human mitochondrial genome, Evo 2 produced 16-kilobase sequences containing the expected set of protein-coding genes, tRNAs, and rRNAs in the correct syntenic order. Prompted with a 10.5-kilobase segment of *Mycoplasma genitalium*, the model generated nearly complete 580-kilobase genomes, with approximately 70% of the predicted genes sharing significant similarity with known protein families—a notable improvement over earlier models. Even for eukaryotic genomes, such as a 330-kilobase region of yeast chromosome III, Evo 2 generated sequences with biologically realistic gene and tRNA densities[1].

The research team also made significant progress in improving the model's interpretability by training sparse autoencoders on Evo 2's internal representations. These unsupervised analyses uncovered distinct features that correspond to well-characterized biological concepts. For instance, one feature is strongly activated in prophage regions of *Escherichia coli* and in CRISPR spacers, indicating that the model has learned an association between phage-derived sequences and bacterial immune elements. Other features capture protein secondary structure, exon-intron boundaries, and transcription factor binding motifs. Notably, the exon-associated features generalized to the woolly mammoth genome, demonstrating cross-species transferability.

One of the most practically ambitious applications of Evo 2 is the design of DNA sequences with predefined chromatin accessibility patterns. By coupling the generative model with a beam search algorithm guided by an ensemble of chromatin accessibility predictors (Enformer and Borzoi)[2,3], the researchers designed several thousand-base-pair sequences, which were then synthesized and introduced into mouse and human cells. In mouse embryonic stem cells, they engineered "Morse code" patterns of accessible and inaccessible chromatin, with experimentally measured accessibility profiles matching the intended patterns at an area under the receiver operating characteristic curve (AUROC) of 0.92–0.95. In human cell lines, Evo 2 enabled the design of both cell-type-specific and shared chromatin accessibility patterns across HEK293T and K562 cells, with over 90% of designs exhibiting the desired profiles.

Evo 2 and its training data are fully open source—a deliberate decision to accelerate research, paired with careful consideration of safety implications. The team excluded sequences from eukaryote-infecting viruses from the training set, and experiments show that the model performs poorly on human viral sequences in both prediction and generation tasks, reducing the risk of accidental misuse. Ancestry bias analyses further revealed that Evo 2's predictions do not systematically disadvantage non-European populations, addressing a well-known limitation of many existing variant effect predictors.

Evo 2 builds directly on its predecessor, Evo 1, which was trained exclusively on prokaryotic genomes[2]. The expansion to include eukaryotic genomes and the scaling of the context window to 1 million base pairs unlock a wide range of applications, from interpreting noncoding genetic variation in human disease to designing synthetic genomes with customized regulatory properties. The combination of a generalist foundation model with inference-time guidance—using predictive models to steer sequence generation—establishes a flexible paradigm for biological design, one that can be adapted to any research domain for which a predictive model exists.

As the authors note, significant work remains to be done. While Evo 2's generated sequences are biologically realistic by multiple metrics, their functional validity in living cells has not yet been demonstrated. The computational cost of guided generation remains high, and integrating experimental feedback via fine-tuning or reinforcement learning could further boost efficiency. Nevertheless, Evo 2 represents a pivotal step toward a unified model of biology—one that can read, interpret, and write genomic blueprints across the full diversity of life.

Author contributions

The author confirms sole responsibility and reviewed the results and approved the final version of the manuscript.

Data availability

Data sharing is not applicable to this article as no datasets were generated or analyzed during the current study.

Funding

No funding was received to assist with the preparation of this manuscript.

Competing interests

The author declare that they have no conflict of interest.

Author details

School of Environment & Environmental Research Institute, South China Normal University, Guangzhou, Guangdong 510006, China

References

[1] Brixi G, Durrant MG, Ku J, Naghipourfar M, Poli M, et al. 2026. Genome modelling and design across all domains of life with Evo 2. *Nature* 652:1349–1361

[2] Nguyen E, Poli M, Durrant MG, Kang B, Katrekar D, et al. 2024. Sequence modeling and design from molecular to genome scale with Evo. *Science* 386:ead9336

[3] Avsec Ž, Agarwal V, Visentin D, Ledsam JR, Grabska-Barwinska A, et al. 2021. Effective gene expression prediction from sequence by integrating long-range interactions. *Nature Methods* 18:1196–1203

Copyright: © 2026 by the author(s). Published by NEW HORIZON PRESS LIMITED, TSIM SHA TSUI, HK, China. This is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY 4.0) License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.