Genome Modeling And Design Across All Domains Of Life With Evo 2 · Concept Sketch

Genome-scale signal aggregation framing PRS vs. foundation model granularity.

Genome Modeling And Design Across All Domains Of Life With Evo 2

Authors: Garyk Brixi, Matthew G. Durrant, Jerome Ku, Michael Poli, Greg Brockman, Daniel Chang, Gabriel A. Gonzalez, Samuel H. King, David B. Li, Aditi T. Merchant, Mohsen Naghipourfar, Eric Nguyen, Chiara Ricci-Tam, David W. Romero, Gwanggyu Sun, Ali Taghibakshi, Anton Vorontsov, Brandon Yang, Myra Deng, Liv Gorton, Nam Nguyen, Nicholas K. Wang, Etowah Adams, Stephen A. Baccus, Steven Dillmann, Stefano Ermon, Daniel Guo, Rajesh Ilango, Ken Janik, Amy X. Lu, Reshma Mehta, Mohammad R.K. Mofrad, Madelena Y. Ng, Jaspreet Pannu, Christopher

Ré, Jonathan C. Schmok, John St. John, Jeremy Sullivan, Kevin Zhu, Greg Zynda, Daniel Balsam, Patrick Collison, Anthony B. Costa, Tina Hernandez-Boussard, Eric Ho, Ming-Yu Liu, Thomas McGrath, Kimberly Powell, Dave P. Burke, Hani Goodarzi, Patrick D. Hsu, Brian L. Hie

Year: 2025

Venue: bioRxiv preprint

---

1. Classification

- **Domain Category:**
  - **Genomics FM.** Evo 2 is a large-scale foundation model trained purely on DNA sequences from genomes across all domains of life, used for prediction and generation tasks that span molecular, genomic, and epigenomic levels.

- **FM Usage Type:**
  - **Core FM development.** The paper introduces a new large biological foundation model (Evo 2) together with a new multi-hybrid architecture (StripedHyena 2), training recipe, and evaluation suite, rather than just applying an existing FM.

- **Key Modalities:**
  - DNA sequence (genomic, organelle, and metagenomic).
  - RNA and protein effects (accessed via DNA-centric modeling and downstream assays, not as separate input modalities).
  - Epigenomic profiles (chromatin accessibility) used as downstream design targets via external predictive models (Enformer, Borzoi).

2. Executive Summary

This paper presents Evo 2, a large-scale biological foundation model that learns directly from DNA sequences across all domains of life to support both prediction and generation of genomic functions. The authors train 7B and 40B parameter models on 9.3 trillion nucleotides with context windows up to 1 million base pairs, enabling the model to reason over entire genomes rather than short fragments. Evo 2 is evaluated on a wide range of tasks, including zero-shot prediction of mutational effects on proteins, noncoding RNAs, and organismal fitness, as well as clinical variant effect prediction for human genes like BRCA1/2. They show that Evo 2 achieves competitive or state-of-the-art performance, especially for noncoding and splice variants, without task-specific fine-tuning and while remaining alignment-free (no multiple sequence alignments). Using mechanistic interpretability tools, they reveal that the model's internal features correspond to biologically meaningful concepts such as exons, introns, transcription factor motifs, protein secondary structures, and prophage regions. Finally, the paper demonstrates genome-scale DNA generation and guided "generative epigenomics," where Evo 2 is steered at inference time by epigenomic predictors to design sequences with specified chromatin accessibility patterns. Overall, Evo 2 is positioned as a generalist genomic language model and design engine that can underpin many downstream tasks in computational biology.

3. Problem Setup and Motivation

- **Scientific / practical problem**
  - How to build a single machine learning model that can **understand and design genomic DNA** across the full tree of life, from bacteria to humans, at scales ranging from short motifs to entire genomes.
  - The model should support **zero-shot mutational effect prediction** (how mutations affect fitness or function) across

proteins, RNAs, and noncoding regions, without relying on multiple sequence alignments or task-specific training.

- It should also enable **genome-scale sequence generation**, including organelle genomes and full chromosomes, and support **controllable design** of higher-level properties such as chromatin accessibility.

- **Why this is hard**
  - **Data scale and diversity:** Genomes vary widely in size and composition (small, compact prokaryotic genomes vs. large, intron-rich eukaryotic genomes with extensive noncoding and repetitive DNA). Capturing useful patterns across this diversity requires massive, carefully curated datasets.

  - **Long-range dependencies:** Many genomic functions (e.g., regulatory interactions, chromatin organization) depend on relationships spread over tens of kilobases to megabases, far beyond the context lengths of typical sequence models.

  - **Multiple biological "modalities" along the central dogma:** DNA variation influences RNA, protein, and organismal phenotypes; a useful model must implicitly encode patterns across these levels while only seeing DNA sequences.

  - **Noncoding and regulatory elements:** Noncoding variants and regulatory regions (enhancers, splice sites, chromatin features) are complex and "fuzzy," making them harder to model than coding regions where the genetic code provides a more direct mapping to function.

  - **Safety and biosecurity:** Large biological foundation models could in principle be misused (e.g., in viral design), so the training data and evaluation strategy must be carefully structured to limit harmful capabilities while retaining broad utility.

4. Data and Modalities

- **Datasets used**
  - **OpenGenome2**: a large, curated genomic dataset totaling **8.84–9.3 trillion nucleotides**, strongly expanding the earlier OpenGenome dataset.
    - Representative **prokaryotic genomes** (bacteria and archaea), expanded from earlier GTDB releases.
    - **Eukaryotic reference genomes** from NCBI (16,000+ genomes), focusing on primary assemblies and non-nuclear contigs.
    - **Organelle genomes** (mitochondria and others), plus **metagenomic** sequencing data.
    - **Function-focused subsets** around coding genes (windows near genes and other functional regions) to enrich training for downstream tasks.
  - **Evaluation datasets** (not all named explicitly in the excerpt, but outlined conceptually):
    - Deep mutational scanning (DMS) datasets for proteins and noncoding RNAs.
    - Human mRNA stability data (for decay rates).
    - Variant effect prediction datasets: ClinVar, SpliceVarDB, BRCA1/2 saturation mutagenesis.
    - Organismal fitness and gene essentiality datasets for bacteria, phage, and human lncRNAs.
- **Modalities**
  - **Primary input modality:** DNA sequence, tokenized at **single-nucleotide resolution**.
  - **Implicit biological modalities captured through tasks:** protein function, RNA function, noncoding regulatory function, and organismal fitness (all inferred from DNA variation).

- ◦ **Epigenomic modality:** chromatin accessibility tracks, used via **external predictors** (Enformer and Borzoi) for guided design, not as training inputs for Evo 2.
- **Preprocessing / representation**
  - ◦ DNA sequences are **byte-tokenized** and presented as long contiguous sequences, with a **context length up to 1 million tokens** during midtraining.
  - ◦ Training uses **sequence packing** and a reweighted cross-entropy loss that downweights repetitive regions to better calibrate model likelihoods between repetitive and non-repetitive DNA.
  - ◦ Lowercase annotations for repeats are used early in pretraining, then removed later so the model learns stable representations of repetitive vs non-repetitive regions.
  - ◦ Special tokens (e.g., stitch tokens and phylogenetic tags) are used to condition the model and are masked out in the loss.

5. Model / Foundation Model

- **Model Type**
  - ◦ Evo 2 is an **autoregressive genomic language model** built on **StripedHyena 2**, a **multi-hybrid architecture** that combines input-dependent convolutions with attention.
  - ◦ It is analogous to large language models (LLMs) in NLP but operates on nucleotide sequences and is optimized for extremely long contexts.
- **Is it a new FM or an existing one?**
  - ◦ **New FM.** Evo 2 extends the earlier Evo 1 models with a new architecture, larger parameter counts, vastly more training data, and a much longer context window (up to 1M bases).

- **Key components and innovations**
  - ◦ **StripedHyena 2 architecture:**
    - ▪ Uses multiple kinds of **input-dependent convolution operators** (Hyena-SE, Hyena-MR, Hyena-LI) interleaved with attention, arranged in a "striped" pattern.
    - ▪ Designed to balance **quality and efficiency**, giving higher throughput than Transformers and earlier StripedHyena variants, especially at long sequence lengths.
    - ▪ Achieves up to **3× speedup** at 1M context compared to optimized Transformers at 40B scale.
  - ◦ **Multi-phase training:**
    - ▪ **Pretraining phase** at shorter context lengths (1024→8192 tokens), enriched for genic windows and high-information regions to learn core biological features efficiently.
    - ▪ **Midtraining phase** for **context extension** to 1M tokens using rotary positional embeddings with positional interpolation and base frequency scaling, plus more whole-genome sequences.
  - ◦ **Model sizes and configuration:**
    - ▪ Evo 2 **40B**: ~40.3B parameters, 50 layers, hidden size ~8192, trained on **9.3T tokens**.
    - ▪ Evo 2 **7B**: ~6.5B parameters, 32 layers, trained on **2.4T tokens**.
    - ▪ Evo 2 **1B base**: smaller 1.1B variant trained on 1T tokens (also released).
  - ◦ **Loss and training tricks:**
    - ▪ Reweighted cross-entropy that reduces emphasis on repetitive DNA.
    - ▪ Context-parallel training infrastructure (Savanna) with 3D parallelism (data, tensor, context) and mixed precision (FP8 in some components).
    - ▪ A new **needle-in-a-haystack evaluation** to test long-context retrieval at up to 1M bases using a categorical Jacobian-based retrieval score.
- **Training setup (as far as available)**

Aspect	Details
Objective	Autoregressive next-token prediction on byte-tokenized DNA
Parameters	1.1B, 6.5B, and 40.3B variants
Tokens seen	Up to 9.3T tokens for Evo 2 40B
Context length	8,192 in pretraining, up to 1,048,576 ($\approx$1M) in midtraining
Architecture	StripedHyena 2 multi-hybrid (convolutions + attention + RoPE)
Optimizer	AdamW with cosine LR decay
Infrastructure	Custom stack (Savanna) with DeepSpeed, GPT-NeoX, Transformer Engine
Availability	Open-source model weights, training and inference code, OpenGenome2

6. Multimodal / Integration Aspects (If Applicable)

This paper does **not** focus on multimodal integration in the sense of combining heterogeneous data modalities like imaging, clinical features, and genomics into a single model or fusion pipeline. Evo 2 operates on DNA sequence alone and learns representations that implicitly capture

information relevant to proteins, RNAs, and organismal fitness; epigenomic aspects (chromatin accessibility) enter only through external predictive models used to score generated sequences.

That said, Evo 2 is highly relevant to multimodal integration pipelines such as those outlined in the **Integration Baseline Plan**: - Evo 2 can serve as a **genomics encoder** that produces compact sequence embeddings for downstream integration with other modalities using **late fusion** (e.g., concatenating Evo 2-derived features with imaging or clinical features for logistic regression or gradient-boosted trees). - Its strong zero-shot performance and robust variant scores align with the plan's emphasis on **modality-specific modeling first**, followed by disciplined late integration and careful evaluation (e.g., AUROC/AUPRC with CIs, bootstrapping or DeLong tests). - Mechanistic interpretability features (e.g., exon, splice, motif features) could act as **interpretable genomic covariates** in CCA or partial correlation analyses, fitting nicely into integration strategies that emphasize interpretability before moving to heavier fusion models.

7. Experiments and Results

- **Tasks / benchmarks**
  - **Zero-shot mutational effect prediction** across proteins, RNAs, and genomes from multiple domains of life, using changes in Evo 2 likelihood when introducing mutations.
  - **Human variant effect prediction (VEP):**
    - Predicting pathogenic vs benign variants in **ClinVar** for both coding and noncoding variants, including indels and other non-SNVs.
    - Predicting splice-altering variants using **SpliceVarDB**, split into exonic and intronic variants.
    - Predicting functional consequences of variants in **BRCA1** and **BRCA2** from saturation mutagenesis/scanning datasets.

- ◦ **Organismal fitness and gene essentiality:**
    - ▪ Zero-shot prediction of gene essentiality in bacteria and phage using mutational likelihood of premature stop codons.
    - ▪ Prediction of human lncRNA essentiality using scrambled subsequences.
- ◦ **Mechanistic interpretability:**
    - ▪ Training sparse autoencoders over Evo 2 representations to identify latent features corresponding to genomic and protein-level concepts.
- ◦ **Generative tasks:**
    - ▪ **Genome-scale generation** of mitochondrial genomes, minimal bacterial genomes (e.g., M. genitalium), and yeast chromosome-scale DNA.
    - ▪ **Generative epigenomics**: designing sequences with specified chromatin accessibility patterns using Enformer/Borzoi-guided inference-time search.

- • **Baselines**
    - ◦ **Sequence and variant models:** Nucleotide Transformer, Evo 1, and specialized variant effect models such as **AlphaMissense** and **GPN-MSA**.
    - ◦ **For DMS and fitness tasks:** State-of-the-art autoregressive protein LMs and DNA LMs used for mutational effect prediction.
    - ◦ **For generation and structure:** Comparisons to Evo 1 and to natural sequence/structure distributions (e.g., using Pfam, ESMFold, AlphaFold 3).

- • **Key findings**
    - ◦ **Zero-shot mutational effects & biological priors**
        - ▪ Evo 2 likelihood changes correctly reflect fundamental coding properties: strong penalties for mutations in start/stop codons, triplet codon periodicity, and decreased penalties at wobble positions.
        - ▪ The model distinguishes different genetic codes (e.g., standard vs mycoplasma vs ciliate) and requires longer contexts (4–8 kb)

to correctly infer some species-specific codes, highlighting the benefit of long context.

- Likelihood changes correlate with known constraints: non-synonymous, frameshift, and stop-gain mutations are more deleterious than synonymous mutations; deletions in essential RNAs (tRNAs, rRNAs) are more damaging than in intergenic regions.

◦ **DMS and fitness prediction**

- Evo 2's zero-shot likelihoods **correlate well with DMS fitness measurements** across diverse proteins and noncoding RNAs, matching or exceeding state-of-the-art protein LMs, especially for ncRNAs.

- For human mRNAs, Evo 2 likelihoods negatively correlate with mRNA decay rates (higher likelihood ⍰ more stable mRNA), and the 40B model performs better than 7B.

- Evo 2 matches Evo 1 on prokaryotic gene essentiality and **extends** predictive power to eukaryotic noncoding elements like lncRNAs, outperforming baselines.

◦ **Clinical variant prediction**

- In ClinVar, Evo 2 is competitive for coding SNVs and **outperforms baselines for non-SNVs** (indels) and noncoding variants, where many specialized models perform poorly or cannot score variants.

- In SpliceVarDB, Evo 2 achieves top zero-shot performance for both exonic and intronic splice variants.

- For BRCA1/2 datasets, Evo 2 sets or matches state-of-the-art performance, particularly when combining coding and noncoding variants.

- A supervised classifier built on Evo 2 embeddings further improves performance, achieving **AUROCs up to ~0.95** on BRCA1 SNVs and outperforming AlphaMissense and other baselines.

- **Mechanistic interpretability**
  - Sparse autoencoders over Evo 2 representations reveal features aligned with:
    - Prophage and mobile genetic elements.
    - ORFs, intergenic regions, tRNAs, rRNAs.
    - Protein secondary structures (α-helices, β-sheets).
    - Exon–intron architectures (including features that fire at specific exon boundaries).
    - Promoter motifs and transcription factor binding-like patterns in human genome regions.
  - These features generalize to out-of-training genomes, e.g., annotating exon–intron structure in the **woolly mammoth** genome.

- **Genome-scale generation**
  - Evo 2 generates mitochondrial genomes with correct counts and arrangements of coding, tRNA, and rRNA genes, with diverse but structurally plausible proteins.
  - It generates **minimal bacterial genomes** (M. genitalium scale) where ~70% of predicted genes have significant Pfam hits, a large improvement over Evo 1.
  - Yeast chromosome-scale generation yields sequences with recognizable genes, introns, promoters, and tRNAs, with proteins that resemble natural yeast proteins in sequence and structure.
  - Viral protein generation from human-infecting viruses remains poor by design, reflecting intentional data exclusion and risk mitigation.

- **Generative epigenomics and inference-time scaling**
  - Using Enformer and Borzoi to score chromatin accessibility, a beam-search-based inference-time procedure produces sequences whose predicted accessibility matches desired patterns (e.g., open vs closed regions).

- Increasing inference-time compute (wider beam search, more sampled chunks) shows a **log-linear improvement** in design success (higher AUROC for distinguishing open vs closed regions), demonstrating **inference-time scaling laws** for a biological design task.
- The method is expressive enough to encode Morse-code messages in chromatin accessibility peaks, while maintaining natural sequence statistics (e.g., dinucleotide distributions).

8. Strengths, Limitations, and Open Questions

- **Strengths**
  - **Scale and coverage:** Evo 2 is trained on one of the largest curated genomic datasets to date, spanning all domains of life and multiple genomic contexts, and achieves a 1M-token context window.
  - **Generalist capabilities:** A single model supports diverse tasks: zero-shot mutational effect prediction, clinical variant scoring, genome-scale generation, and guided epigenomic design.
  - **Strong performance on hard regimes:** Evo 2 is particularly strong for noncoding, splice, and indel variants, where many existing models struggle.
  - **Mechanistic interpretability:** The use of sparse autoencoders reveals interpretable features that align with biological concepts, improving trust and offering new tools for discovery and annotation.
  - **Open release:** Model weights, training and inference code, and the OpenGenome2 dataset are released under an open license, enabling reproducibility and community-driven extensions.
  - **Safety-aware design:** The training corpus deliberately excludes pathogenic viral genomes for eukaryotic hosts, and empirical tests show degraded performance in that domain, demonstrating concrete risk mitigation.

- **Limitations**
  - **DNA-only input:** Despite implicitly touching proteins, RNAs, and epigenomics, Evo 2 only ingests DNA; it does not explicitly integrate multi-omic modalities (e.g., expression, proteomics) or other data types like imaging or clinical variables.

  - **Reliance on external models for function-specific design:** Generative epigenomics depends on Enformer and Borzoi for scoring; Evo 2 itself is not directly conditioned on chromatin states or other annotations.

  - **Dataset and species bias:** Although the training set is huge, it is still limited by what genomes are available and curated, and some domains (e.g., viruses infecting humans) are intentionally underrepresented.

  - **Compute and infrastructure requirements:** Training and inference at 40B scale with 1M context is expensive and requires specialized infrastructure (Savanna, Vortex), limiting who can retrain or significantly modify the model.

  - **Interpretability coverage:** Even with SAEs, only a subset of latent features is mapped to interpretable biological concepts; many features likely encode higher-order or mixed patterns that remain to be understood.

- **Open Questions and Future Directions:**
  - How can Evo 2 embeddings be integrated with **population-scale human genetic variation** (e.g., biobanks) to further improve pathogenicity prediction and identify subtle regulatory effects?

  - Can mechanistic interpretability tools (e.g., feature steering, activation clamping) be used to **control specific biological properties** during generation, such as alternative splicing or tissue-specific regulation?

  - What are the best strategies to combine Evo 2 with other modalities (transcriptomics, epigenomics, imaging) in **late fusion** or **two-tower architectures**, in line with the integration baseline plan?

- How robust are Evo 2's predictions across underrepresented species or genomic contexts, and what additional data curation or training strategies would reduce biases?

- Can inference-time design frameworks be generalized beyond chromatin accessibility to optimize for other complex properties (e.g., protein–protein interaction networks, gene circuit behavior) while maintaining safety?

9. Context and Broader Impact

Evo 2 sits at the frontier of **foundation models for genomics**, analogous to large language models in NLP but operating directly on DNA across the tree of life. Compared to earlier genomic models and protein LMs, Evo 2 pushes three main frontiers at once: **scale of data**, **model size and architecture**, and **context length**, enabling it to capture patterns from local motifs up to genome-scale organization. Its success on zero-shot variant effect prediction and generative tasks shows that **alignment-free, DNA-only models** can be powerful generalists, complementing or even competing with highly specialized models like AlphaMissense. For students interested in multimodal integration, Evo 2 can be viewed as a **genome encoder**: its embeddings and interpretable features are natural inputs to late-fusion or contrastive multimodal pipelines that combine genomics with imaging, behavior, or clinical data, consistent with the integration baseline plan's emphasis on strong per-modality modeling. More broadly, Evo 2 demonstrates how ideas from large-scale LMs (scaling laws, long-context training, inference-time scaling, mechanistic interpretability) can be translated into biology, opening up a path toward "virtual cell" models that integrate genomics with epigenomic and transcriptomic layers.

10. Key Takeaways (Bullet Summary)

- Evo 2 is a **large-scale genomic foundation model** trained on ~9.3T nucleotides from genomes spanning all domains of life, with context windows up to 1M base pairs.

- The model uses a new **StripedHyena 2 multi-hybrid architecture**, combining input-dependent convolutions and attention to achieve efficient training and inference at long context lengths.

- Evo 2 supports **zero-shot mutational effect prediction** for proteins, RNAs, and noncoding regions, with likelihood changes that align with known biological constraints and genetic codes.

- On clinical variant tasks (ClinVar, SpliceVarDB, BRCA1/2), Evo 2 is competitive with or better than specialized models, especially for noncoding, splice, and non-SNV variants.

- Evo 2 embeddings can power **supervised classifiers** that achieve state-of-the-art performance on clinically important tasks such as BRCA1 variant classification.

- Mechanistic interpretability with sparse autoencoders reveals **latent features corresponding to exons, introns, motifs, protein secondary structure, and prophages**, and these features generalize to unseen genomes.

- The model can **generate genome-scale DNA sequences** (mitochondrial genomes, minimal bacterial genomes, yeast chromosomes) that resemble natural sequences in structure and function.

- By combining Evo 2 with epigenomic predictors and beam-search-style inference-time search, the authors demonstrate **controllable "generative epigenomics"** and the first inference-time scaling laws in a biological design setting.

- Evo 2 is released fully open (weights, training and inference code, data), offering a powerful **foundation for downstream genomics, variant interpretation, and design tasks**, while incorporating explicit risk mitigation for viral domains.

- For a new grad student, Evo 2 is a key reference for how to build, evaluate, and interpret **DNA-based foundation models**, and a natural starting point for projects in variant effect prediction, genome design, and multimodal integration.