Genome modeling and design across all domains of life with Evo 2

Garyk Brixi[*,1,2,3], Matthew G. Durrant[*,1,2], Jerome Ku[*,1,2], Michael Poli[*,2,3,5], Greg Brockman[**,2,6,§], Daniel Chang[**,1,2,3], Gabriel A. Gonzalez[**,1,2], Samuel H. King[**,1,2,3], David B. Li[**,1,2,3], Aditi T. Merchant[**,1,2,3], Mohsen Naghipourfar[**,1,2,7], Eric Nguyen[**,2,3], Chiara Ricci-Tam[**,1,2], David W. Romero[**,2,4], Gwanggyu Sun[**,1,2], Ali Taghibakshi[**,2,4], Anton Vorontsov[**,2,4], Brandon Yang[**,2,6], Myra Deng[8], Liv Gorton[8], Nam Nguyen[8], Nicholas K. Wang[8], Etowah Adams[9], Stephen A. Baccus[3], Steven Dillmann[3], Stefano Ermon[3], Daniel Guo[1,3], Rajesh Ilango[1], Ken Janik[4], Amy X. Lu[7], Reshma Mehta[6], Mohammad R.K. Mofrad[7], Madelena Y. Ng[3], Jaspreet Pannu[3], Christopher Ré[3], Jonathan C. Schmok[1], John St. John[4], Jeremy Sullivan[1], Kevin Zhu[7], Greg Zynda[4], Daniel Balsam[8,10], Patrick Collison[1,10], Anthony B. Costa[4,10], Tina Hernandez-Boussard[3,10], Eric Ho[8,10], Ming-Yu Liu[4,10], Thomas McGrath[8,10], Kimberly Powell[4,10], Dave P. Burke[‡,1,2,10], Hani Goodarzi[‡,1,2,10,11], Patrick D. Hsu[‡,†,1,2,7,10], Brian L. Hie[‡,†,1,2,3,10]

[1]Arc Institute; [2]Core Contributor, Evo 2 Team; [3]Stanford University; [4]NVIDIA; [5]Liquid AI; [6]Independent Researcher; [7]University of California, Berkeley; [8]Goodfire; [9]Columbia University; [10]Senior Contributor, Evo 2 Team; [11]University of California, San Francisco

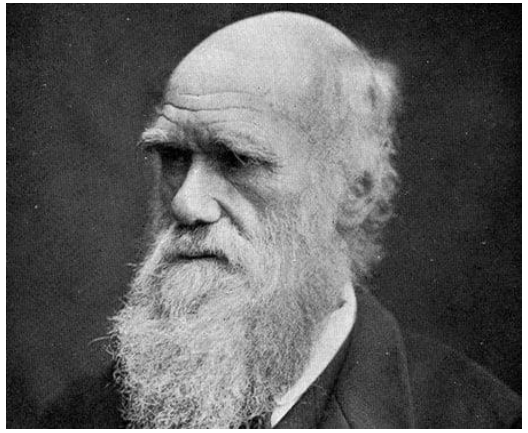
Why has Evo2 been controversial?

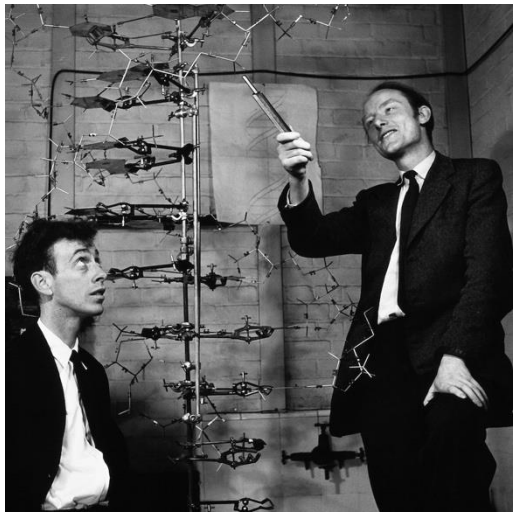
Extreme training requirements

- “Evo 2 was trained for **several months** on the NVIDIA DGX Cloud AI platform via AWS, utilizing over **2,000** NVIDIA **H100 GPUs** and bolstered by **collaboration with NVIDIA** researchers and engineers.” - Arc Institute blog post
- Roughly **$10M** training

Braggadocious framing

- “Biological research scales from molecules to systems to organisms, seeking to understand and design functional components across all domains of life (**Darwin**, 1859; **Mendel**, 1866; Dobzhansky, 1951) … All domains of life express complex functions from DNA sequences (**Watson and Crick**, 1953; Nirenberg and Matthaei, 1961), yet genomic content and length vary dramatically across organisms.”





Difficulty of use

- Evo2-40B weights take up 80GB in GPU RAM
- H100 node required to compile the code and run inference
- Can take hours of inference time to process a single sample

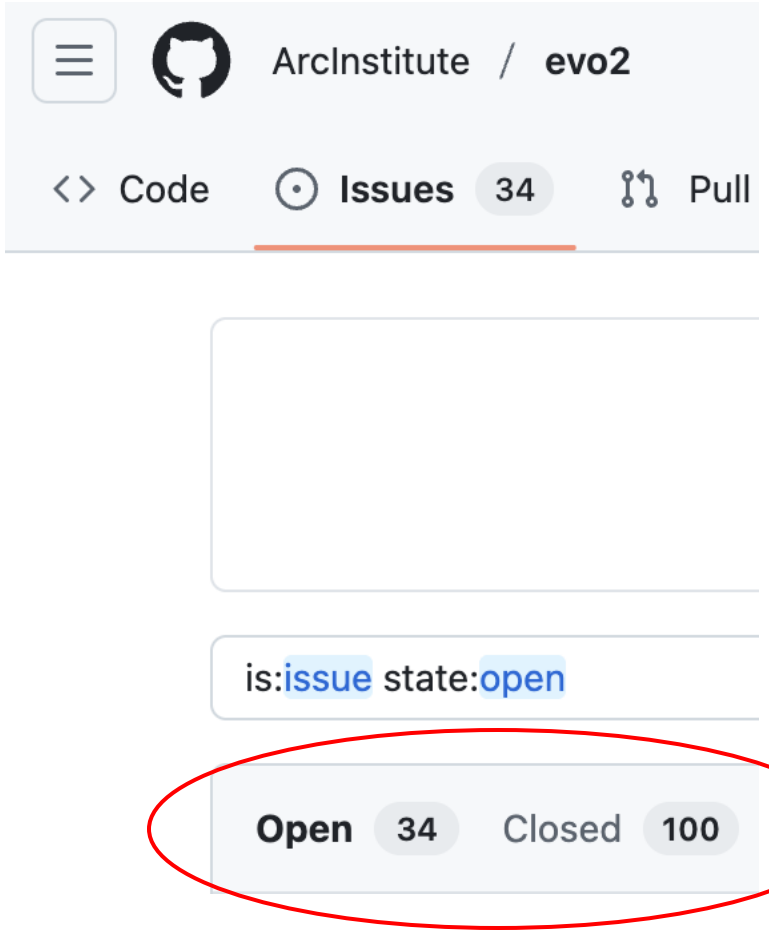
Nvidia H100 NVL GPU Computing Processor

$29,500.00

Pre-owned

eBay & more

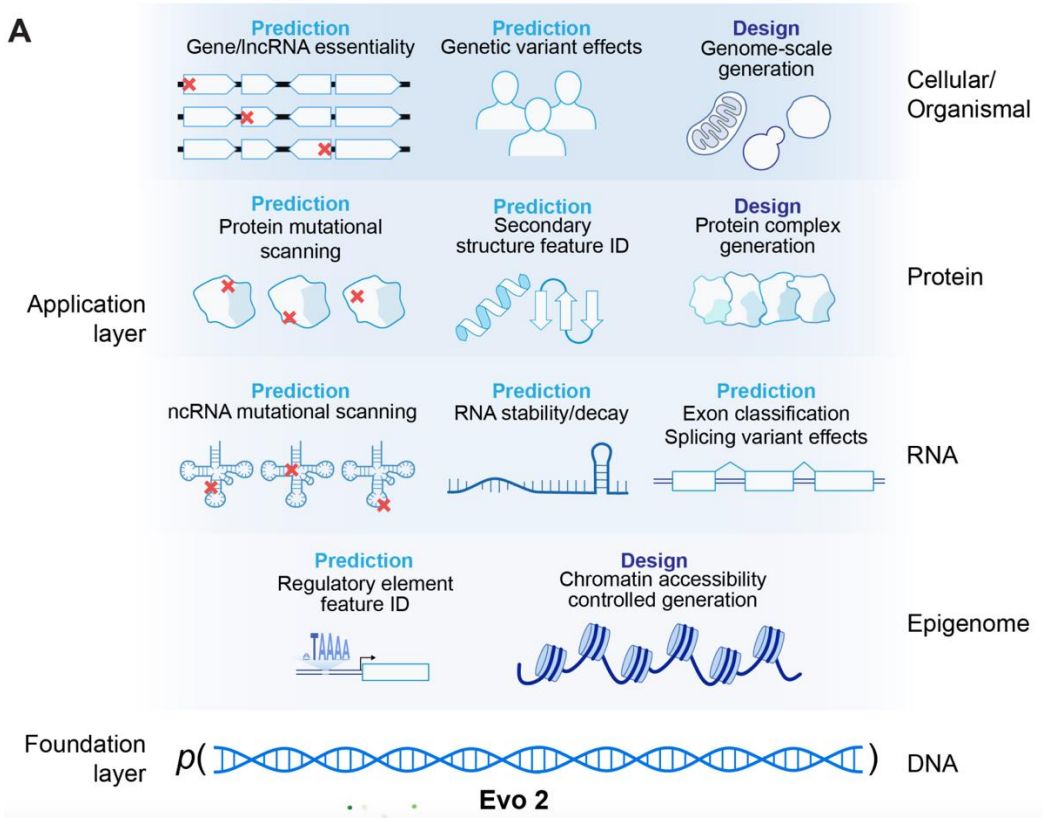
4.0 ★★★★☆ (1)



What’s unique about Evo2?

Intro section summary

- Advances in:
  - Model architecture
  - Data curation
  - Training and inference infrastructure
  - Inference-time compute
- Mechanistic interpretability
- Focus on genome generation



Evo2 is BIG

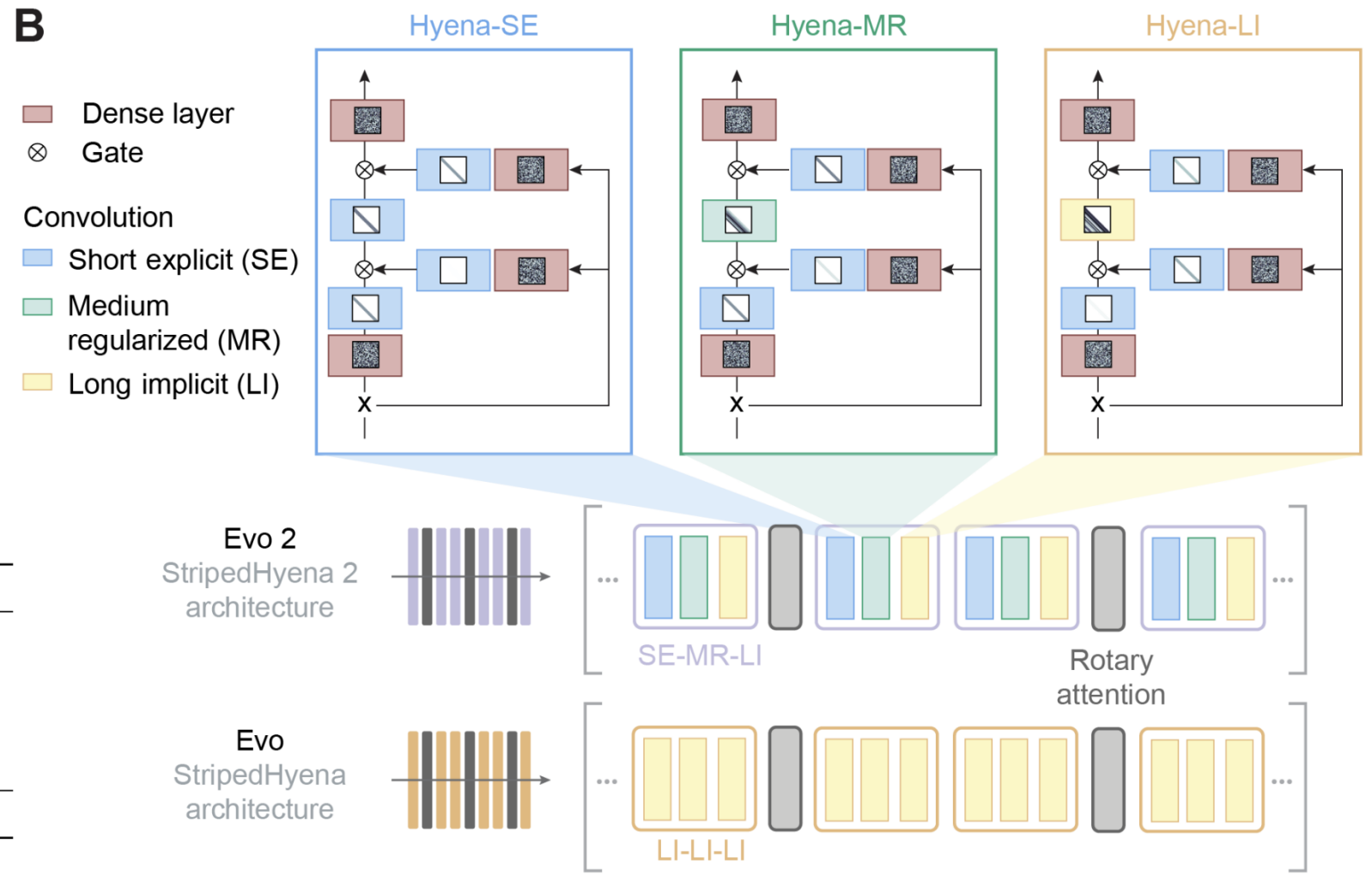
Model	Parameters (M)	Training Bases (B)	Context window
Evo2-40B	**40000**	**9300**	**1M**
Evo2-7B	**7000**	**2400**	**1M**
Evo	7000	300	131K
NT-1000G	2500	174	12K
Borzoi	186	6	524K
DNABERT-2	117	30	1k
HyenaDNA	6.6	3	1M

Model architecture

- Convolutions find features
- Attention gives features context
- Early layer features are simple, like kmer motifs
- Late layer features are complex, like gene-specific CREs

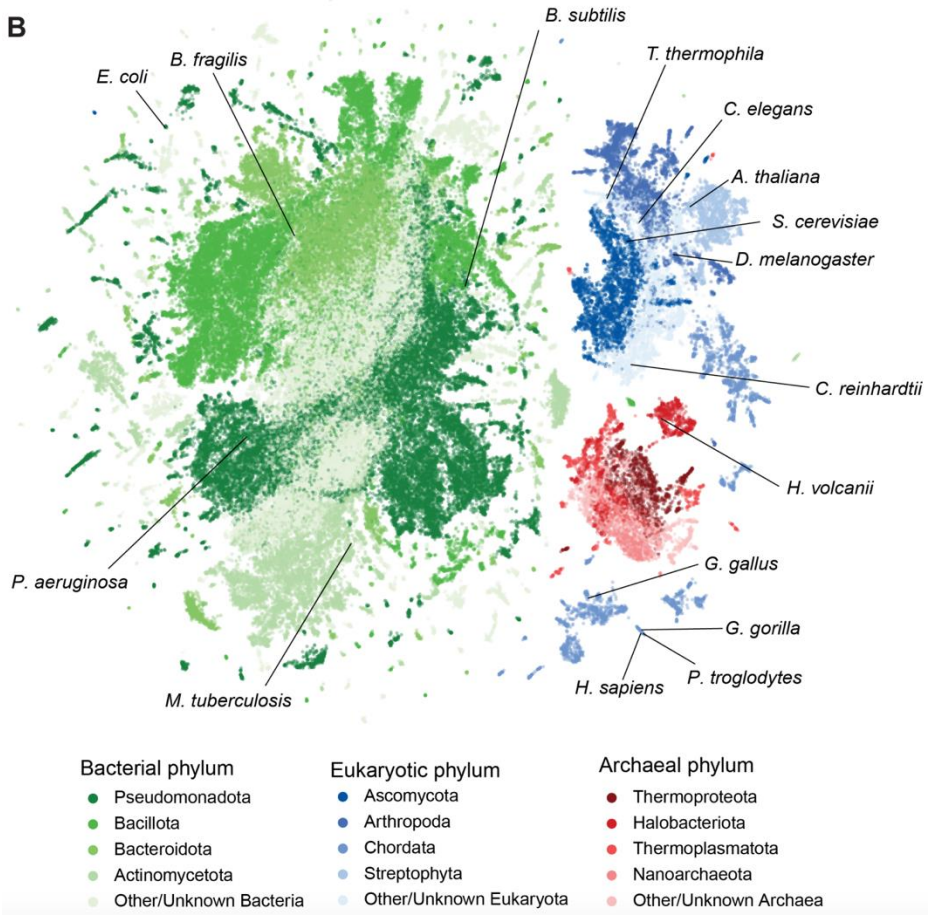
- StripedHyena 2
- Intentionally patterned series of convolutions and attention layers
- Mix of 7, 128, and full-length convolutions

	Evo 2 40B	Evo 2 7B	Evo 2 1B base
Parameters	40.3B	6.5B	1.1B
Total Layers	50	32	25
Hidden Size	8,192	4,096	1,920
FFN Size	22,528	11,264	5,120
Num Heads	64	32	15
Total Tokens	9.3T	2.4T	1T



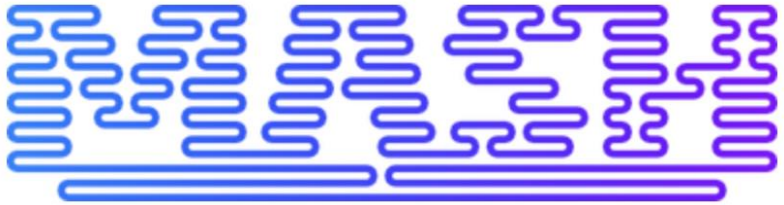
Training data

- OpenGenome2 dataset
- 113k prokaryotic genomes (357B bp)
- 15,032 eukaryotic genomes (6.98T bp)
- Non-redundant metagenomes (854B bp)
- Euk. organelle genomes (2.82B bp)
- Euk. coding sequences (602B bp)
- 8.8T bases total

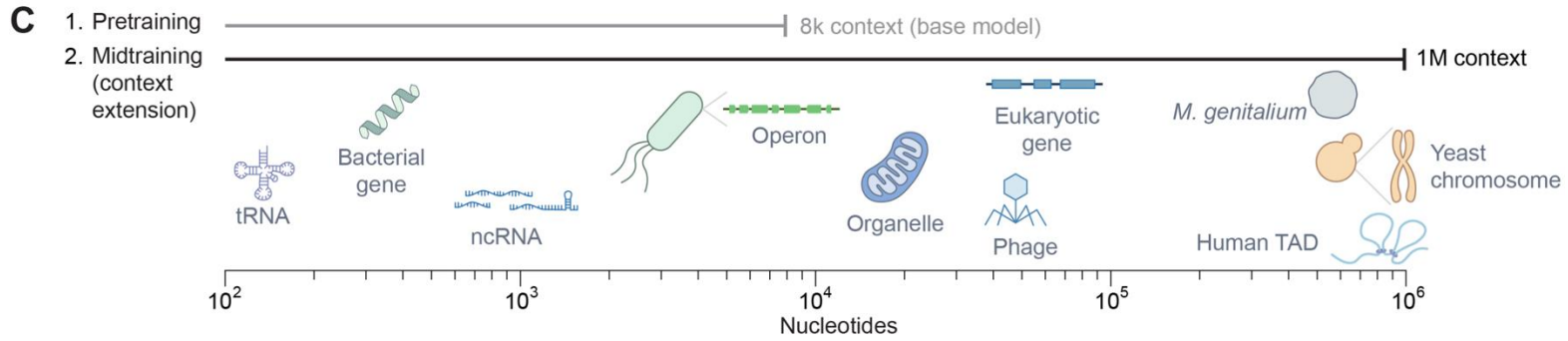


Training data curation

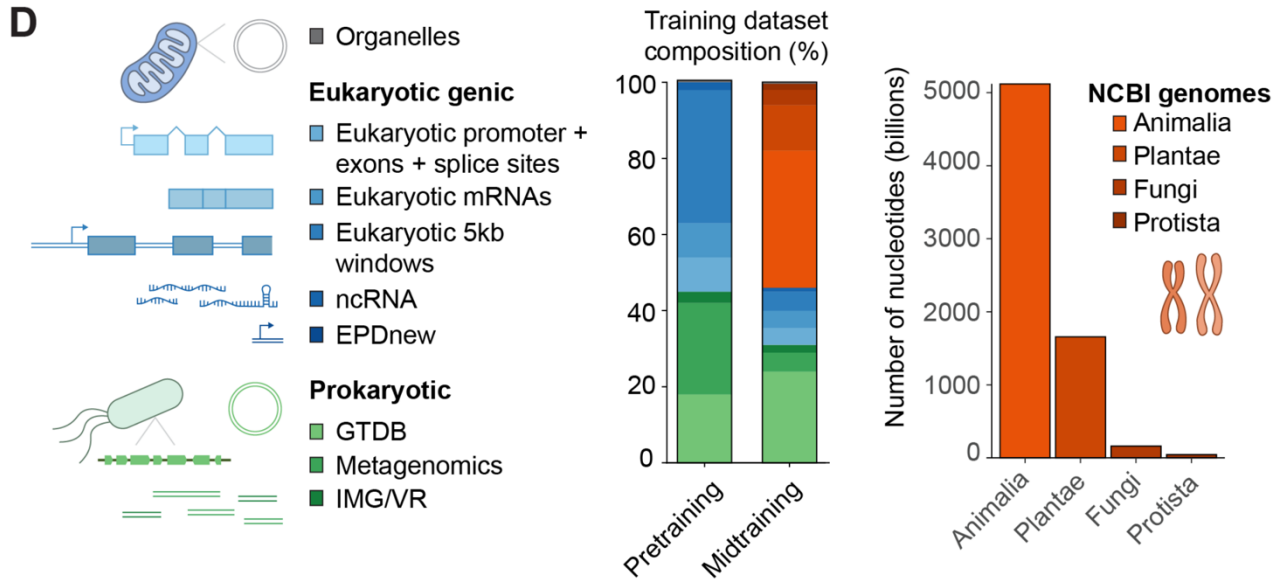
- Exclude viruses that infect eukaryotes
- Remove atypical, metagenome-assembled, and large multi-isolate eukaryotic genomes
- Remove eukaryotic genomes with >99% MASH similarity
- Remove metagenomic sequences with >10% protein clusters already seen in set
- Discard contigs that are too short or have too many ambiguous bases
- Generally, remove similar short contigs

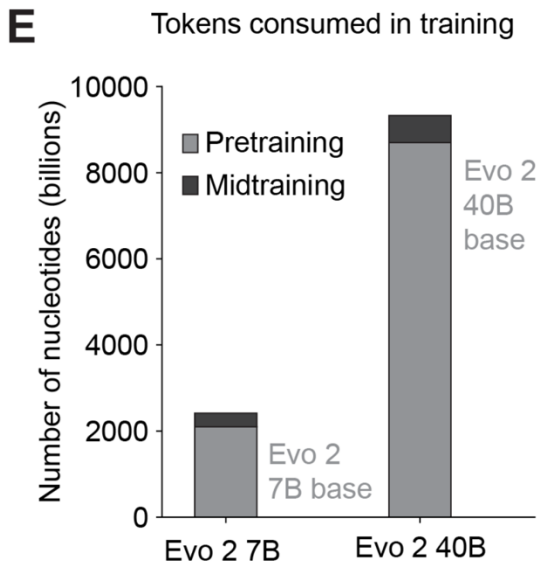


Two-phase pre-training



- First stage: 8192bp context window, train on euk. genes + prok.s
- Second stage: extend to 1Mbp context window, train on the tree of life
- This approach mirrors the SOTA in training natural language models





Training considerations

- Reverse complemented 50% of sequences
- Stitched together short sequences to reach long context windows
  - Phylogenetic tags inserted every 131kb “to help condition the model”

```
|D__BACTERIA;P__PSEUDOMONADOTA;C__GAMMAPROTEOBACTERIA;
O__ENTEROBACTERALES;F__ENTEROBACTERIACEAE;G__ESCHERICHIA;
S__ESCHERICHIA|
```

- Weighted loss function to prioritize non-repetitive regions
  - Loss has 10% of the impact on model weights in repeat regions

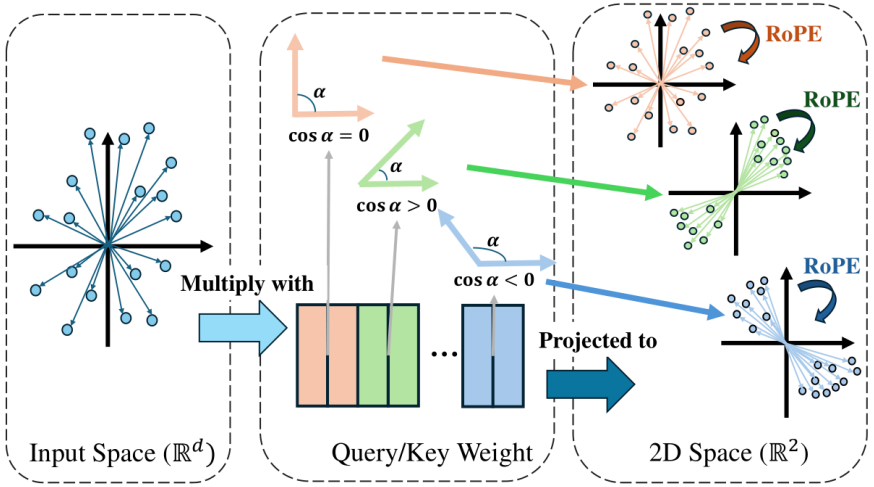
$$\ell_{\text{wCE}} = \frac{1}{Z} \sum_t w_t \ell_{\text{CE}}(t)$$

$$w_t = \begin{cases} 0.1 & \text{if position } t \text{ is in repetitive region} \\ 1.0 & \text{otherwise} \end{cases}$$

$$Z = 0.1\, N_{\text{repeat}} + N_{\text{non\_repeat}}$$

Midtraining: context extension

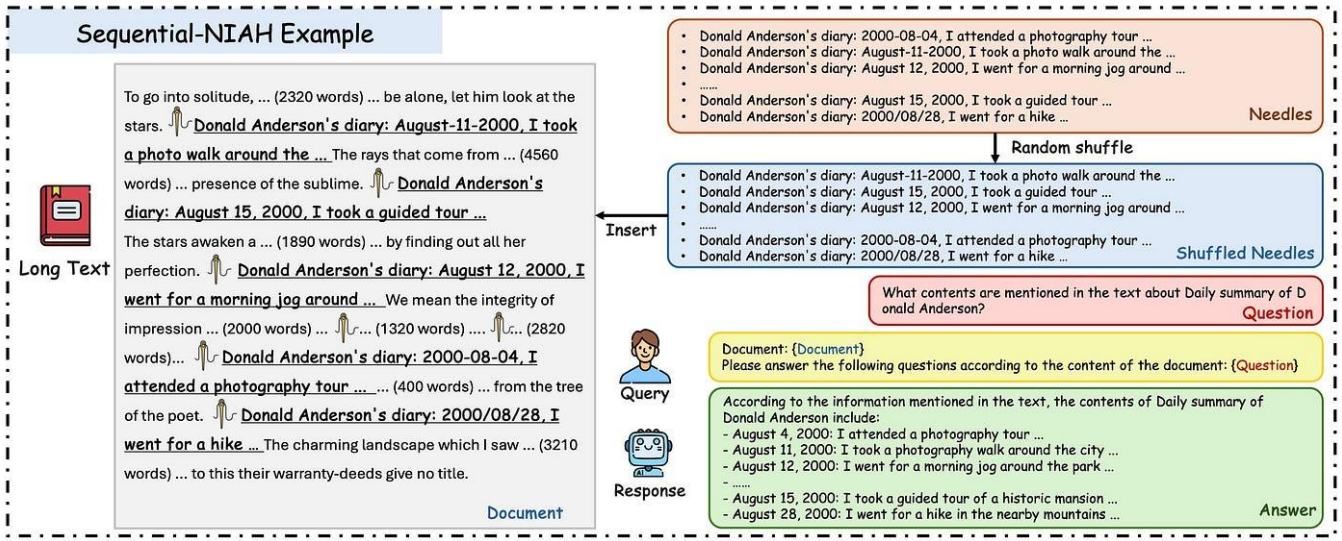
Context Length	Base Frequency	Scale Factor	Evo 2 7B Training Tokens	Evo 2 40B Training Tokens
32.8K	1.0×10^{6}	4	50B	-
65.5K	1.0×10^{7}	8	50B	-
131.1K	1.0×10^{8}	16	50B	200B
262.1K	1.0×10^{9}	32	50B	200B
524.3K	1.0×10^{10}	64	50B	-
1048.6K	1.0×10^{11}	128	50B	200B
			300B	600B

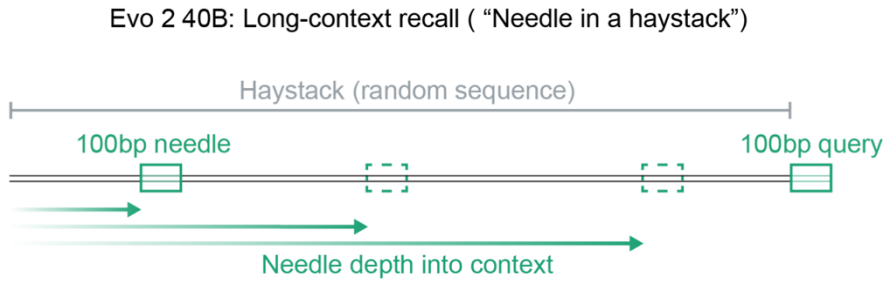


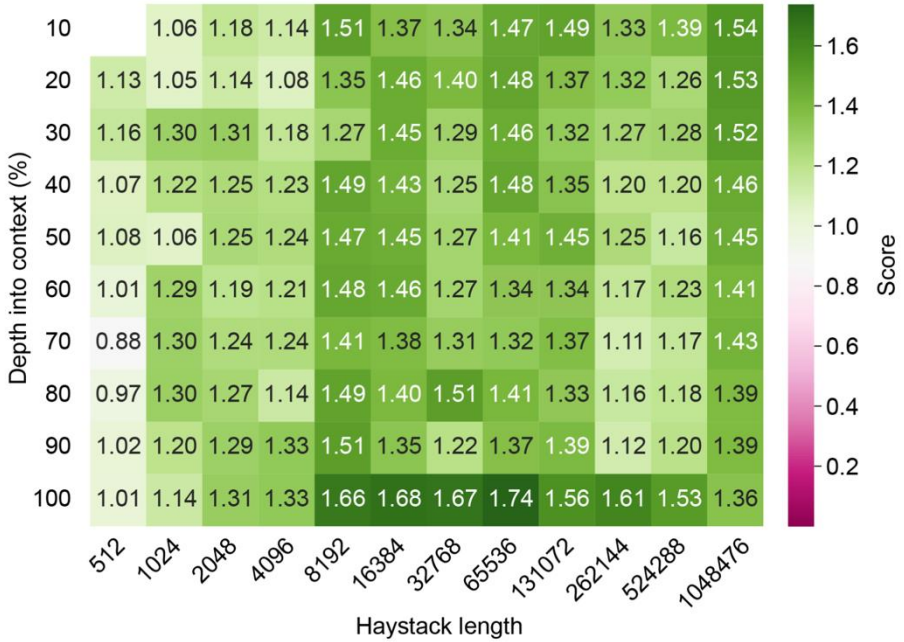
- Gradually increase the sequence length of training samples
- Scale positional encoding according to sequence length

Midtraining: needle-in-a-haystack evaluation

- Do we learn the entire context window?
- Two identical 100bp sequences: one somewhere in the seq, one at end
- Test if changes to the first affect predictions of the second

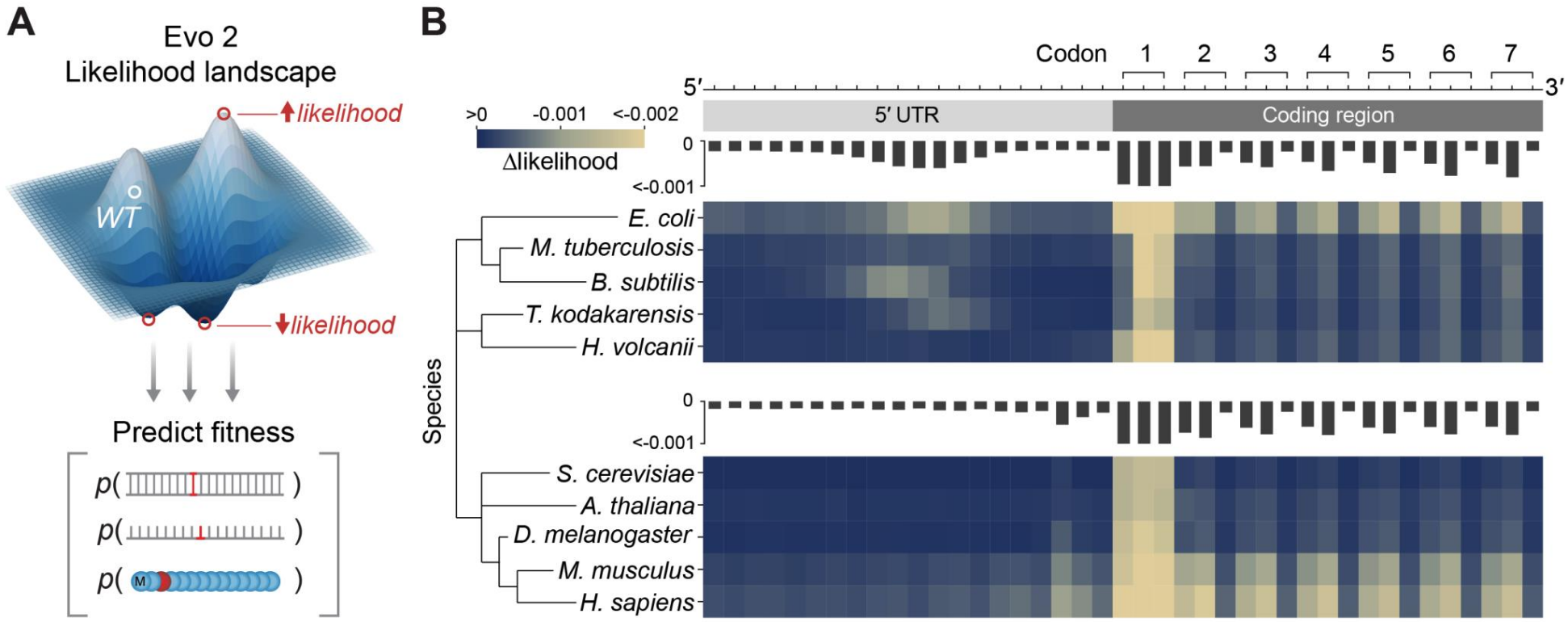






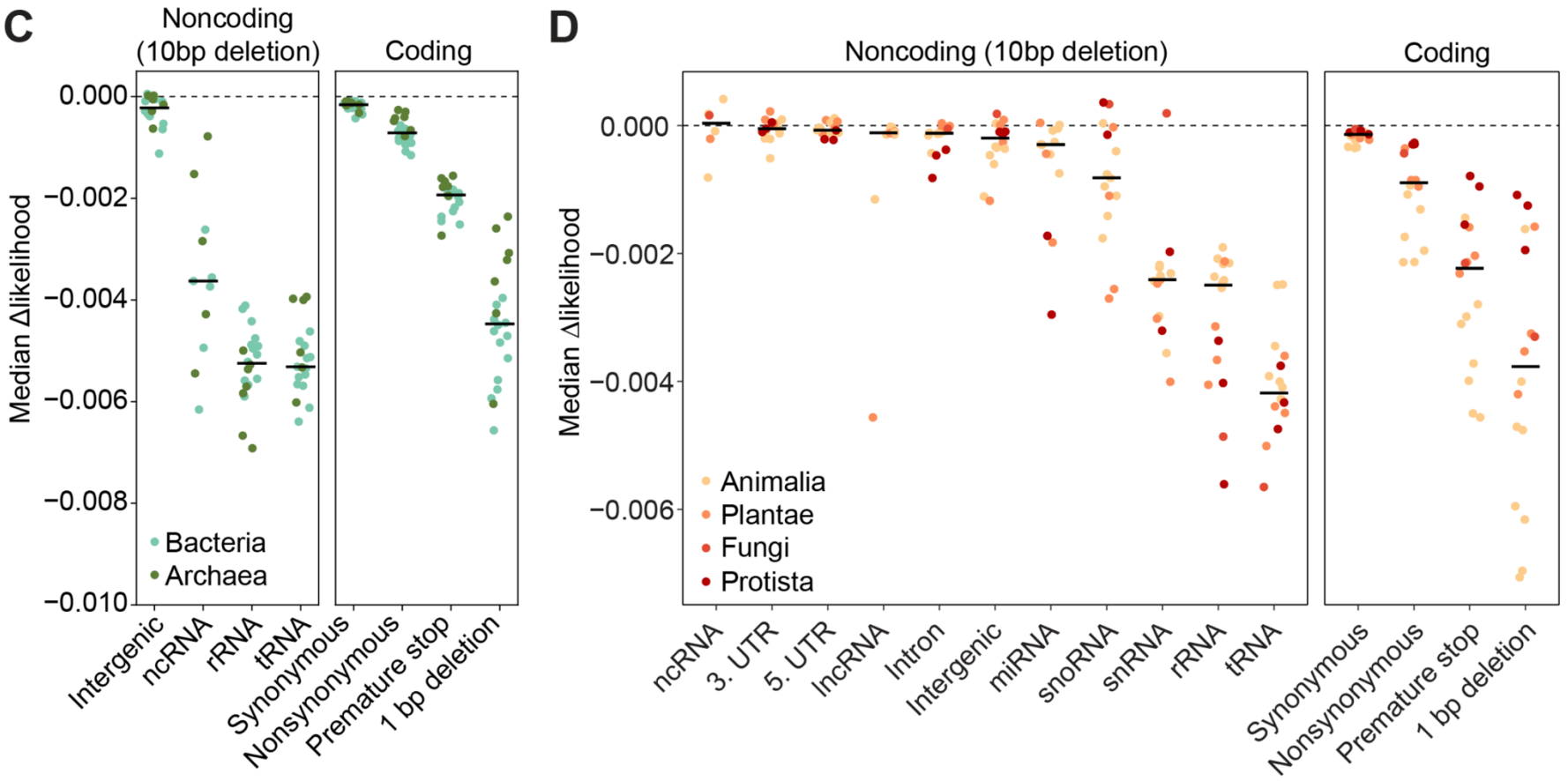
What does Evo2 learn during pre-training?

Evo2 learns coding syntax with no labels



- Mutate bases at different positions and measure change in sequence likelihood
- Learns that mutations in the start codon are very impactful / unlikely
- Learns that mutations in the 3rd codon base (wobble position) are less impactful

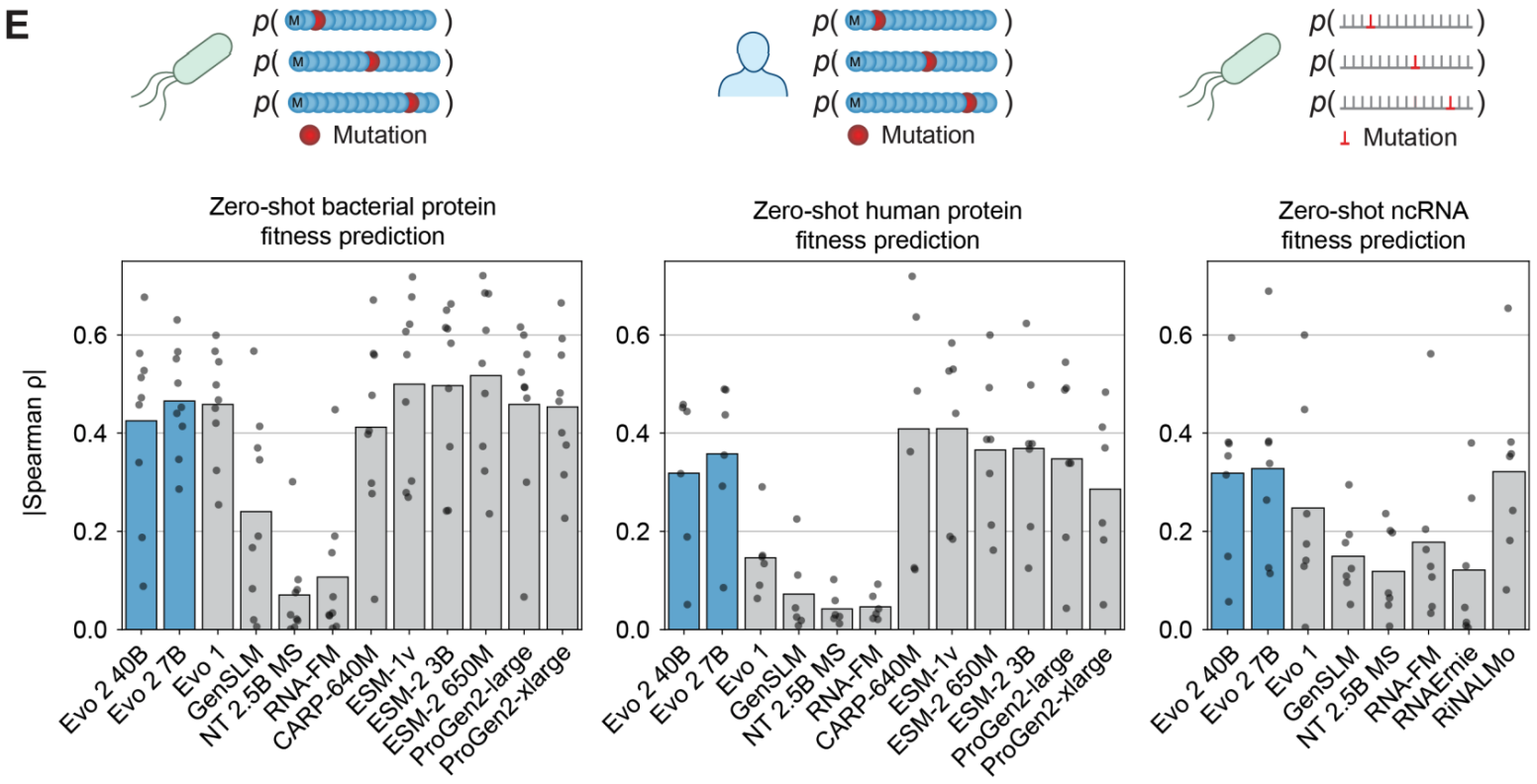
Evo2 learns mutation significance w/ no labels



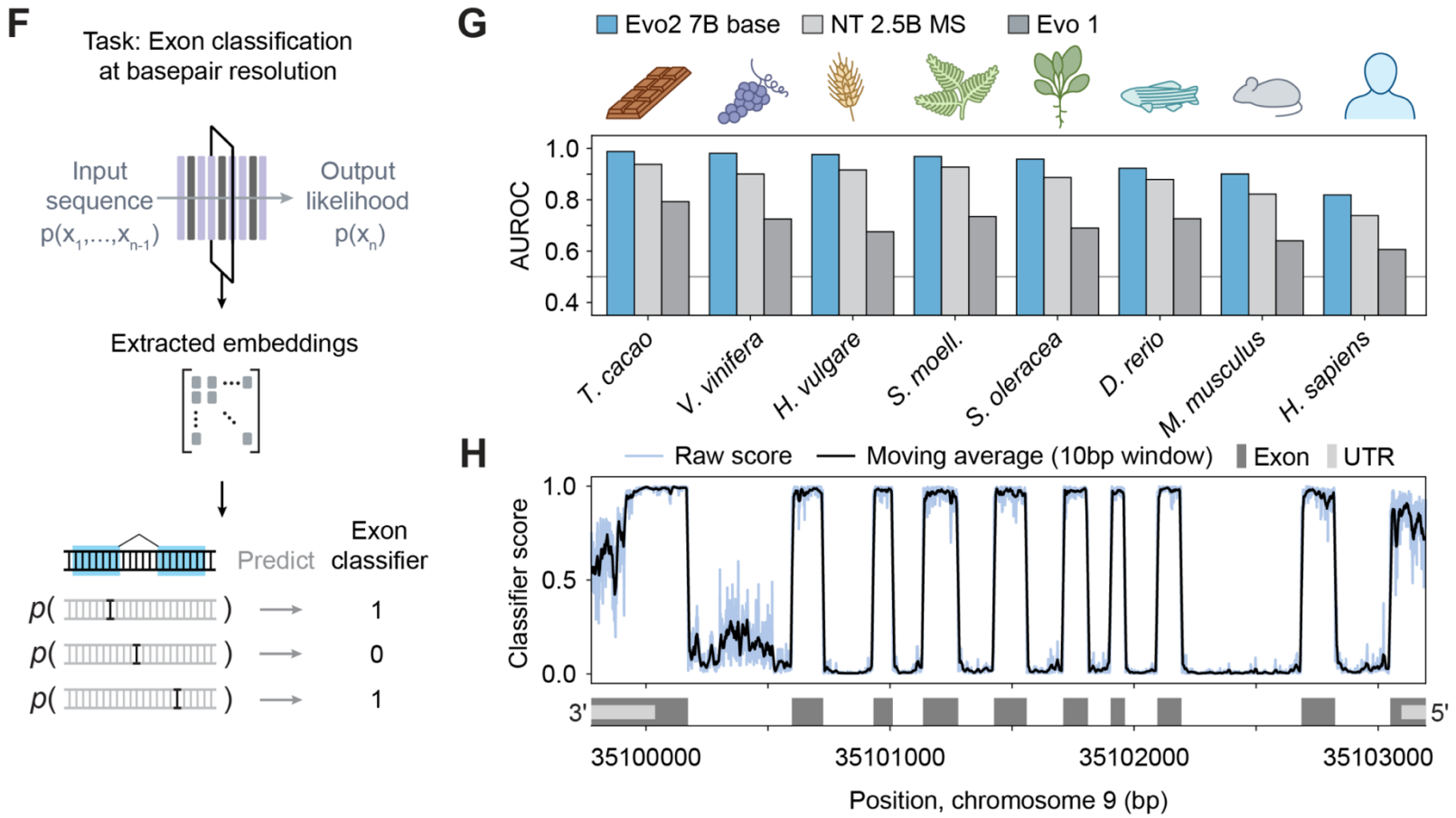
- Implicit, zero-shot understanding that some mutations are more impactful than others

Evo2 predictions match experimental values

- Y axis: correlation between Evo2 predicted sequence likelihood and Deep Mutational Scanning (DMS) measurements
- Evo2 beats leading DNA LMs and is competitive with leading protein LMs



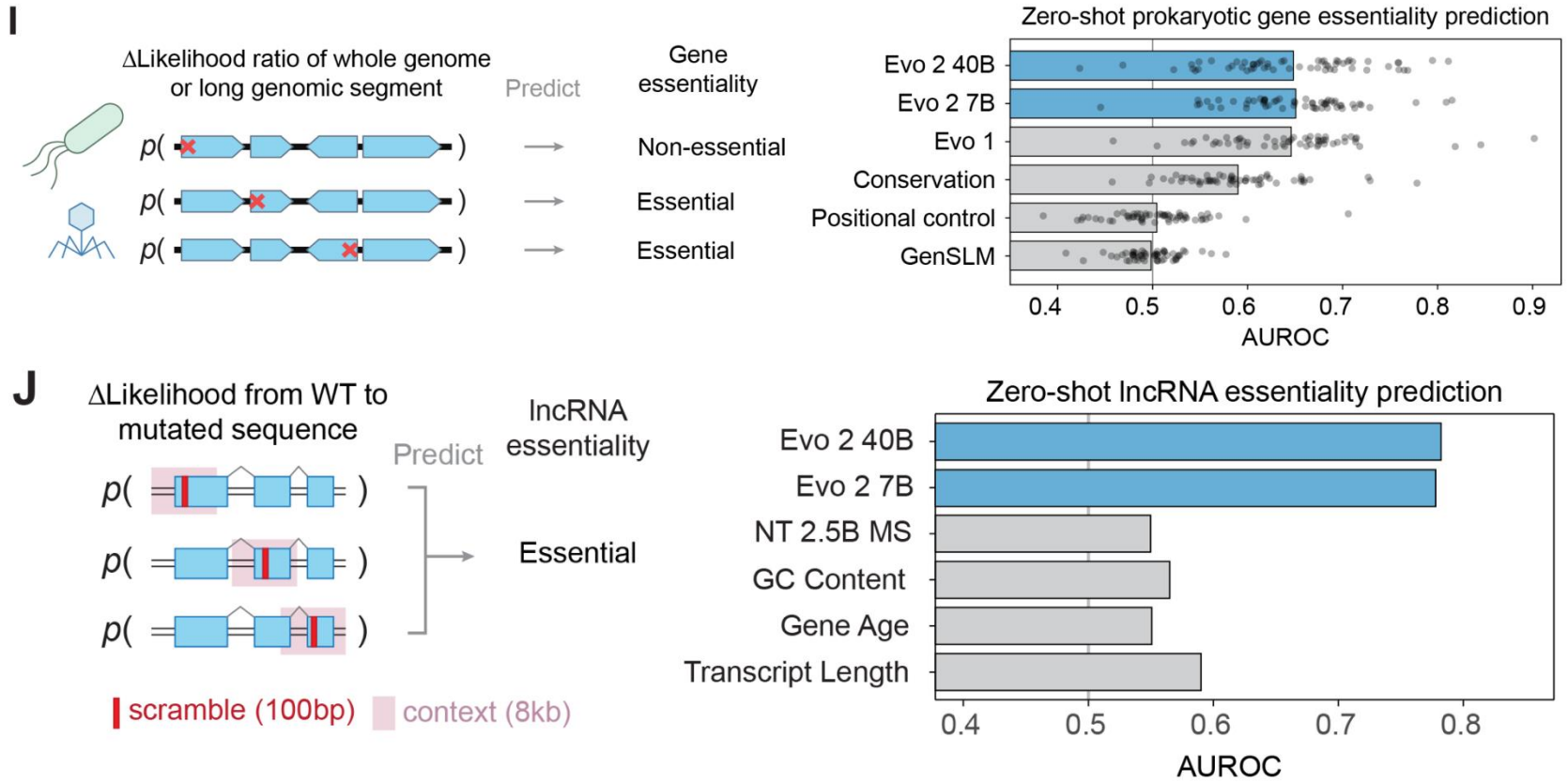
Evo2 learns splicing with no labels



- Classifier fit to Evo2 embeddings knows exons vs. introns without any supervised training
- Does this better than other leading DNA LMs
- Could annotate new assemblies without any further training

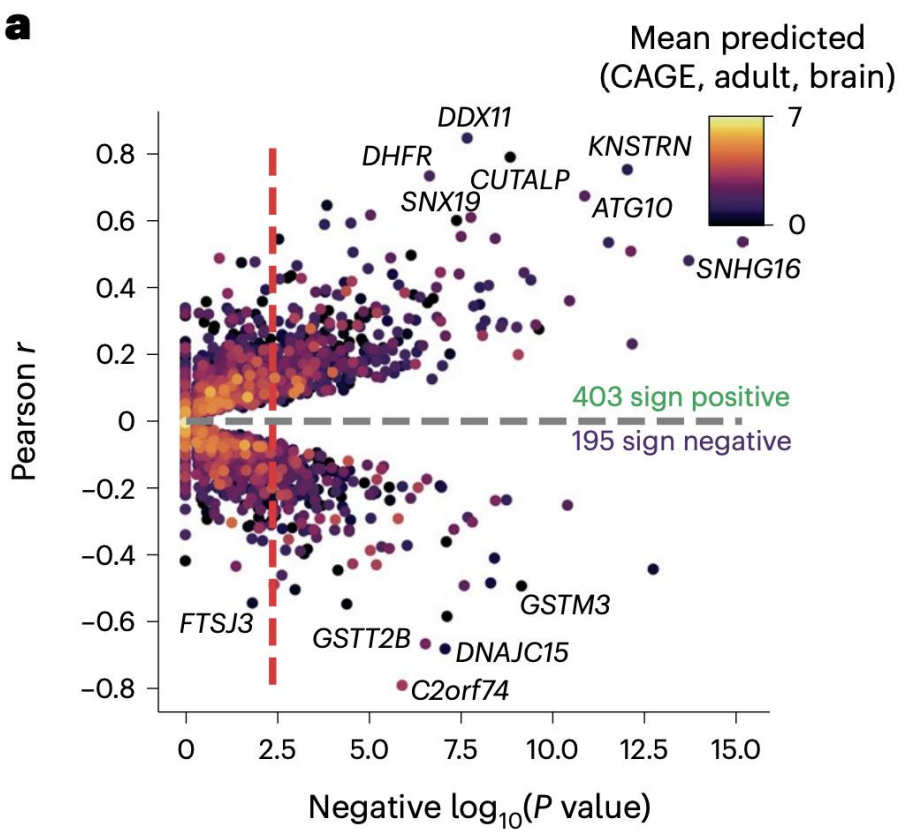
Evo2 learns gene essentiality with no labels

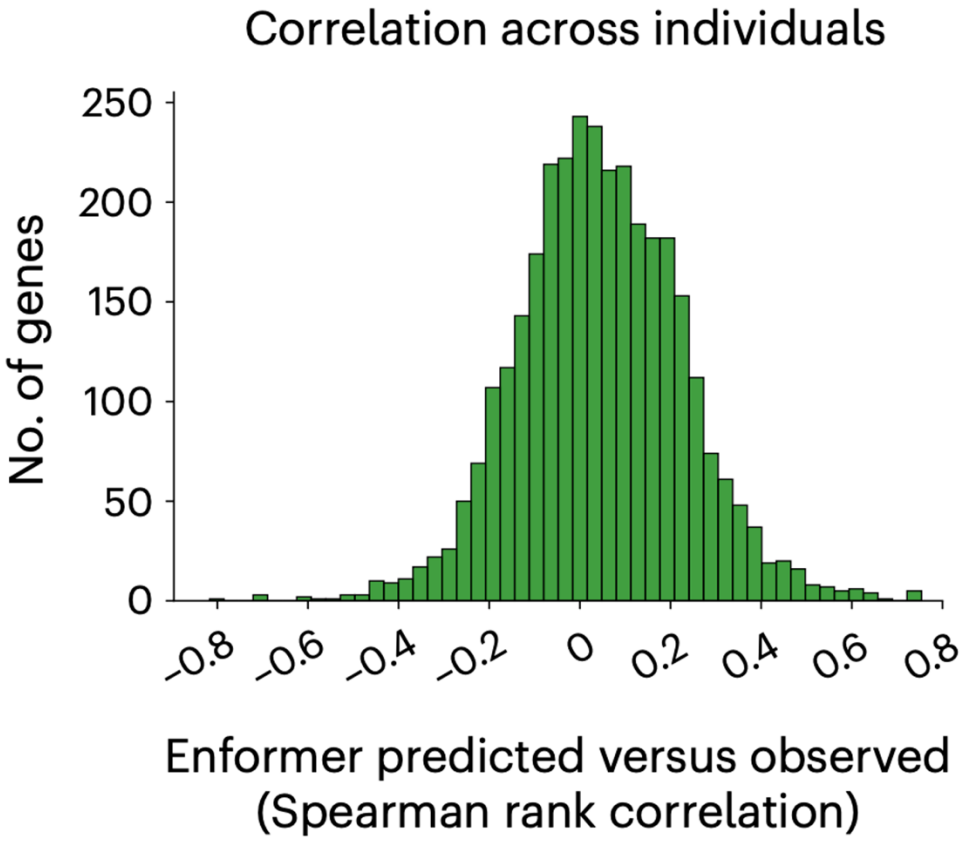
- Insert premature stop codons into genes, measure resulting likelihood of the entire genome
- Tells us how important each gene is to the organism
- Can implicitly determine which genes are essential to life
- Same experiment repeated for human lncRNAs



Can Evo2 do human variant effect prediction?

Variant effect prediction is hard for GLMs

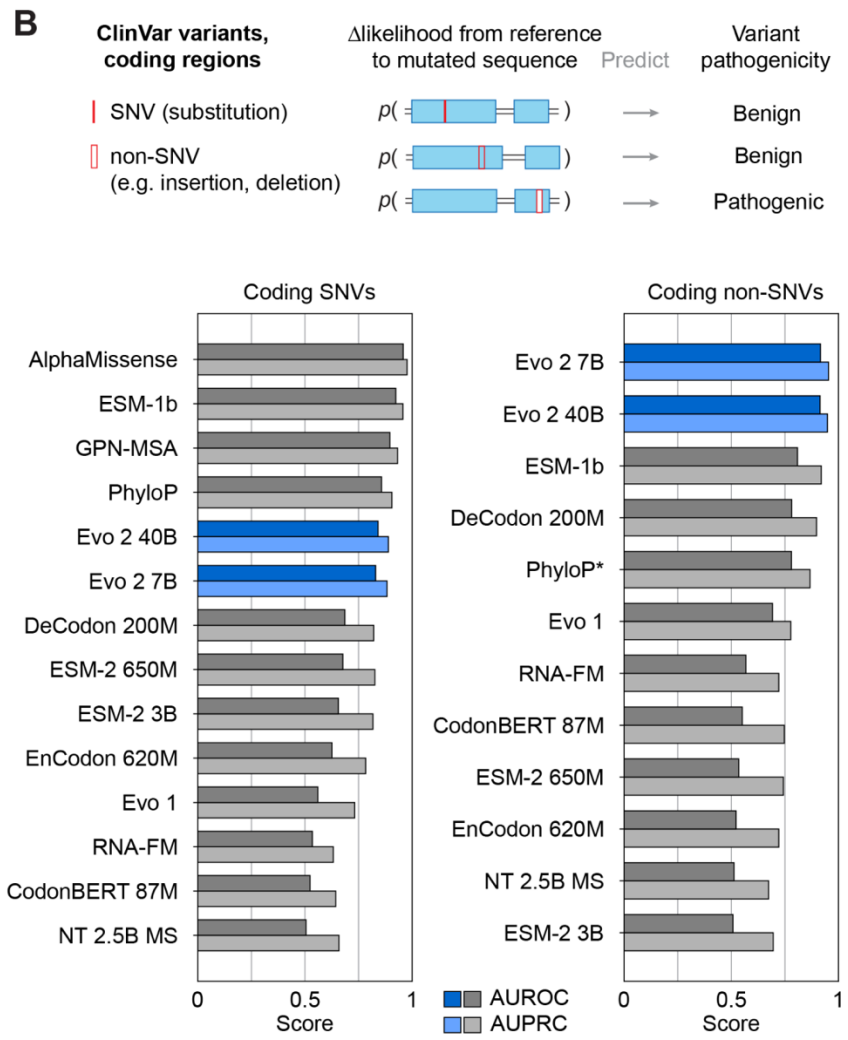


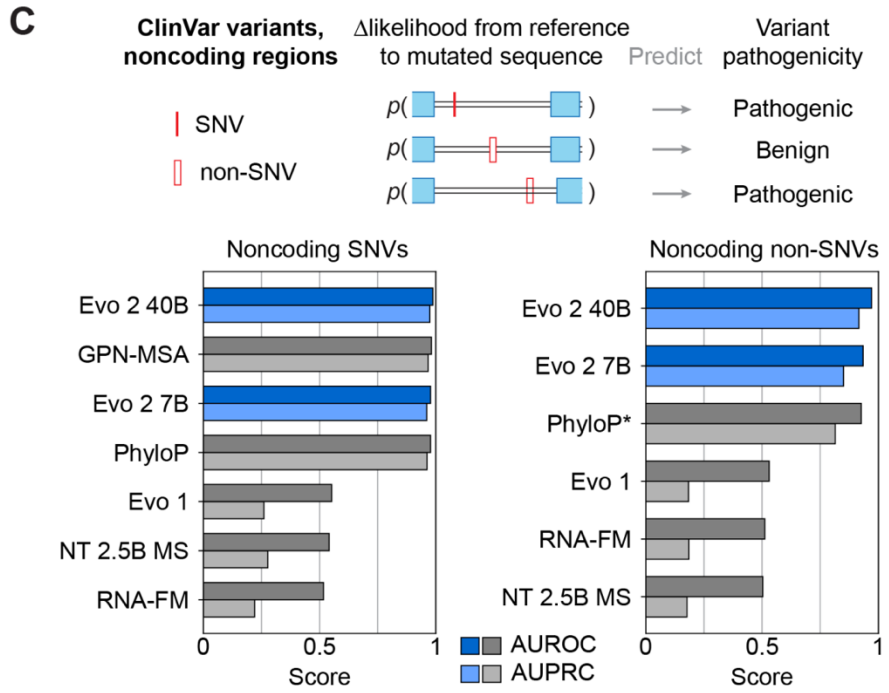


- Arguably the defining task of any genomic language model
- Substandard performance so far in the field

Huang et al. *Nature Genetics* (2023).

Good Evo2 zero-shot performance on ClinVar

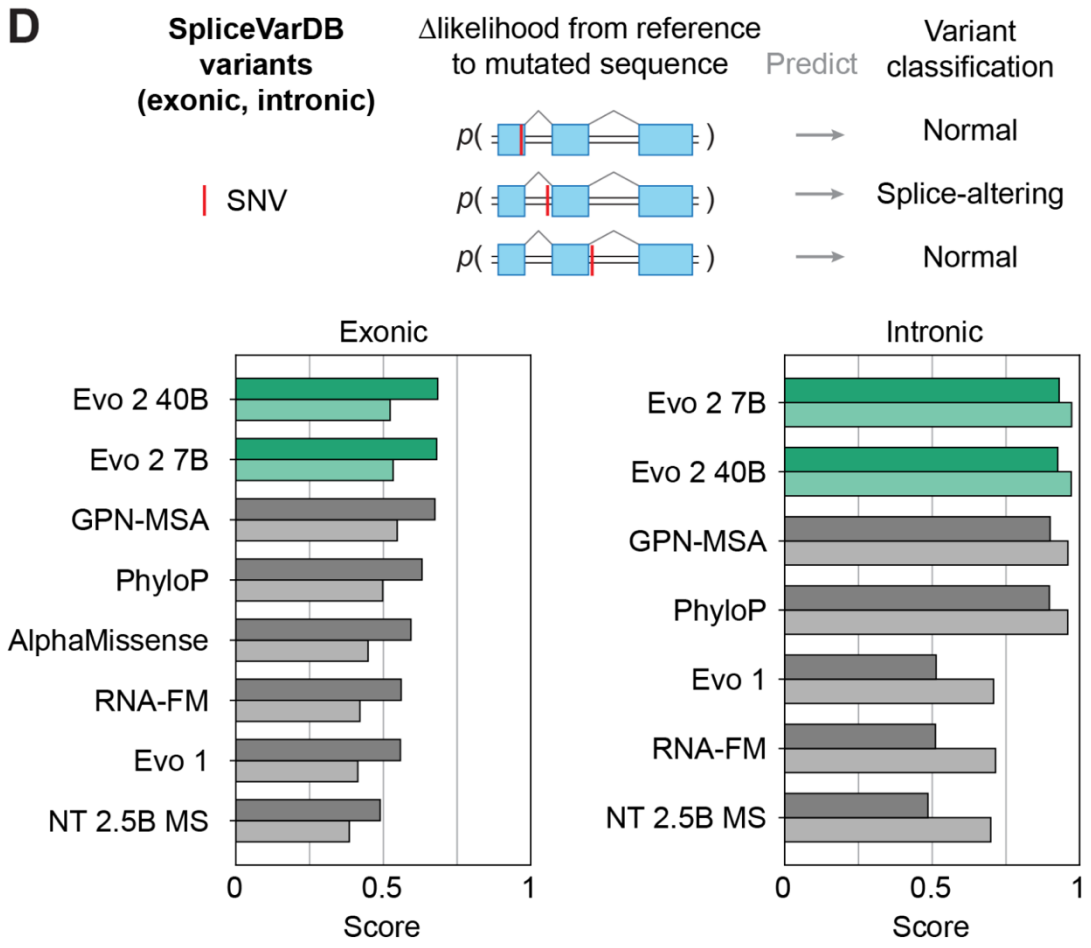




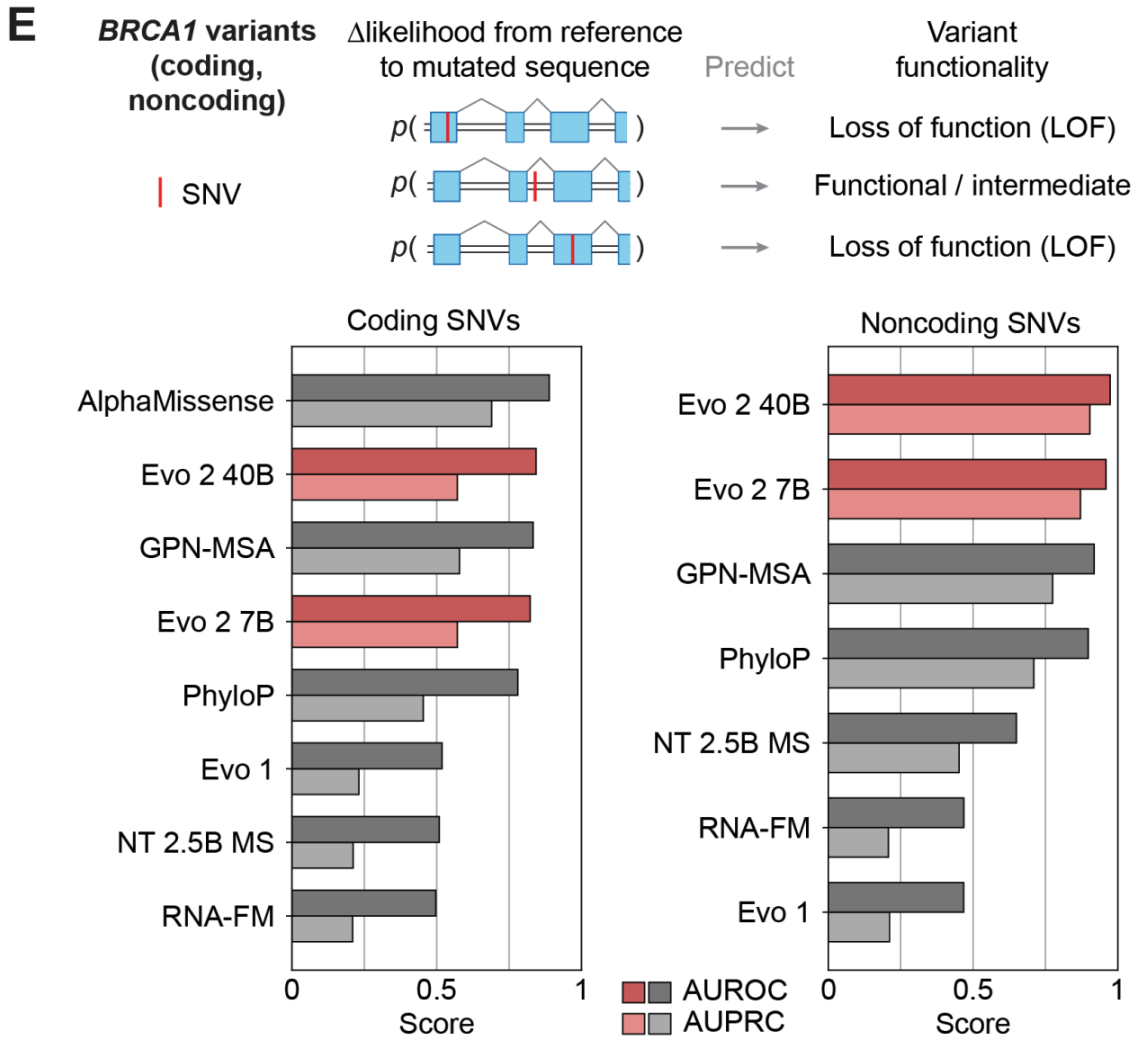
- Uses sequence likelihood as a proxy for variant effect
  - Loss-of-function variants tend to be pathogenic
- 15k coding, 38k noncoding variants
- Competes with protein LMs and exceeds DNA LMs

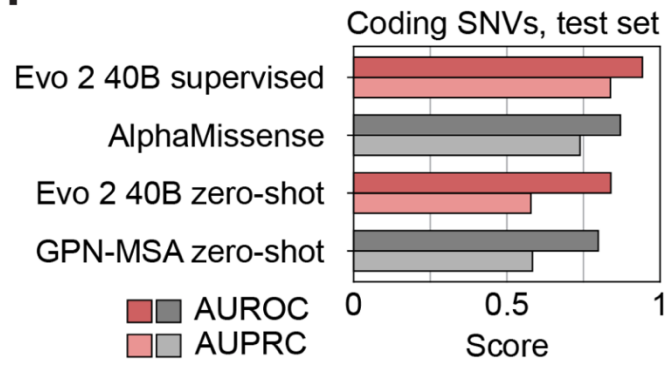
Evo2 also does well on SpliceVarDB

- SpliceVarDB is a database of experimentally validated splicing effects
- Evo2 excels on both exonic and intronic variants



Good performance on BRCA variants

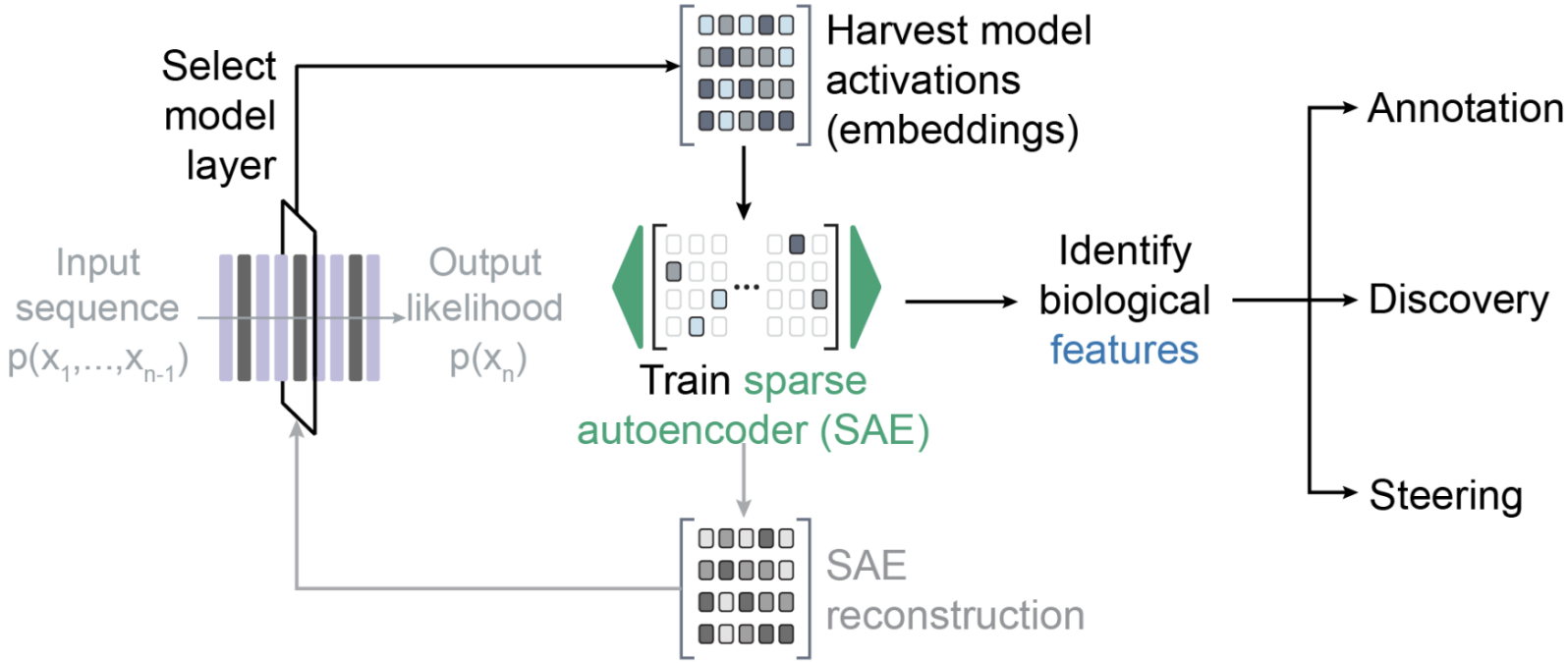




- Sets the high-water mark on BRCA variant classification among GLMs
- Especially strong performance on noncoding SNVs
- Real-world applicability for modeling disease

What is Evo2 doing under the hood?

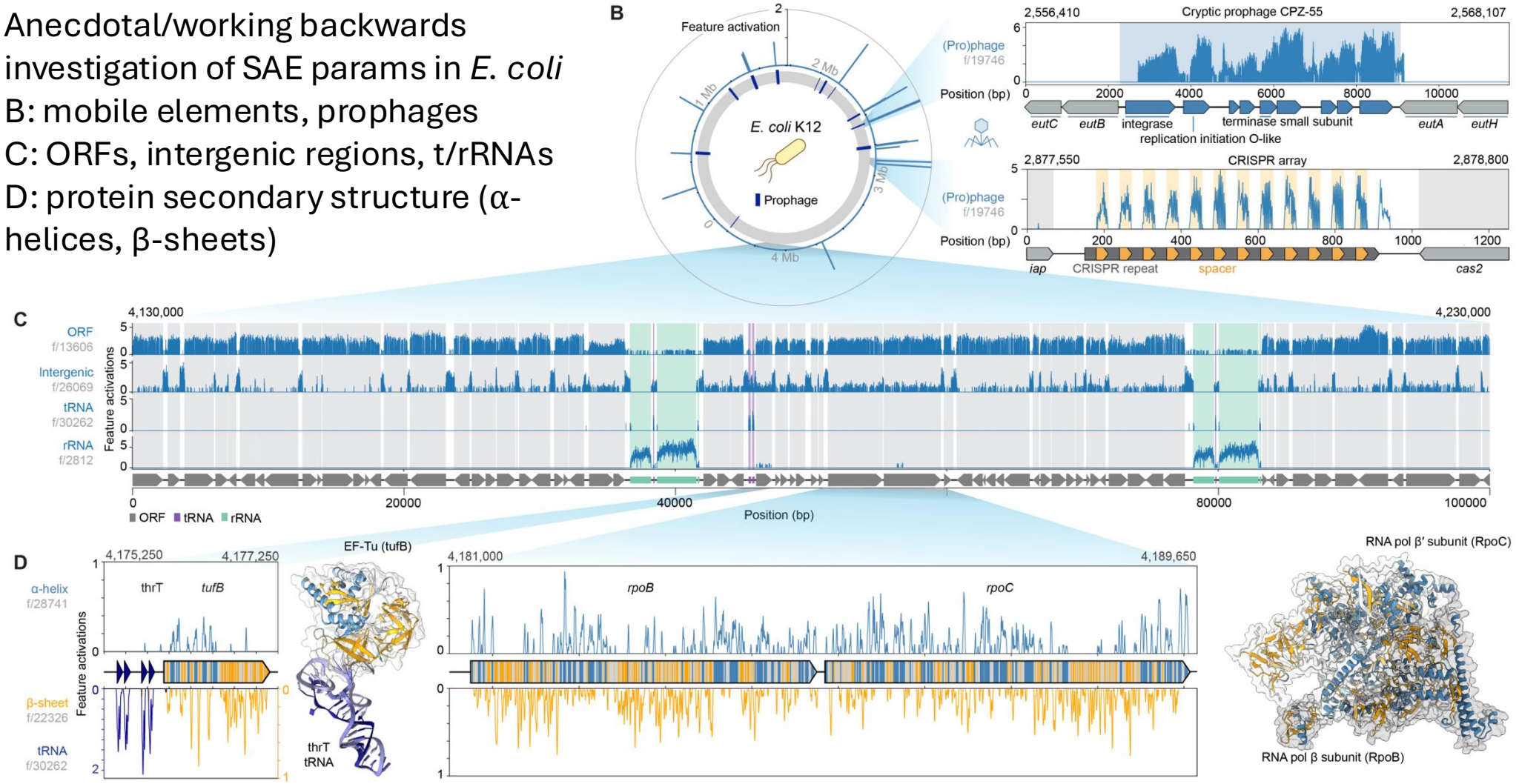
Interpretability by fitting a sparse autoencoder



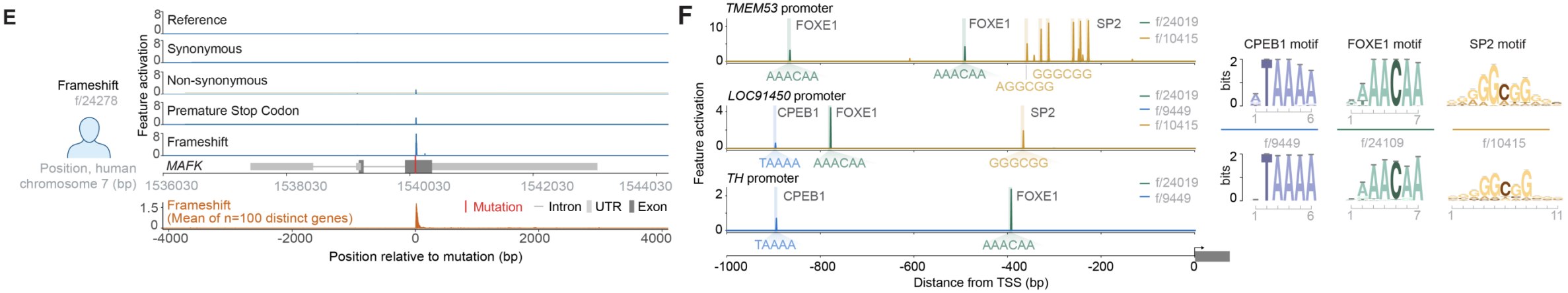
- Train a simple model to predict the outputs of the complex model
- Study the parameters of the simple model to estimate how the complex model operates
- In this case, fit a Batch-TopK Sparse AutoEncoder to Evo2's layer 26

Evo2 learns DNA content and structure

- Anecdotal/working backwards investigation of SAE params in *E. coli*
- B: mobile elements, prophages
- C: ORFs, intergenic regions, t/rRNAs
- D: protein secondary structure (α-helices, β-sheets)



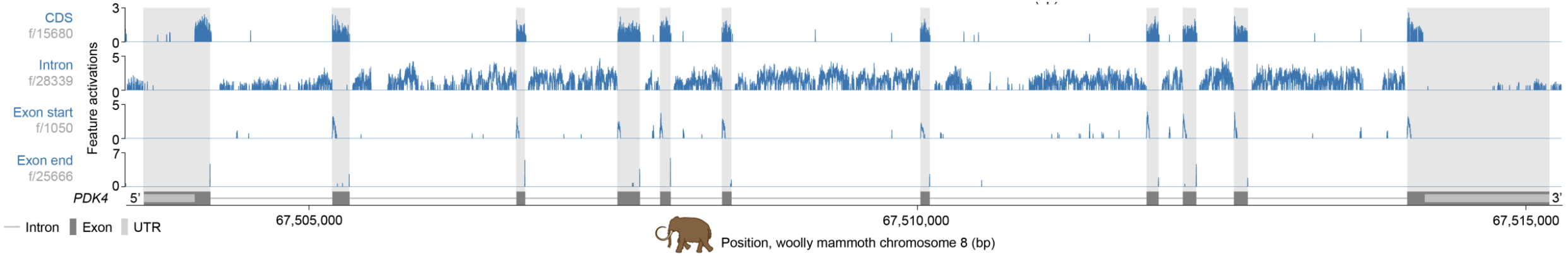
Interpretability for human predictions



- Fit a different sparse autoencoder to predictions on eukaryotes
- Features for frameshifts, premature stop codons imply mutational severity
- TF binding motif-specific features

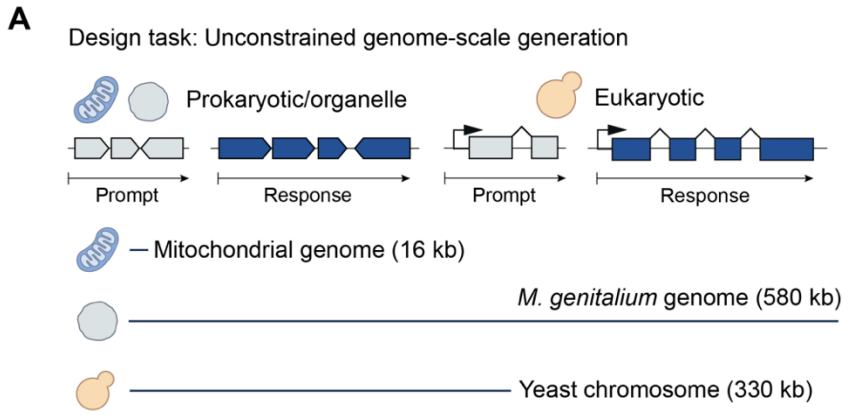
Interpretable features can annotate genomes

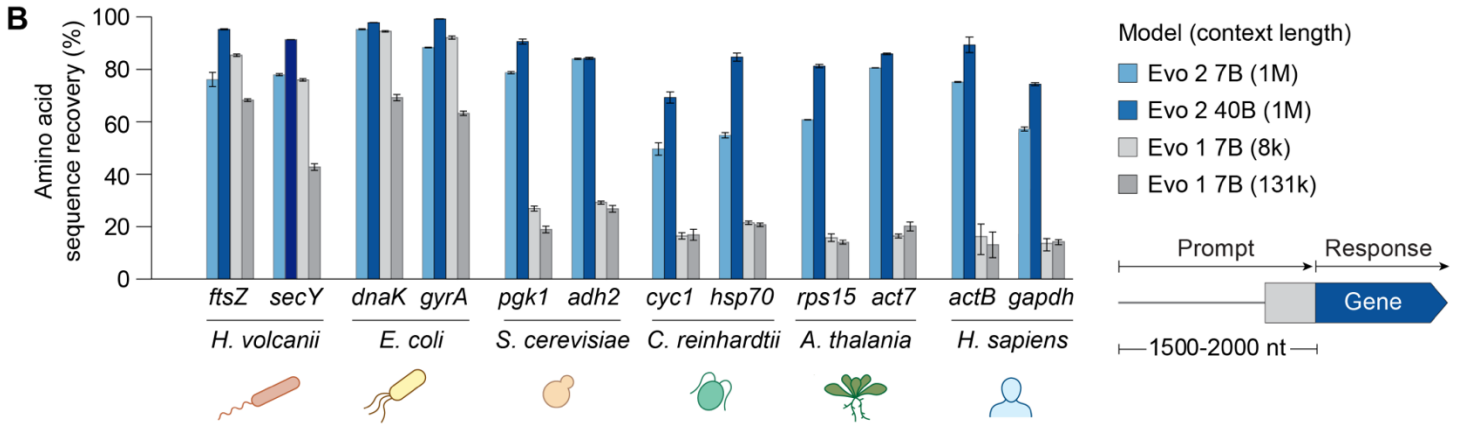
- Use four features to annotate the woolly mammoth genome
- Introns, exons, and coding regions



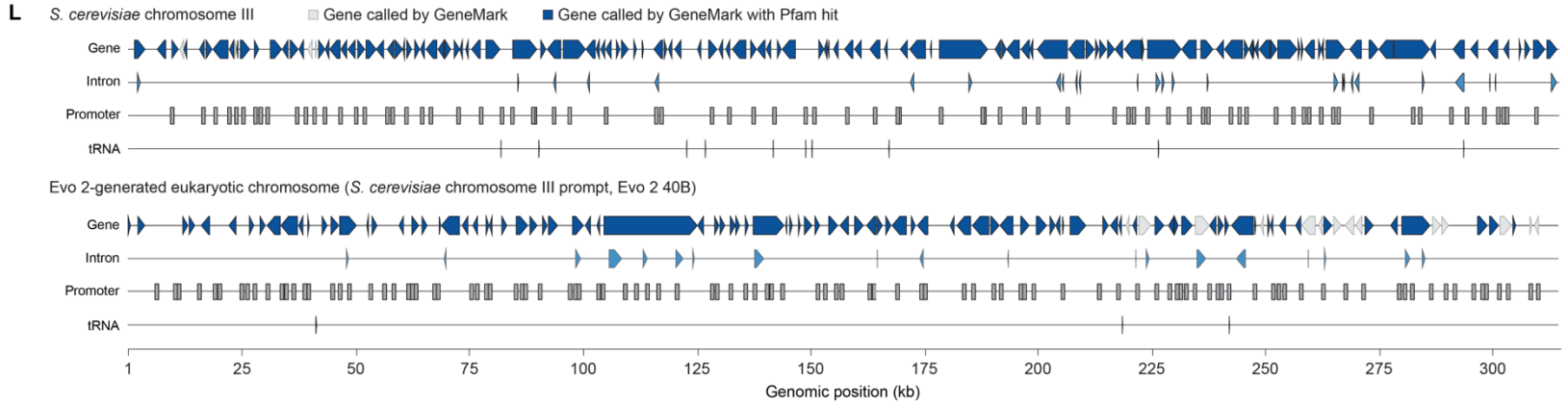
Genome sequence generation

Prokaryotic generation

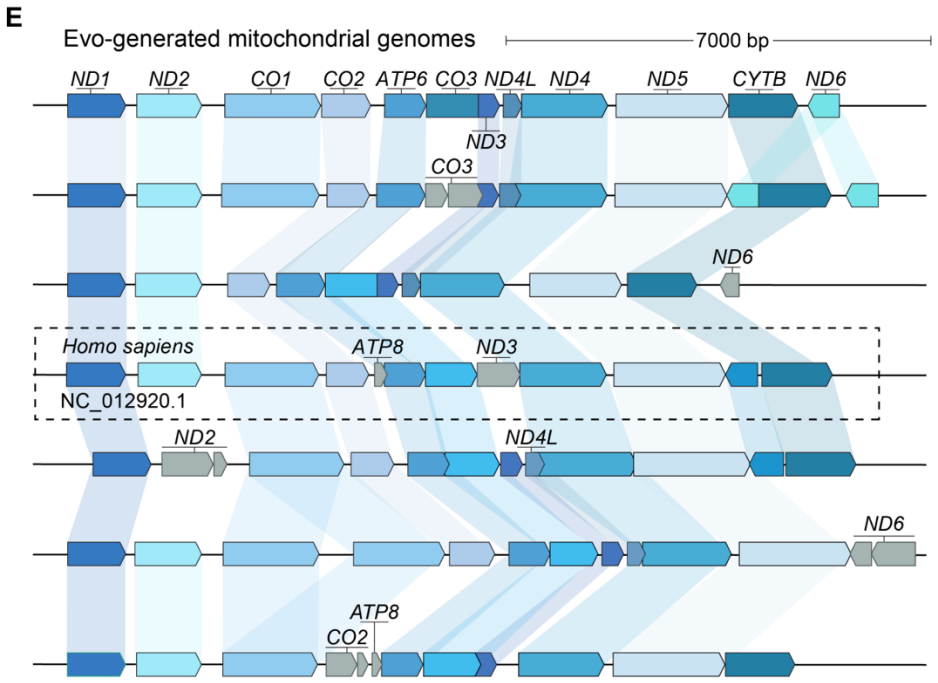


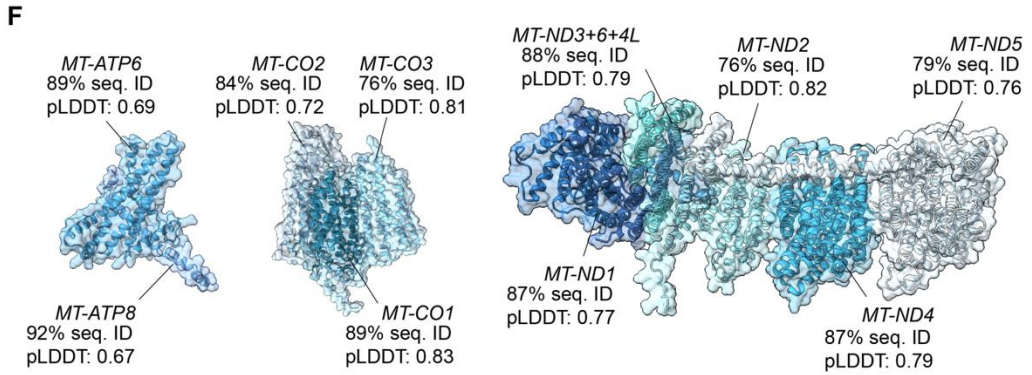


- Evo2 is a generative model
- Prompt with some upstream sequence, generate downstream
- These results could be improved with future advancements



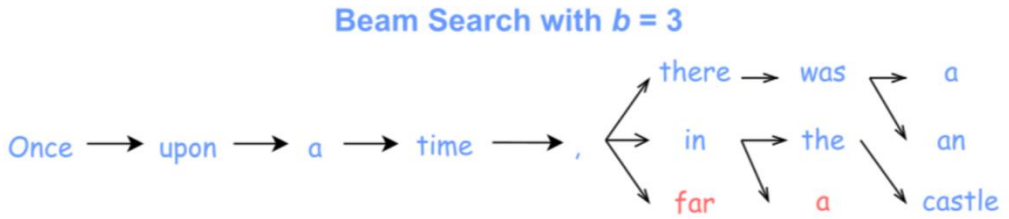
Human mitochondrial generation



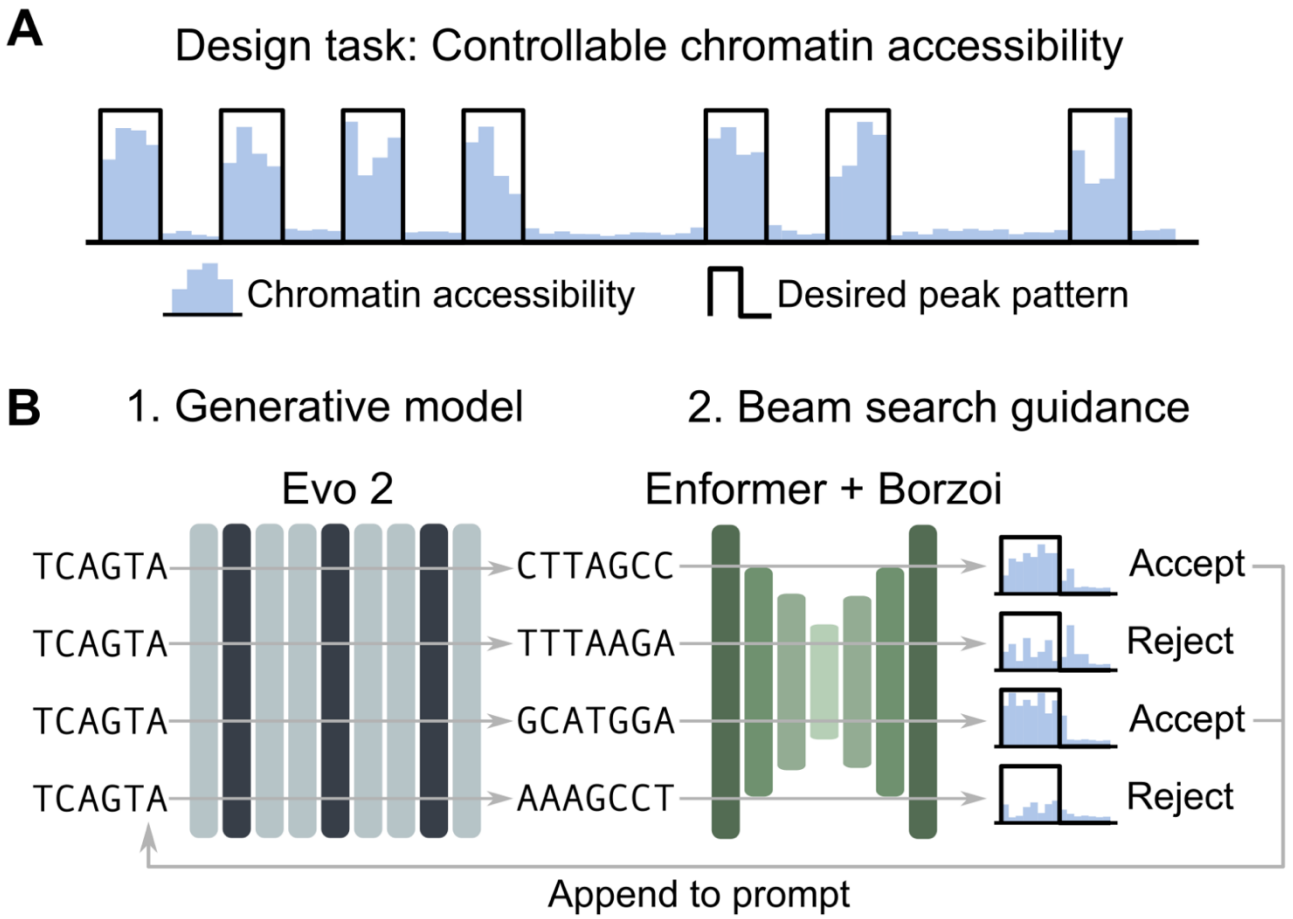


- Mitochondrial generation has variation, while still preserving synteny to human reference
- Proposed MT proteins are alike those naturally occurring

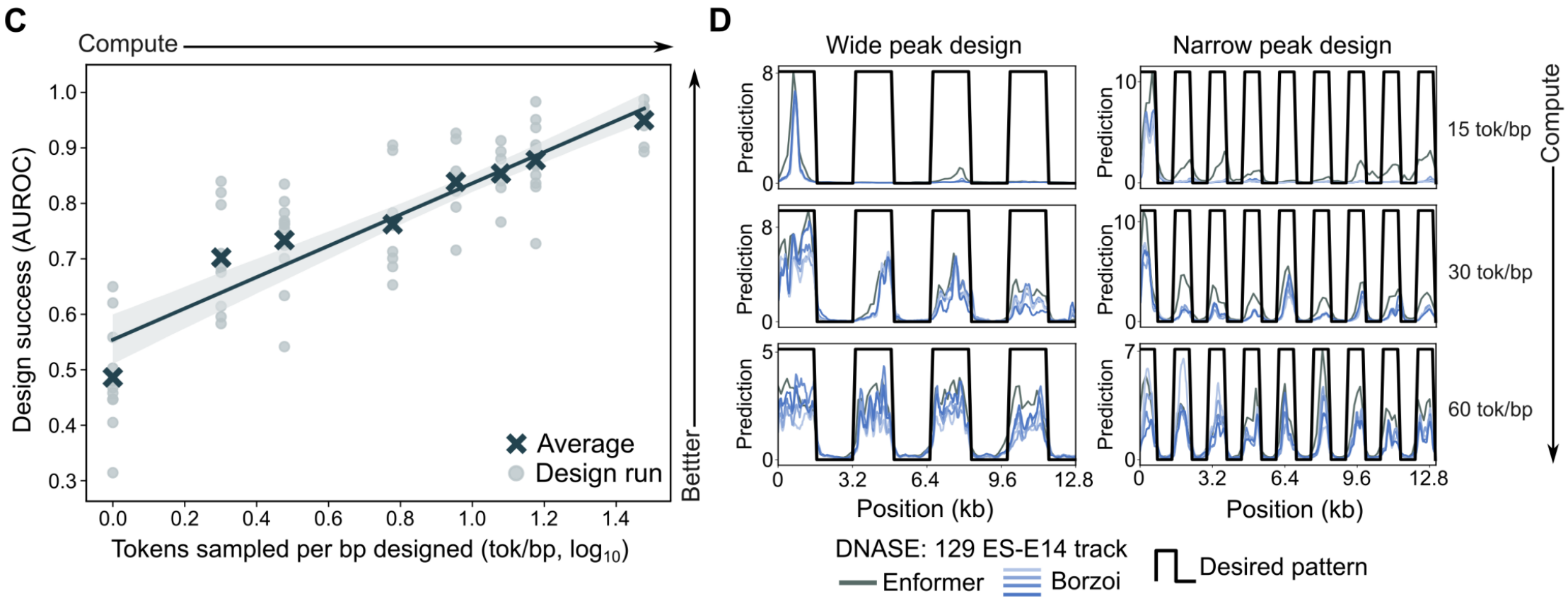
Eukaryotic generation



- Much harder task than prokaryotic generation due to epigenetics
- Goal: generate a sequence that obeys a given chromatin profile
- Generate 128bp chunks, check each for validity based on Enformer+Borzoi profiling, continue with best segment
- Iteratively build successful sequence with live feedback



More inference compute = more performance



Summary and thoughts

Major takeaways

- Strong performance on variant pathogenicity classification
- Trains without any labels or alignment information
- Ability to generate long sequences from different phylogenies
- Interpretability analysis to understand model workings
- Reasoning during inference
- Open-source code and data

Editorial critiques

- Few results on supervised fine-tuning – does it not work?
- Noticeable lack of classification tasks
  - Suggests lower performance than the field
- Autoregressive/generative models seem suboptimal for genomics
  - Unclear applications of generative tasks
  - Tradeoff of losing bidirectional information
- Enormous size makes it hard to use for the average researcher

Editorial praise

- Innovative methodology
  - Architecture can handle ultralong sequences
  - Training regime mirrors SOTA in natural language
  - Creative ways to minimize damage from repeat regions
- Self-supervised is the way to go
  - Learn terabytes of sequence
  - Variation across the tree of life is a proxy for human variation, enabling good zero-shot performance on variant effect prediction
  - Not limited by the availability of labelled data
  - One of the only models to train on the tree of life