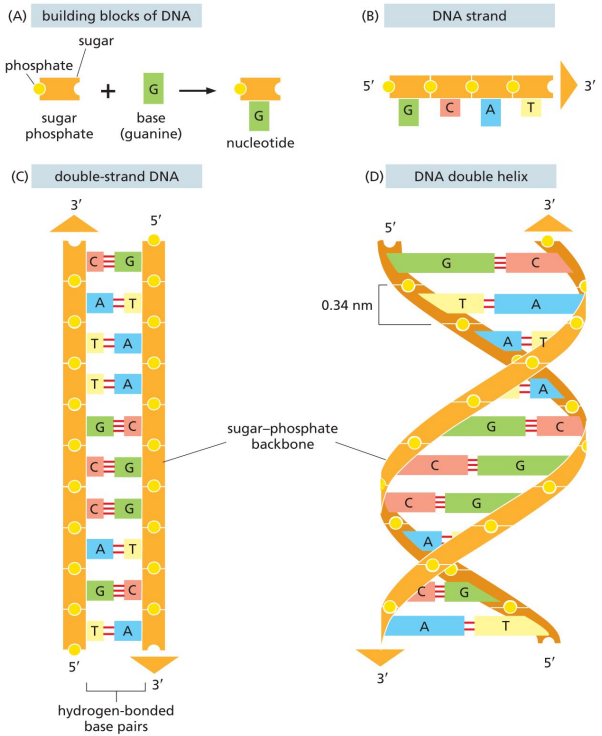
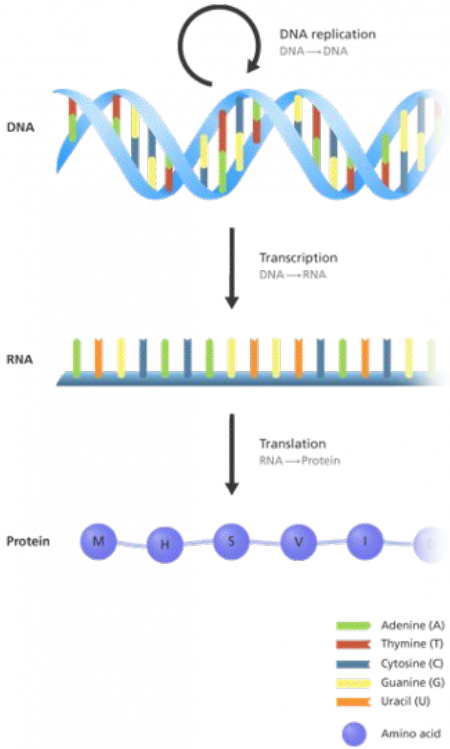
Sequence modeling and design from molecular to genome scale with Evo

Presented by Xingyu Chen and Cong Yu Fang

DNA molecules

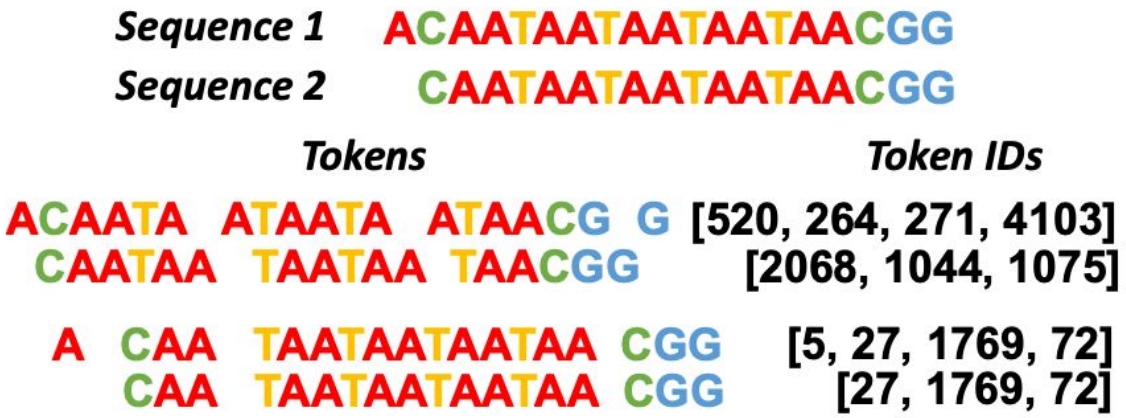




https://biomedguide.com/biology/central-dogma-of-biology/
https://nerd.wwnorton.com/ebooks/epub/ecb6/EPUB/content/5.1-chapter05.xhtml

DNA tokenization

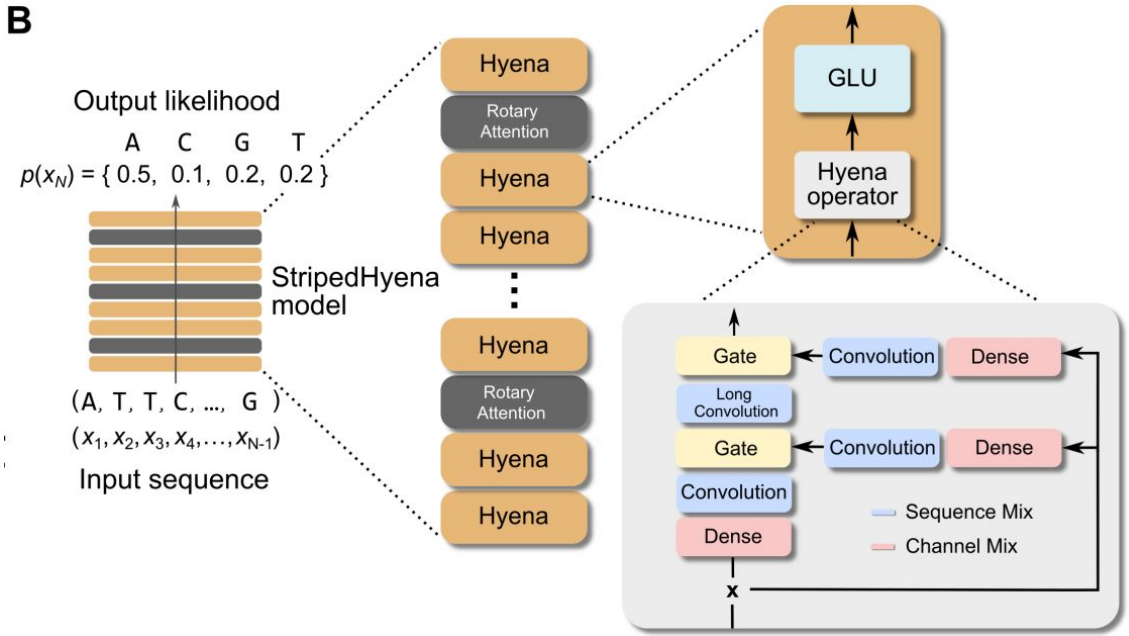
- K-mer
  - each contiguous k-length genome segment is considered as a token.
- Byte Pair Encoding
  - iteratively merges the most frequent pairs of characters to create a vocabulary of subword units
- Single-nucleotide
  - A, G, C, T



Zhou et al. DNABERT2. ICLR. 2024

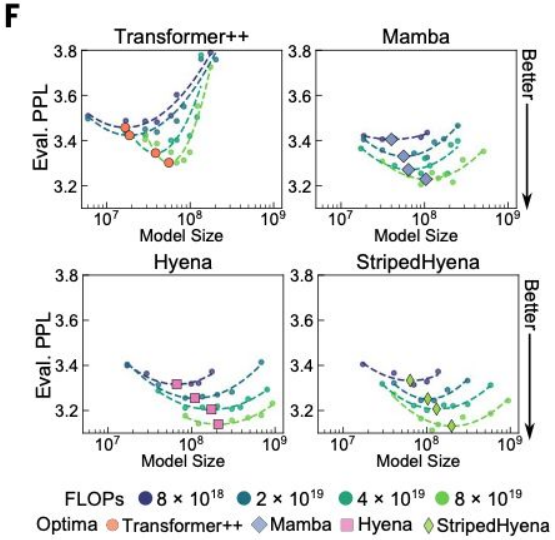
Evo: hybrid of Hyena and attention

- Handle long context
- 29 Hyena layers and 3 attention layers
- Long convolution, i.e., filter size = input length



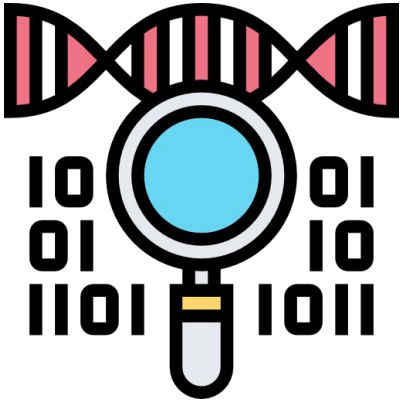
Training

- Context length:
  - Stage 1: 8k
  - Stage 2: 131k
- Data:
  - Bacterial and archaeal genomes from the Genome Taxonomy Database
  - Curated prokaryotic viruses from the IMG/VR v4 database
  - Plasmid sequences from the IMG/PR database
- Scaling Law



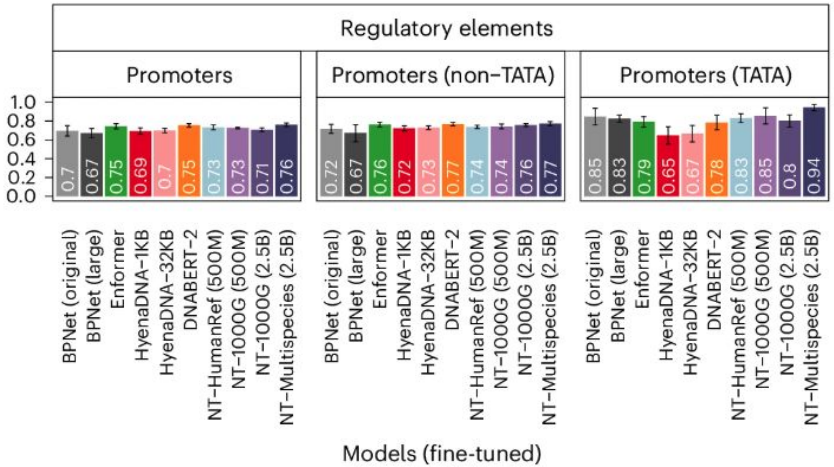
Evaluation

- Previous works
- Does Evo understand biology
- Bio sequence generation by Evo



Evaluation: does it understand biology?

- Previous works: transfer learning for downstream task
  - Probing, e.g., take the embedding from a pretrained model, train a head (linear, CNN, etc.) for the task
  - Finetuning the entire model + a task-specific head



Dalla-Torre et al., Nucleotide Transformer: building and evaluating robust foundation models for human genomics. Nature Methods. 2024.

Evaluation: does it understand biology?

- Previous works' success may come from high-dim embedding and/or finetuning

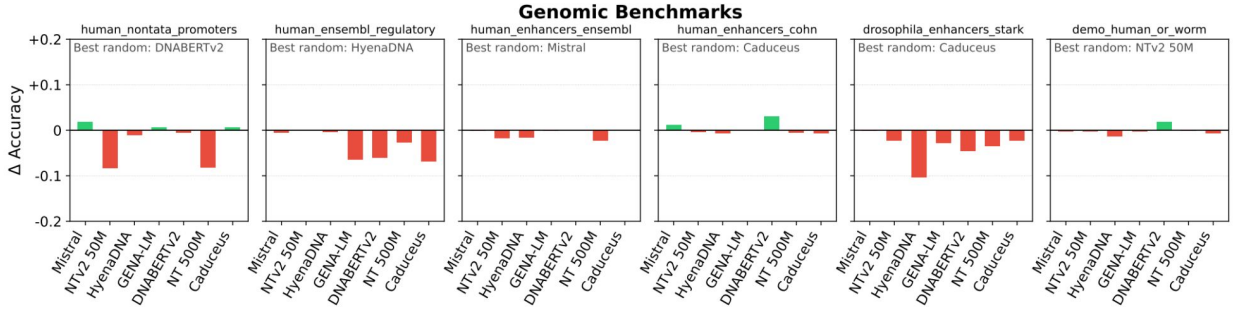


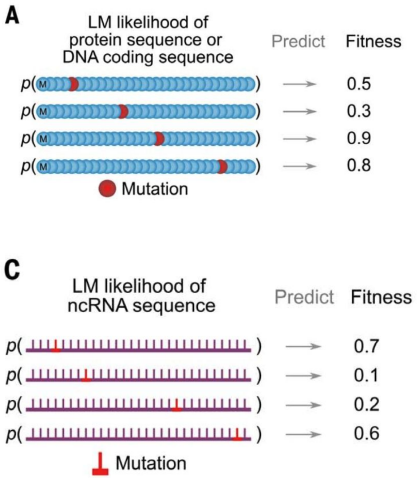
Figure 2: **Difference of performance between pretrained models and the best random model on NT Benchmark.** For each task, we finetuned each model, starting from both pretrained and randomly initialized weights. Green bars indicate the advantage of pretrained models, and red bars indicate the advantage of the best random model. The best random model consistently outperforms several pretrained ones on each task, highlighting the inefficiency of current pretraining approaches in genomics. In most cases, the best random model is Caduceus which has only 8M parameters, yet it has better performance than much bigger pretrained models such as NT 500M, GENA-LM, DNABERTv2, NTv2 50M, and Mistral.

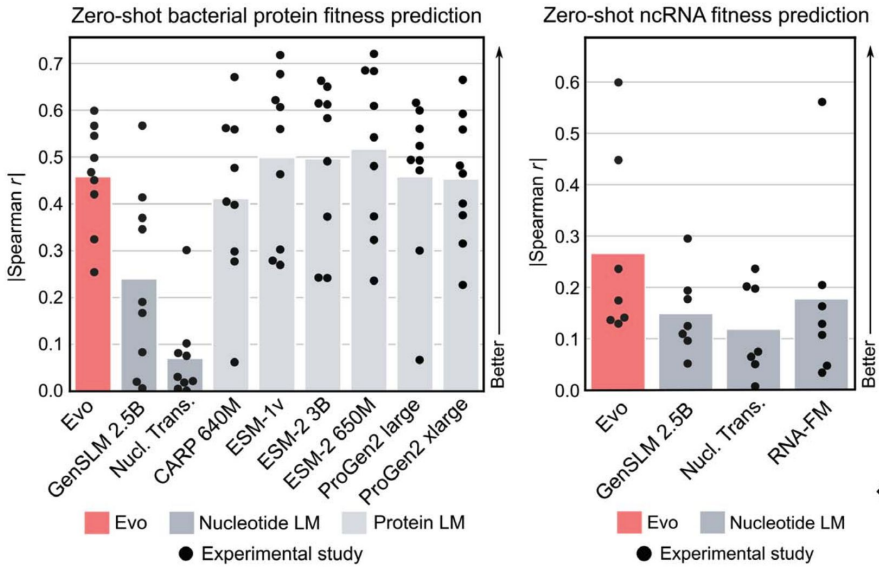
Vishniakov et al., GENOMIC FOUNDATIONLESS MODELS: PRETRAINING DOES NOT PROMISE PERFORMANCE. 2024.

Evaluation: does it understand biology?

How does a change in DNA sequence affect protein and RNA functions?

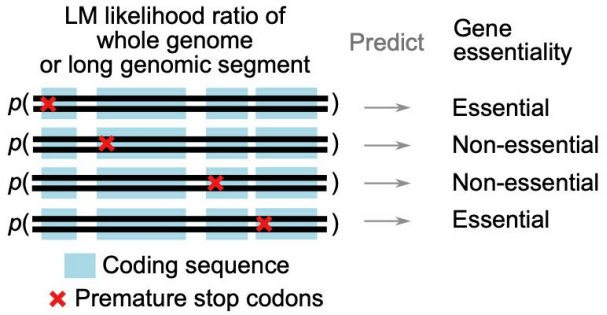
- **Fitness** measures how well the mutated sequence performs its biological function
- Deep mutational scanning

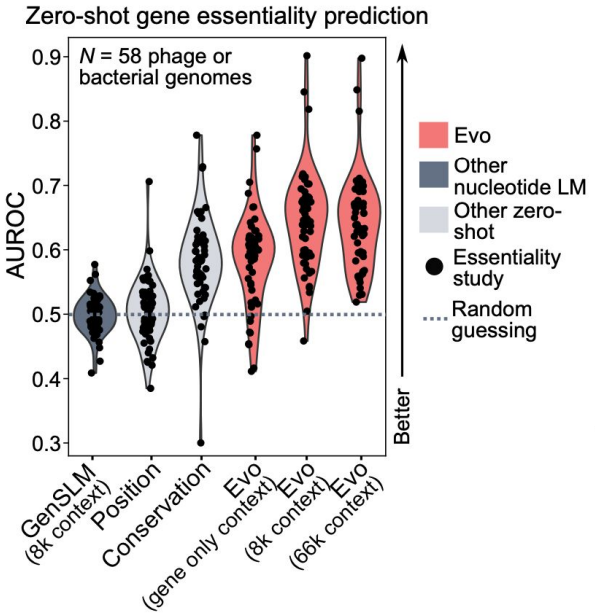




Evaluation: does it understand biology?

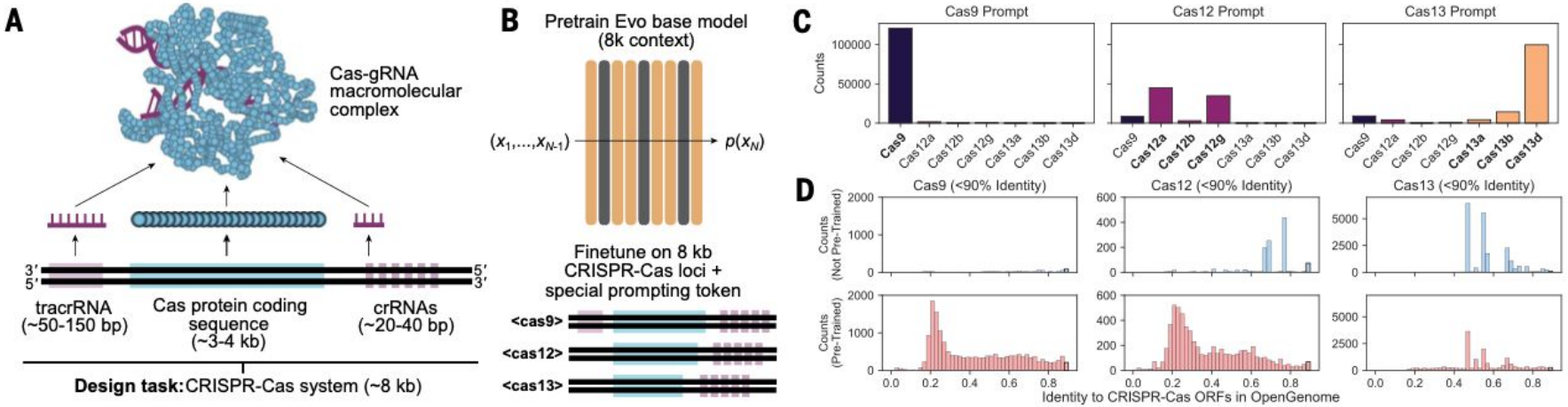
- Essential genes are genes required for a cell or an organism to survive





Generation of functional DNA sequences

- Crispr-cas: a gene cutting system



Insights

Pros:

- Handle long genome sequences
- Demonstrate that Evo “understands” some biology

Cons:

- Performance has large variance; cannot reproduce some of the results since evaluation data is not released
- “Pseudo” multimodal evaluation: things like post-translational modifications, non-canonical amino acids are not directly reflected in the genome
- The 131k-context model sometimes perform worse than the 8k version

Aside:

- Prior work HyenaDNA: context length of 1M. But no comparison showed
- Unsure why it outperforms previous works (e.g., context length, Single-nucleotide resolution, training data?)

Thank You For Listening !

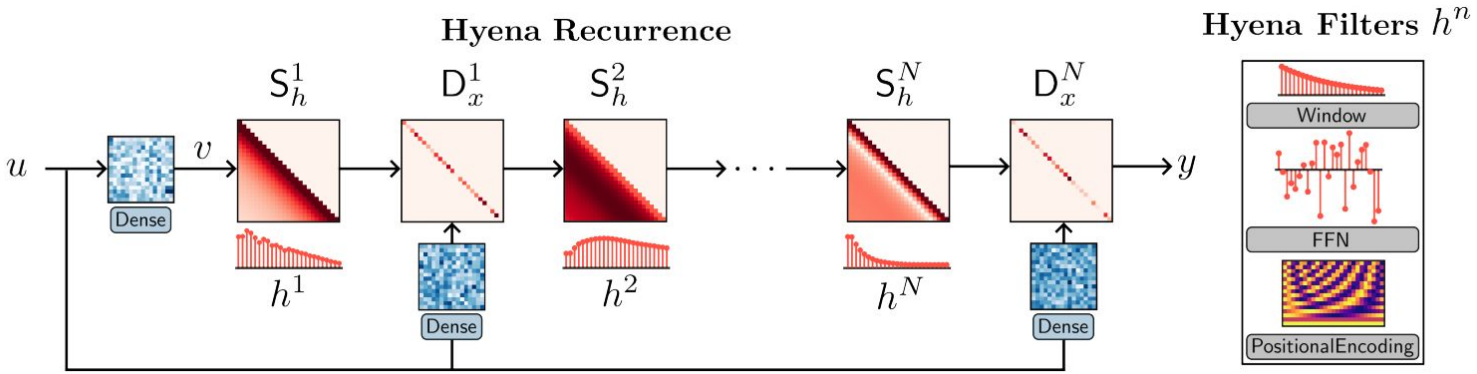
Any Questions?

Hyena Operator

$$\mathcal{F}[f * g] = \mathcal{F}[f]\,\mathcal{F}[g].$$

complexity of the general convolution algorithm: O(N^2)

O(nlog(n)) complexity (in Fourier domain)



Poli et al., Hyena Hierarchy: Towards Larger Convolutional Language Models. 2023