Sequence modeling and design from molecular to genome scale with Evo

March 4th, Deep Learning Reading Group

Outline

1. Introduction
    a. Previous work
    b. Why LLMs in genomics?
    c. Evo
2. Data Collection
3. Methods
    a. Model architecture
4. Results
5. Discussion
    a. Limitations and challenges
    b. Ethical considerations
    c. Evo 2
6. Questions

Introduction

Previous work

Previous work in the field has been:

- Focused on modality-specific models specialized to proteins, coding sequences, RNA, or regulatory DNA
- Limited to the design of single molecules, simple complexes, or short DNA sequences
- Due to Transformer architecture, constrained to short context lengths and tokens without single-nucleotide resolution

Why LLMs in genomics?

LLMs can uncover patterns in DNA, enabling functional predictions

They can analyze large datasets and complex interactions

Future opportunities involving LLMs:

- Gene-editing
- Disease diagnostics
- Synthetic biology

Evo

Evo is a foundation LLM designed to interpret and generate DNA sequences at various biological scales (from nucleotide to genome level)

- ➔ Predicts molecular interactions
- ➔ Generates genetic sequences
- ➔ Analyzes genomic variation

Created by researchers at the Arc Institute, an independent non-profit organization focused on biomedical research

- ➔ Collaborators include Stanford, UC Berkeley, and UCSF

Data Collection

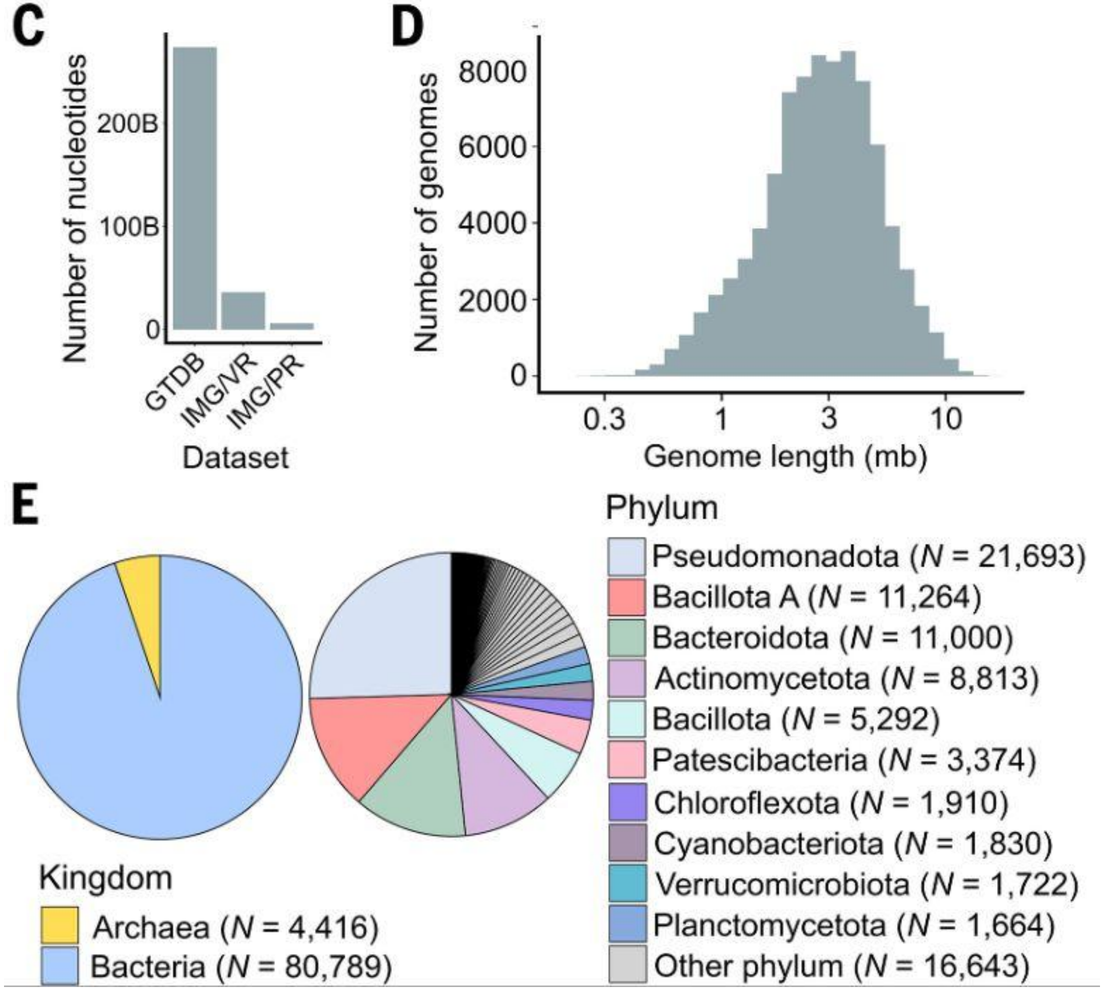
Data Collection

The OpenGenome pre-training dataset was compiled from three sources:

1. Bacterial and archaeal genomes from the Genome Taxonomy Database
2. Curated prokaryotic viruses from the IMG/VR v4 database
3. Plasmid sequences from the IMG/PR database

300 billion nucleotide tokens in total, 100x more data than HyenaDNA

Excluded viral genomes that infect eukaryotes

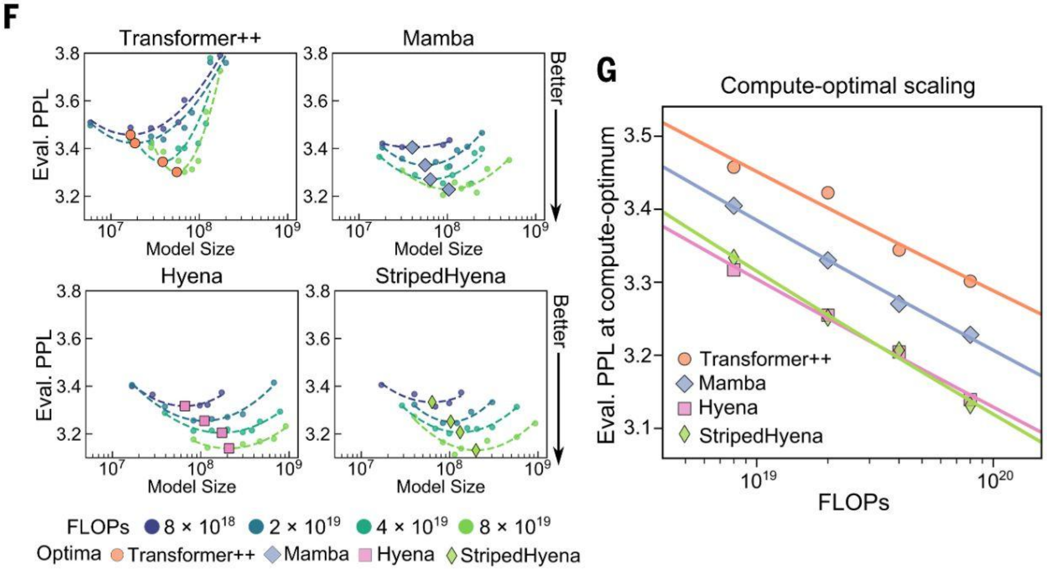

C
Number of nucleotides
200B
100B
0
GTDB
IMG/VR
IMG/PR
Dataset
D
Number of genomes
8000
6000
4000
2000
0
0.3
1
3
10
Genome length (mb)
E
Kingdom
Archaea (N = 4,416)
Bacteria (N = 80,789)
Phylum
Pseudomonadota (N = 21,693)
Bacillota A (N = 11,264)
Bacteroidota (N = 11,000)
Actinomycetota (N = 8,813)
Bacillota (N = 5,292)
Patescibacteria (N = 3,374)
Chloroflexota (N = 1,910)
Cyanobacteriota (N = 1,830)
Verrucomicrobiota (N = 1,722)
Planctomycetota (N = 1,664)
Other phylum (N = 16,643)

Methods

Model architecture

More than 300 models were trained across four architectures:

- ➔ Transformer++
- ➔ Mamba
- ➔ Hyena
- ➔ StripedHyena



Perplexity: a measure of

next token prediction quality

Model architecture

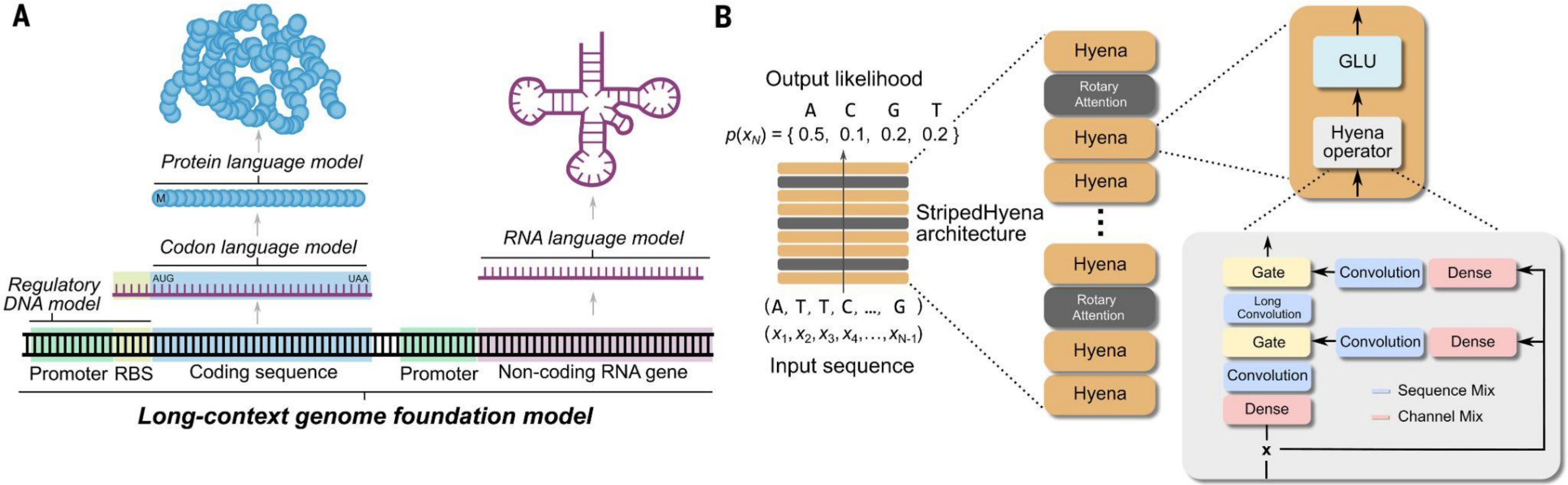
StripedHyena architecture:

➔ 32 blocks at a model width of 4096 dimensions

The model is a hybrid of:

➔ 29 layers of data-controlled convolutional operators (hyena layers), interleaved with

➔ 3 layers (10%) of multihead attention equipped with rotary position embeddings (RoPEs)


A
Protein language model
M
Codon language model
AUG
UAA
Regulatory DNA model
RNA language model
Promoter
RBS
Coding sequence
Promoter
Non-coding RNA gene
Long-context genome foundation model
B
Output likelihood
A C G T
$p(x_N)$ = { 0.5, 0.1, 0.2, 0.2 }
StripedHyena architecture
(A, T, T, C, ..., G)
$(x_1, x_2, x_3, x_4, \ldots, x_{N-1})$
Input sequence
Hyena
Rotary Attention
Hyena
Hyena
⋮
Hyena
Rotary Attention
Hyena
Hyena
GLU
Hyena operator
Gate
Convolution
Dense
Long Convolution
Gate
Convolution
Dense
Convolution
Dense
x
Sequence Mix
Channel Mix

Model architecture

Final Evo model:

➔ StripedHyena architecture

➔ 7 billion parameters

➔ Context length: 131,072 tokens (single nucleotide tokenization)

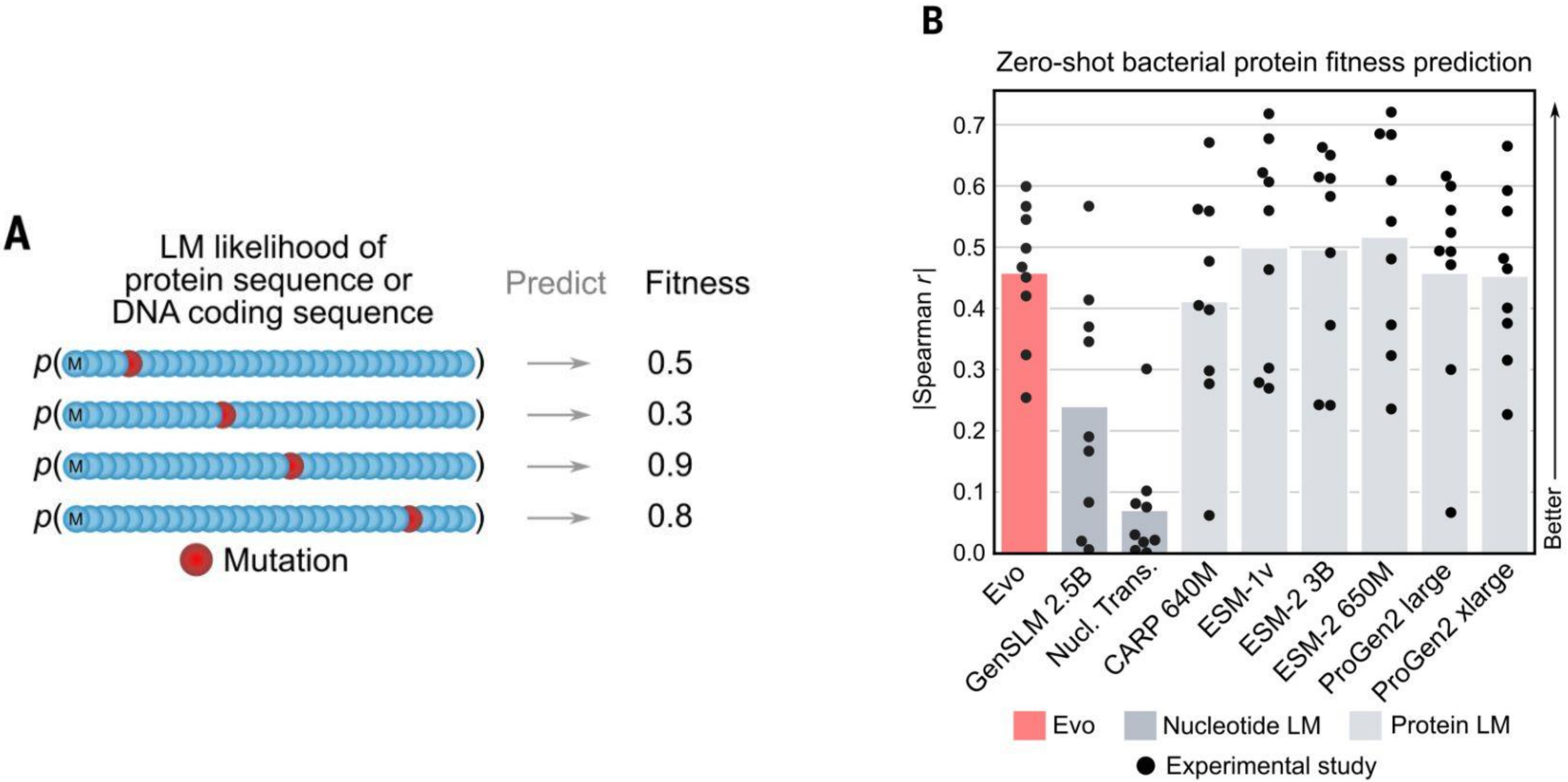
➔ 1000x larger than HyenaDNA

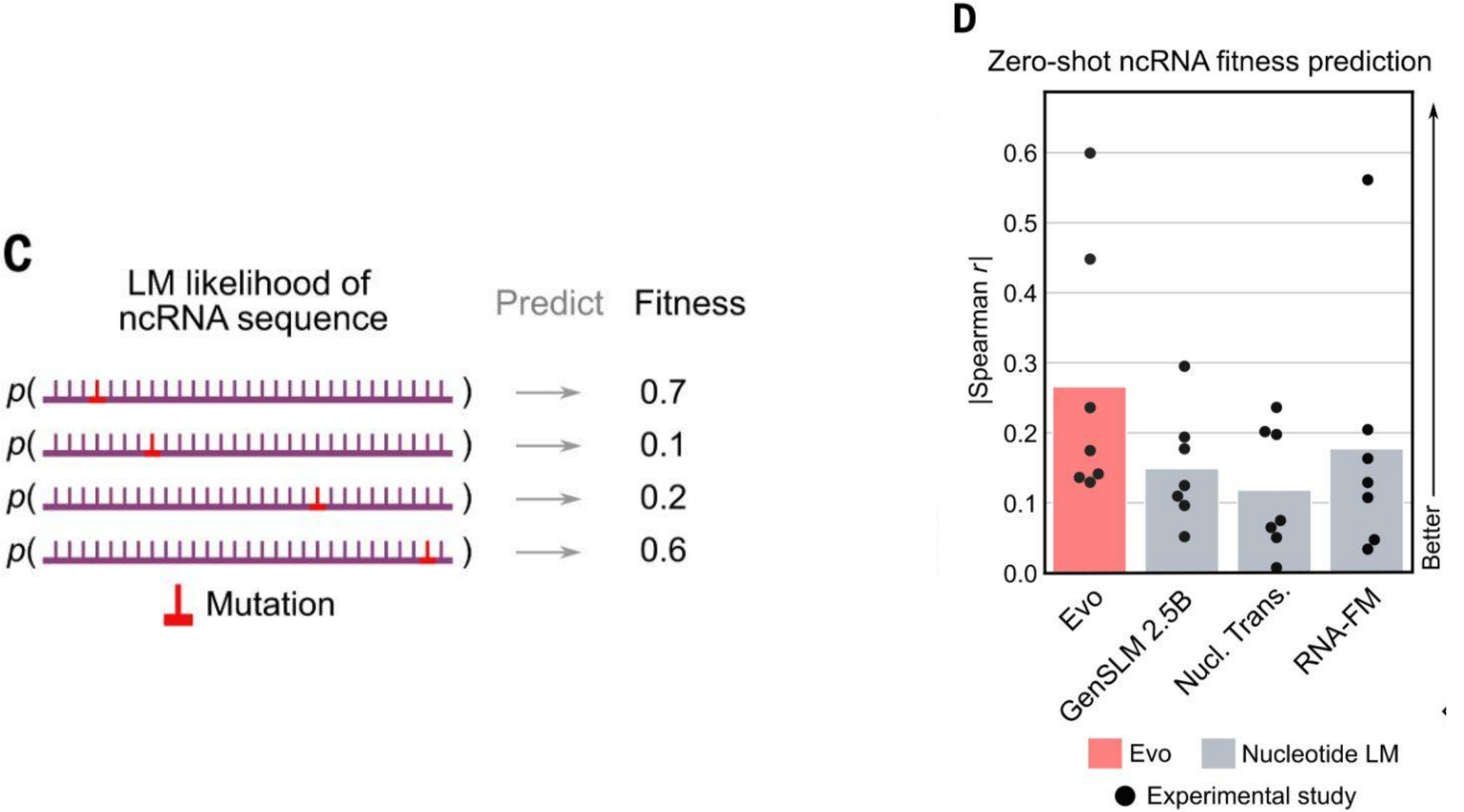
Results

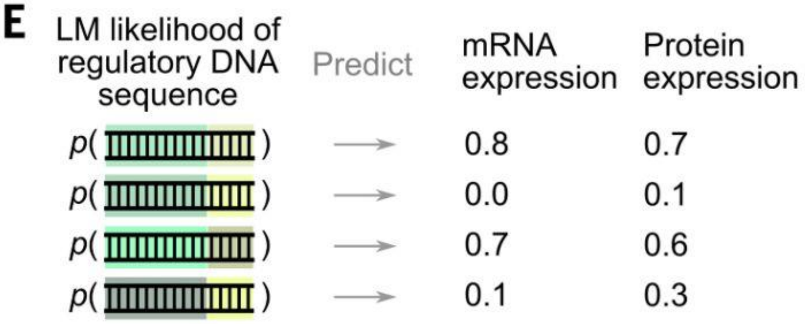
Evo learns across DNA, RNA, and protein modalities

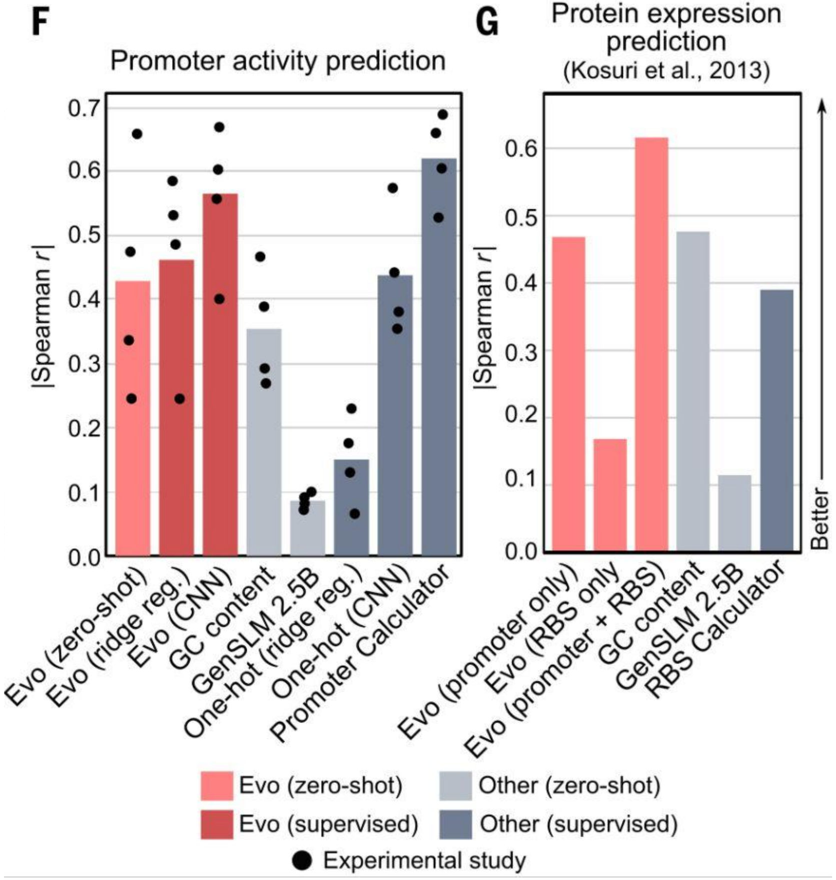
First, the zero-shot performance of the model was evaluated on several biologically relevant tasks:

1. Predicting mutational effects on protein function
2. Predicting mutational effects on ncRNA function
3. Predicting activity of regulatory DNA


A
LM likelihood of protein sequence or DNA coding sequence
Predict
Fitness
p(M)
0.5
0.3
0.9
0.8
Mutation
B
Zero-shot bacterial protein fitness prediction
|Spearman r|
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
Better
Evo
GenSLM 2.5B
Nucl. Trans.
CARP 640M
ESM-1v
ESM-2 3B
ESM-2 650M
ProGen2 large
ProGen2 xlarge
Evo
Nucleotide LM
Protein LM
Experimental study


C
LM likelihood of ncRNA sequence
Predict
Fitness
p( )
0.7
p( )
0.1
p( )
0.2
p( )
0.6
Mutation
D
Zero-shot ncRNA fitness prediction
|Spearman r|
0.0
0.1
0.2
0.3
0.4
0.5
0.6
Better
Evo
GenSLM 2.5B
Nucl. Trans.
RNA-FM
Evo
Nucleotide LM
Experimental study

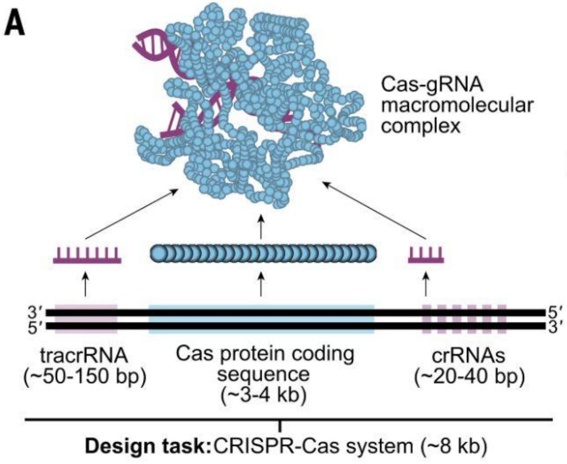

E
LM likelihood of regulatory DNA sequence
Predict
mRNA expression
Protein expression
p( )
0.8
0.7
p( )
0.0
0.1
p( )
0.7
0.6
p( )
0.1
0.3


F
Promoter activity prediction
|Spearman r|
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
Evo (zero-shot)
Evo (ridge reg.)
Evo (CNN)
GC content
GenSLM 2.5B
One-hot (ridge reg.)
One-hot (CNN)
Promoter Calculator
G
Protein expression prediction
(Kosuri et al., 2013)
|Spearman r|
0.0
0.1
0.2
0.3
0.4
0.5
0.6
Better
Evo (promoter only)
Evo (RBS only)
Evo (promoter + RBS)
GC content
GenSLM 2.5B
RBS Calculator
Evo (zero-shot)
Evo (supervised)
Other (zero-shot)
Other (supervised)
● Experimental study

Generative design of CRISPR-Cas molecular complexes

Evo was fine-tuned on a dataset of 82,430 genomic loci with 8 kb-length genomic sequences containing CRISPR-Cas systems

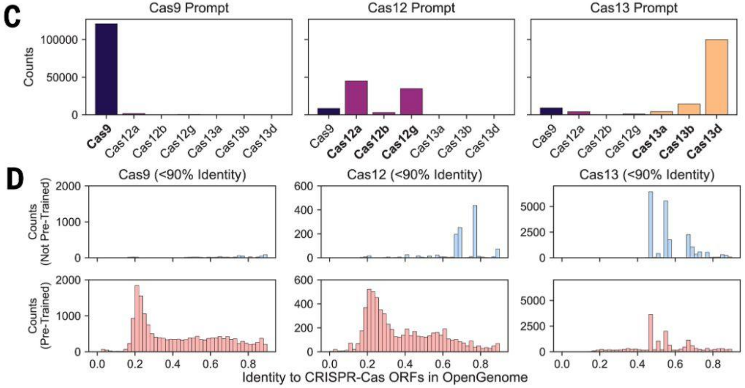
➔ CRISPR-Cas systems comprise >=1 CRISPR ncRNAs and >=1 Cas proteins

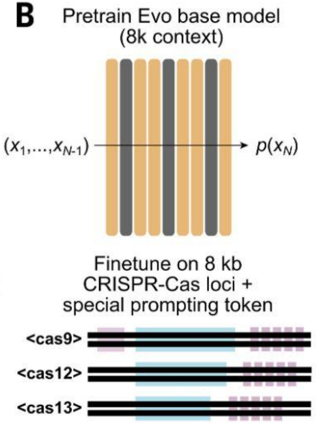

A
Cas-gRNA macromolecular complex
3′
5′
5′
3′
tracrRNA (~50-150 bp)
Cas protein coding sequence (~3-4 kb)
crRNAs (~20-40 bp)
Design task: CRISPR-Cas system (~8 kb)

Generative design of CRISPR-Cas molecular complexes

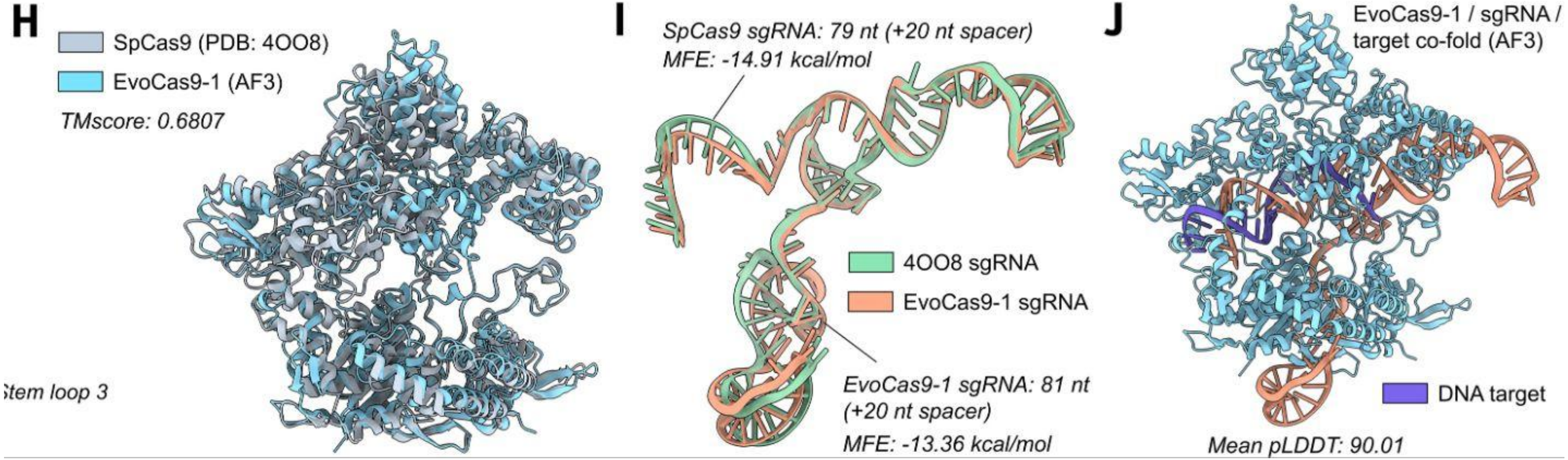
To evaluate the quality of Cas generation:

- ➔ Compared generated Cas proteins to canonical proteins
- ➔ Evaluated AlphaFold2 structure predictions against canonical structures





B
Pretrain Evo base model (8k context)
$(x_1,\ldots,x_{N-1})$
$p(x_N)$
Finetune on 8 kb CRISPR-Cas loci + special prompting token
<cas9>
<cas12>
<cas13>


H
SpCas9 (PDB: 4OO8)
EvoCas9-1 (AF3)
TMscore: 0.6807
tem loop 3
I
SpCas9 sgRNA: 79 nt (+20 nt spacer)
MFE: -14.91 kcal/mol
4OO8 sgRNA
EvoCas9-1 sgRNA
EvoCas9-1 sgRNA: 81 nt
(+20 nt spacer)
MFE: -13.36 kcal/mol
J
EvoCas9-1 / sgRNA / target co-fold (AF3)
DNA target
Mean pLDDT: 90.01

Generative design of CRISPR-Cas molecular complexes

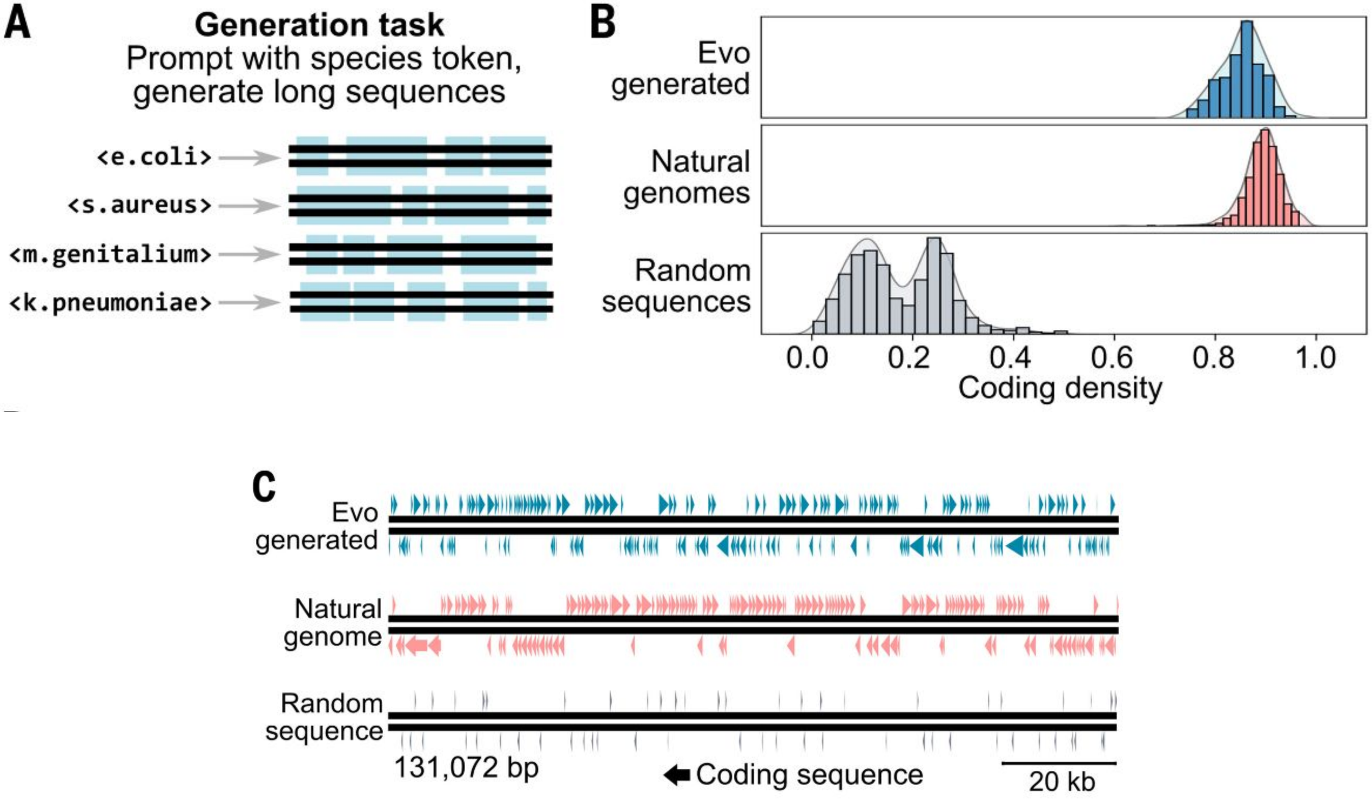
To evaluate the quality of Cas generation:

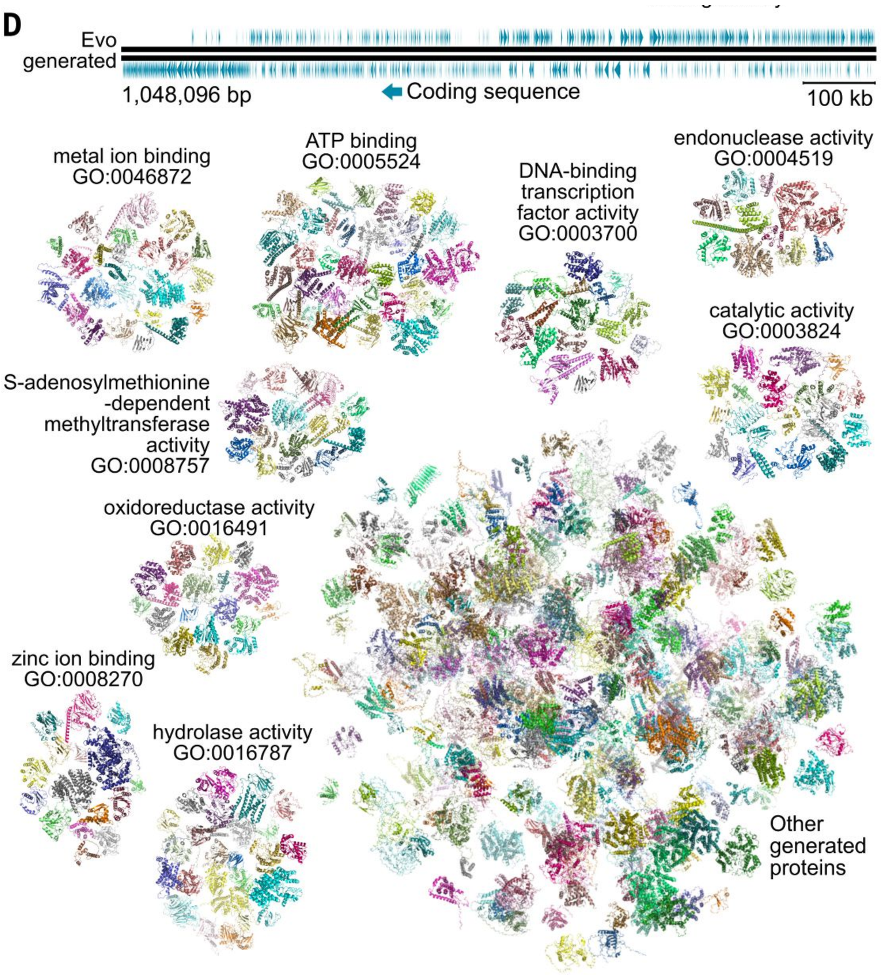
- ➔ Tested viable systems experimentally, focusing on Cas9 as metric
- ➔ ~2 million Evo-generated sequences for Cas9 loci
- ➔ Filtered to 11 systems with robust test scores

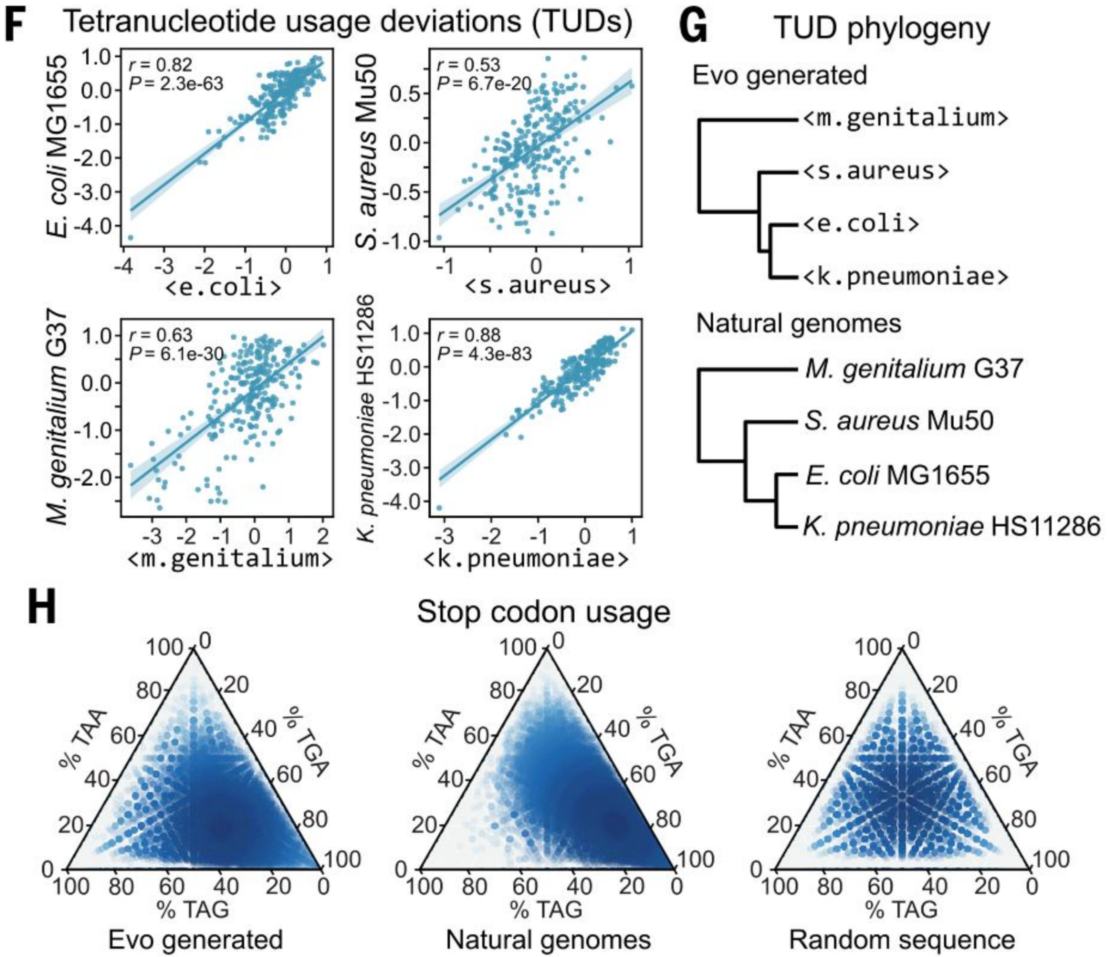
Generating DNA sequences at genome scale

The model generated bacterial genomes using species-level tokens:

- ➔ The smallest “minimal” bacterial genomes are ~580 kb in length
- ➔ To evaluate similarity between the generated sequences and natural genomes, CheckM was used
    - ◆ CheckM: a tool designed to assess the quality of bacterial DNA sequenced from nature


A
Generation task
Prompt with species token, generate long sequences
<e.coli>
<s.aureus>
<m.genitalium>
<k.pneumoniae>
B
Evo generated
Natural genomes
Random sequences
0.0
0.2
0.4
0.6
0.8
1.0
Coding density
C
Evo generated
Natural genome
Random sequence
131,072 bp
Coding sequence
20 kb


D
Evo
generated
1,048,096 bp
Coding sequence
100 kb
metal ion binding
GO:0046872
ATP binding
GO:0005524
DNA-binding
transcription
factor activity
GO:0003700
endonuclease activity
GO:0004519
catalytic activity
GO:0003824
S-adenosylmethionine
-dependent
methyltransferase
activity
GO:0008757
oxidoreductase activity
GO:0016491
zinc ion binding
GO:0008270
hydrolase activity
GO:0016787
Other
generated
proteins



F Tetranucleotide usage deviations (TUDs)

G TUD phylogeny

H Stop codon usage

Discussion

Limitations and challenges

Evo was only trained on prokaryotic and phage genomes

➔ A larger model and more computing power would be required to include eukaryotic genomes

Generated sequences are to some extent “hallucinated”

➔ Requires that large outputs are filtered computationally

➔ *“[Evo generated samples] represent a “blurry image” of a genome that contains key characteristics but lacks the finer-grained details typical of natural genomes”*

Chatbot LLMs can be easily corrected, but this can’t be

Ethical considerations

The ethical considerations for genomic LLMs are potentially even greater than those for classic LLMs like ChatGPT

Competent genomic LLMs could enable:

- ➔ Advances in gene-editing technology
- ➔ Creation of biohazards using synthetic biology
- ➔ Development of bioweapons

Summary

Evo is a genomic foundation model with:

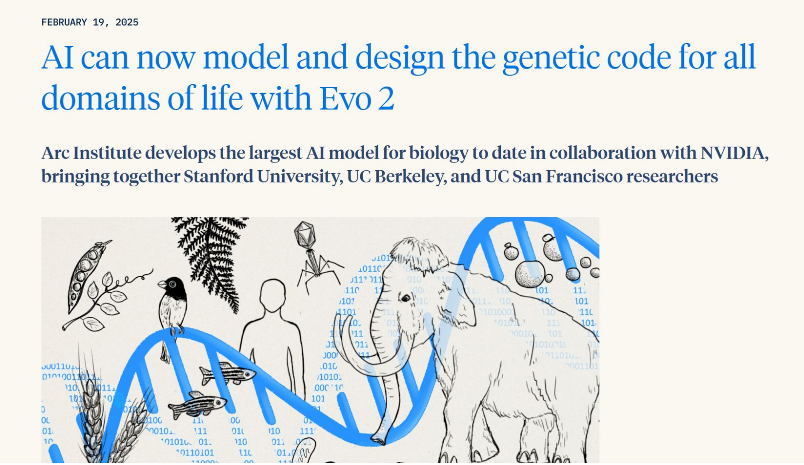
- Prokaryotic, phage, and plasmid data; StripedHyena architecture
- 131k context width, single-nucleotide resolution
- 100x more data, 1000x larger than HyenaDNA

The model was evaluated using:

- Zero-shot functional predictions
- Generative design of CRISPR-Cas molecular complexes
- DNA sequence generation at genome scale

Evo 2

- ➔ Preprint was released Feb. 19, 2025
- ➔ Includes human, animal, plant, and other eukaryotic genomes



FEBRUARY 19, 2025

AI can now model and design the genetic code for all domains of life with Evo 2

Arc Institute develops the largest AI model for biology to date in collaboration with NVIDIA, bringing together Stanford University, UC Berkeley, and UC San Francisco researchers

Questions?