Hands-on workshop on genomic language models

Guillaume GAUTREAU (CRCN), MaIAGE unit, StatInfOmics team

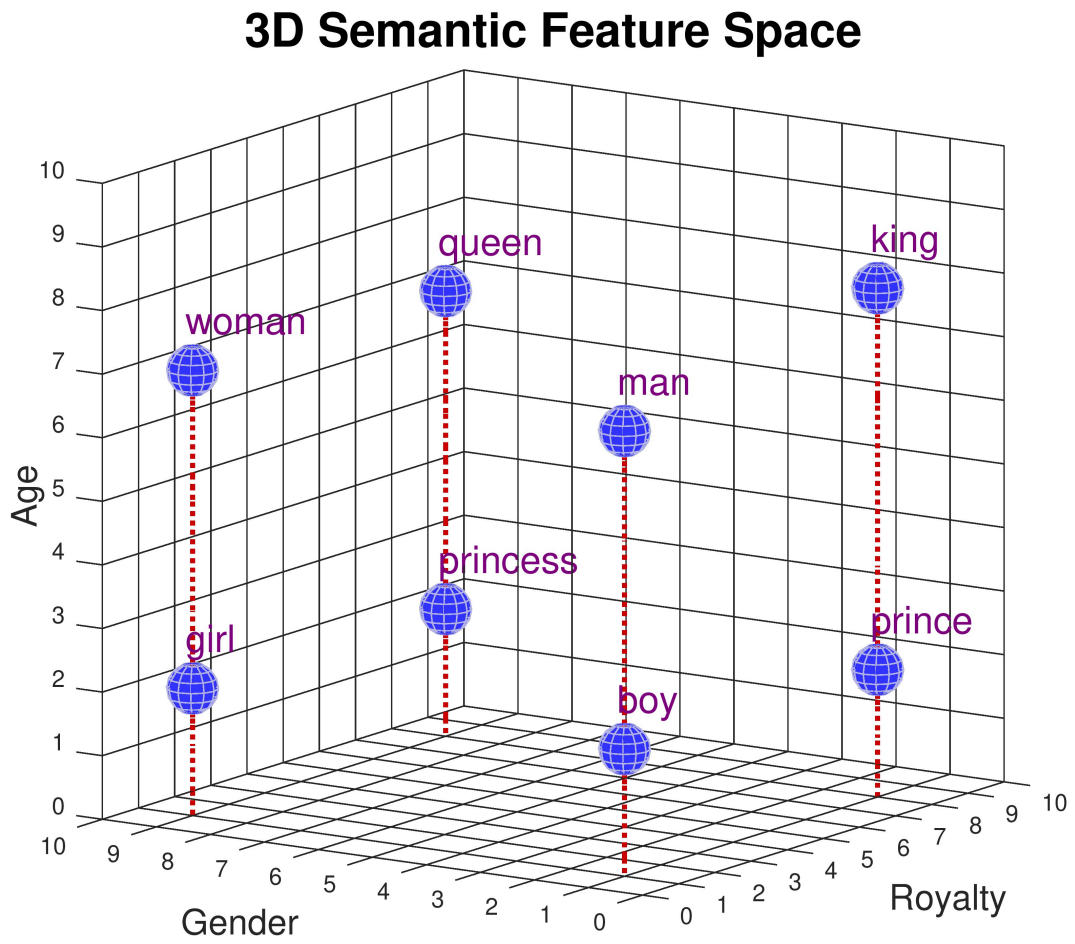
12/06/2026

RÉPUBLIQUE FRANÇAISE
Liberté
Égalité
Fraternité

INRAE

MaIAGE

Token embeddings in a multidimensional space



Transformer architecture « Attention Is All You Need »

Attention Is All You Need

Ashish Vaswani*
Google Brain
avaswani@google.com

Noam Shazeer*
Google Brain
noam@google.com

Niki Parmar*
Google Research
nikip@google.com

Jakob Uszkoreit*
Google Research
usz@google.com

Llion Jones*
Google Research
llion@google.com

Aidan N. Gomez* †
University of Toronto
aidan@cs.toronto.edu

Łukasz Kaiser*
Google Brain
lukaszkaiser@google.com

Illia Polosukhin* ‡
illia.polosukhin@gmail.com

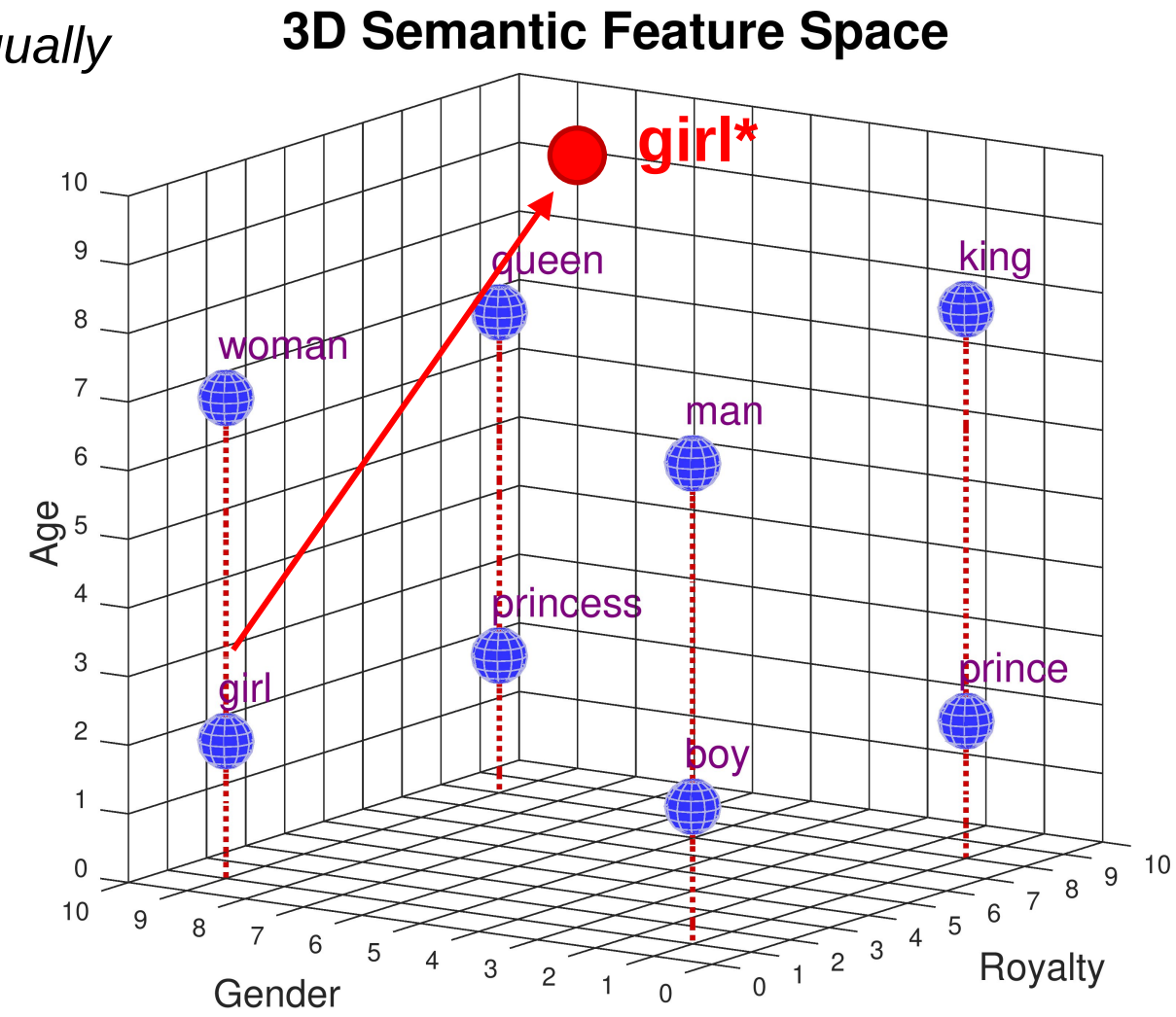
Abstract

The dominant sequence transduction models are based on complex recurrent or convolutional neural networks that include an encoder and a decoder. The best performing models also connect the encoder and decoder through an attention mechanism. We propose a new simple network architecture, the Transformer, based solely on attention mechanisms, dispensing with recurrence and convolutions entirely. Experiments on two machine translation tasks show these models to be superior in quality while being more parallelizable and requiring significantly less time to train. Our model achieves 28.4 BLEU on the WMT 2014 English-to-German translation task, improving over the existing best results, including ensembles, by over 2 BLEU. On the WMT 2014 English-to-French translation task, our model establishes a new single-model state-of-the-art BLEU score of 41.8 after training for 3.5 days on eight GPUs, a small fraction of the training costs of the best models from the literature. We show that the Transformer generalizes well to other tasks by applying it successfully to English constituency parsing both with large and limited training data.

- 2017 and already on track to become one of the most cited articles (>240k)
- Design for translation purposes

Token embeddings updated through the self-attention mechanism

In-context learning, look at everything, but not equally



LLM have thousands of dimensions

Next token prediction

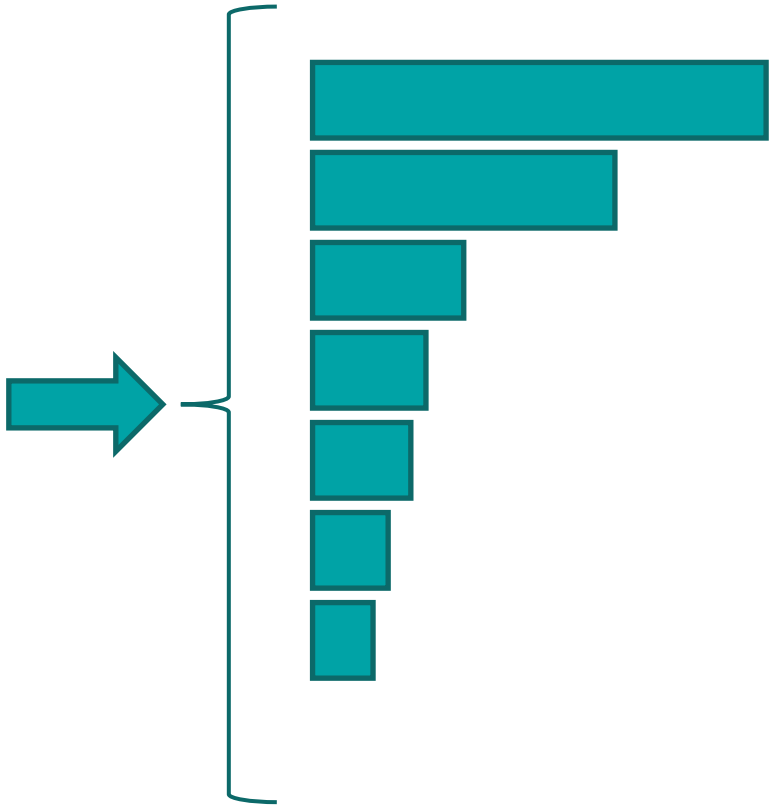
an elderly girl wearing a crown .....

Next token prediction

an elderly girl wearing a crown .....

Next token prediction

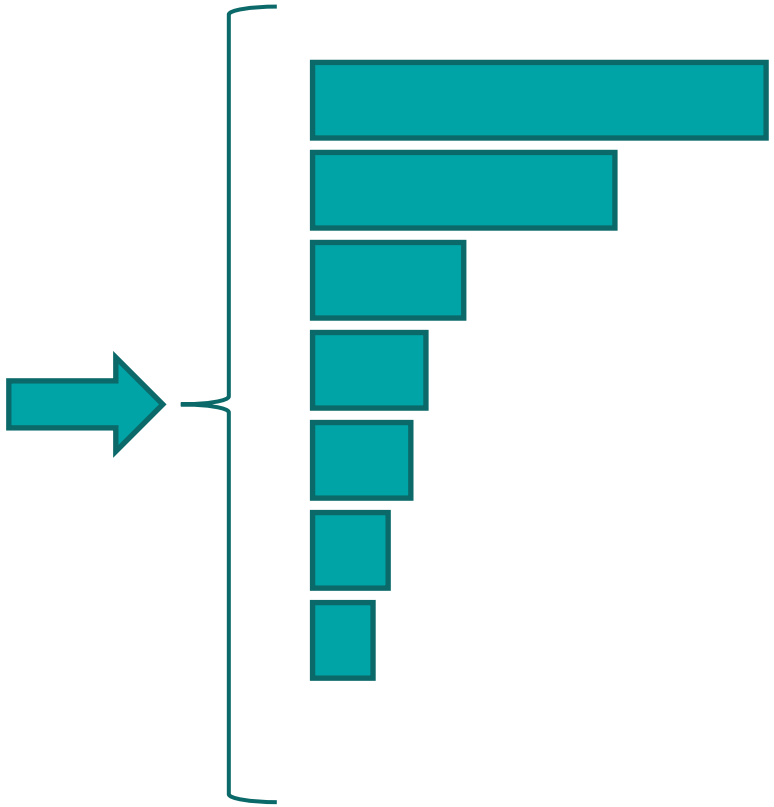
an elderly girl wearing a crown .....



was(15%)
said (10%)
loves (5%)
played (3.75%)
wrote (3%)
goes (2.5%)
thinks (2%)

Next token prediction

an elderly girl wearing a crown .....

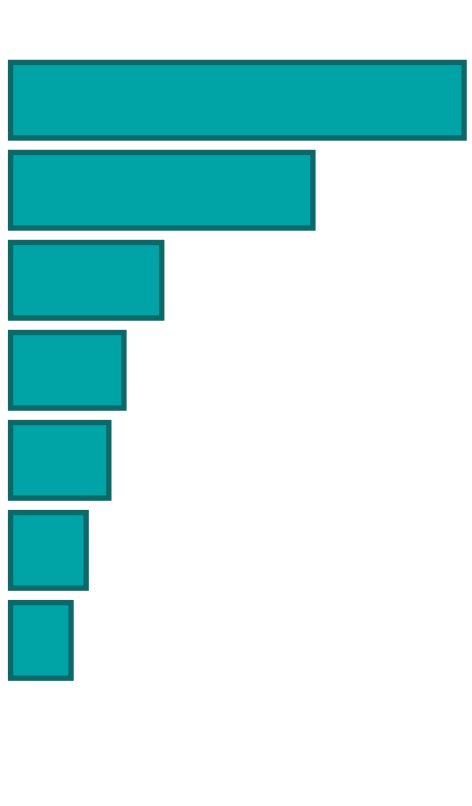


was(15%)
said (10%)
loves (5%)
played (3.75%)
wrote (3%)
goes (2.5%)
thinks (2%)

wrote

Next token prediction are biased

an elderly **boy** wearing a crown .....

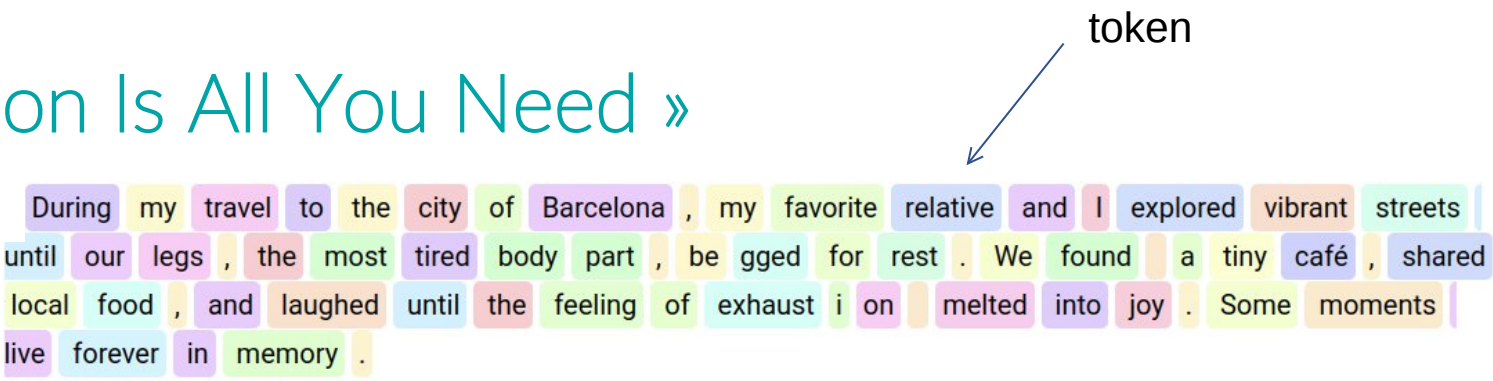


decided (15%)
fought (10%)
went (5%)
became (3.75%)
with (3%)
invaded (2.5%)
killed (2%)

fought

« Attention Is All You Need »

Given a prompt:



Ruler, 2016
https://www.ruder.io/
word-embeddings-1/

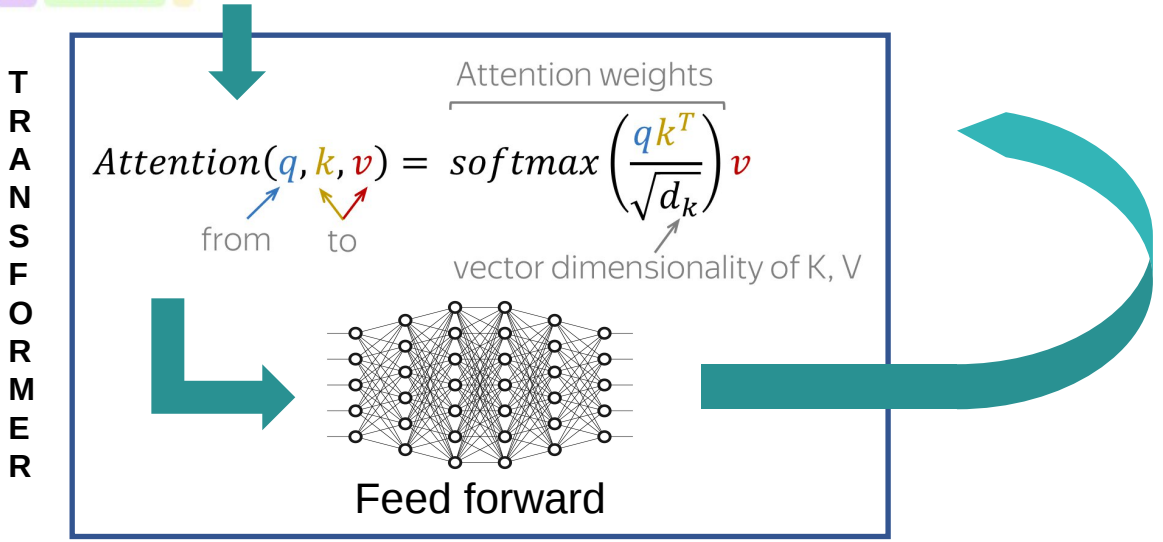
« Attention Is All You Need »

token

Given a prompt:

During my travel to the city of Barcelona , my favorite relative and I explored vibrant streets until our legs , the most tired body part , be gged for rest . We found a tiny café , shared local food , and laughed until the feeling of exhaust i on melted into joy . Some moments live forever in memory .

Billions of trained weights in the model



- Passes through dozens of transformer layers.
- Each layer captures more and more information about the links between tokens

« Attention Is All You Need »

token

Given a prompt:

During my travel to the city of Barcelona , my favorite relative and I explored vibrant streets until our legs , the most tired body part , be gged for rest . We found a tiny café , shared local food , and laughed until the feeling of exhaust i on melted into joy . Some moments live forever in memory .

Billions of trained weights in the model

TRANSFORMER

Attention weights

$$Attention(q, k, v) = softmax\left(\frac{qk^T}{\sqrt{d_k}}\right)v$$

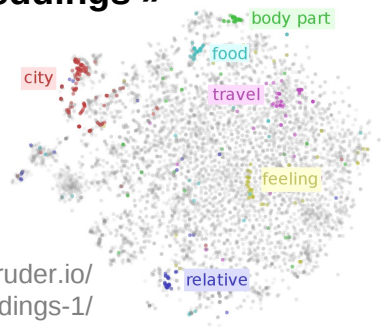
from (q) to (k)

vector dimensionality of K, V

Feed forward

- Passes through dozens of transformer layers.
- Each layer captures more and more information about the links between tokens

Tokens represented in a multidimensional space
« embeddings »



Ruder, 2016
https://www.ruder.io/word-embeddings-1/

Encoder-only

Encoder-Decoder

Lors de mon voyage dans la ville de Barcelone , mon parent préféré et moi avons parcour u des rues anim ées jusqu ' à ce que nos jambes , la partie du corps la plus fatigu ée , récla ment du repos . Nous avons trouvé un petit café , partagé des spécial ités locales et ri jusqu ' à ce que le sentiment d ' é pu isement se transforme en joie . Certains instants restent à jamais grav és en mémoire .

Translate into another language

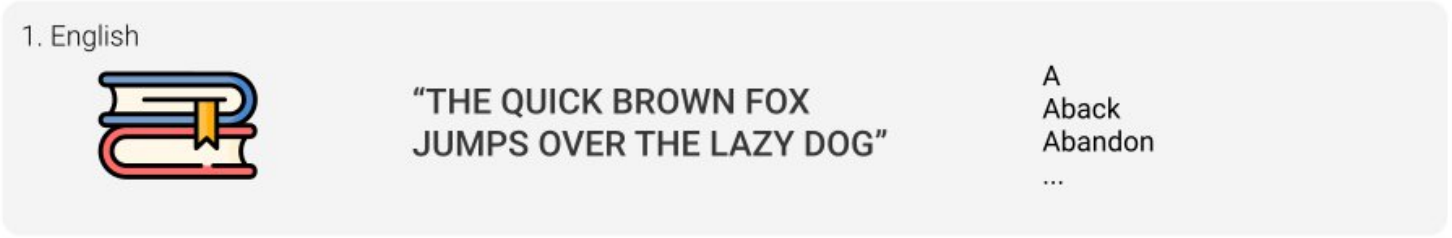
Decoder-only

The sun dipped below the rooftops , casting golden light on our table , as music played softly and strangers smiled passing by .

Predicts next tokens

Large Language of Life Models (LLLM) as suggested by Eric Topol

LLM



- BioMedLM
- PubMedBERT
- BioBERT
- ChatGPT/Claude/Gemini...

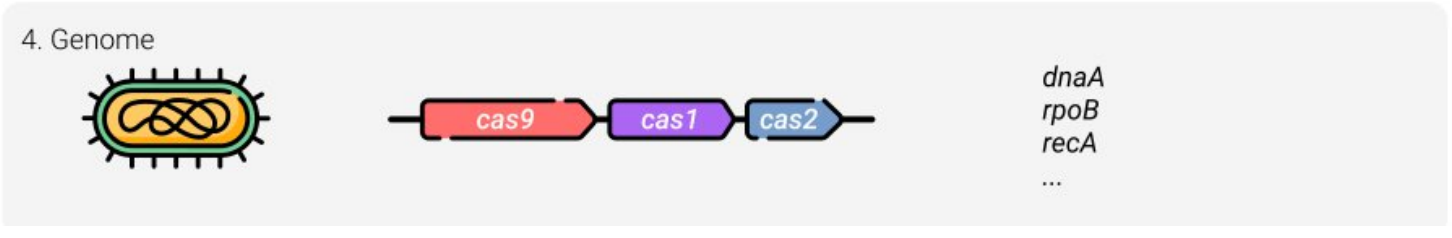
gLM

- DNABERT ⇨ 2 ⇨ S
- Nucleotide Transformer
- **Evo 1 ⇨ 1.5 ⇨ 2**

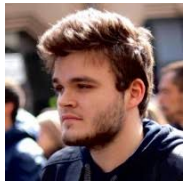
pLM

- ESM 1 ⇨ 2 ⇨3
- ProteinBERT
- ProtMamba
- ProtT5

gLM



- gLM
- Bacformer / panBART
- ANR project PanGAIMiX

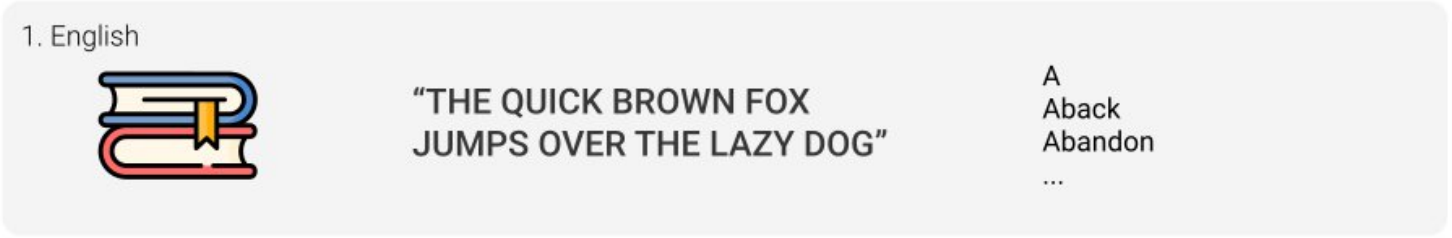


Meriem Youssef Gaspar Roy

Rannon et al. 2025
https://doi.org/10.48550/arXiv.2506.02212

Large Language of Life Models (LLM) as suggested by Eric Topol

LLM



- BioMedLM
- PubMedBERT
- BioBERT
- ChatGPT/Claude/Gemini...

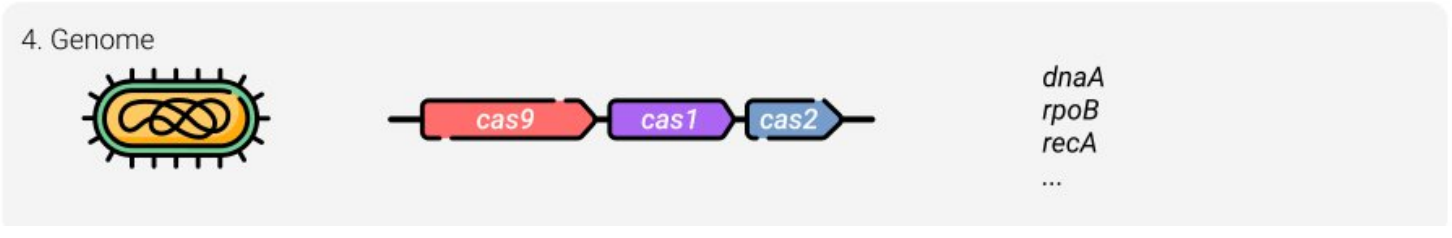
gLM

- DNABERT ⇨ 2 ⇨ S
- Nucleotide Transformer
- **Evo 1 ⇨ 1.5 ⇨ 2**

pLM

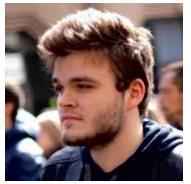
- ESM 1 ⇨ 2 ⇨3
- ProteinBERT
- ProtMamba
- ProtT5

gLM



- gLM
- Bacformer / panBART
- ANR project PanGAIMiX

Meriem Youssef



Gaspar Roy

Rannon et al. 2025
https://doi.org/10.48550/arXiv.2506.02212

Evo (November 2024)

Science
JOURNALS AAAS

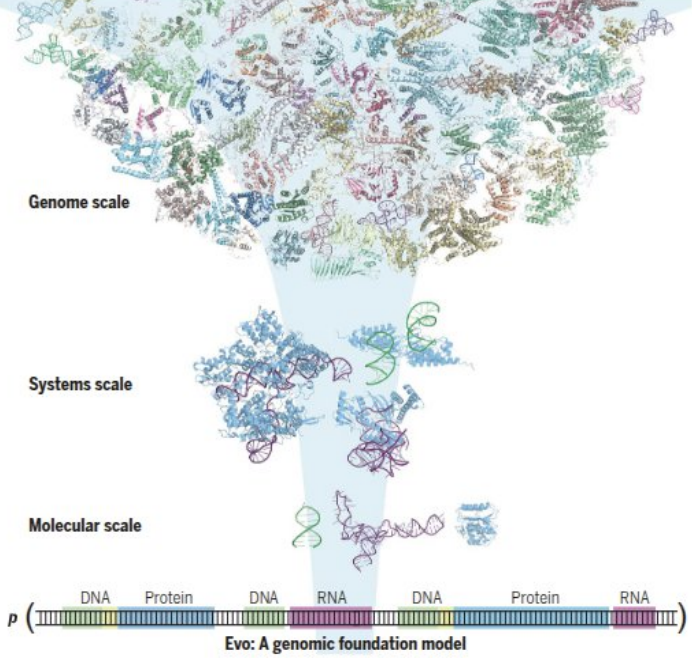
RESEARCH ARTICLE

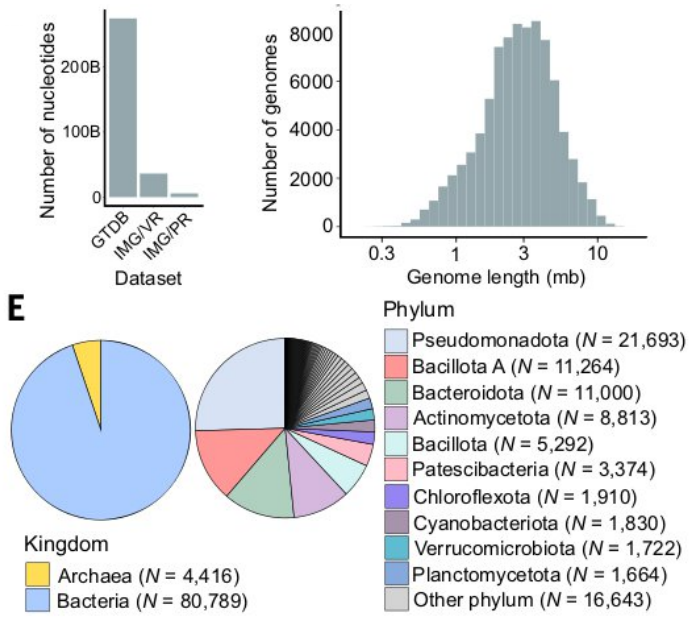
GENERATIVE GENOMICS

Sequence modeling and design from molecular to genome scale with Evo

Eric Nguyen[1,2]†, Michael Poli[3,4]†‡, Matthew G. Durrant[1]†, Brian Kang[1,2]†, Dhruva Katrekar[1]†, David B. Li[1,2]†, Liam J. Bartie[1], Armin W. Thomas[5], Samuel H. King[1,2], Garyk Brixi[1,6], Jeremy Sullivan[1], Madelena Y. Ng[7], Ashley Lewis[8], Aaron Lou[3], Stefano Ermon[3,9], Stephen A. Baccus[10], Tina Hernandez-Boussard[8], Christopher Ré[3], Patrick D. Hsu[1,11]*, Brian L. Hie[1,5,12]*

The genome is a sequence that encodes the DNA, RNA, and proteins that orchestrate an organism's function. We present Evo, a long-context genomic foundation model with a frontier architecture trained on millions of prokaryotic and phage genomes, and report scaling laws on DNA to complement observations in language and vision. Evo generalizes across DNA, RNA, and proteins, enabling zero-shot function prediction competitive with domain-specific language models and the generation of functional CRISPR-Cas and transposon systems, representing the first examples of protein-RNA and protein-DNA codesign with a language model. Evo also learns how small mutations affect whole-organism fitness and generates megabase-scale sequences with plausible genomic architecture. These prediction and generation capabilities span molecular to genomic scales of complexity, advancing our understanding and control of biology.

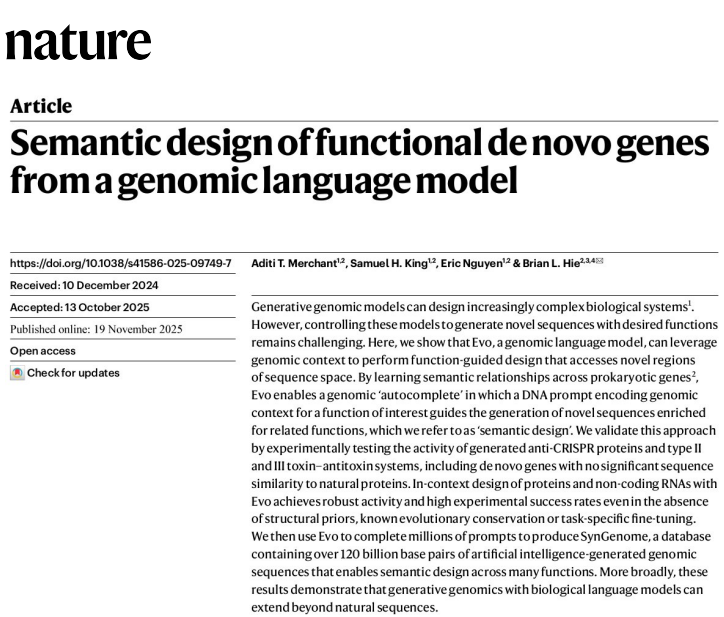




Model size	7B
Context size (# of tokens/ bases)	**single-nucleotide resolution: 131Kb or 8Kb**
Architecture	**Decoder only: Striped Hyena**
Training dataset	**Prokaryotic genomes / plasmids / phages**
Training dataset size	**315 billions**
Training cost (estimation by myself at market price)	**70k$**

Evo 1.5

- Same architecture but extending the pretraining dataset of Evo 1 (8k) **+50%**
- Still on prokaryote/phage
- from 315 billion tokens (75,000 iterations) to 470 billion tokens (112,000 iterations)
- Only an 8Kb version



nature

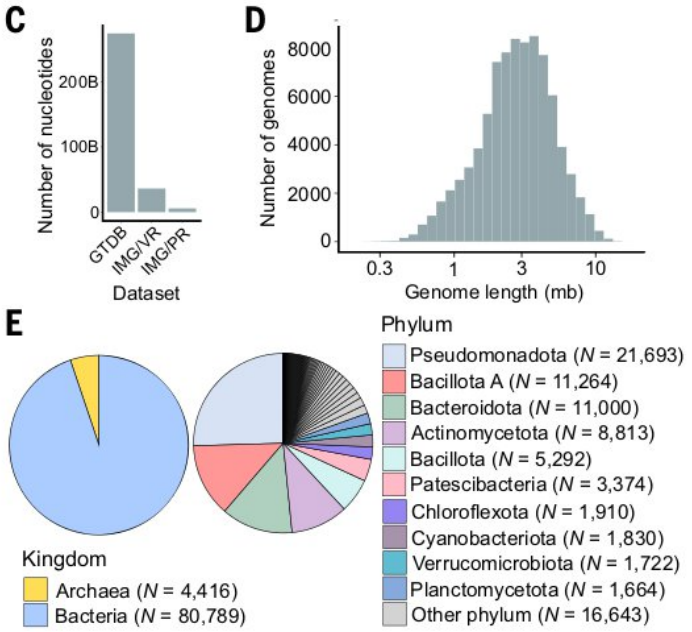
Article

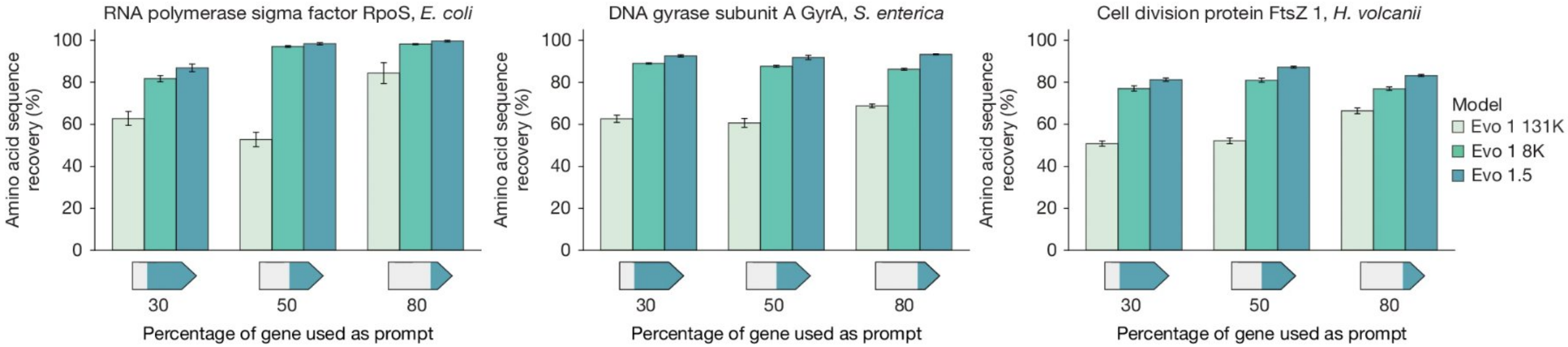
Semantic design of functional de novo genes from a genomic language model

https://doi.org/10.1038/s41586-025-09749-7
Received: 10 December 2024
Accepted: 13 October 2025
Published online: 19 November 2025
Open access
Check for updates

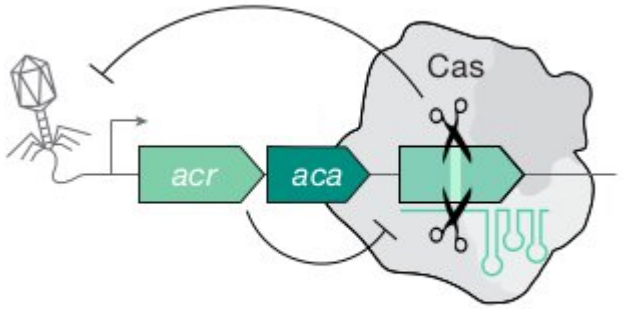
Aditi T. Merchant[1,2], Samuel H. King[1,2], Eric Nguyen[1,2] & Brian L. Hie[2,3,4]

Generative genomic models can design increasingly complex biological systems[1]. However, controlling these models to generate novel sequences with desired functions remains challenging. Here, we show that Evo, a genomic language model, can leverage genomic context to perform function-guided design that accesses novel regions of sequence space. By learning semantic relationships across prokaryotic genes[2], Evo enables a genomic 'autocomplete' in which a DNA prompt encoding genomic context for a function of interest guides the generation of novel sequences enriched for related functions, which we refer to as 'semantic design'. We validate this approach by experimentally testing the activity of generated anti-CRISPR proteins and type II and III toxin–antitoxin systems, including de novo genes with no significant sequence similarity to natural proteins. In-context design of proteins and non-coding RNAs with Evo achieves robust activity and high experimental success rates even in the absence of structural priors, known evolutionary conservation or task-specific fine-tuning. We then use Evo to complete millions of prompts to produce SynGenome, a database containing over 120 billion base pairs of artificial intelligence-generated genomic sequences that enables semantic design across many functions. More broadly, these results demonstrate that generative genomics with biological language models can extend beyond natural sequences.

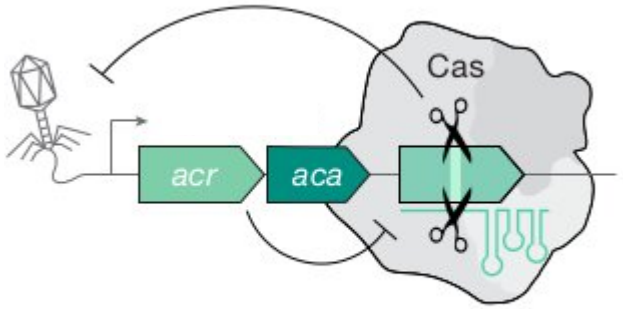


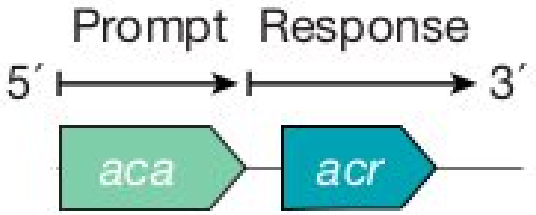


Evo generates novel functional anti-CRISPR proteins

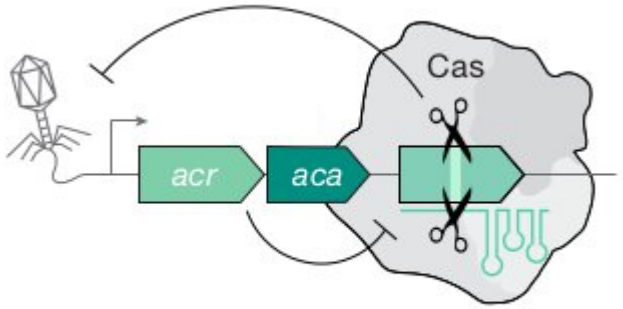


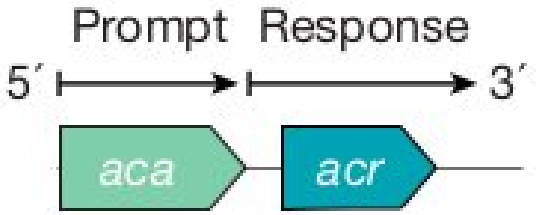
Evo generates novel functional anti-CRISPR proteins

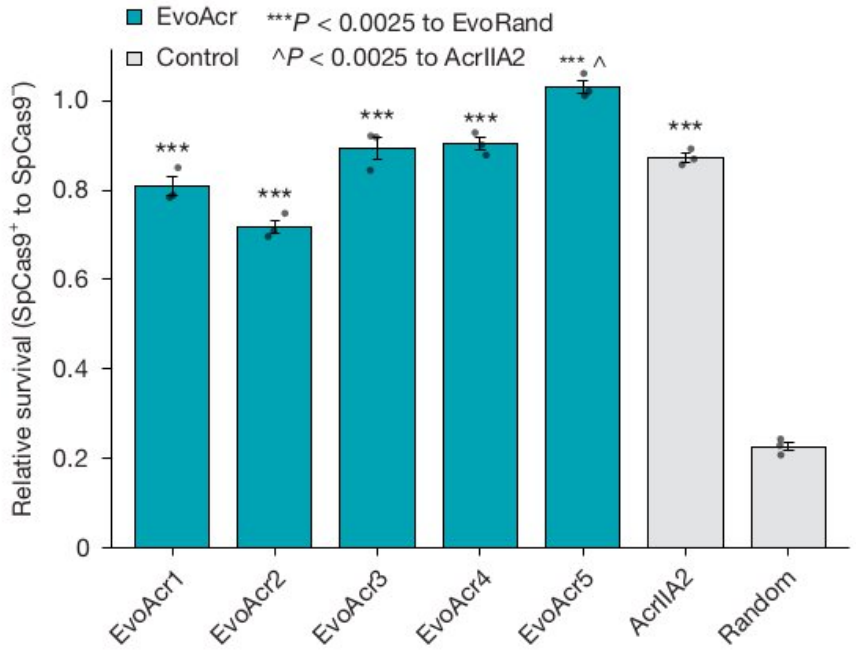


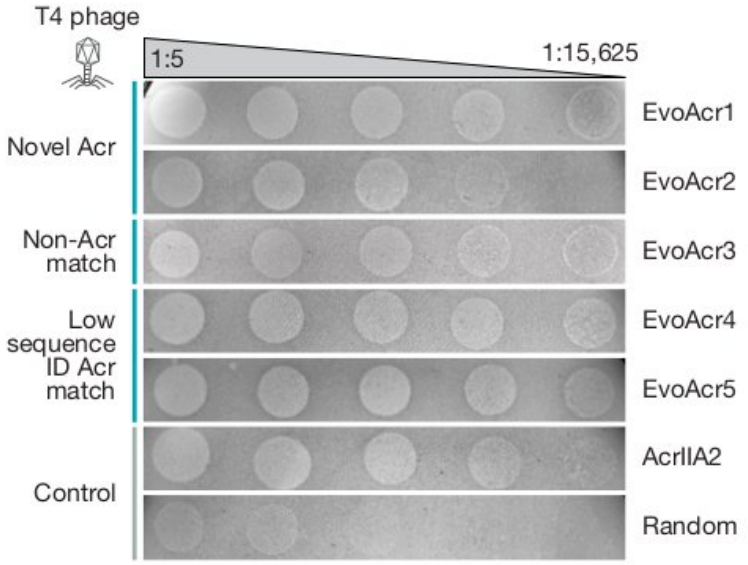


Evo generates novel functional anti-CRISPR proteins

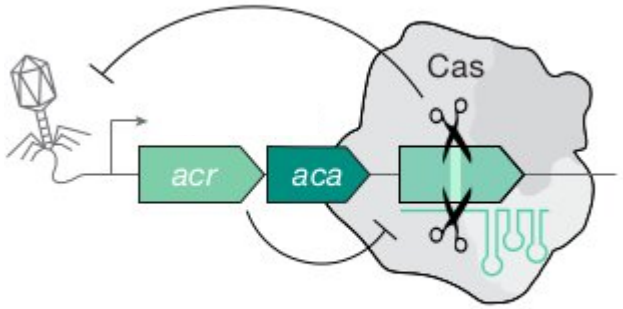


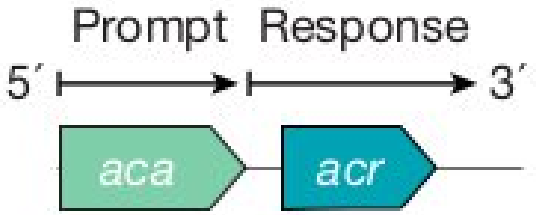


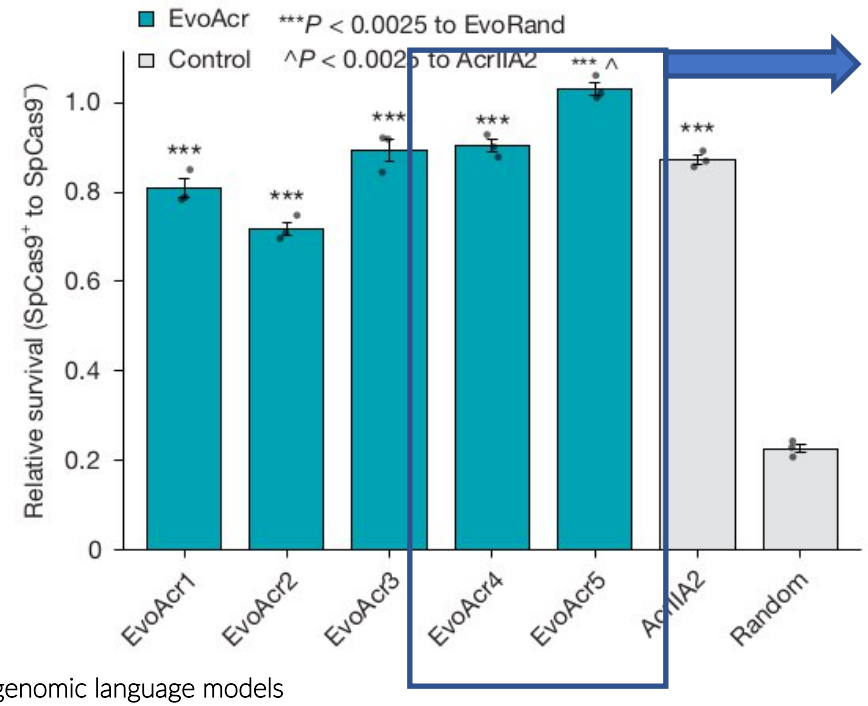


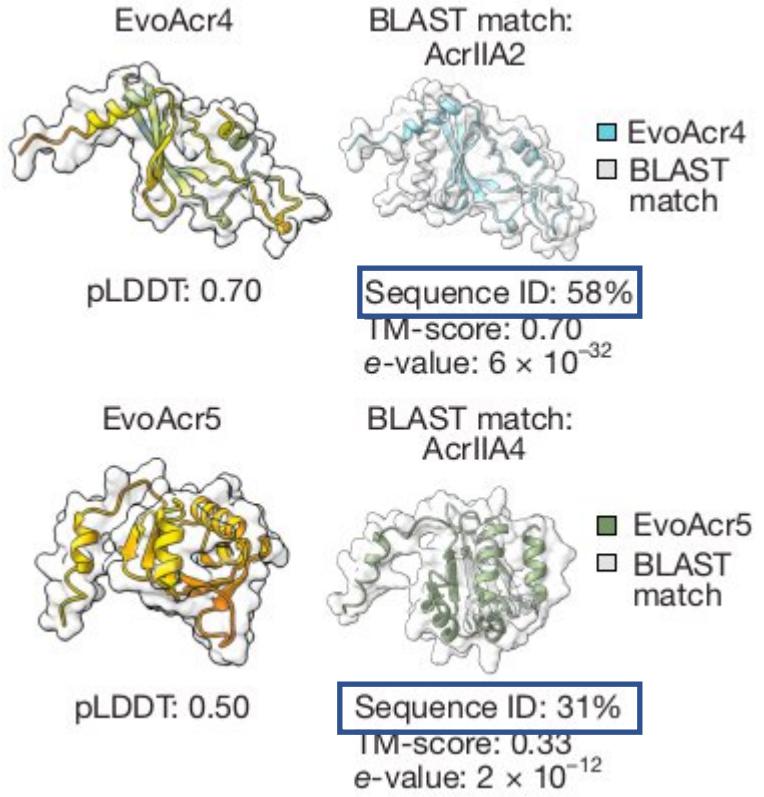


Evo generates novel functional anti-CRISPR proteins

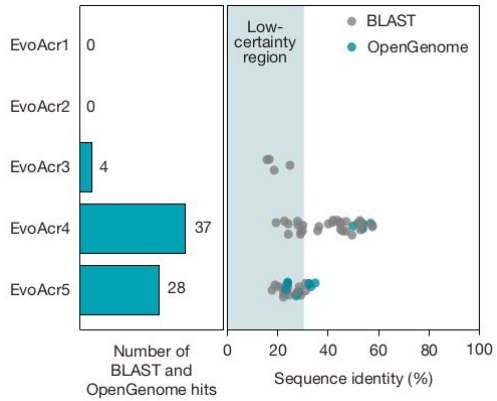


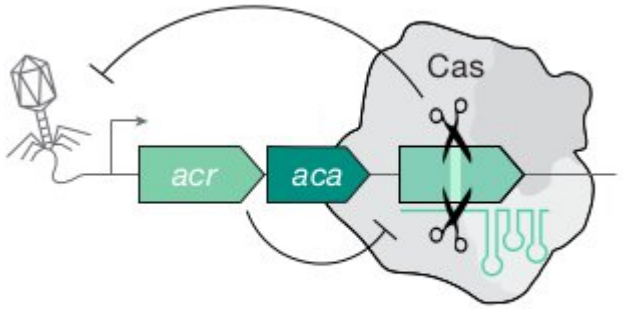


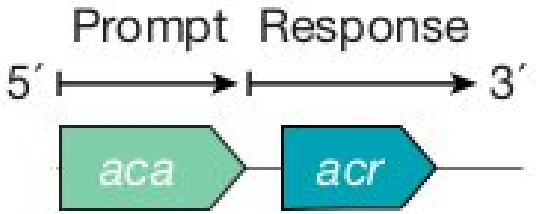




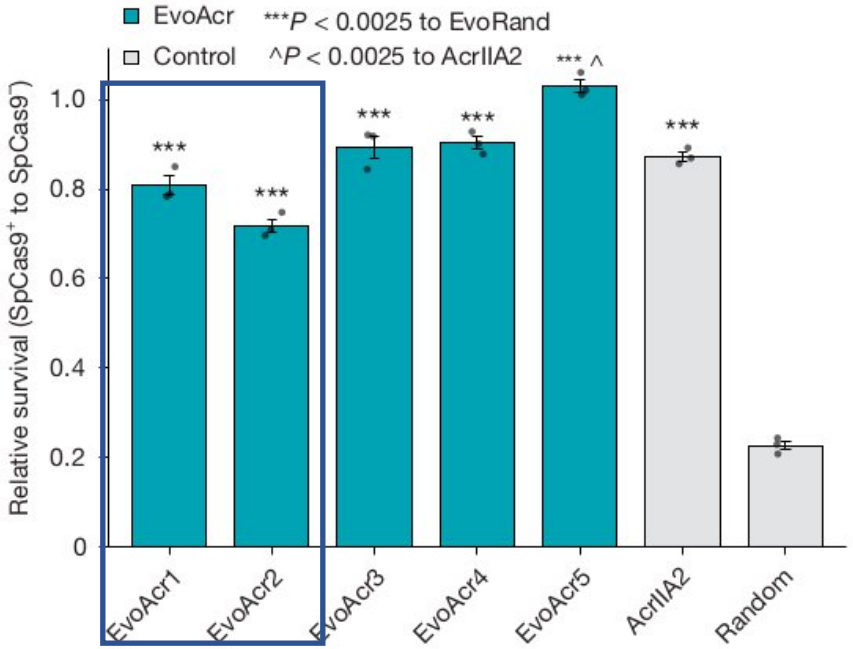
Evo generates novel functional anti-CRISPR proteins

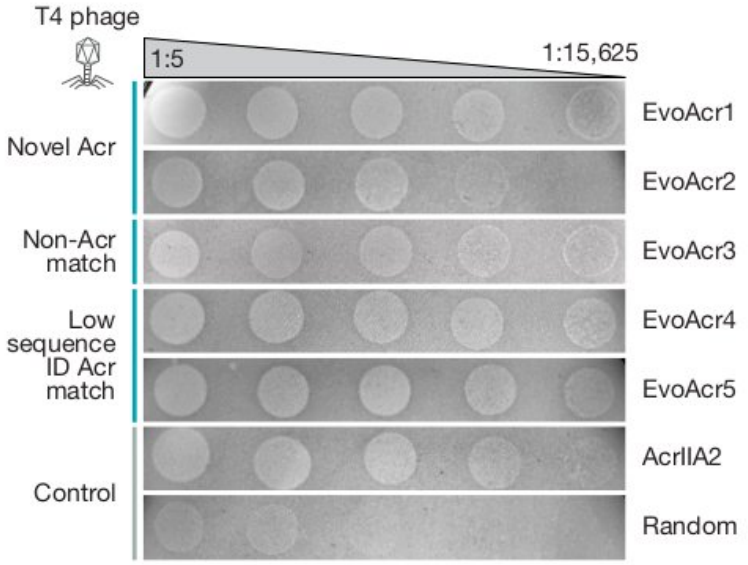






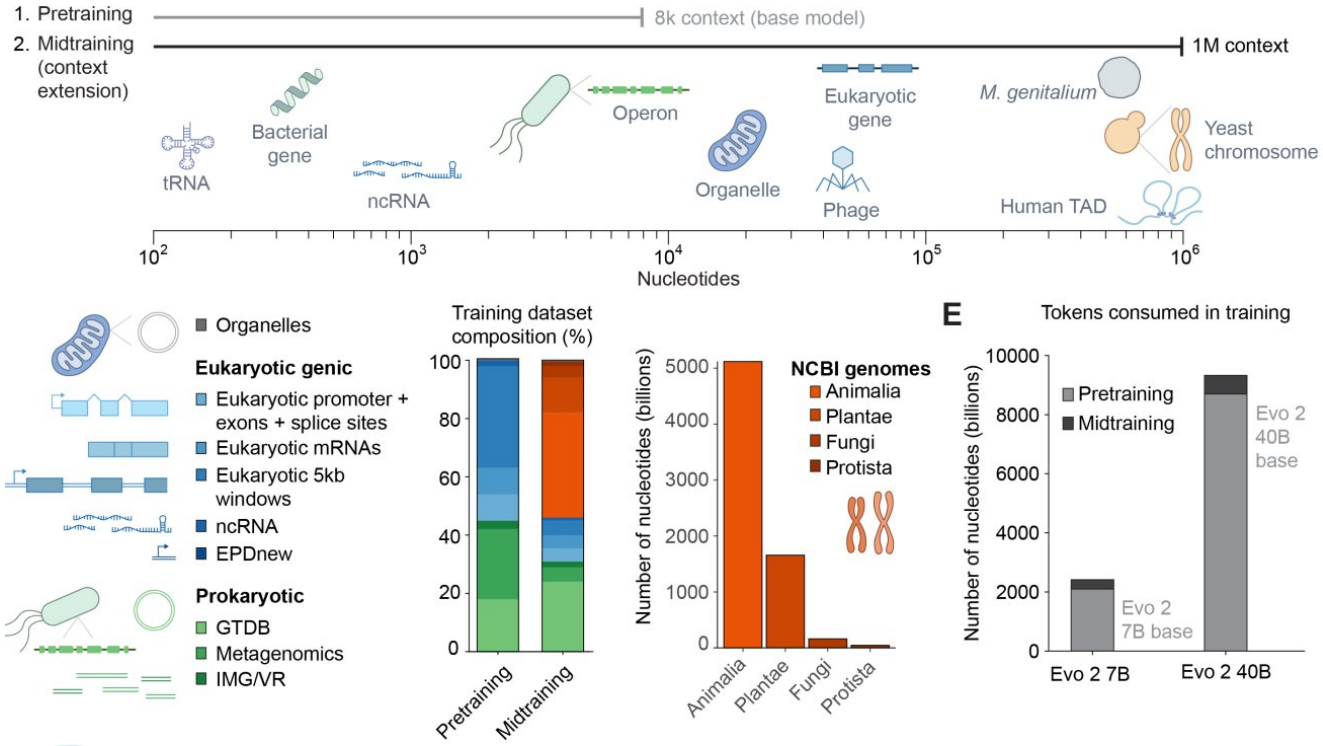
- **No homology with any known gene!**
- **Original creation from AI!**





Evo2

- As Evo1, built to predict the next token at single-nucleotide resolution
- Uses an improved version 2 of Striped Hyena
- Train on all available kinds of DNA (9.3 trillion DNA base pairs)
- Context size goes to megabase scale
- Training cost ~5-10 M$ (GPU + electricity only) ! 150×Alphafold2
- Developed in direct collaboration with NVIDIA/Amazon

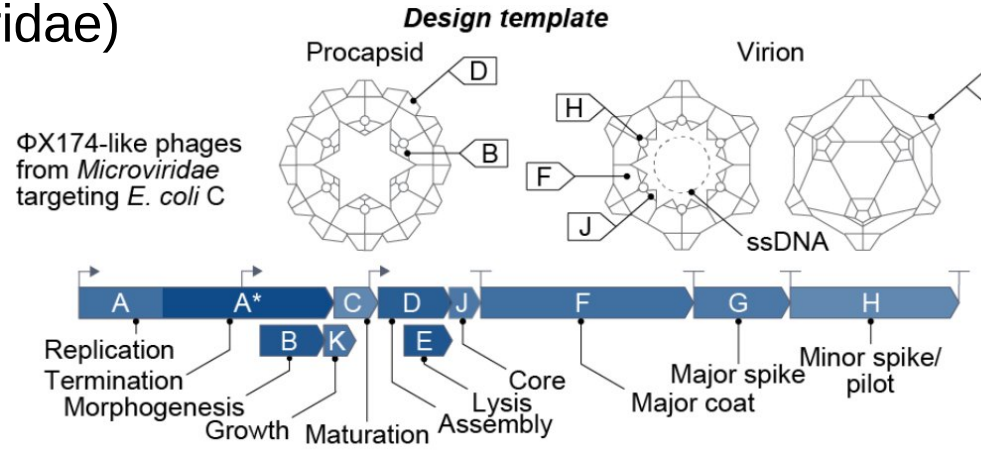


	Evo1	Evo 1.5	Evo 2
Publication	November 2024	November 2025	**pre-print**
Model size	7B	7B	**1B, 7B, 40B**
Context size (# of tokens/ bases)	8 Kb or 131 Kb	8 Kb	**1 Mb**
Architecture	Striped Hyena	Striped Hyena	**Striped Hyena 2**
Training dataset	Prokaryotic genomes / plasmids / phages	Updated Evo1 database	**Evo1 + All NCBI Eukaryotes + Metagenome (MGnify)**
Training dataset size	315 billions	470 billions	**9300 billions**
Training cost (estimation by myself at market price)	70k$	105k$	**6M$**

The claim: first working gLM-generated genome

One of the easiest case scenario but still :

- Phage ΦX174 (Microviridae)
- Target E. coli cells
- Abundant in databases
- Very small: ~5Kb



pre-print on BioRxiv (09/2025) :

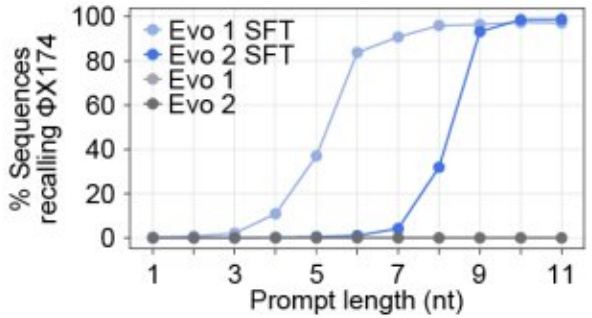
Generative design of novel bacteriophages with genome language models

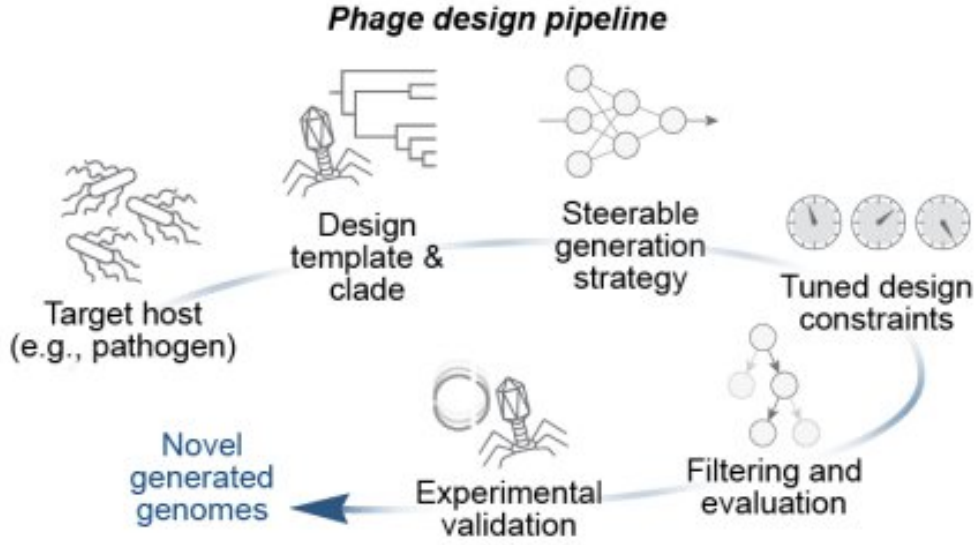
Samuel H. King[1,2], Claudia L. Driscoll[1,3], David B. Li[1,2], Daniel Guo[1,4], Aditi T. Merchant[1,2], Garyk Brixi[1,5], Max E. Wilkinson[6], and Brian L. Hie[1,3,7,*]

[1]Arc Institute, Palo Alto, CA, USA
[2]Department of Bioengineering, Stanford University, Stanford, CA, USA
[3]Department of Chemical Engineering, Stanford University, Stanford, CA, USA
[4]Department of Computer Science, Stanford University, Stanford, CA, USA
[5]Department of Genetics, Stanford University, Stanford, CA, USA
[6]Structural Biology Program, Memorial Sloan Kettering Cancer Center, New York, NY, USA
[7]Stanford Data Science, Stanford University, Stanford, CA, USA

- Zero-shot inference doesn't work

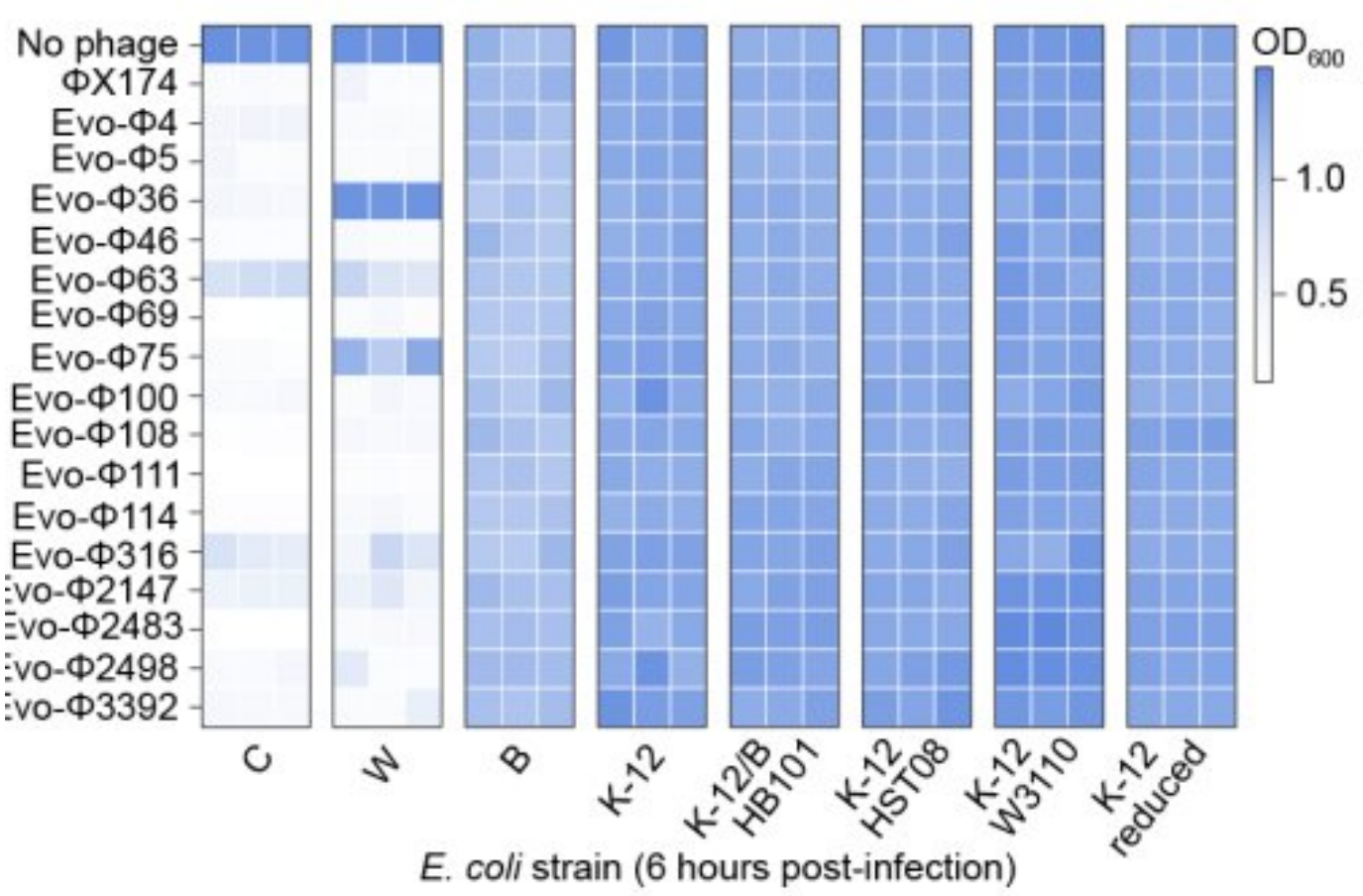
⇨Supervised fine-tuning (SFT) strategy on Microviridae genomes with soft prompting to introduce special token to 'control' identity during inference





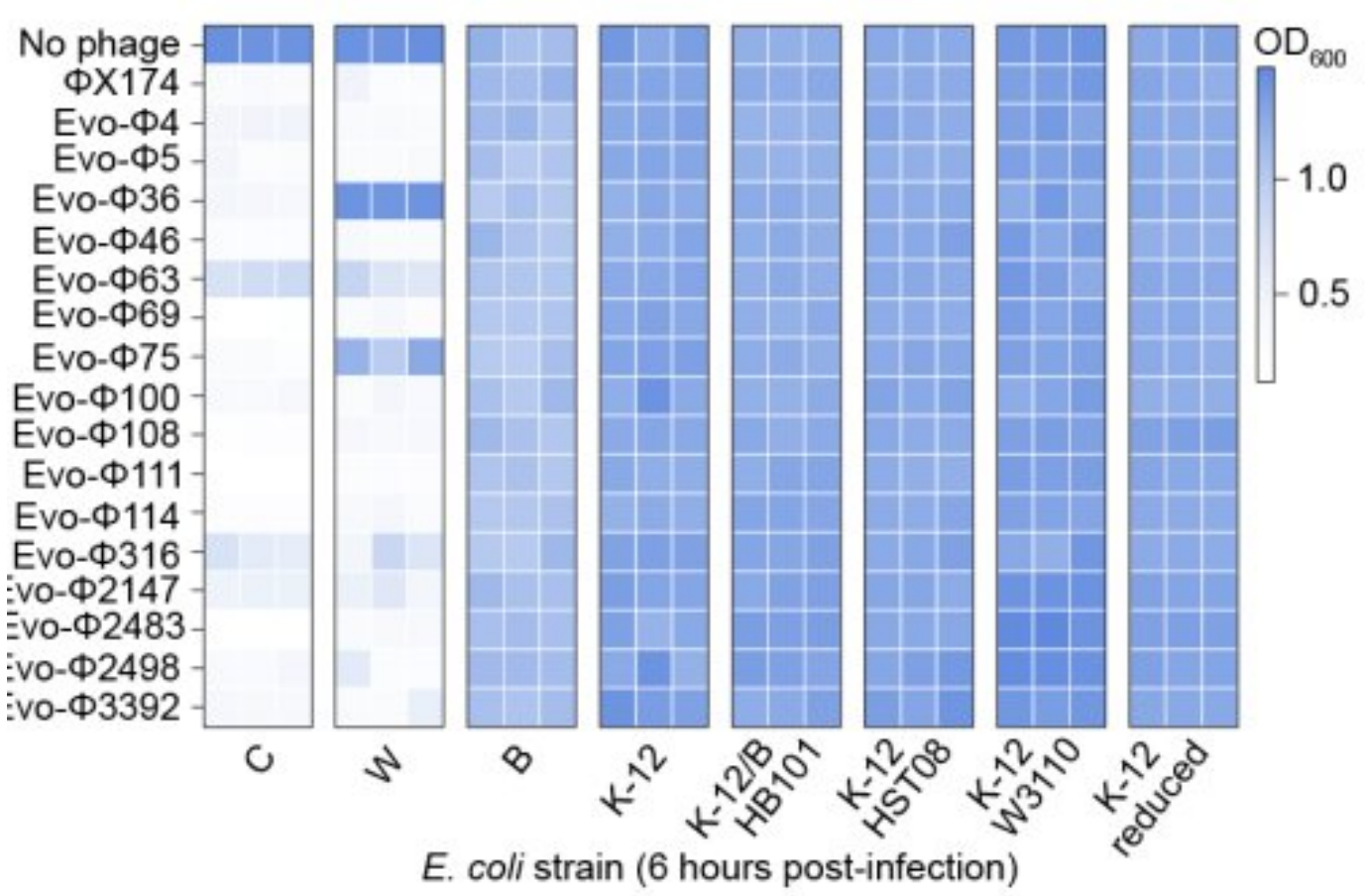
The claim: first working AI-generated genome

Evo-generated phages actually lyze E. coli genomes!

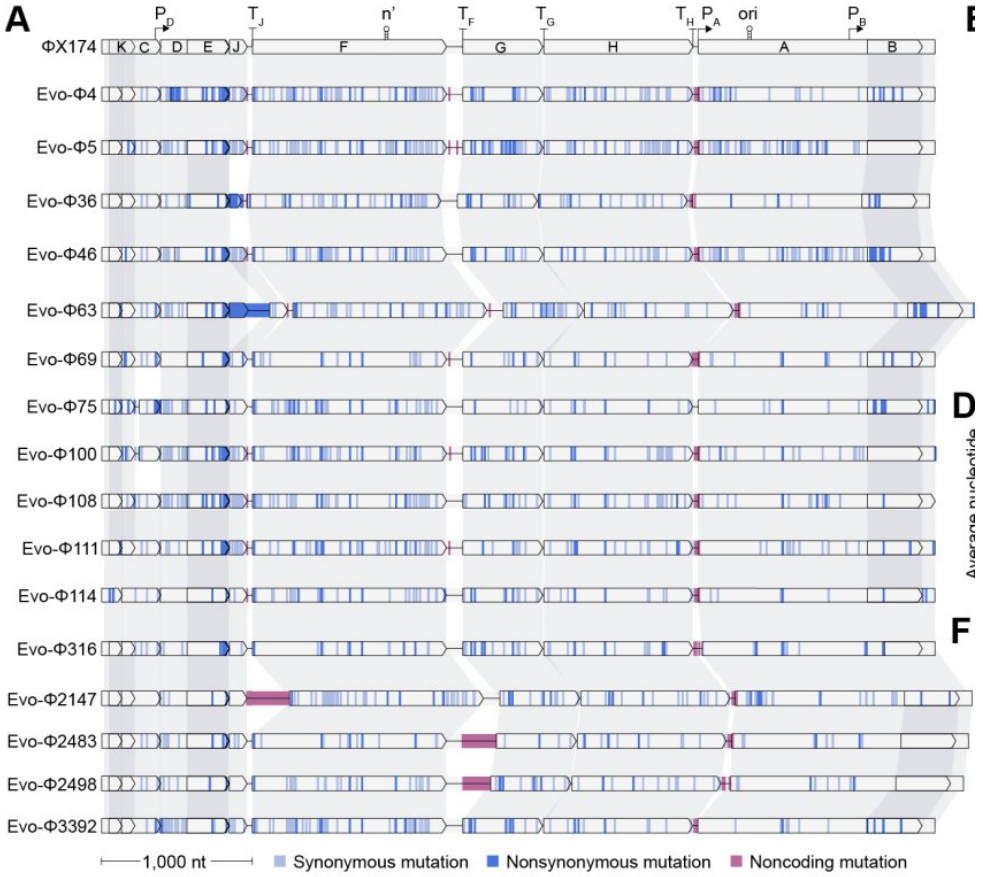


The claim: first working AI-generated genome

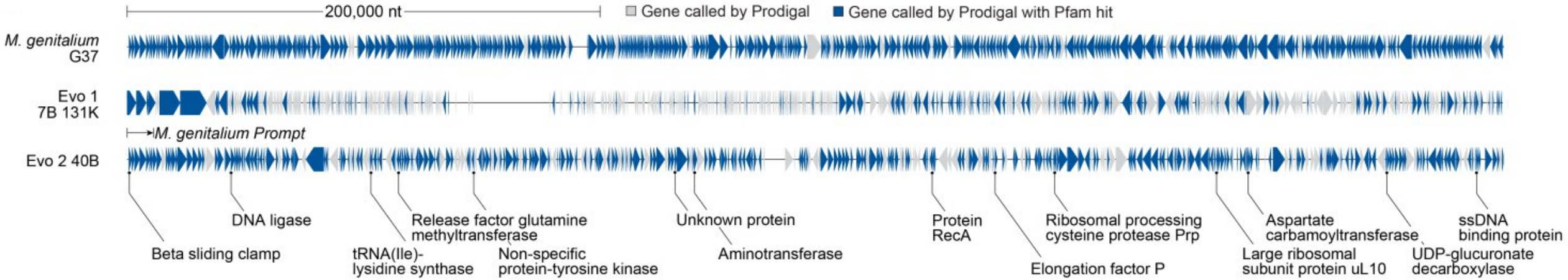
Evo-generated phages actually lyze E. coli genomes!



...but even if modifications were introduced that were not so trivial, still it generates something **not totally new** compared to the actual WT phage...



Generative applications: DNA design up to bacterial genome scale



- Synthetic design: generate proteins, enzymes, operons, pathways, systems
- Promoter engineering: generate inducible / tunable promoters
- Anonymization: generate synthetic genomes preserving signal
- Benchmarking: generate realistic synthetic datasets, diversity normalization
- Pangenome compression/normalization : generate normalized representations
- Assembly gap filling (MAGs)
- Data augmentation
- Any other idea?

but are there actually working genomes?

Sequence Generation and Alignment Analysis with Evo2

Sequence Generation and Alignment Analysis with Evo2

This notebook demonstrates how to generate biological sequences using the Evo2 model and analyze them using Biopython alignments.

Setup and Dependencies

First, let's import our required libraries and set up our environment. Note you need Jupyter to run notebooks.

In [4]:

```
import os
import argparse
import csv
from pathlib import Path
from typing import List, Optional, Tuple
import numpy as np
import torch
import torch.nn.functional as F
from Bio import pairwise2
from Bio.pairwise2 import format_alignment
from Bio.Seq import Seq

from evo2 import Evo2

# Set random seeds for reproducibility
torch.manual_seed(42)
torch.cuda.manual_seed(42)
```

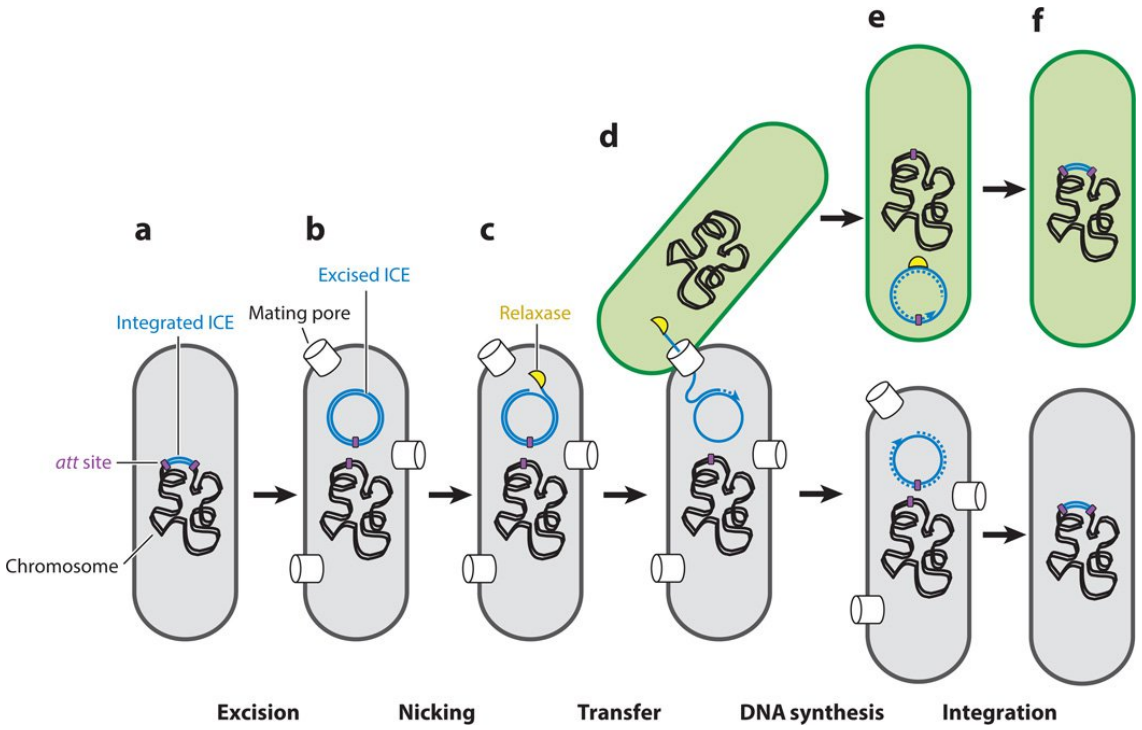
https://github.com/ArcInstitute/evo2/blob/main/notebooks/generation/generation_notebook.ipynb

INRAE

Evo: transfer learning

A strategy exemplified on ICEs/IMEs.

Integrative Conjugative Elements/Integrative Mobilizable Elements



Element type	Max size between two sequential SPs	Combinations of SPs
ICE	≤ 100 CDS	
IME	≤ 10 CDS	
Conjugation module	≤ 100 CDS	
Mobilizable element	≤ 10 CDS	

Relaxase Coupling VirB4 Integrase

Methods exist to detect them (e.g., ICEScreen), but they fail to generalize outside their reference database and to delineate elements precisely.

Firmidata annotations

A total of 98 ICEs and 148 IMEs were manually annotated from Bacillota genomes

Clostridioides difficile 630 NC_009089.1
Clostridioides difficile R20291 NC_013316.1
Dehalobacterium formicoaceticum DMC NZ_CP022121.1
Enterococcus faecalis V583 NC_004668.1
Enterococcus faecium ISMMS_VRE_1 NZ_CP012430.1
Faecalibacterium prausnitzii A2-165 NZ_CP048437.1
Lacticaseibacillus paracasei LOCK919 NC_021721.1
Lactococcus lactis IO-1 NC_020450.1
Listeria monocytogenes SLCC2378 NC_018585.1
Lachnoclostridium phocaeense Marseille-P3177T NZ_LT635479.1
Lachnoclostridium sp. YL32 NZ_CP015399.2
Roseburia hominis A2-183 NC_015977.1
Streptococcus agalactiae 09mas018883 NC_021485.1
Streptococcus agalactiae GD201008-001 NC_018646.1
Streptococcus agalactiae NEM316 NC_004368.1
Streptococcus anginosus C1051 NC_022244.1
Streptococcus constellatus subsp. pharyngis C1050 NC_022238.1
Streptococcus constellatus subsp. pharyngis C232 NC_022236.1
Streptococcus dysgalactiae subsp. equisimilis ATCC 12394 NC_017567.1
Streptococcus dysgalactiae subsp. equisimilis RE378 NC_018712.1

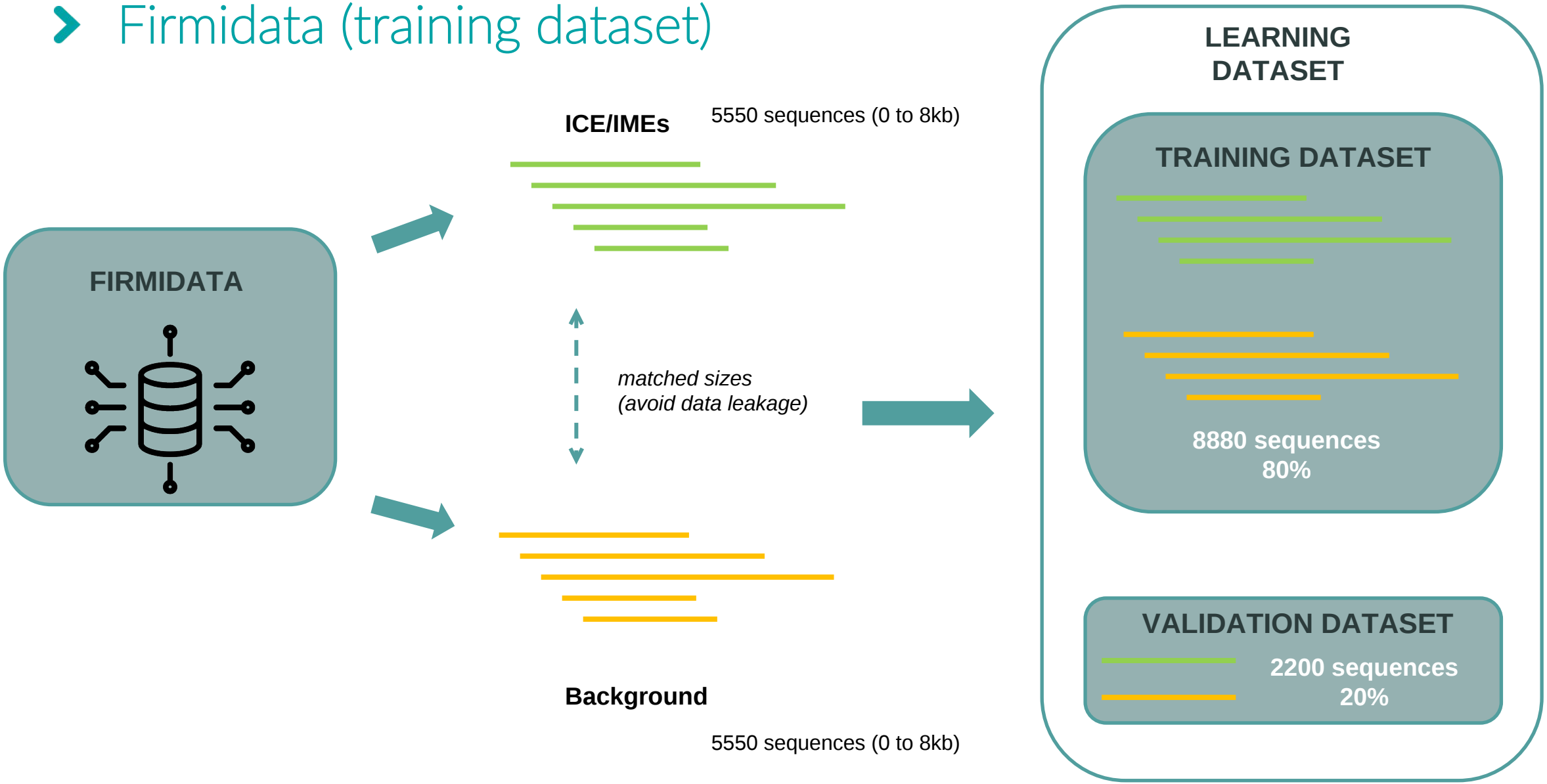
DATA NOTE Open Access

FirmiData: a set of 40 genomes of Firmicutes with a curated annotation of ICEs and IMEs

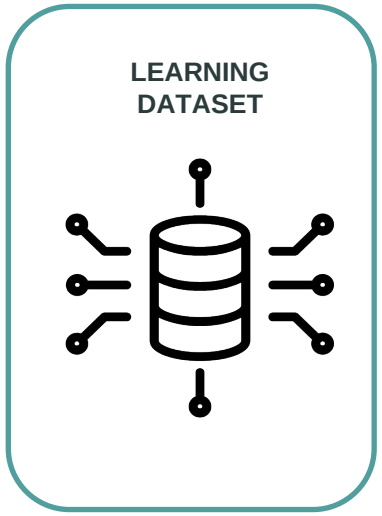
Gérard Guédon[1], Julie Lao[1,2], Sophie Payot[1], Thomas Lacroix[2], Hélène Chiapello[2] and Nathalie Leblond-Bourget[1*]

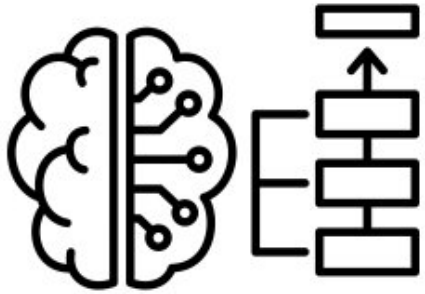
Streptococcus equi subsp. zooepidemicus MGCS10565 NC_011134.1
Streptococcus equi subsp. zooepidemicus ATCC35246 NC_017582.1
Streptococcus gallolyticus ATCCBAA-2069 NC_015215.1
Streptococcus parasanguinis ATCC15912 NC_015678.1
Streptococcus pneumoniae SPN034183 NC_021028.1
Streptococcus pneumoniae P1031 NC_012467.1
Streptococcus pyogenes MGAS2096 NC_008023.1
Streptococcus pyogenes HKU QMH11M0907901 NZ_AFRY01000001.1
Streptococcus salivarius FDAARGOS_259 NZ_CP020451.2
Streptococcus salivarius ATCC 25975 NZ_CP015283.1
Streptococcus salivarius JF NZ_CP014144.1
Streptococcus suis BM407 NC_012926.1
Streptococcus suis NSUI002 NZ_CP011419.1
Streptococcus suis SC84 NC_012924.1
Streptococcus suis ST1 NC_017950.1
Streptococcus suis T15 NC_022665.1
Streptococcus suis 05ZYH33 NC_009442.1
Staphylococcus epidermidis ATCC12228 NC_004461.1
Streptococcus thermophilus JIM8232 NC_017581.1
Staphylococcus pseudintermedius HKU10-03 NC_014925.1

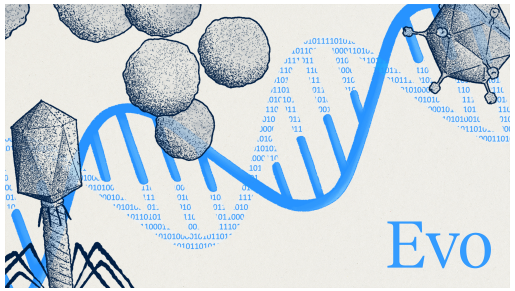
Firmidata (training dataset)

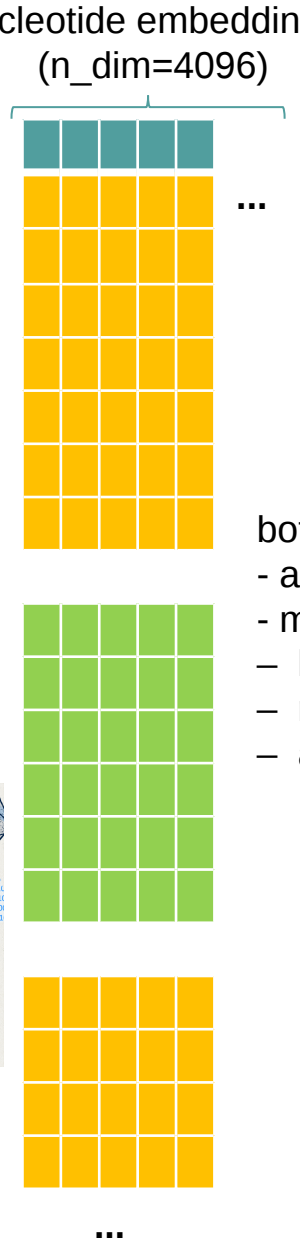


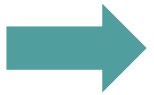
Firmidata transfer learning using Evo1.5





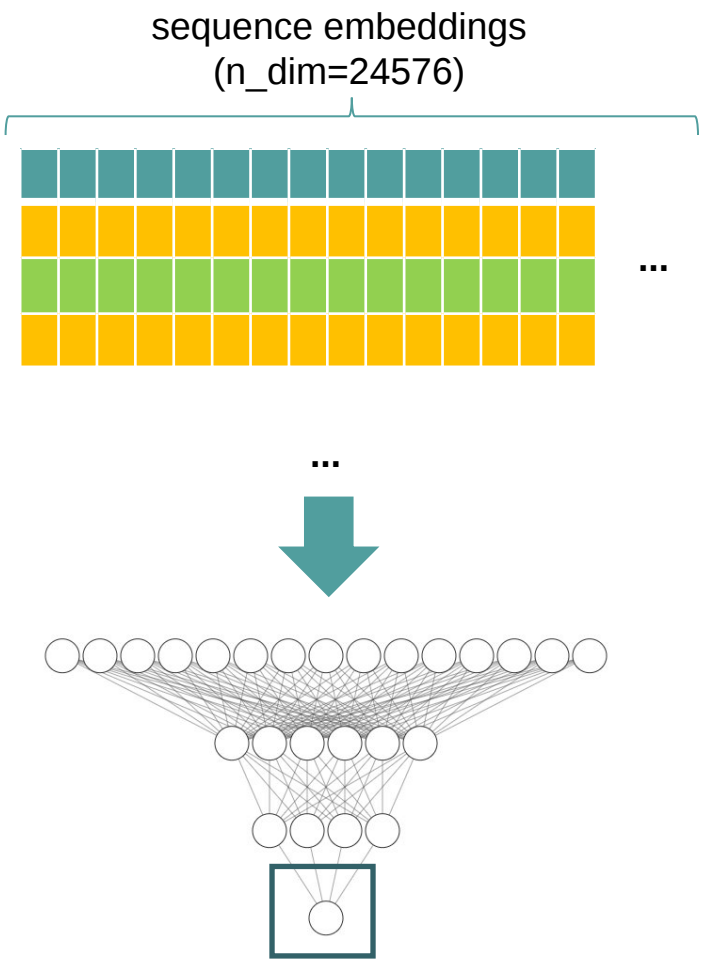






Pooling:

both direction :
- average
- max
– last token
– max
– average



[0–1] probability of being an ICE/IMEs

Firmidata transfer learning using Evo

	precision	recall	f1-score	support
Background	0.94	0.89	0.91	1110
Mobile	0.89	0.94	0.92	1110
accuracy			0.92	2220
macro avg	0.92	0.92	0.92	2220
weighted avg	0.92	0.92	0.92	2220

AUC-ROC: 0.9764
Matrice de Confusion:
[[987 123]
[63 1047]]

Firmidata transfer learning using Evo

	precision	recall	f1-score	support
Background	0.94	0.89	0.91	1110
Mobile	0.89	0.94	0.92	1110
accuracy			0.92	2220
macro avg	0.92	0.92	0.92	2220
weighted avg	0.92	0.92	0.92	2220

AUC-ROC: 0.9764
Matrice de Confusion:
[[987 123]
[63 1047]]

This model also works on another dataset: ICEBerg (198/207 ICEs/IMEs correctly identified).

Ongoing transfer learning experiments using a broader training dataset

Aliou Diallo,
M2 student,
StatInfOmics team

Firmidata + ICEberg experimental: distribution of ICEs and IMEs

Phylum
- Bacillota
- Pseudomonadota
- Bacteroidota
- Mycoplasmatota
- Campylobacterota

Type d'élément
- ICE (n=277)
- IME (n=205)

Source
- Firmidata v2 (n=256)
- ICEberg experimentally verified (n=226)

Ongoing transfer learning experiments using a broader training dataset

Aliou Diallo,
M2 student,
StatInfOmics team

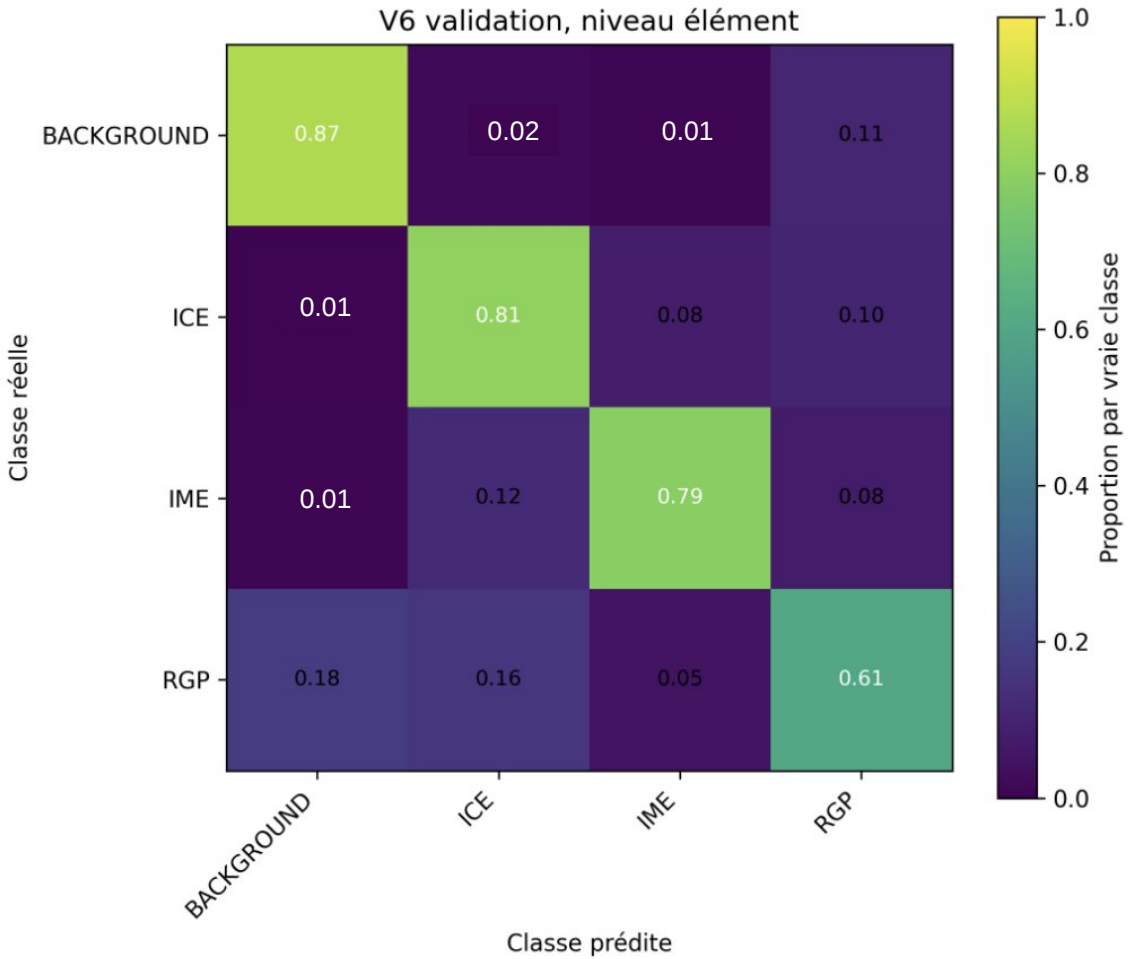
Firmidata + ICEberg experimental: distribution of ICEs and IMEs

external validation

Phylum
- Bacillota
- Pseudomonadota
- Bacteroidota
- Mycoplasmatota
- Campylobacterota

Type d'élément
- ICE (n=277)
- IME (n=205)

Source
- Firmidata v2 (n=256)
- ICEberg experimentally verified (n=226)



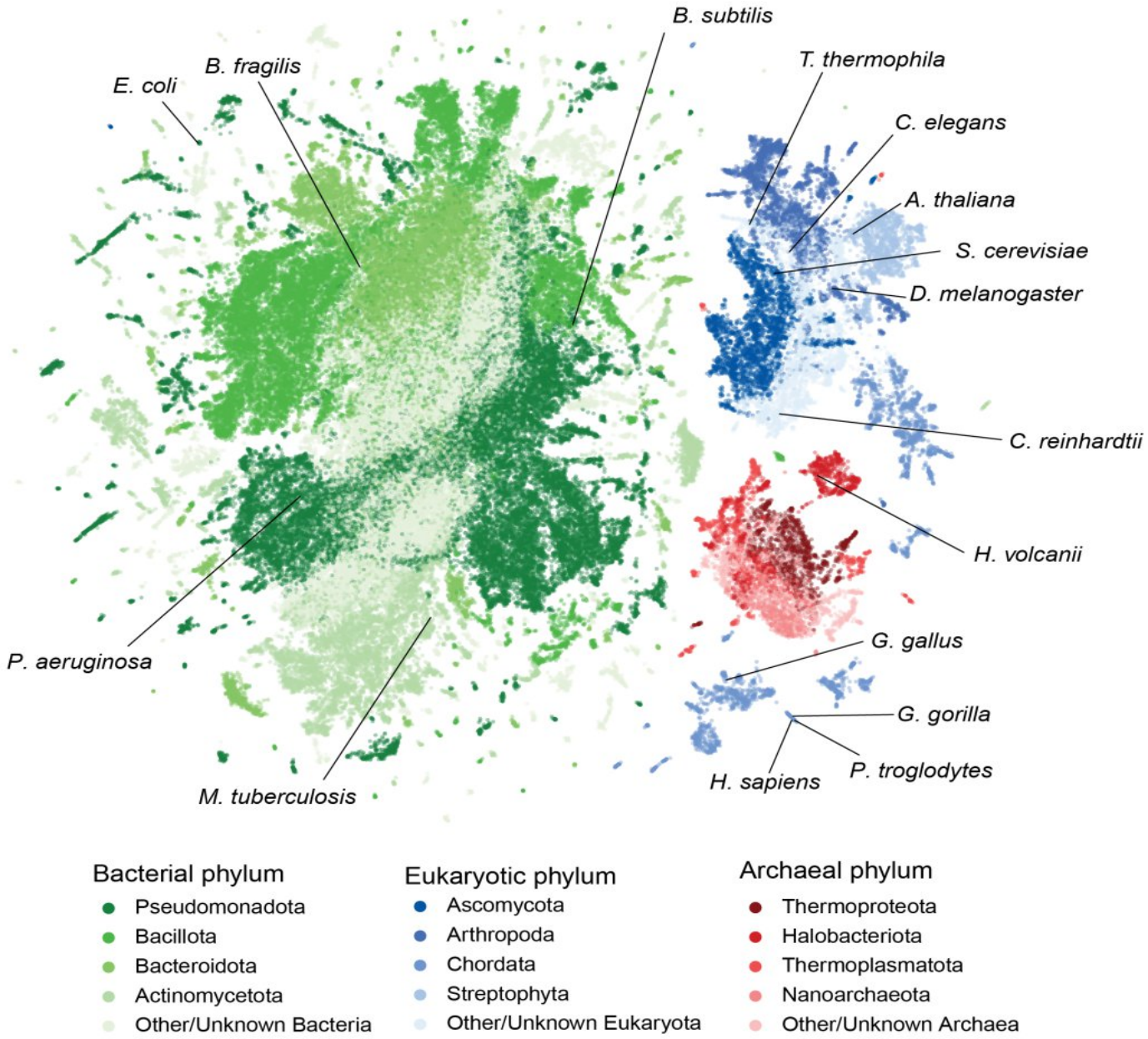
JOBIM 2026 Poster

Evo2: Genome modeling and design across **all domains of life** with Evo 2

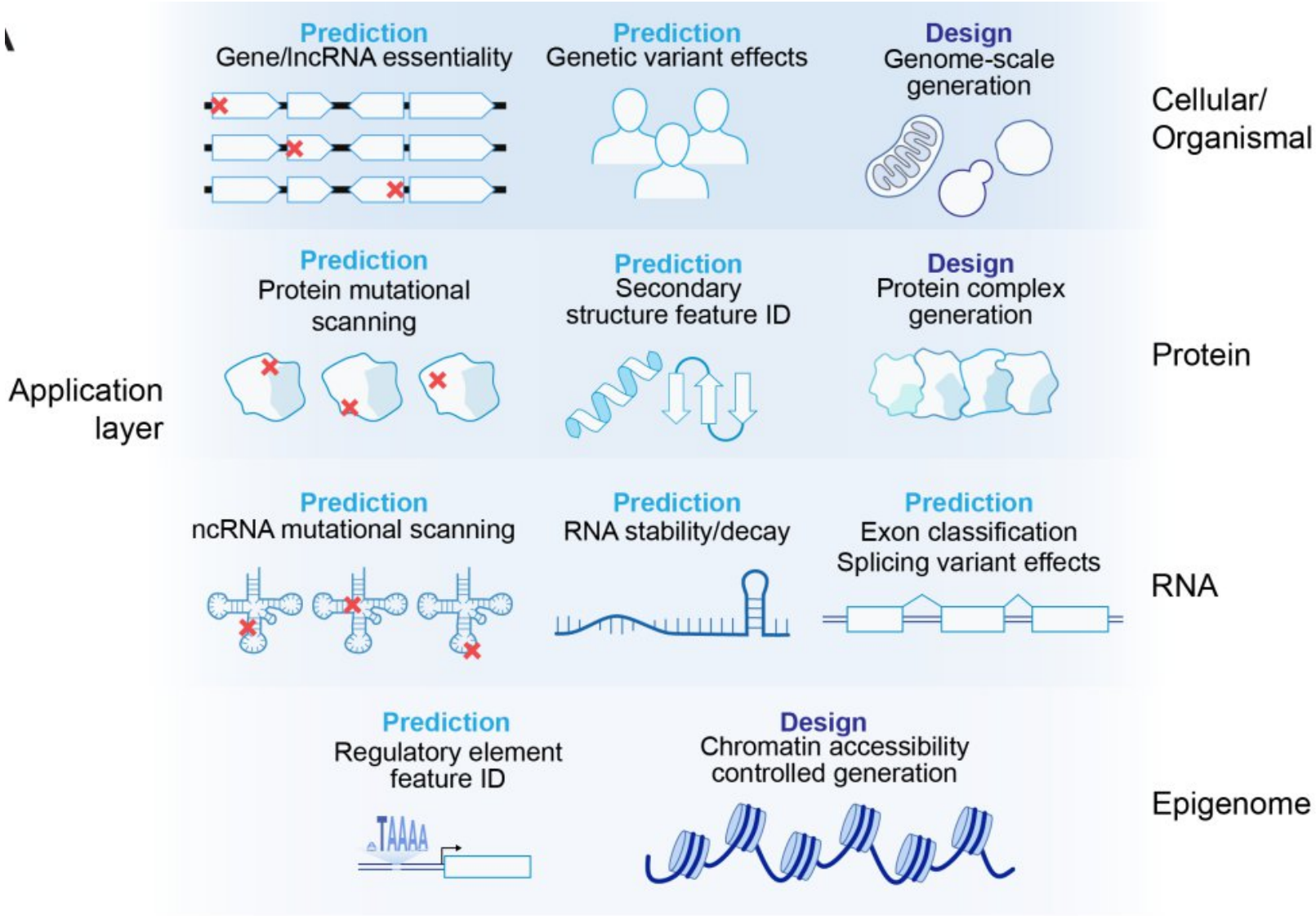
When computational biology changes its scale

Global UMAP of sequence diversity

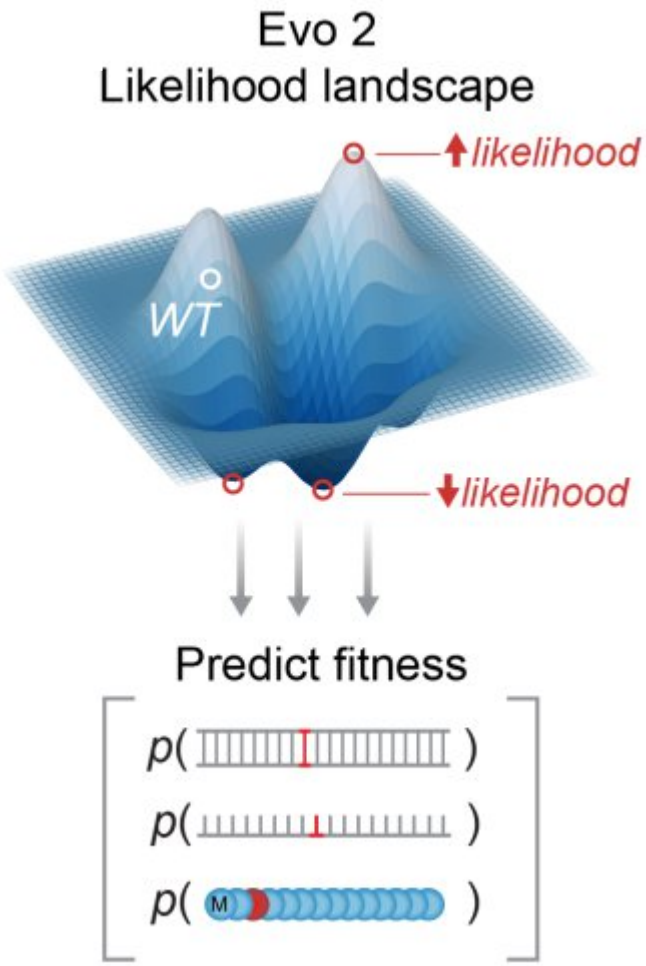
- Depict the great prokaryotic diversity
- The clustering doesn't exactly match the expected phylogeny especially for prokaryotes
(could be a UMAP artifact)

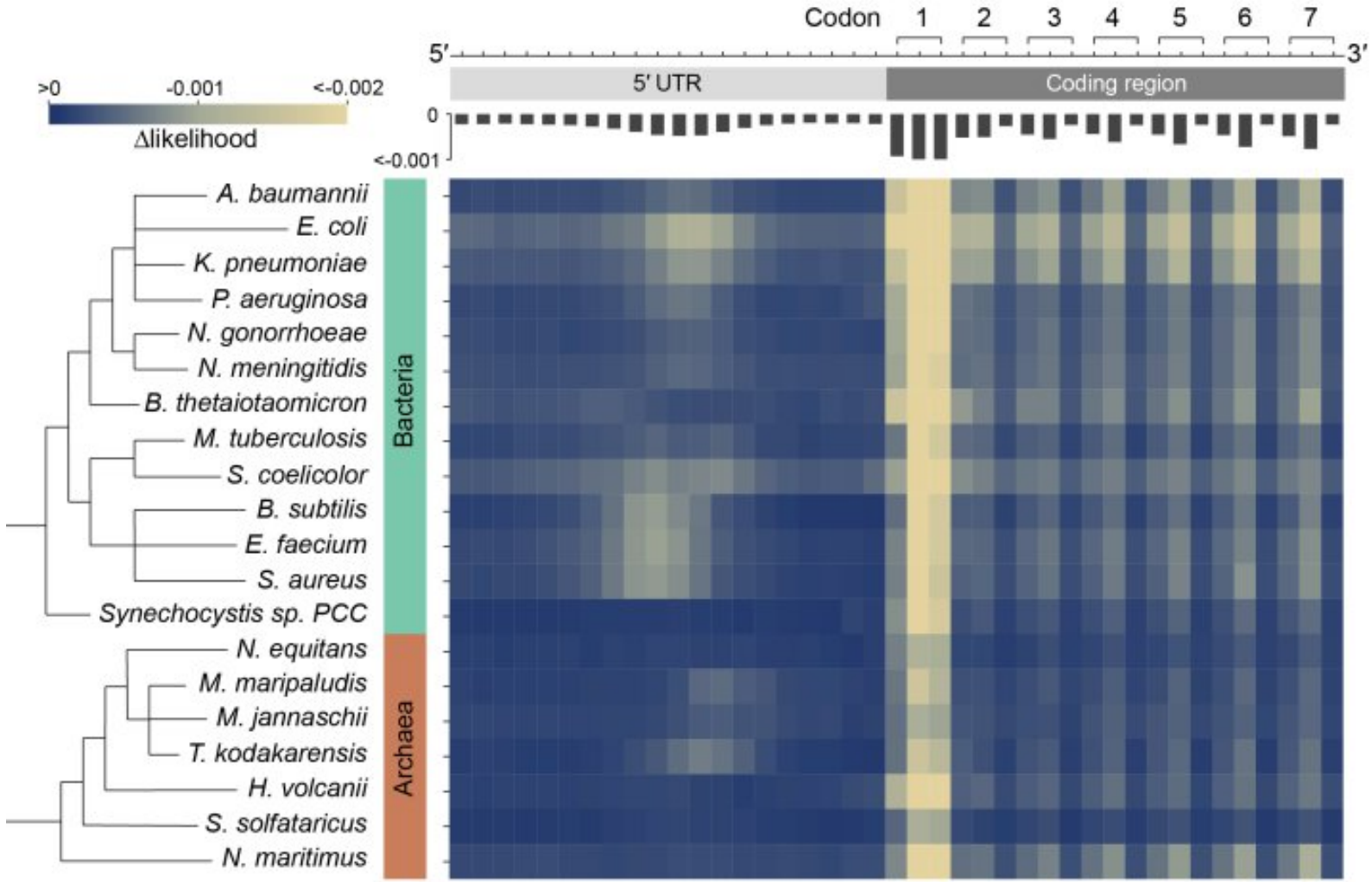


So many application layers



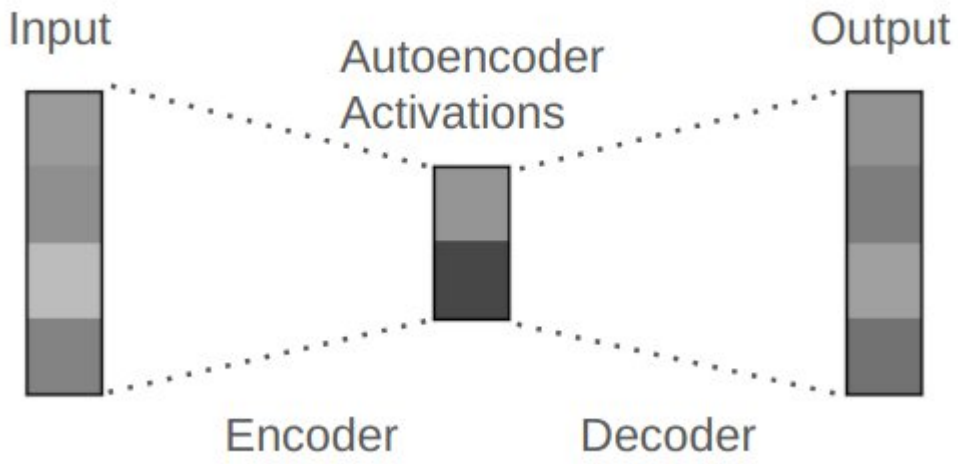
Applications: zero-shot prediction of variant effects

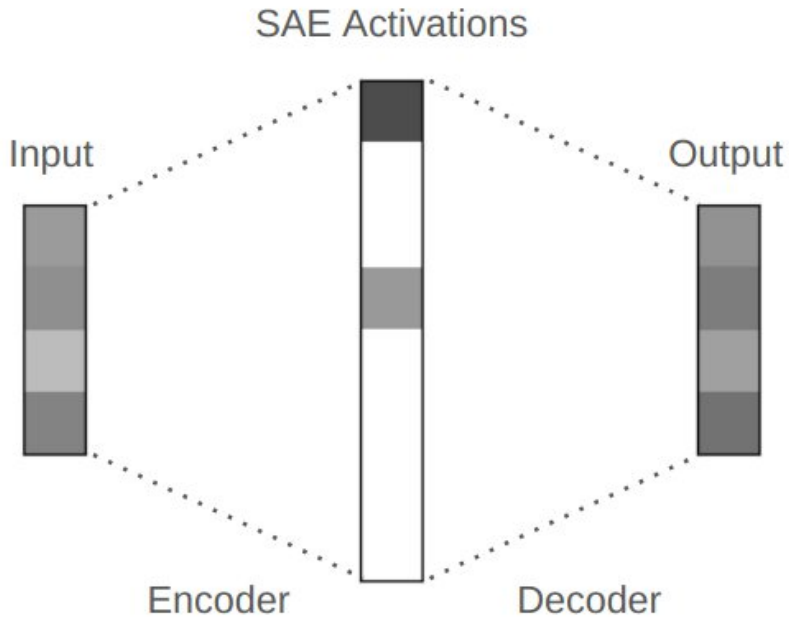




https://adamkarvonen.github.io/machine_learning/2024/06/11/sae-intuitions.html

Towards an interpretable model: Sparse Auto-Encoder (SAE)

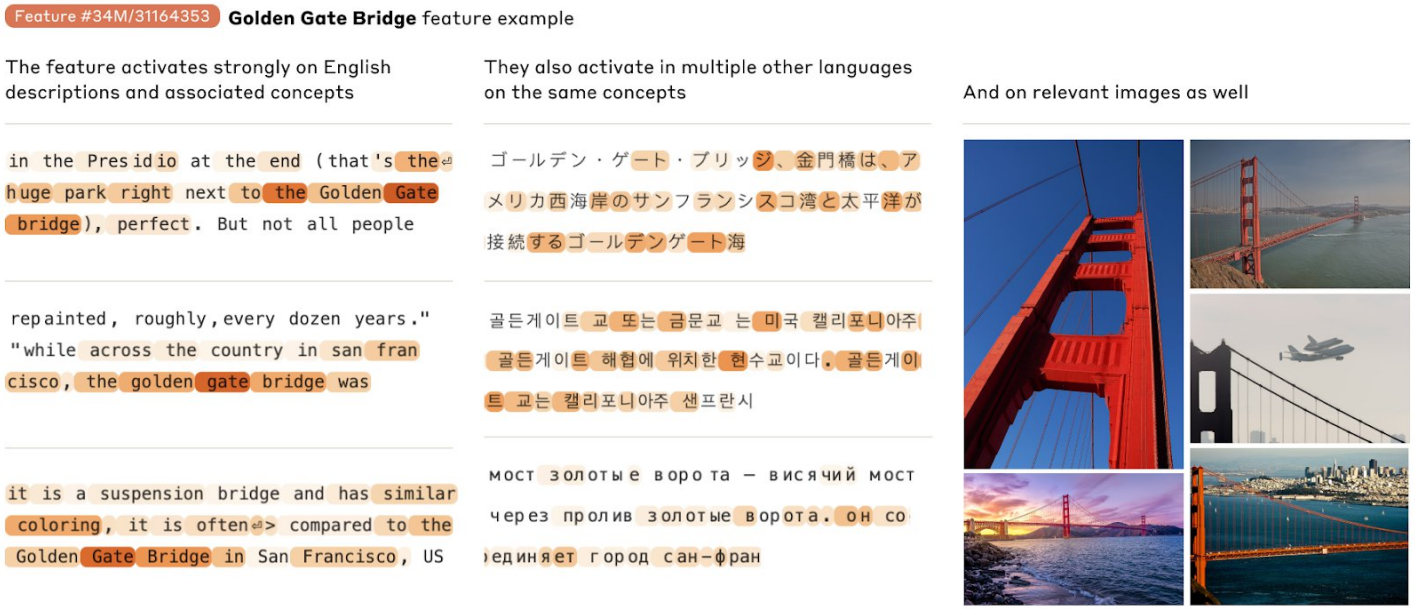


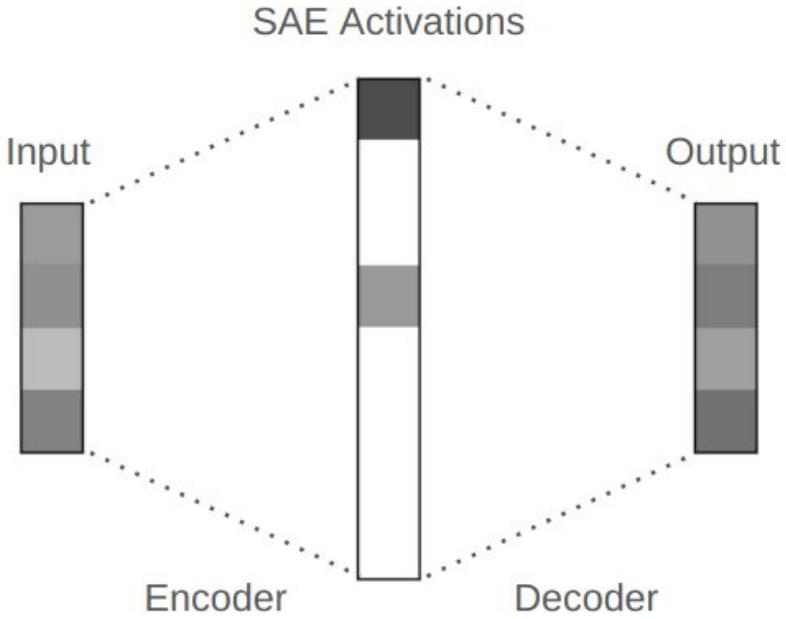


+ sparsity penalty in the loss function

https://adamkarvonen.github.io/machine_learning/2024/06/11/sae-intuitions.html

Towards an interpretable model: Sparse Auto-Encoder (SAE)





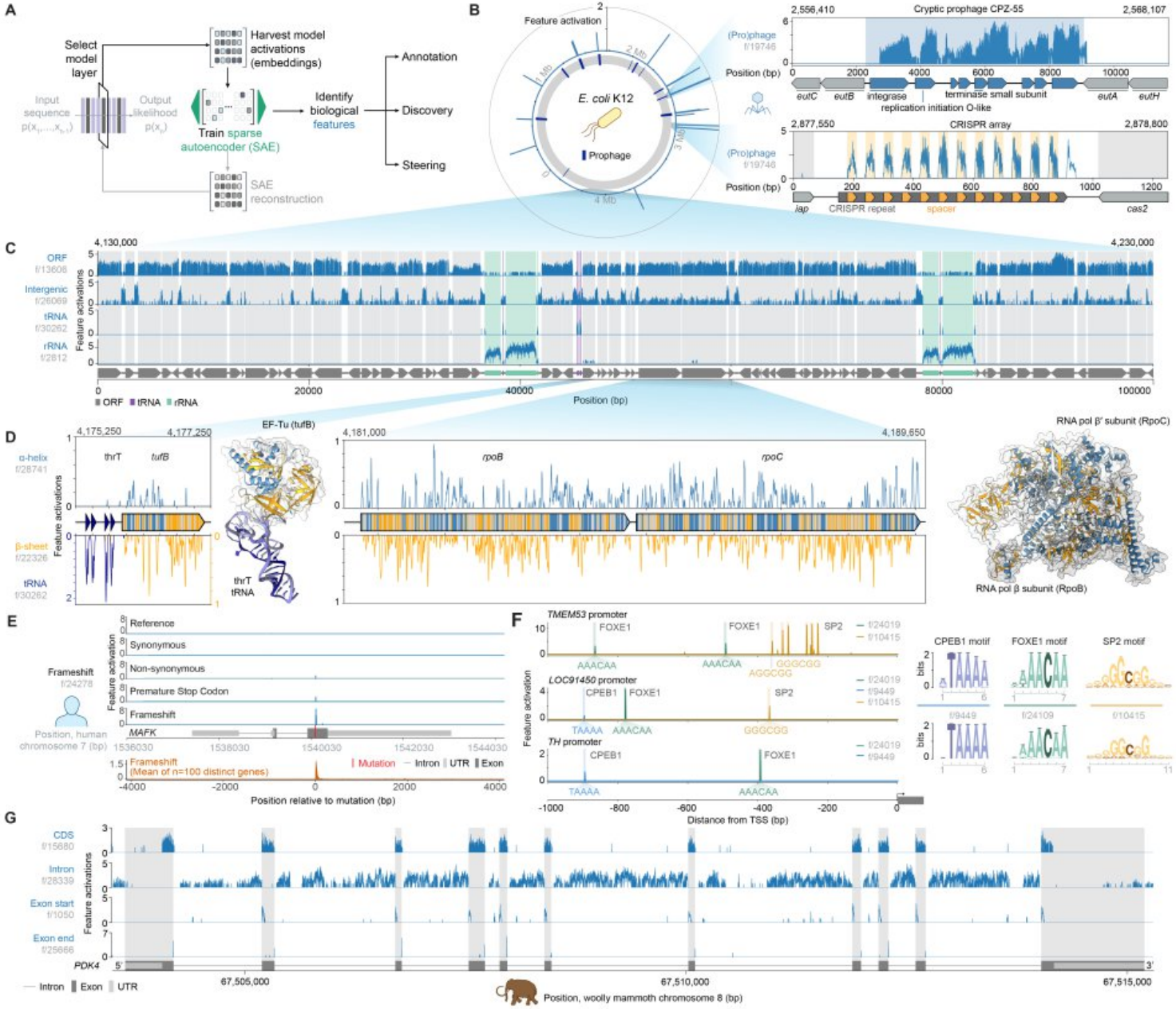
+ sparsity penalty in the loss function

https://adamkarvonen.github.io/machine_learning/2024/06/11/sae-intuitions.html

Towards an interpretable model: Sparse Auto-Encoder (SAE)

Features observed :

- Phage-associated elements
- ORFs and intergenic regions
- tRNAs and rRNAs
- α-helices and β-sheets
- Frameshift mutations
- Transcription-factor binding motifs
- Exon–intron signals (human genomes used to annotate the woolly mammoth genome)

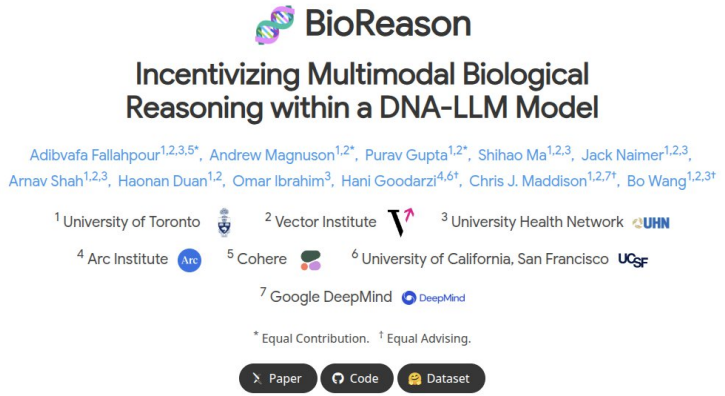


Conclusion

- Openly available, but requires several Nvidia A/H100 GPU to run Evo2 7B/40B (Running Evo 1/1.5 is easier)

→ User-friendly interface available (but limited to small prompts): https://arcinstitute.org/tools/evo/evo-designer

- Independant work on relevant tasks required ! ==> PhD thesis funded on this topic
- Ethical considerations to anticipate (biosecurity, misuses)
  - Bioterrorism (eukaryotic viruses weren't included in Evo2 but: what about fine tuning ?)
  - Fake dataset pollution (SynGenome database of 'generated genomes'), science integrity issues...
- Environmental concern if pre-training get again larger and larger (Evo 3 ?)
- Partially interpretable via Sparse AutoEncoders (explored in the Evo2 article)
- Multimodality : gLM X pPLM (gLM2, LucasOne)
  - LLM x gLM : « Talk to your genome » (hard prompting strategies)
    - ChatNT (NT X LLaMA, published Nature Machine Intelligence)
    - BioReason (Evo2 x Qwen3 CoT/RAG LLM, preprint on Biorxiv)
- DNA Retrieval Augmented Generation / Retrieval Augmented Fine Tuning



Acknowledgements

MaIAGE

+ inspiring chat with

Mahendra Mariadassou
StatInfOmics team leader

Naïa Périnelle,
former M2 student,
now PhD student at INRAE IRHS

Thomas Lacroix, IE
StatInfOmics team

Hélène Chiapello, IR
StatInfOmics and Migale team

Guillaume Kon Kam King, CR
StatInfOmics team

Arnaud Ferré, CR
Bibliome team

Internship funding:

GRADUATE SCHOOL
Informatique et Sciences du Numérique

université PARIS-SACLAY

GPU providers:

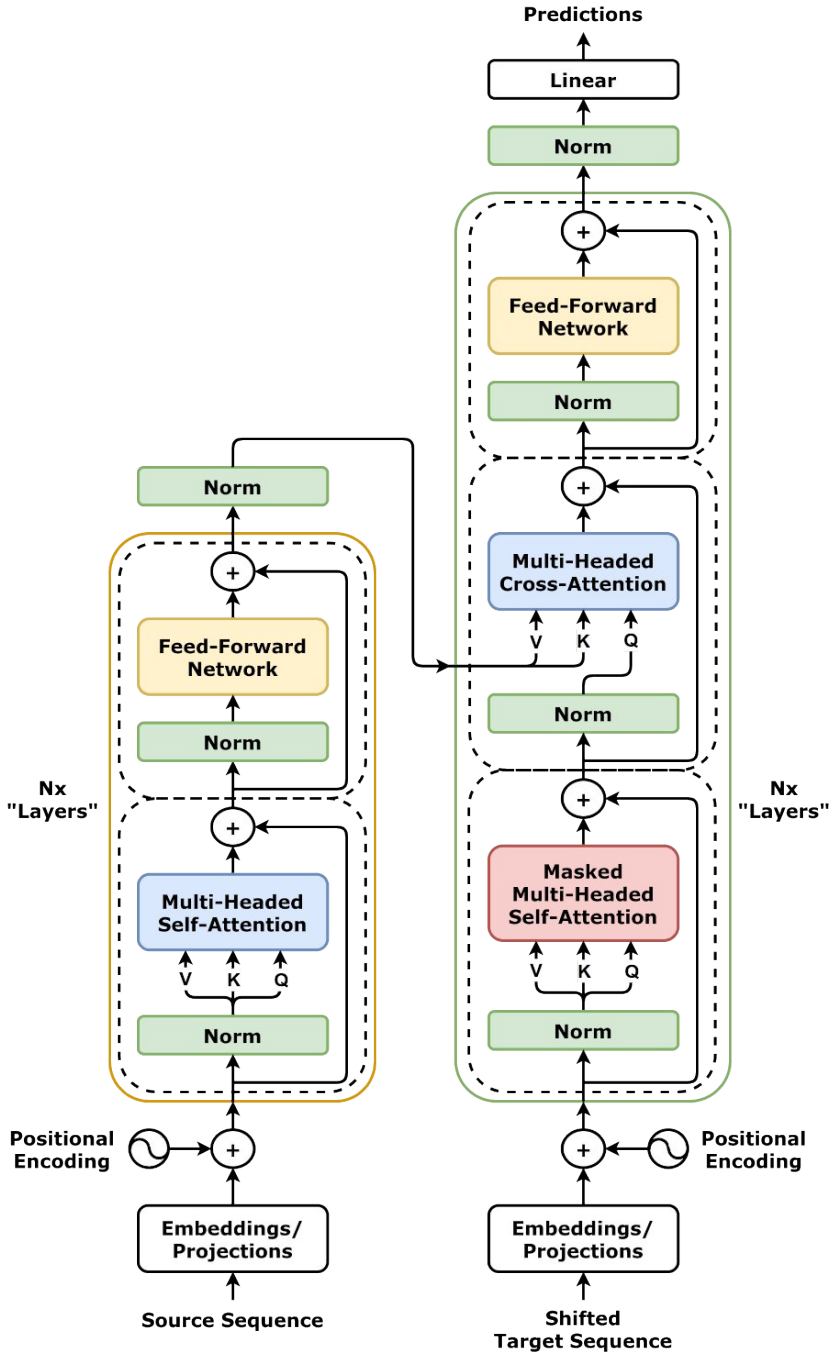
– Evo 1: LabIA (Paris Saclay)

– Evo 2: Jean Zay (IDRIS)

Evo: not (exactly) a transformer

Transformer

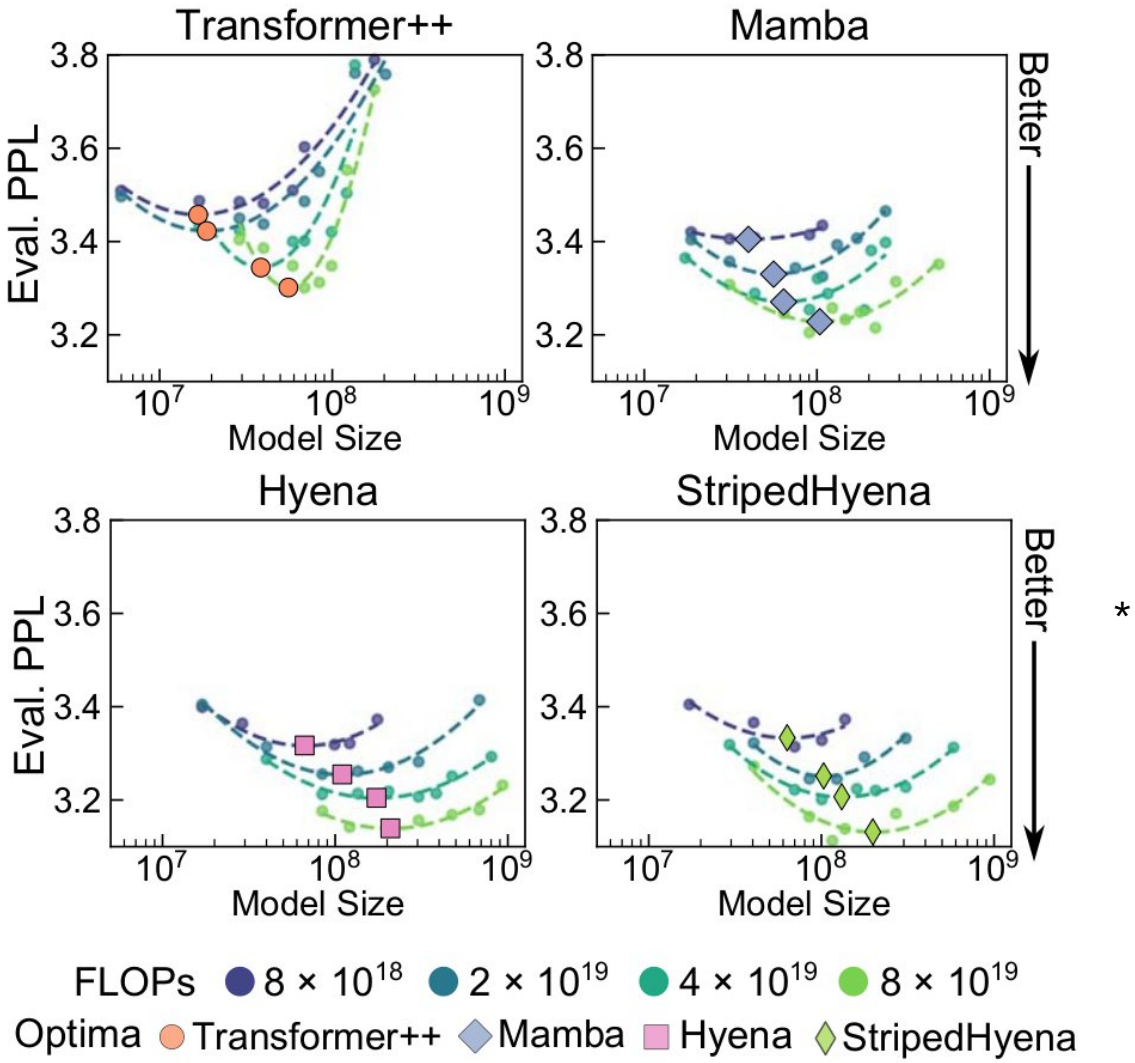
Attention Is All You Need*



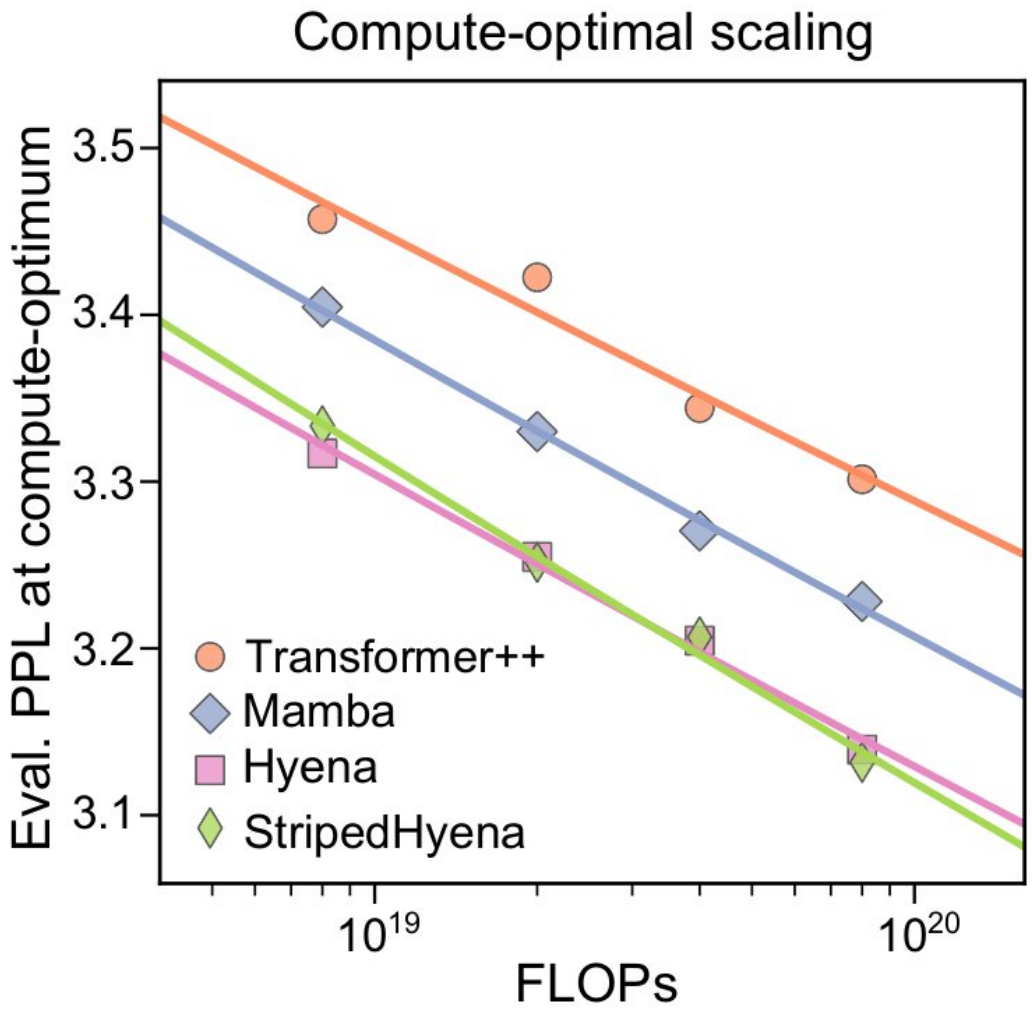
* provided you are NOT trying to process **whole genomes at base scale** without **billion-dollar scale** funding since:

Attention complexity is $O(N^2)$

Evo : not (exactly) a transformer



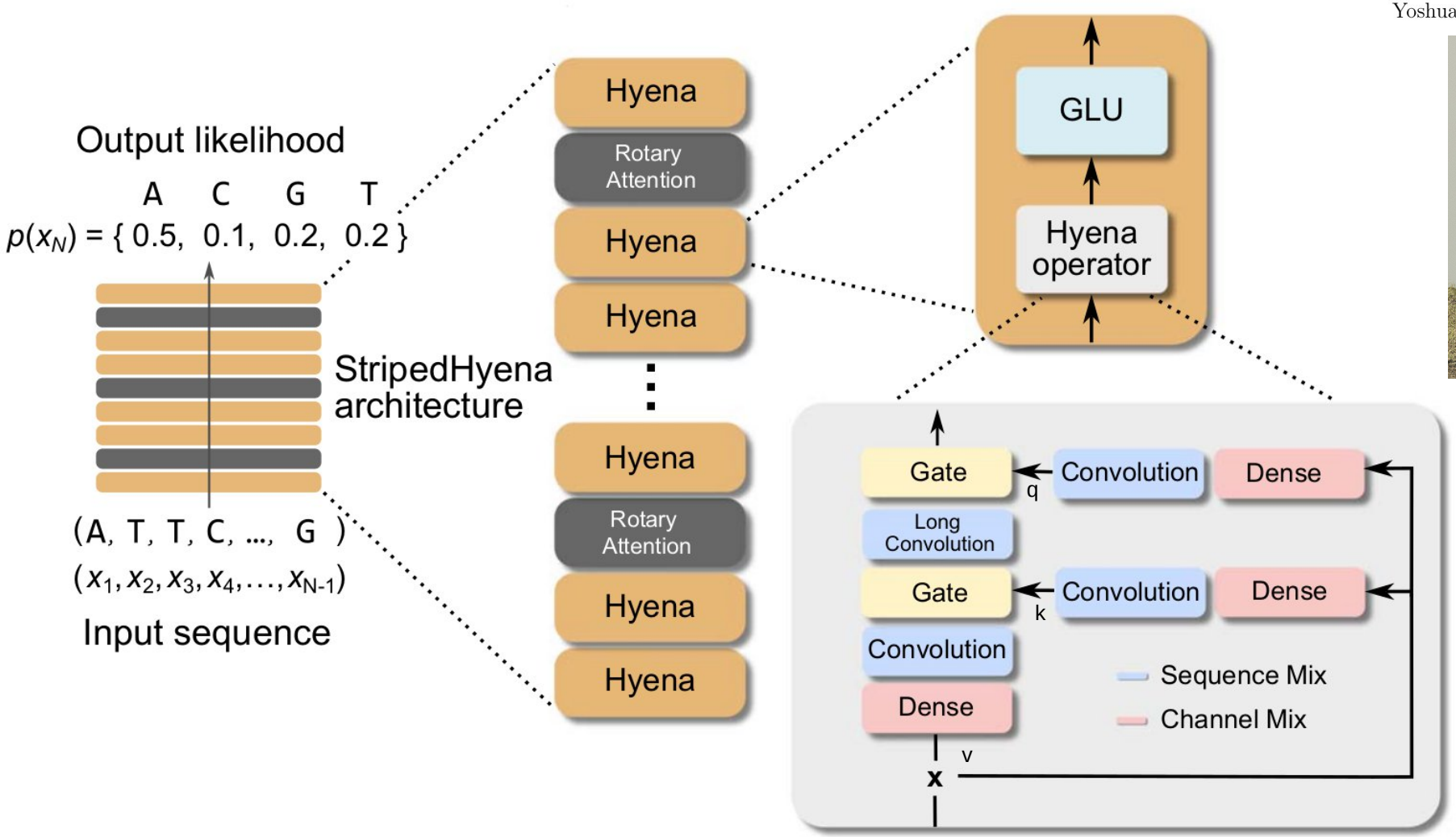
*



Evo is based on Striped Hyena

Hyena Hierarchy:
Towards Larger Convolutional Language Models

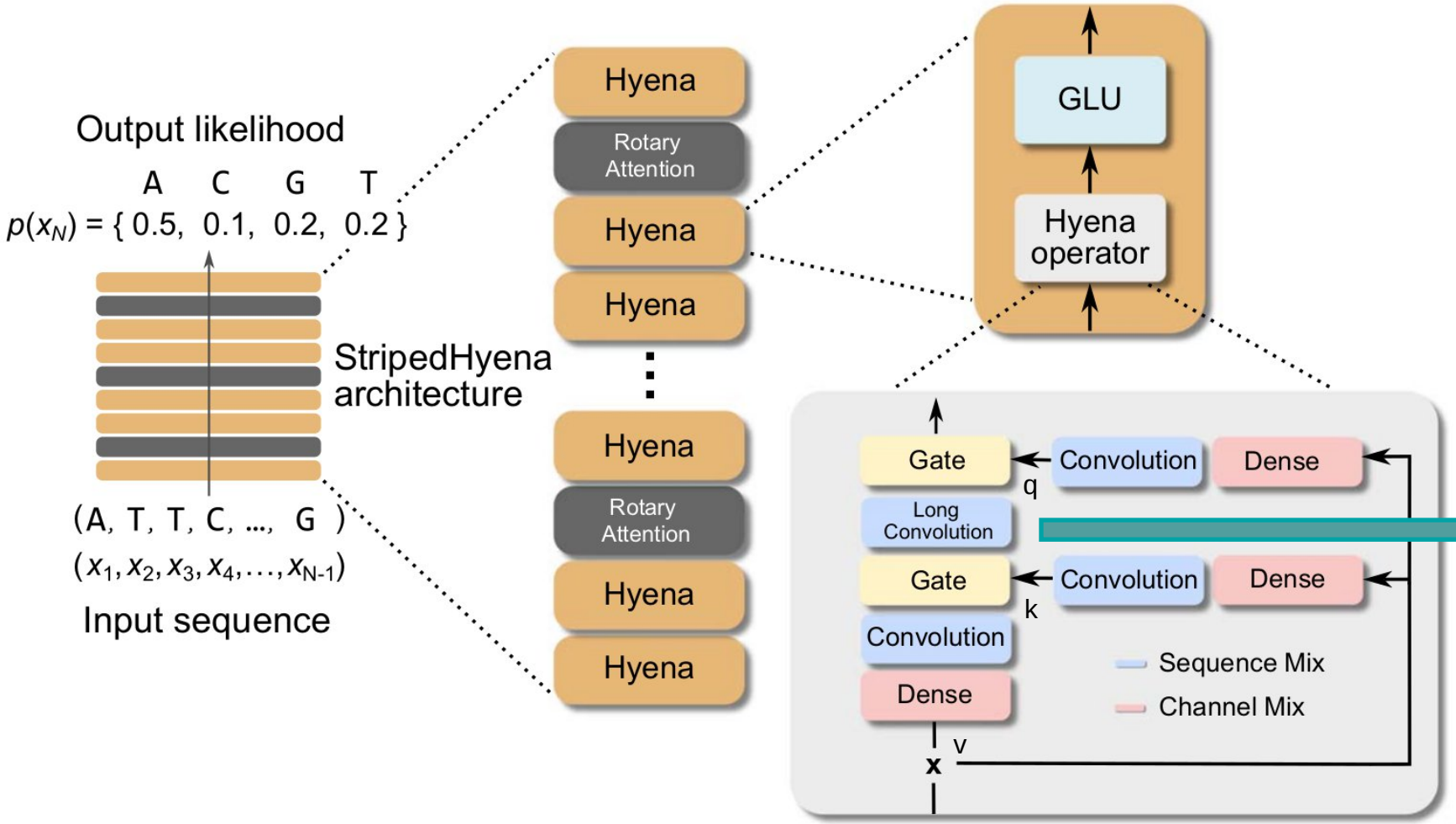
Michael Poli[*,1], Stefano Massaroli[*,2], Eric Nguyen[1,*],
Daniel Y. Fu[1], Tri Dao[1], Stephen Baccus[1],
Yoshua Bengio[2], Stefano Ermon[1,†], Christopher Ré[1,†]



Evo is based on Striped Hyena

Hyena Hierarchy:
Towards Larger Convolutional Language Models

Michael Poli[*,1], Stefano Massaroli[*,2], Eric Nguyen[1,*],
Daniel Y. Fu[1], Tri Dao[1], Stephen Baccus[1],
Yoshua Bengio[2], Stefano Ermon[1,†], Christopher Ré[1,†]



Complexity: $O(Nk)$

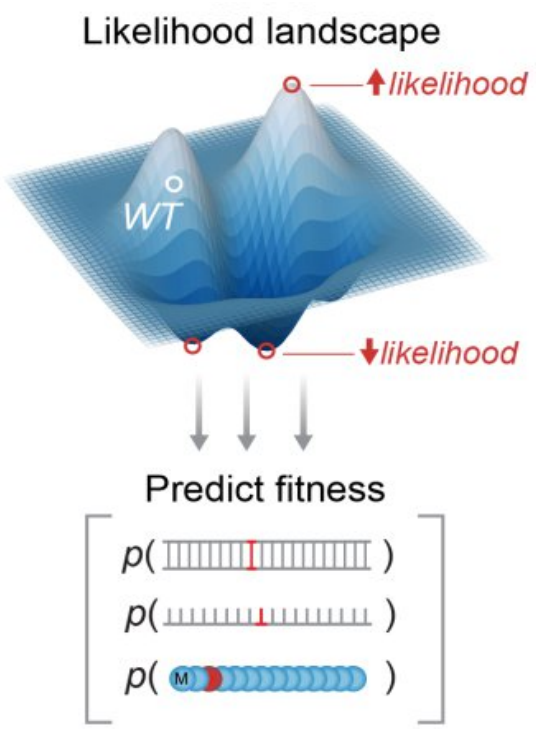
N : # of tokens in window

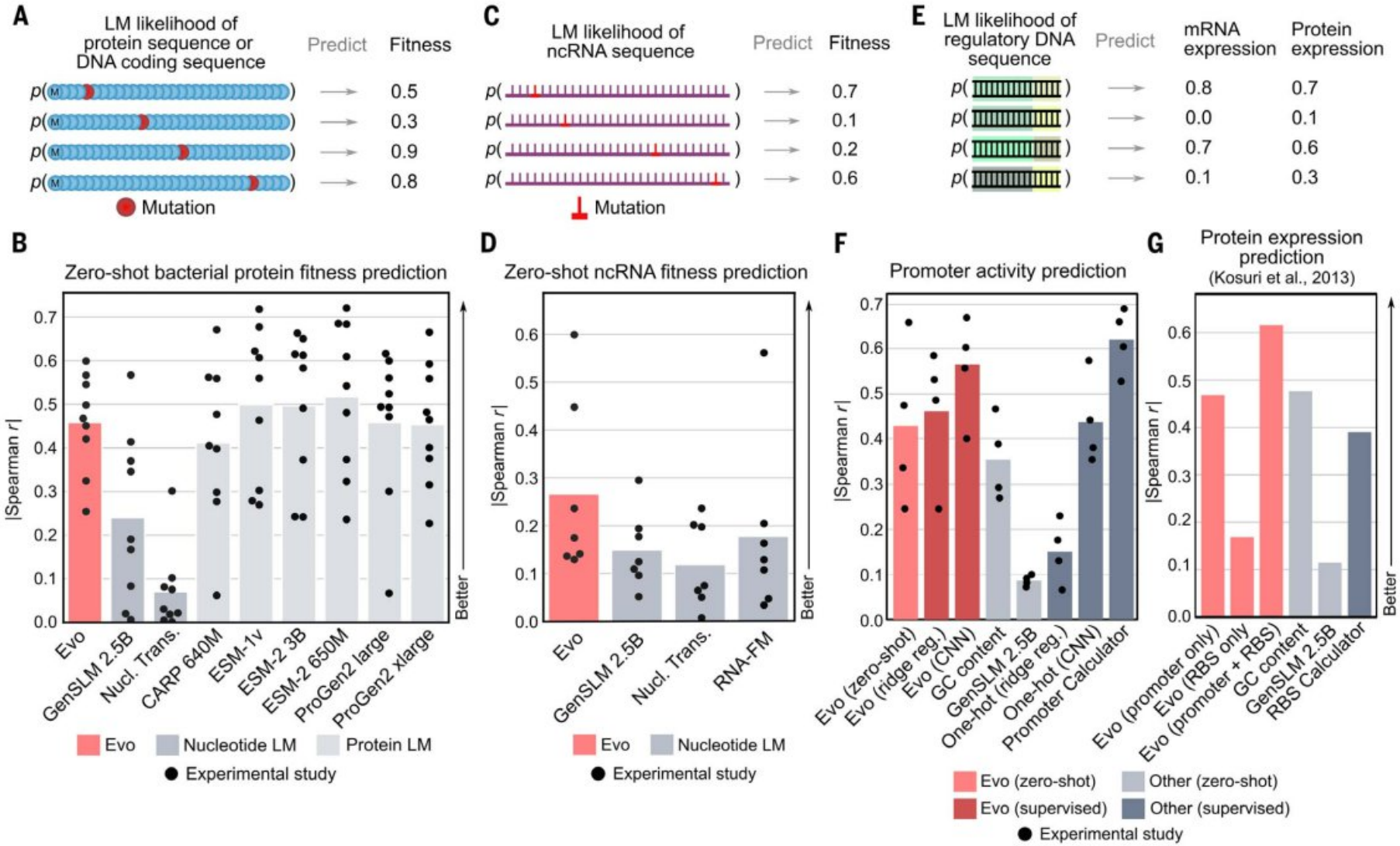
k : convolution kernel size

if $k \to N \Rightarrow$ Complexity: $O(N^2)$

Fast Fourier Transform: $O(N \log(N))$

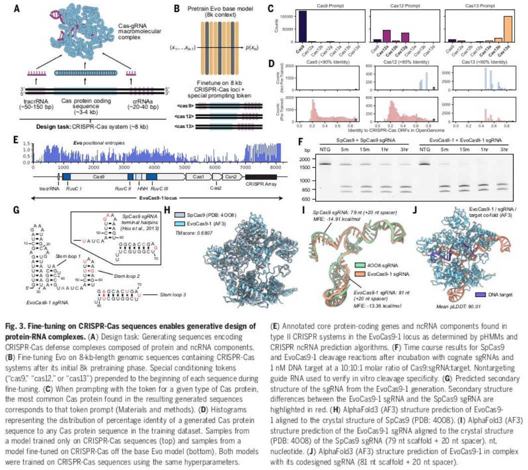
Applications: experimentation-free prediction of sequence effects



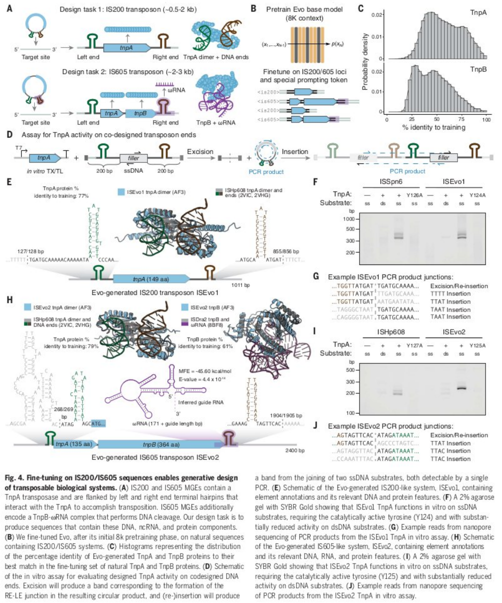


Much more applications

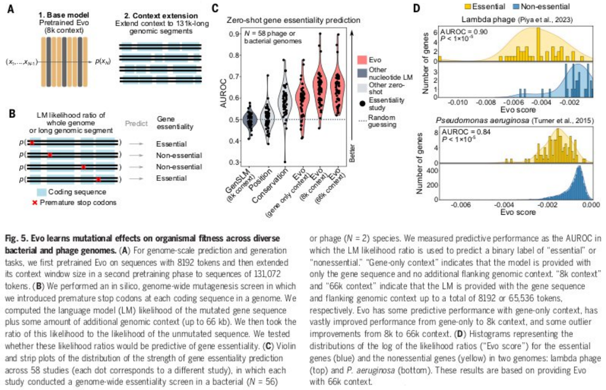
- Generative design of protein-RNA complexes



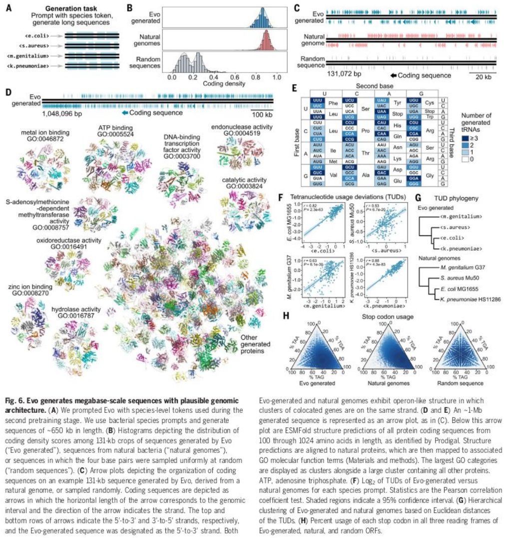
- Generative design of transposable biological systems



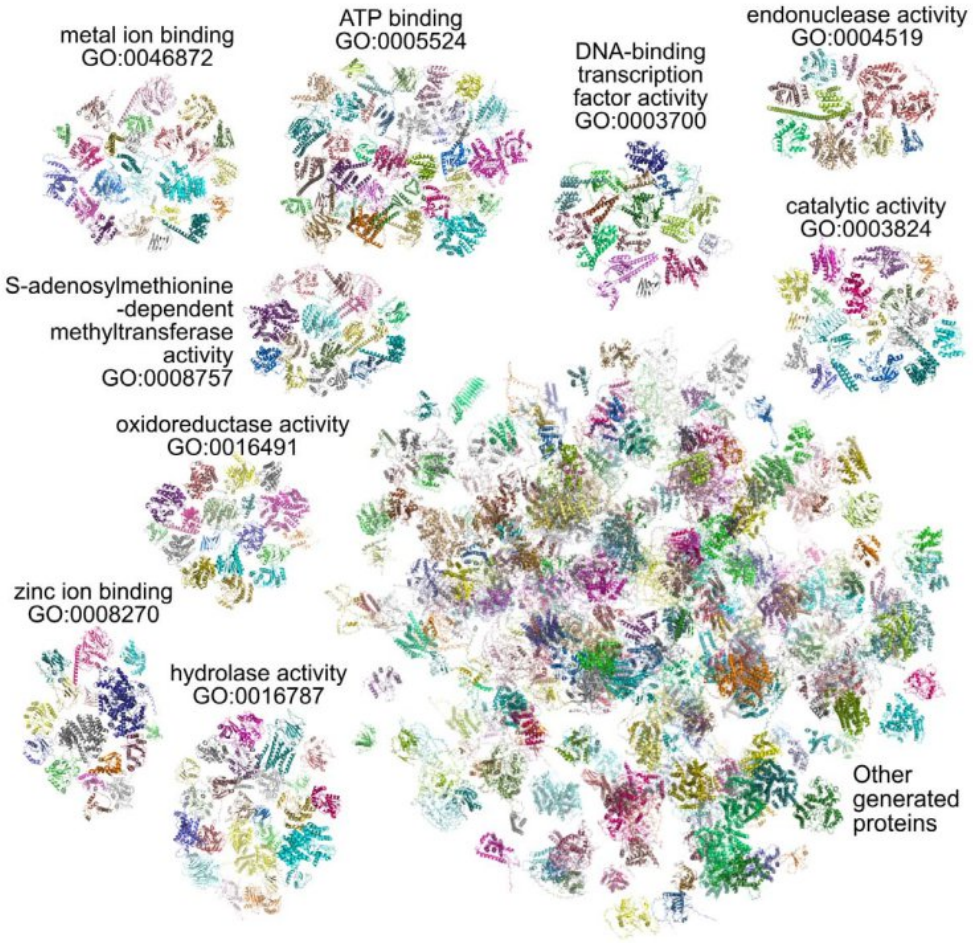
- Prediction of gene essentiality



- Evo generates megabase-scale sequences with plausible genomic architecture



Generated sequence : what about homology with real sequences?



« ESMFold structure predictions of all protein-coding sequences from 100 to 1024 amino acids in length, as identified by Prodigal.

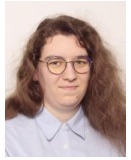
The structure predictions are aligned to natural proteins, which are then mapped to associated GO molecular function terms (Materials and Methods).

The largest GO categories are displayed as clusters, alongside a large cluster containing all other proteins. ATP, adenosine triphosphate. »

⇨ why something so sophisticated without just indicating the %id percent?

Feedback on Evo 1

Evaluation of the credibility of the output



Naïa Périnelle
(M2 student)

Methodology

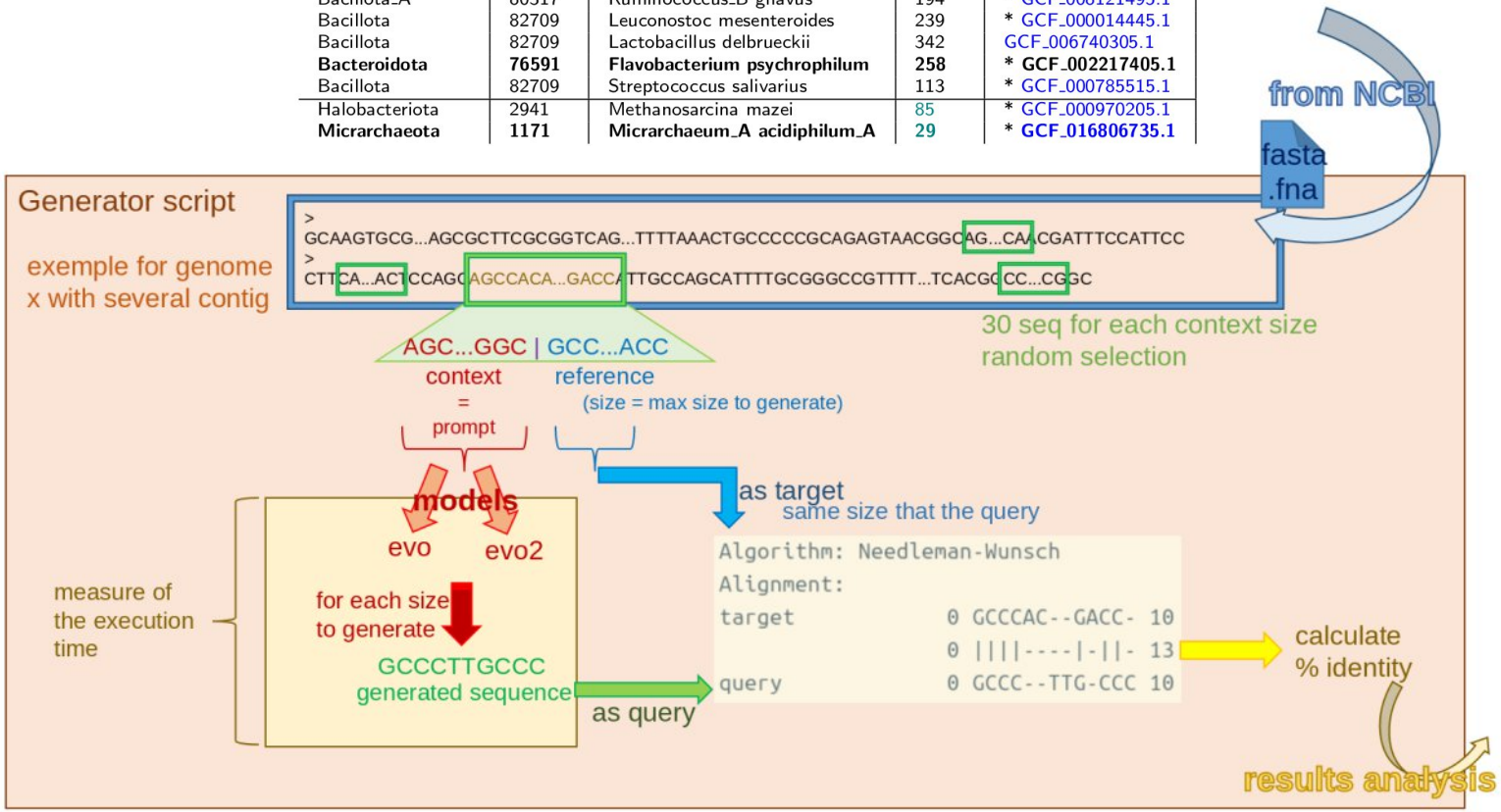
1. Sampling real sequences at random position from 28 genomes

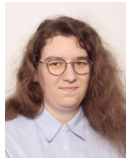
2. Each sequence is split into:
   - a prompt (context)
   - a reference

3. A sequence is generated from the prompt using Evo with the same length as the reference

4. The generated sequence is aligned to the reference
   (Gotoh global alignment, match: 1 ; open gap: -5 ; mismatch/gap extension: -1)

Phylum	p_nbr	Species	s_nbr	Accession number
Acidobacteriota	3175	Geothrix sp903857495	50	* GCA_903857495.1
Actinomycetota	44996	Mycobacterium tuberculosis	7337	* GCF_000195955.2
Bacillota	82709	Staphylococcus aureus	16021	* GCF_001027105.1
Bacillota_A	80317	Clostridioides difficile	2991	* GCF_001077535.1
Bacillota_B	1092	Avidehalobacter sp022797295	133	* GCA_022797295.1
Bacillota_C	2749	Phascolarctobacterium faecium	115	* GCF_003269275.1
Bacillota_I	10814	Faecalibaculum rodentium	403	* GCF_001564455.1
Bacteroidota	76591	UBA7173 sp001689485	1211	* GCA_001689485.1
Campylobacterota	11105	Campylobacter_D jejuni	2873	* GCF_001457695.1
Chloroflexota	4762	UBA9611 sp002746355	30	GCA_950054275.1
Cyanobacteriota	5634	MGBC122484 sp910586855	146	* GCA_910586855.1
Desulfobacterota	4840	Taurinovorans muris	498	* GCF_025232395.1
Myxococcota	1378	Corallococcus exiguus	19	* GCF_009909105.1
Nitrospirota	1001	Nitrospira_A sp900170025	23	* GCF_900170025.1
Patescibacteria	8106	Nanosyncoccus sp948851665	87	GCA_949464065.1
Pseudomonadota	214930	Echerichia coli	38926	* GCF_003697165.2
Spirochaetota	4402	Leptospira interrogans	391	* GCF_900156205.1
Verrucomicrobiota	6636	Akkermansia muciniphila	1061	* GCF_000020225.1
Pseudomonadota	214930	Halomonas elongata	6	* GCF_000196875.2
Bacteroidota	76591	Prevotella copri	29	* GCF_025151535.1
Bacillota_A	80317	Faecalibacterium prausnitzii	124	* GCF_003324185.1
Bacillota_A	80317	Ruminococcus_B gnavus	194	* GCF_008121495.1
Bacillota	82709	Leuconostoc mesenteroides	239	* GCF_000014445.1
Bacillota	82709	Lactobacillus delbrueckii	342	GCF_006740305.1
Bacteroidota	**76591**	**Flavobacterium psychrophilum**	**258**	* **GCF_002217405.1**
Bacillota	82709	Streptococcus salivarius	113	* GCF_000785515.1
Halobacteriota	2941	Methanosarcina mazei	85	* GCF_000970205.1
Micrarchaeota	**1171**	**Micrarchaeum_A acidiphilum_A**	**29**	* **GCF_016806735.1**

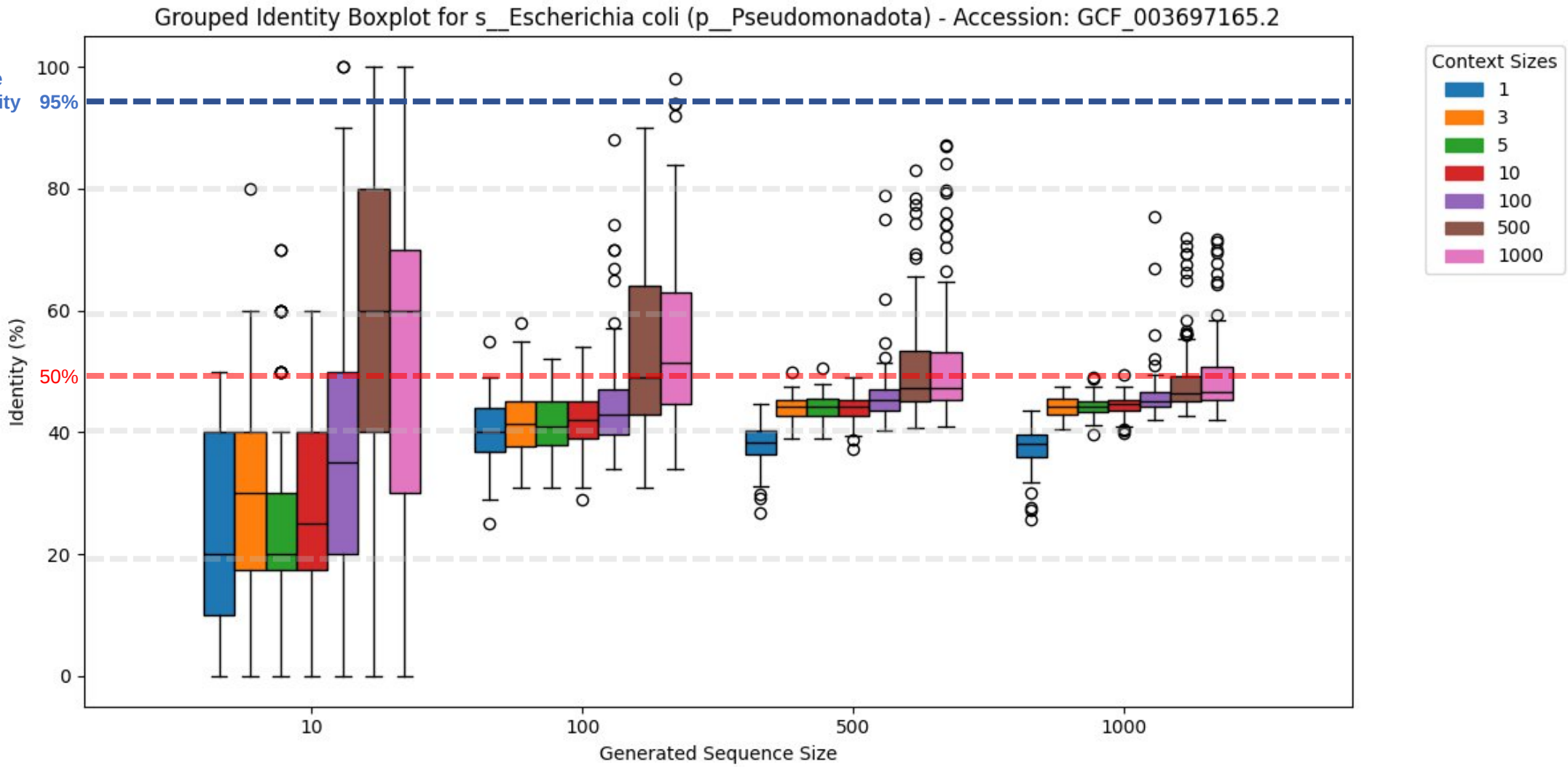


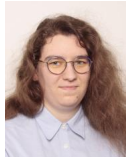


Results Evo 1 131k on E. coli

Naïa Périnelle
(M2 student)

Species Average Nucleotide Identity in Prokaryotes

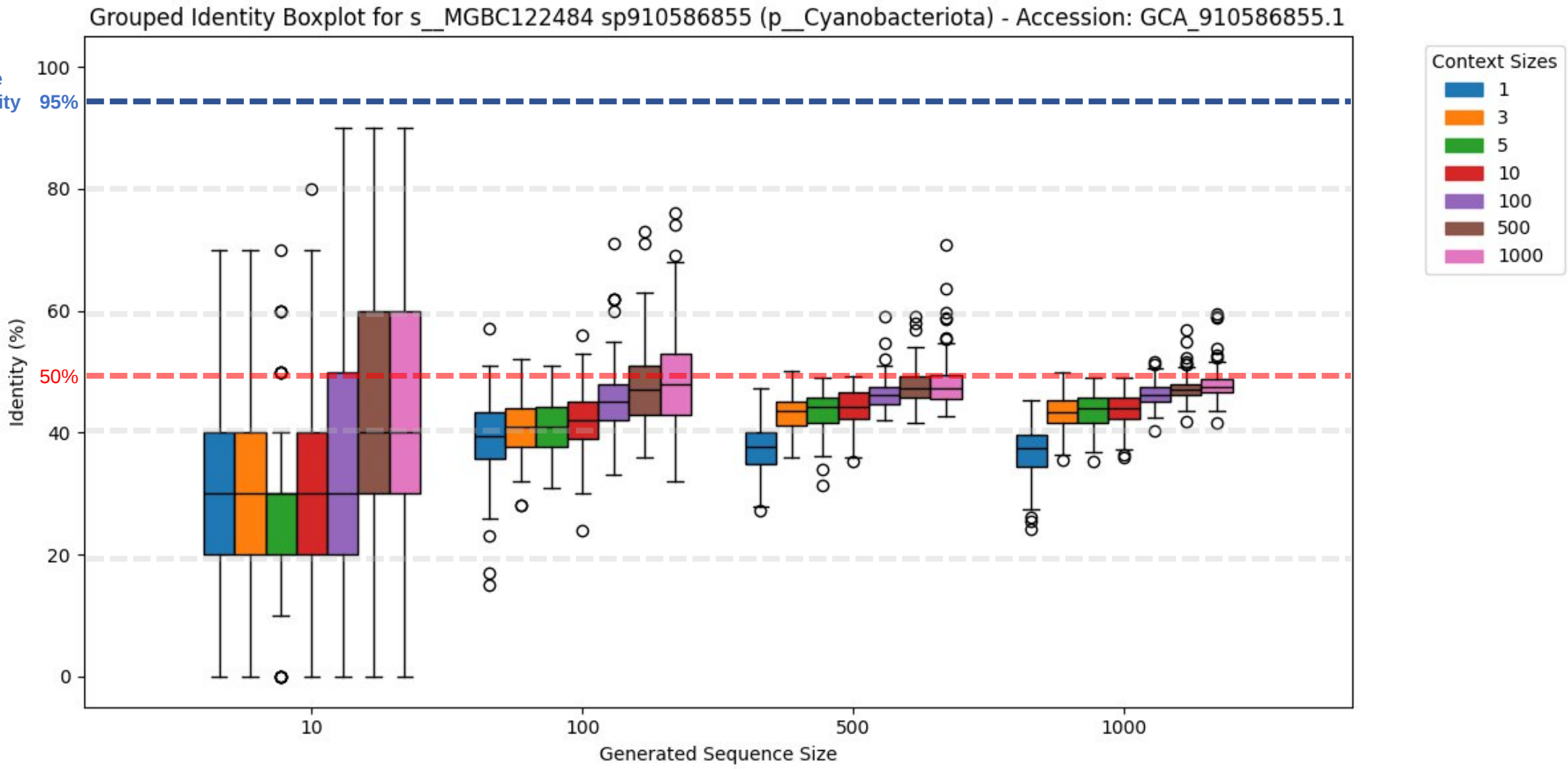


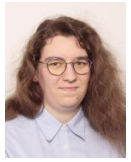


Naïa Périnelle
(M2 student)

Results Evo 1 131k on « unnamed cyanobacteria »

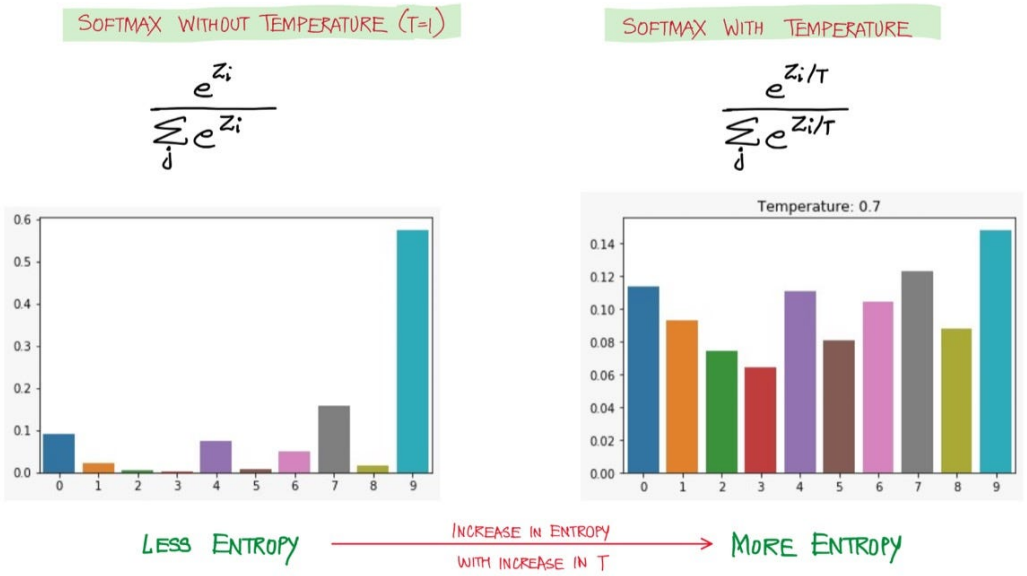
Species Average Nucleotide Identity in Prokaryotes

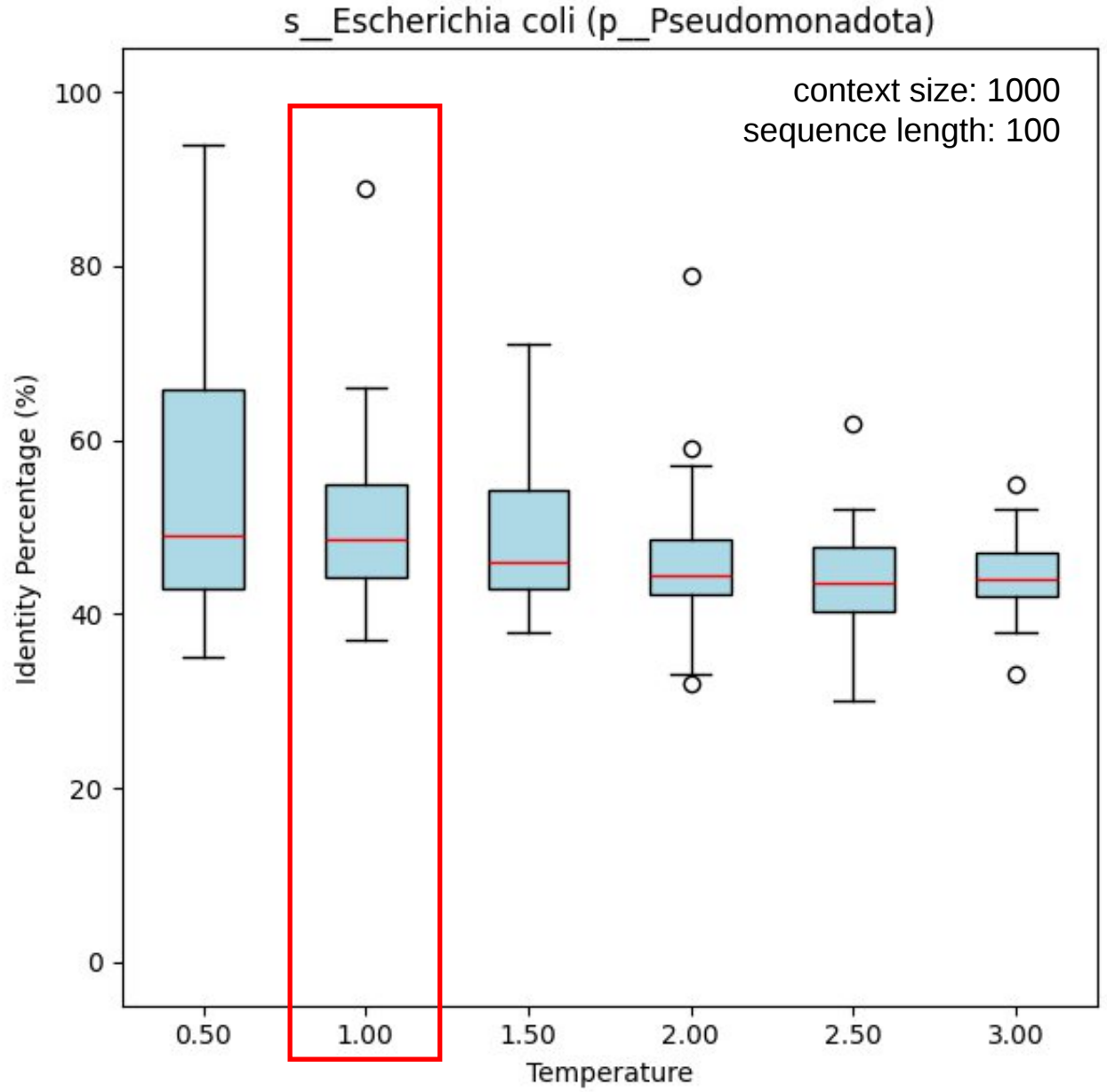




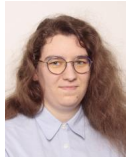
Naïa Périnelle
(M2 student)

Impact of temperature





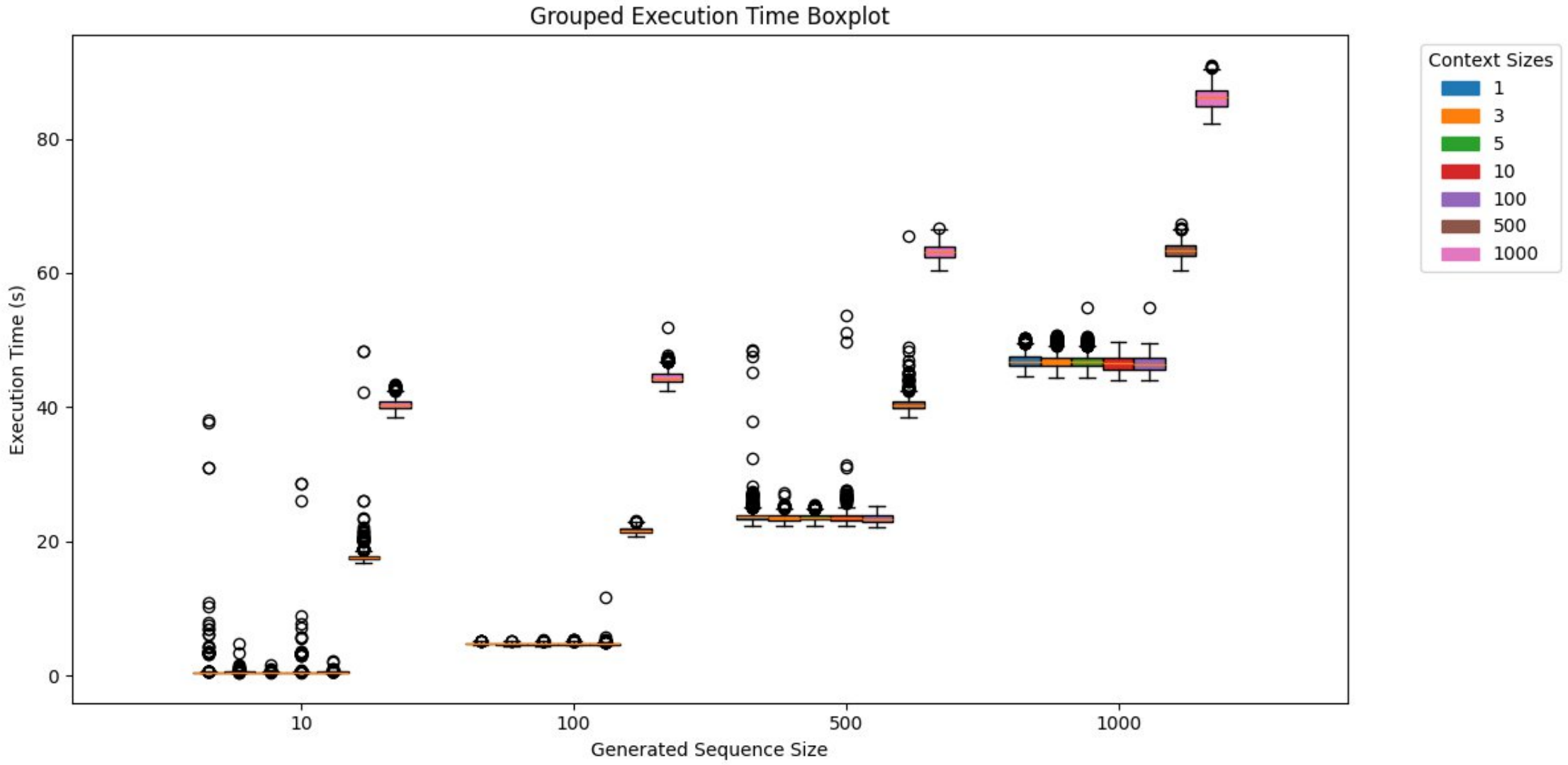
+ top_p=1, top_k=4



Naïa Périnelle
(M2 student)

Execution times

On RTX A6000 (LabIA LISN Univ. Paris-Saclay)



INRAE

Credibility seems to be improved in Evo 1.5

And it leads to new unprecedented forms of life

Evo 1.5

- Same architecture but extending the pretraining dataset of Evo 1 (8k) **+50%**
- Still on prokaryote/phage
- from 315 billion tokens (75,000 iterations) to 470 billion tokens (112,000 iterations)
- Only an 8Kb version

nature

Article

Semantic design of functional de novo genes from a genomic language model

https://doi.org/10.1038/s41586-025-09749-7

Received: 10 December 2024

Accepted: 13 October 2025

Published online: 19 November 2025

Open access

Check for updates

Aditi T. Merchant[1,2], Samuel H. King[1,2], Eric Nguyen[1,2] & Brian L. Hie[2,3,4]

Generative genomic models can design increasingly complex biological systems[1]. However, controlling these models to generate novel sequences with desired functions remains challenging. Here, we show that Evo, a genomic language model, can leverage genomic context to perform function-guided design that accesses novel regions of sequence space. By learning semantic relationships across prokaryotic genes[2], Evo enables a genomic 'autocomplete' in which a DNA prompt encoding genomic context for a function of interest guides the generation of novel sequences enriched for related functions, which we refer to as 'semantic design'. We validate this approach by experimentally testing the activity of generated anti-CRISPR proteins and type II and III toxin–antitoxin systems, including de novo genes with no significant sequence similarity to natural proteins. In-context design of proteins and non-coding RNAs with Evo achieves robust activity and high experimental success rates even in the absence of structural priors, known evolutionary conservation or task-specific fine-tuning. We then use Evo to complete millions of prompts to produce SynGenome, a database containing over 120 billion base pairs of artificial intelligence-generated genomic sequences that enables semantic design across many functions. More broadly, these results demonstrate that generative genomics with biological language models can extend beyond natural sequences.