



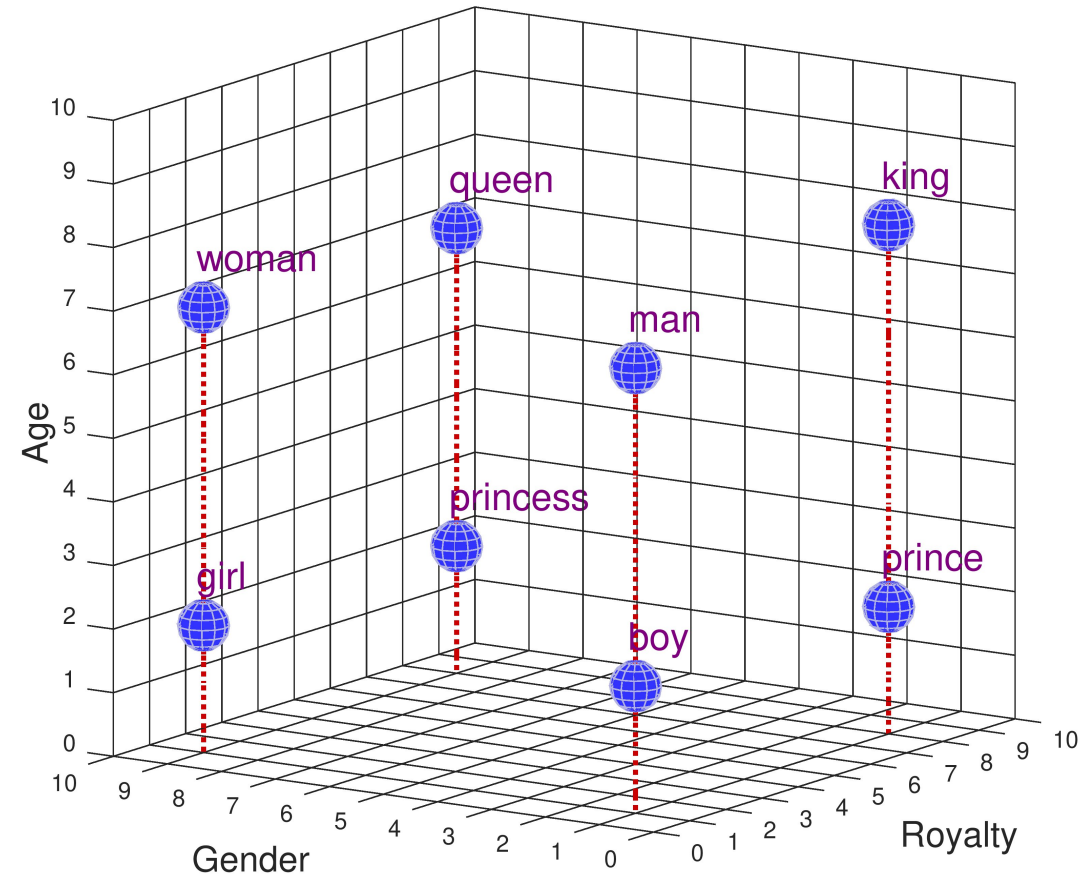
➤ Hands-on workshop on genomic language models

Guillaume GAUTREAU (CRCN), MaIAGE unit, StatInfOmics team

12/06/2026

➤ Token embeddings in a multidimensional space

3D Semantic Feature 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

Lukasz 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

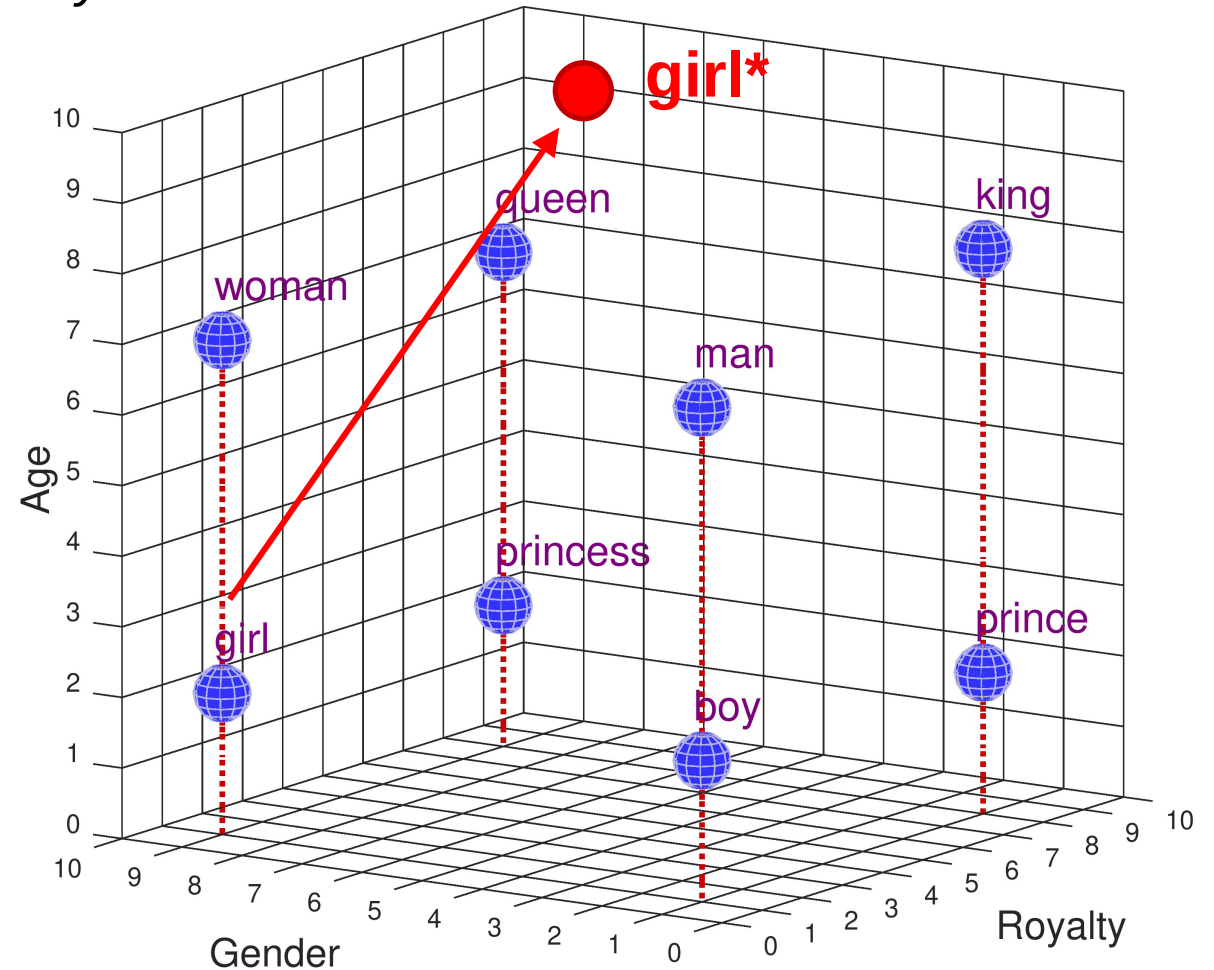


➤ Token embeddings updated through the self-attention mechanism

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



3D Semantic Feature Space



LLM have thousands of dimensions



INRAE

Hands-on workshop on genomic language models

12/06/2026 / Guillaume GAUTREAU / StatInfOmics team / MalAGE unit

➤ Next token prediction

an elderly girl wearing a crown



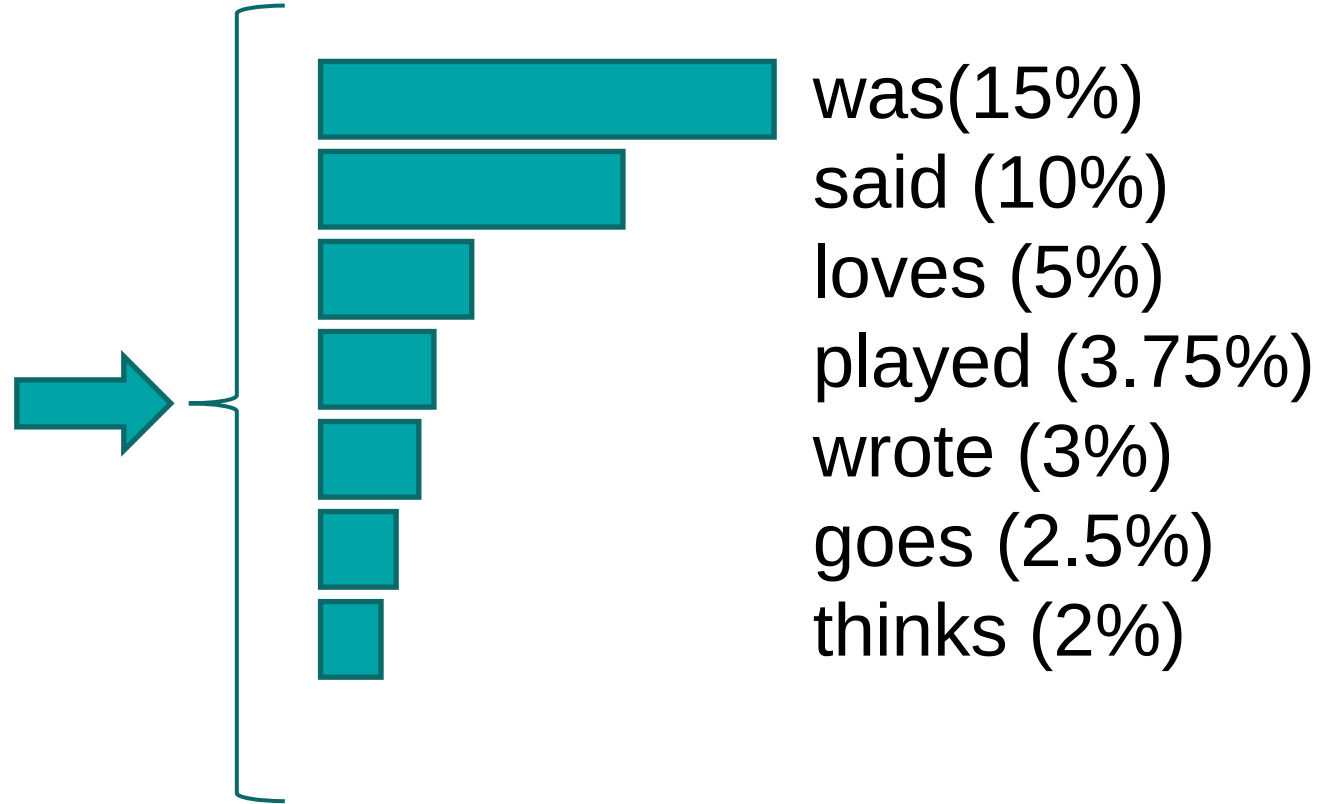
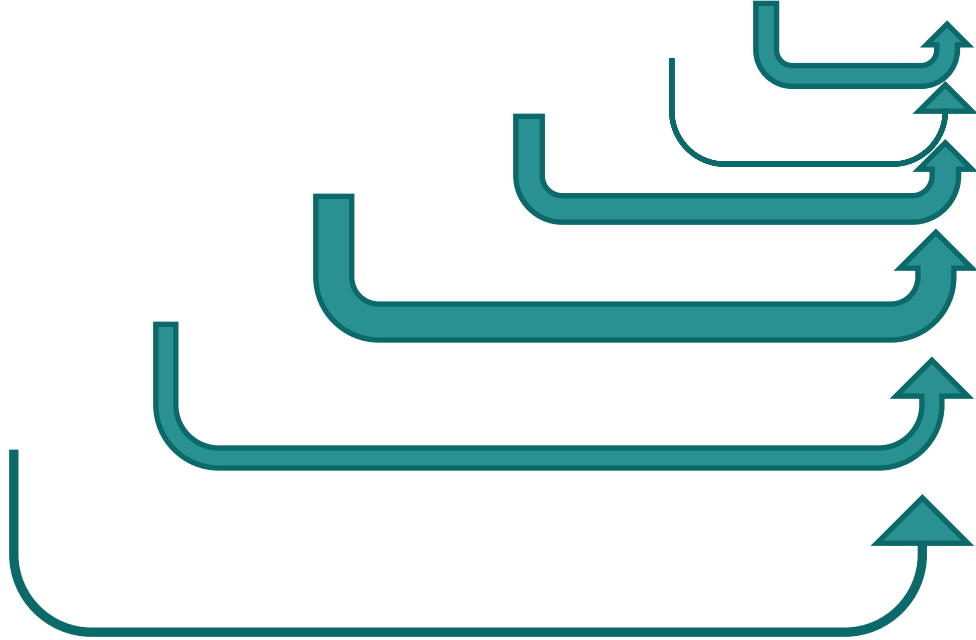
➤ Next token prediction

an elderly girl wearing a crown



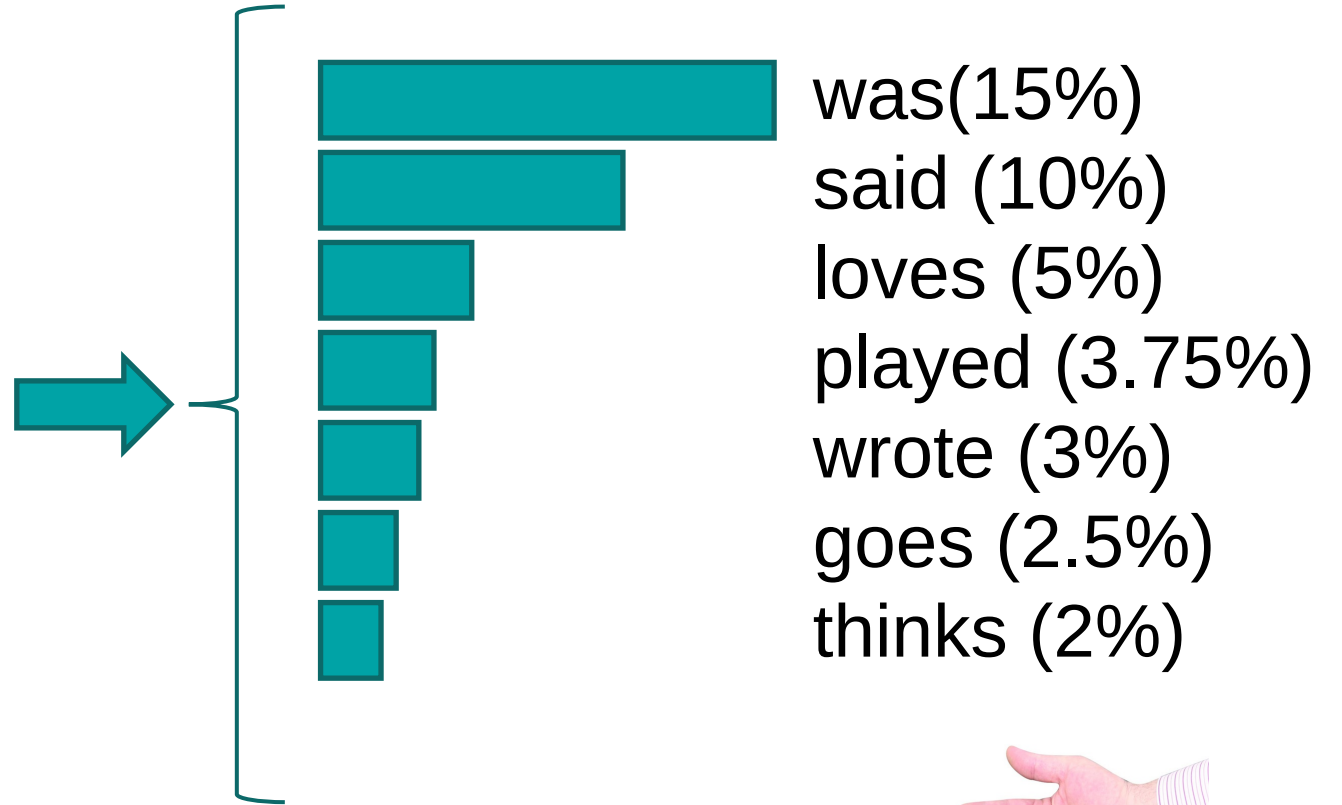
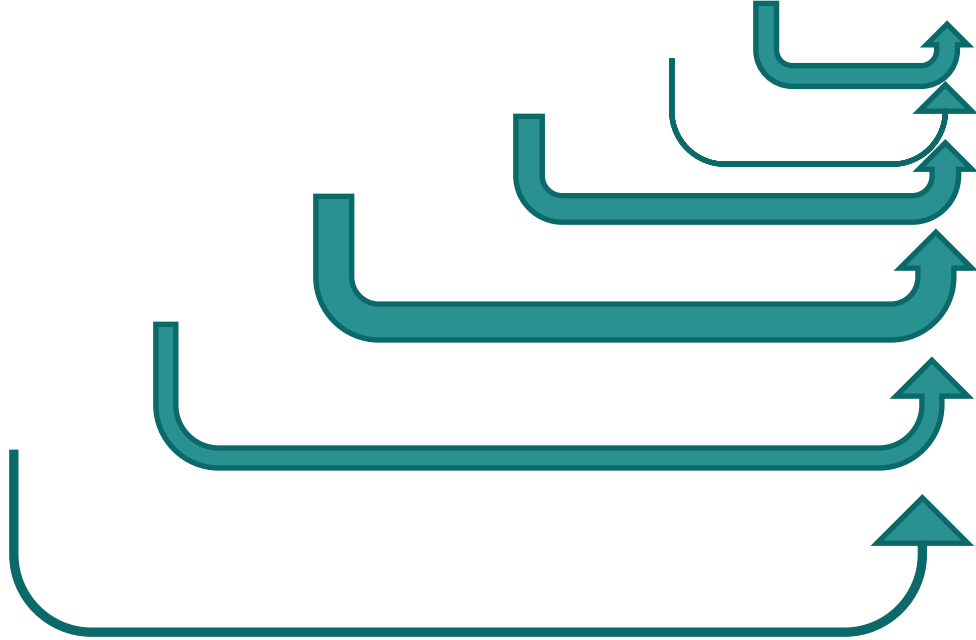
➤ Next token prediction

an elderly girl wearing a crown



➤ Next token prediction

an elderly girl wearing a crown

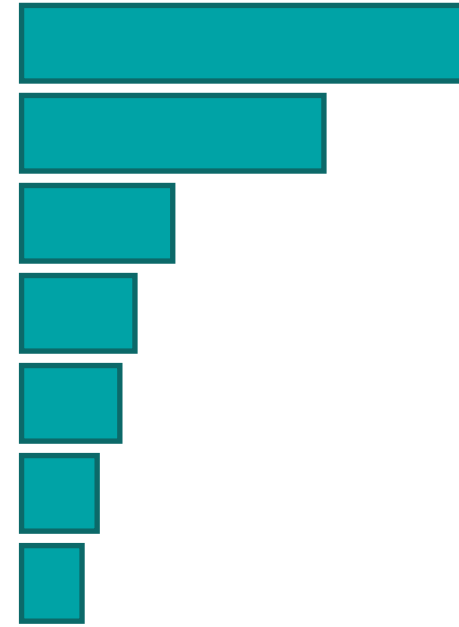
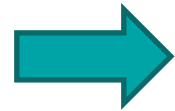


wrote



➤ 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



INRAE

Hands-on workshop on genomic language models

12/06/2026 / Guillaume GAUTREAU / StatInfOmics team / MalAGE unit

» « 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 , begged for rest . We found a tiny café , shared
local food , and laughed until the feeling of exhaustion melted into joy . Some moments
live forever in memory .



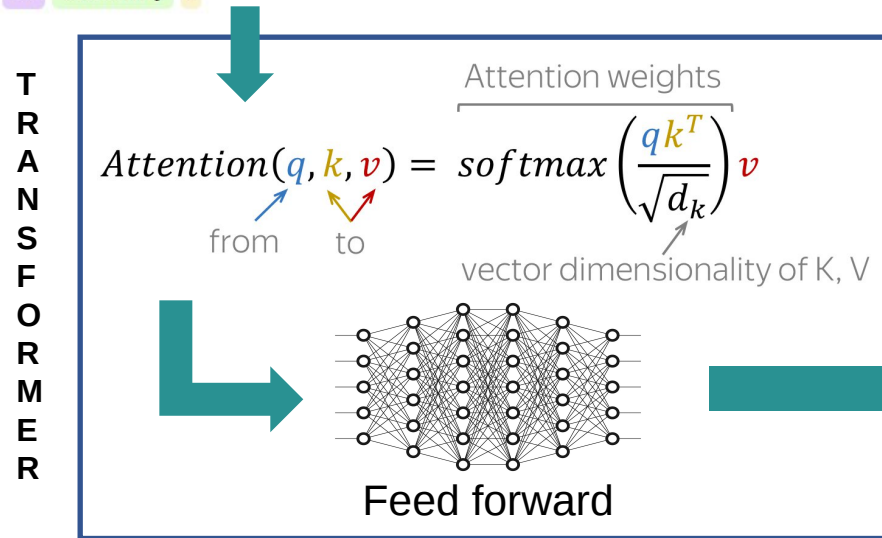
» « Attention Is All You Need »

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

token

Given a prompt:

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 »

Given a prompt:

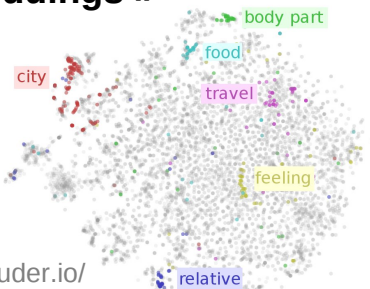
During my travel to the city of Barcelona, my favorite relative and I explored vibrant streets until our legs, the most tired body part, begged for rest. We found a tiny café, shared local food, and laughed until the feeling of exhaustion melted into joy. Some moments live forever in memory.

token

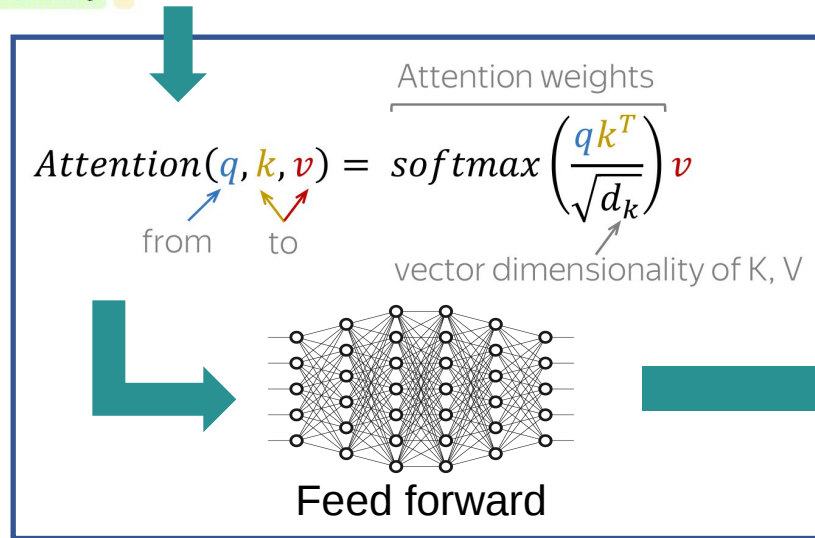
Billions of trained weights in the model

Tokens represented in a multidimensional space « embeddings »

Encoder-only



T
R
A
N
S
F
O
R
M
E
R



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

Encoder-Decoder

Lors de mon voyage dans la ville de Barcelone, mon parent préféré et moi avons parcouru des rues animées jusqu'à ce que nos jambes, la partie du corps la plus fatiguée, réclament du repos. Nous avons trouvé un petit café, partagé des spécialités locales et ri jusqu'à ce que le sentiment d'épuisement se transforme en joie. Certains instants restent à jamais gravés 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

1. English



"THE QUICK BROWN FOX
JUMPS OVER THE LAZY DOG"

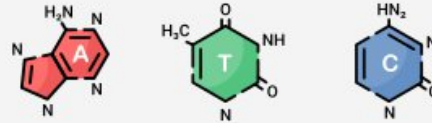
A
Aback
Abandon
...

gLM

2. DNA



AGGACTGGACCT

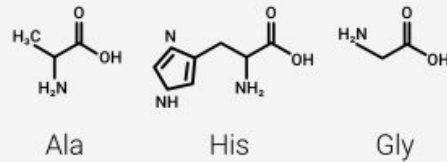


pLM

3. Protein



FYERATIMKHWE



gLM

4. Genome



dnaA
rpoB
recA
...

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

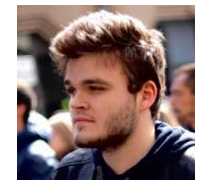
- DNABERT $\Rightarrow 2 \Rightarrow S$
- Nucleotide Transformer
- **Evo 1 \Rightarrow 1.5 \Rightarrow 2**

- ESM 1 \Rightarrow 2 \Rightarrow 3
- ProteinBERT
- ProtMamba
- ProtT5

- gLM
- Bacformer / panBART
- ANR project PanGAIMiX



Meriem Youssef



Gaspar Roy

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

LLM

1. English



"THE QUICK BROWN FOX
JUMPS OVER THE LAZY DOG"

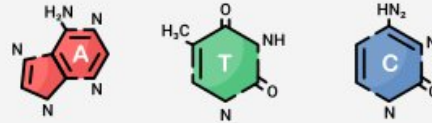
A
Aback
Abandon
...

gLM

2. DNA



AGGACTGGACCT

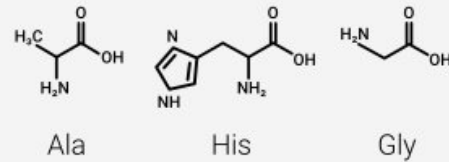


pLM

3. Protein

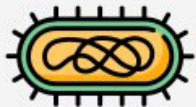


FYERATIMKHWE



gLM

4. Genome



dnaA
rpoB
recA
...

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

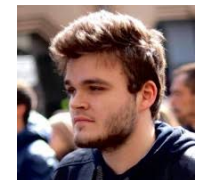
- DNABERT $\Rightarrow 2 \Rightarrow S$
- Nucleotide Transformer
- **Evo 1 \Rightarrow 1.5 \Rightarrow 2**

- ESM 1 \Rightarrow 2 \Rightarrow 3
- ProteinBERT
- ProtMamba
- ProtT5

- gLM
- Bacformer / panBART
- ANR project PanGAIMiX



Meriem Youssef



Gaspar Roy

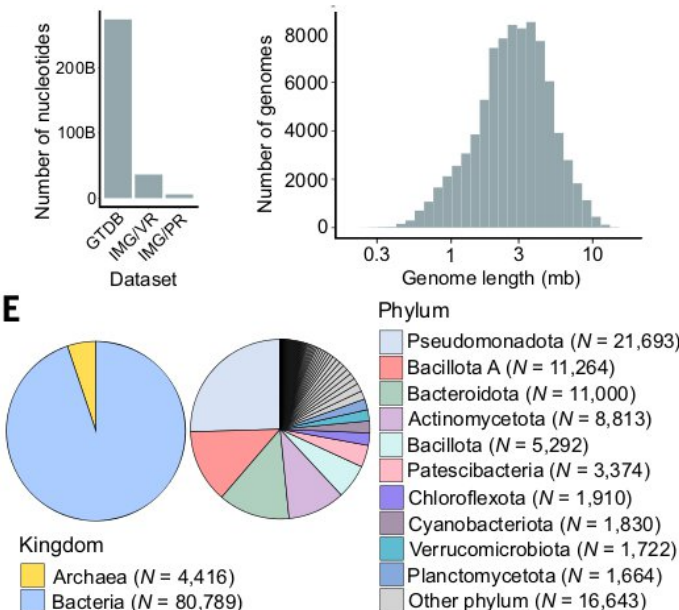
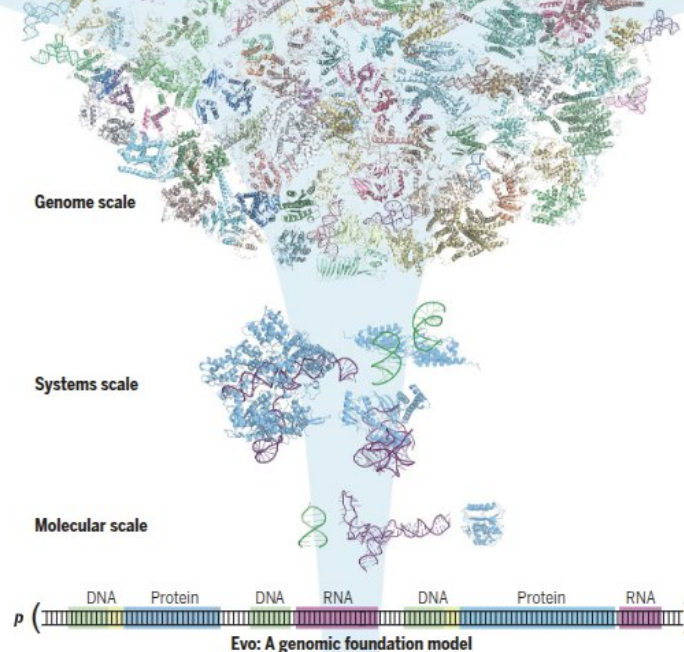
RESEARCH ARTICLE

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. Bartie¹, Armin W. Thomas⁵, Samuel H. King^{1,2}, Garyk Brix^{1,6}, Jeremy Sullivan¹, Madelena Y. Ng⁷, Ashley Lewis⁸, Aaron Lou³, Stefano Ermon^{3,9}, Stephen A. Baccus¹⁰, Tina Hernandez-Boussard⁸, Christopher Ré³, 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

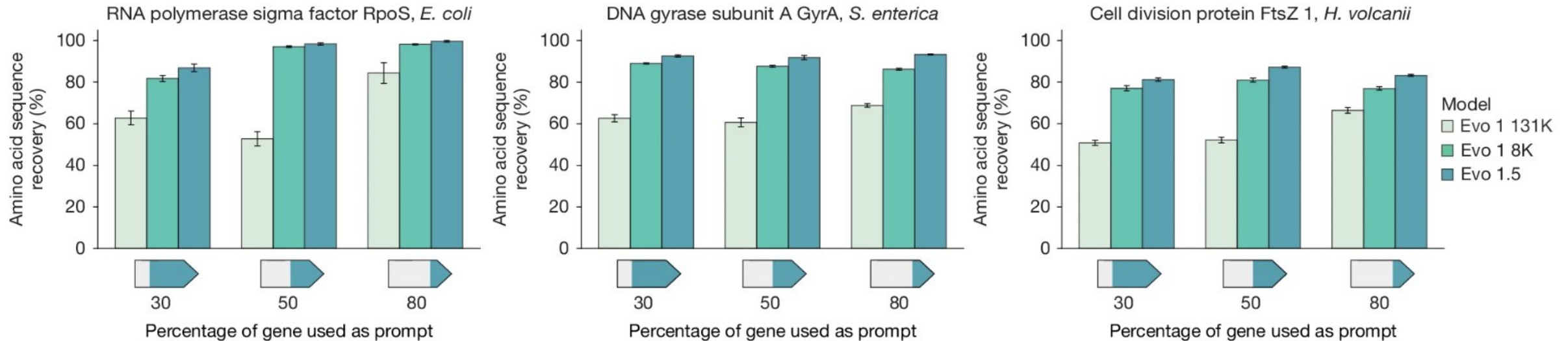
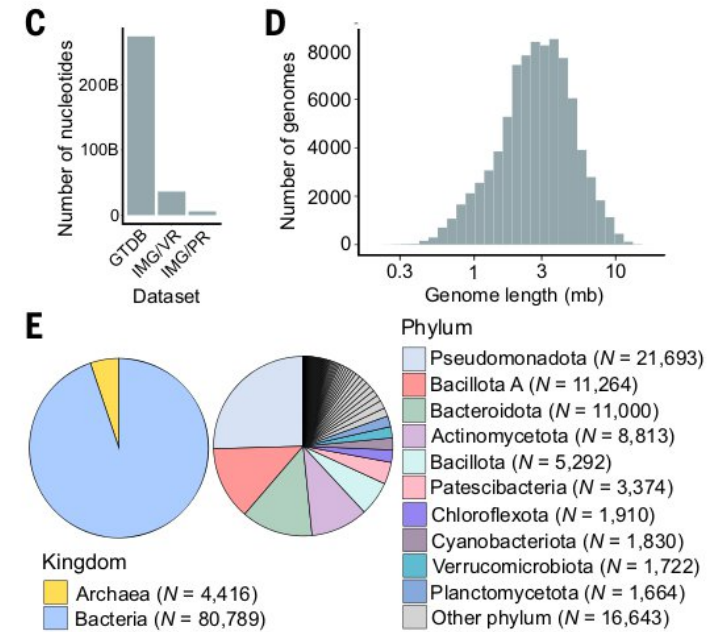
Published online: 19 November 2025

Open access

Check for updates

Aditi T. Merchant^{1,2}, Samuel H. King^{1,2}, Eric Nguyen^{1,2} & Brian L. Hie^{2,4,5}

Generative genomic models can design increasingly complex biological systems¹. 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 genes², 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.

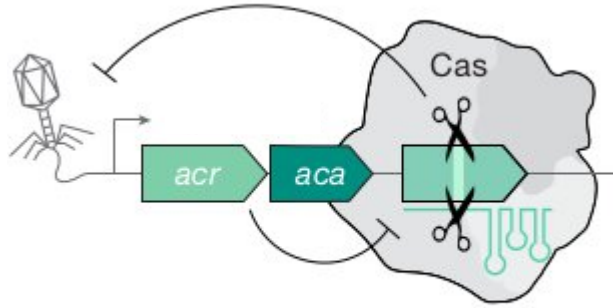


INRAE

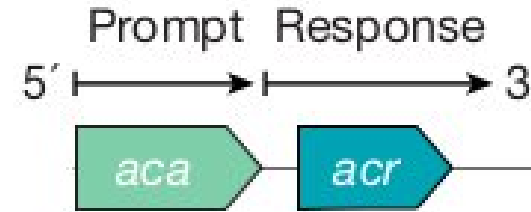
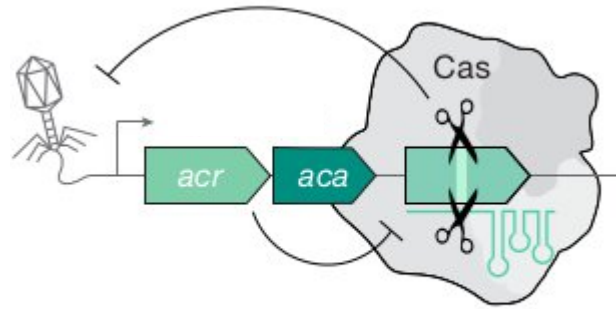
Hands-on workshop on genomic language models

12/06/2026 / Guillaume GAUTREAU / StatInfOmics team / MalAGE unit

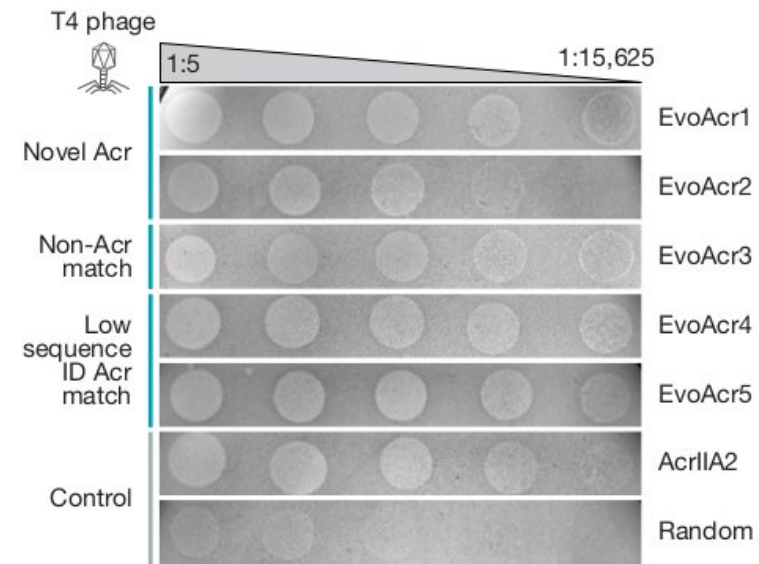
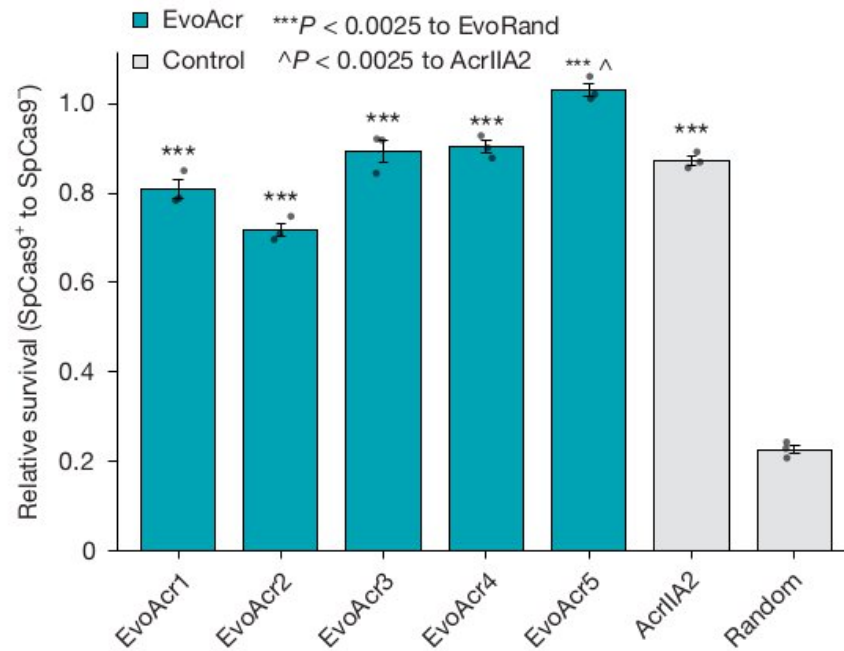
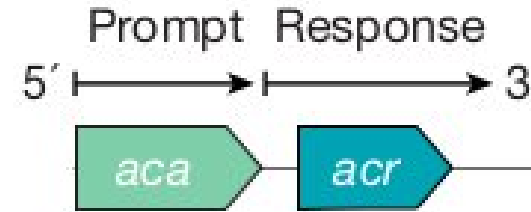
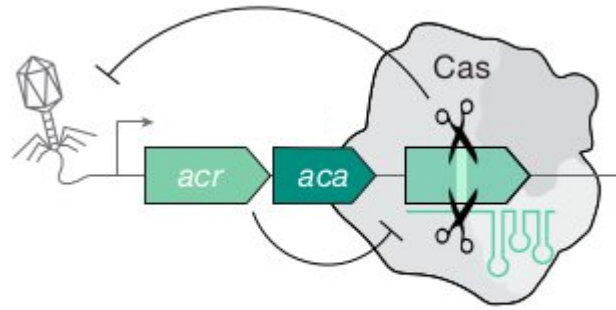
➤ Evo generates novel functional anti-CRISPR proteins



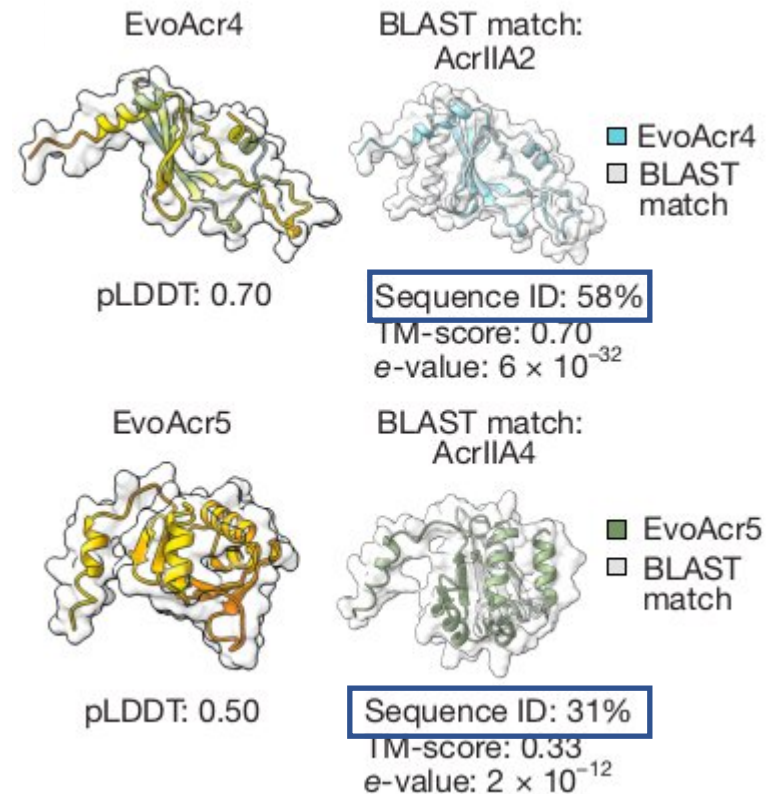
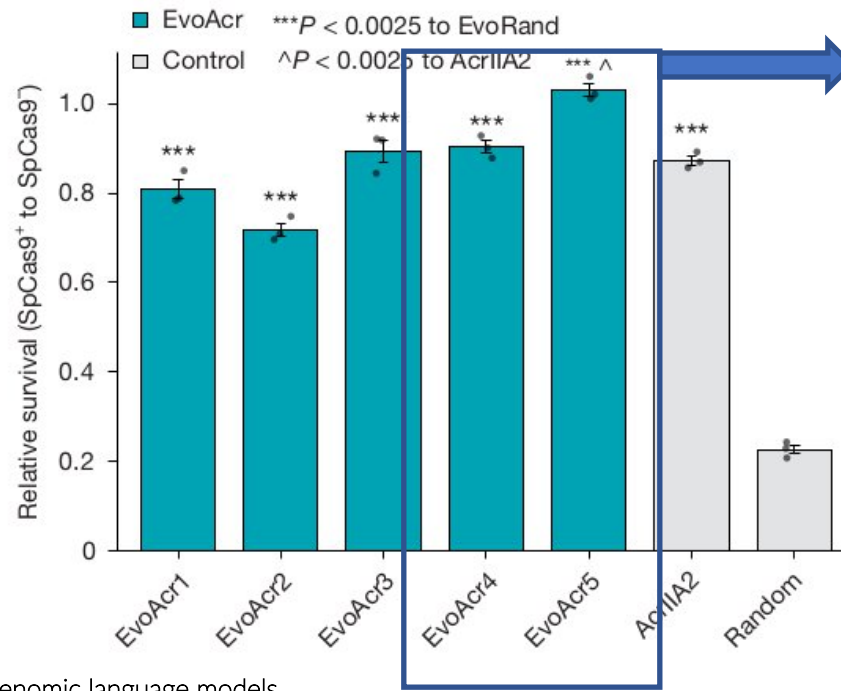
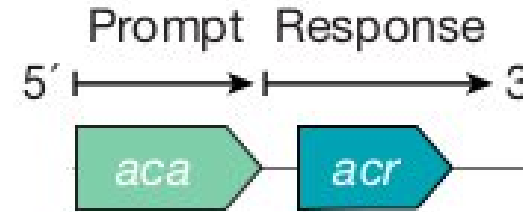
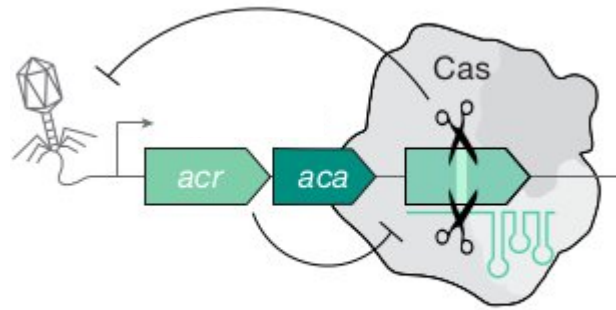
➤ Evo generates novel functional anti-CRISPR proteins



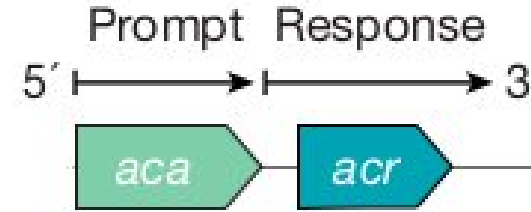
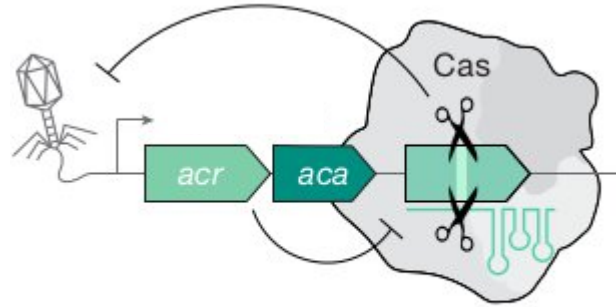
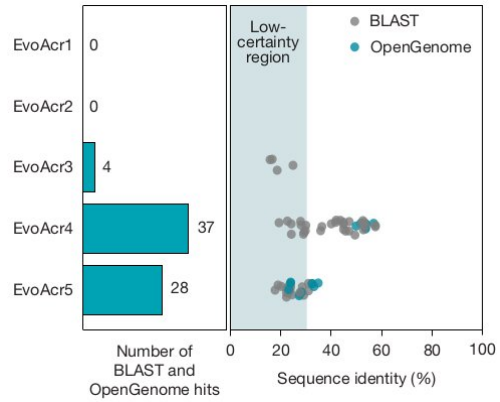
➤ Evo generates novel functional anti-CRISPR proteins



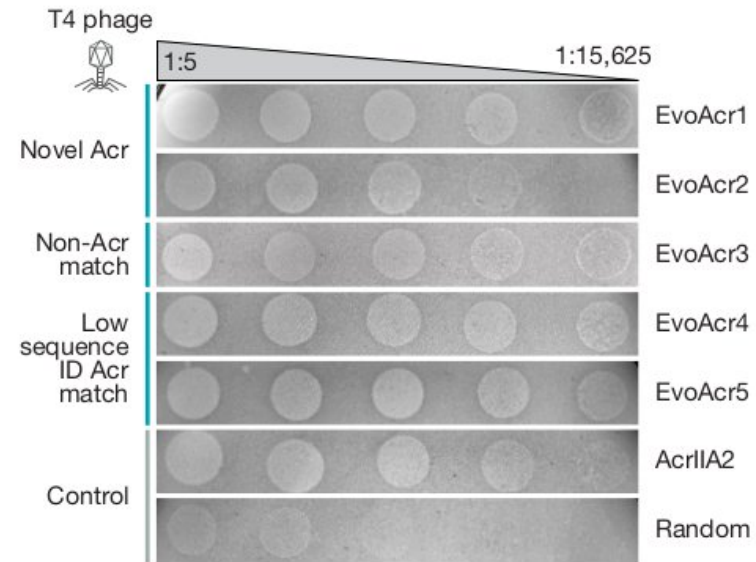
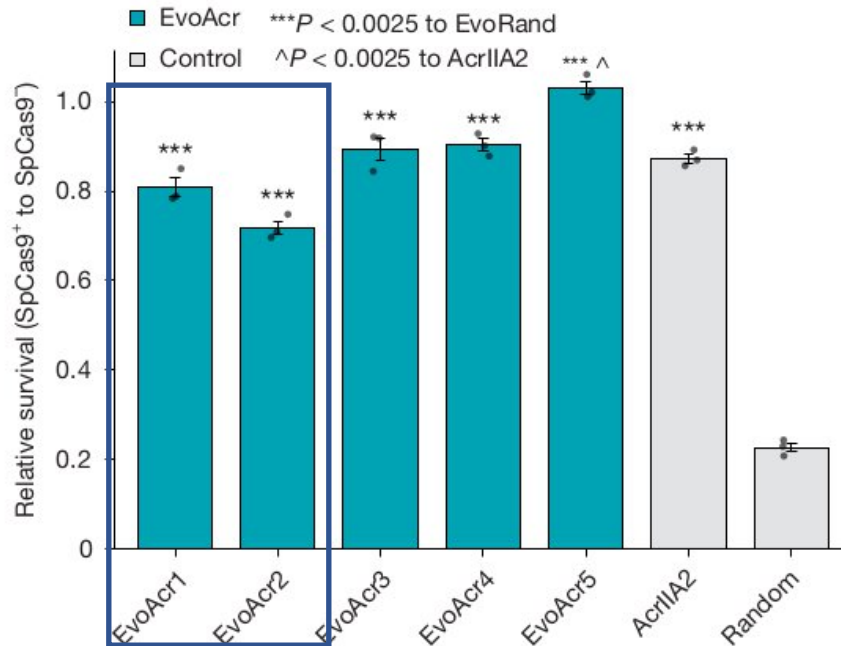
➤ Evo generates novel functional anti-CRISPR proteins



➤ Evo generates novel functional anti-CRISPR proteins

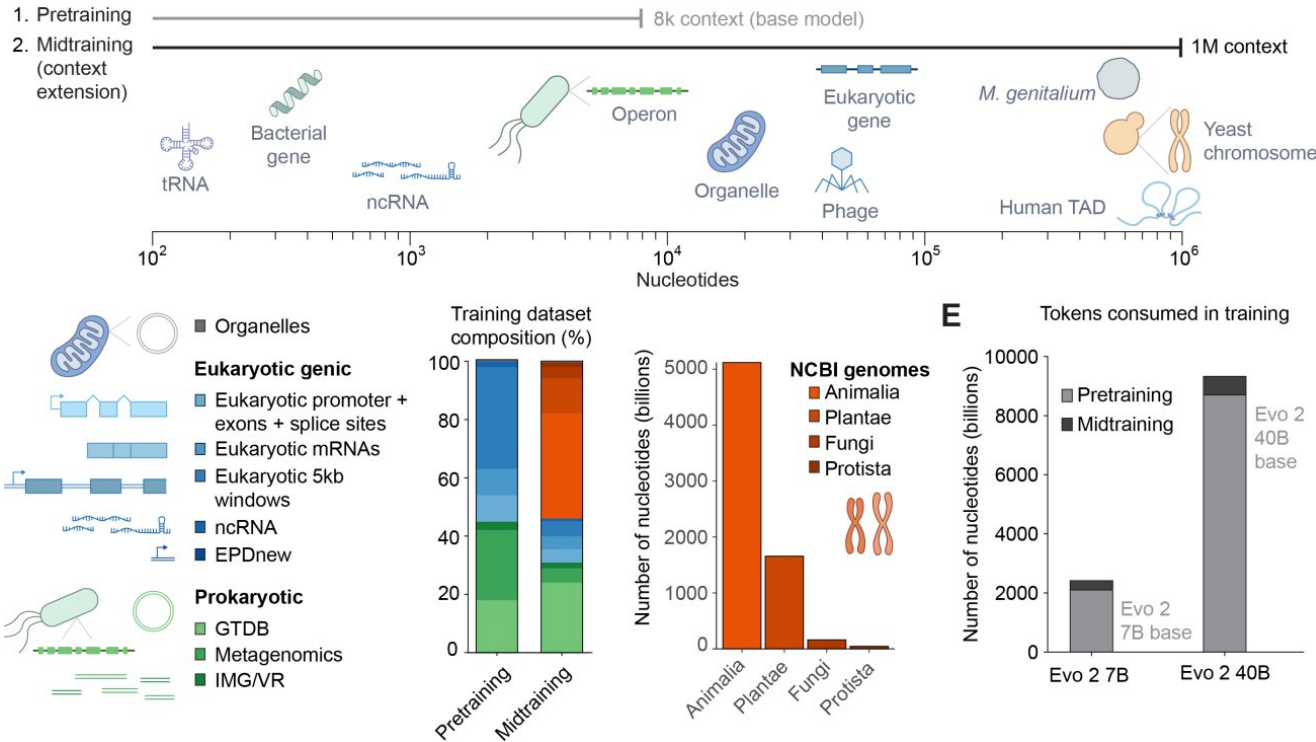


- No homology with any known gene!
- Original creation from AI!



➤ 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

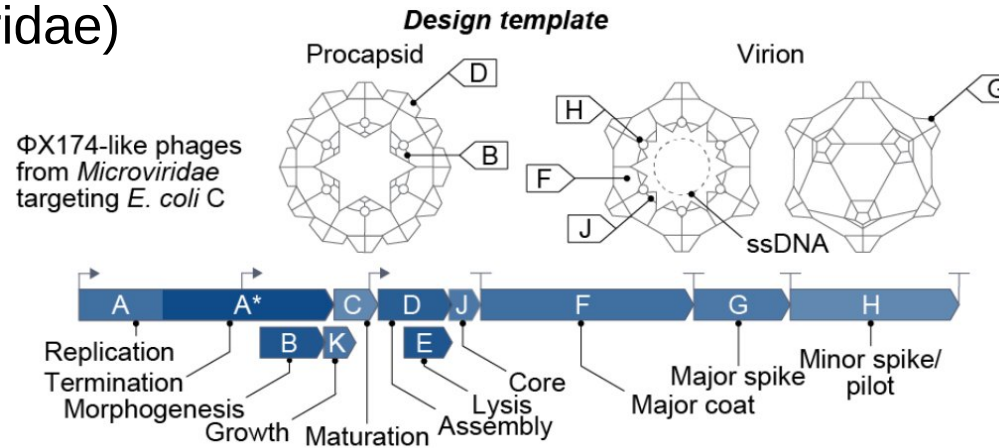


	Evo1	Evo 1.5	Evo 2
Publication	November 2024	November 2025	pre-print
Model size	7B	7B	1B, 7B, 40B
Context size (# of tokens/bases)	8 Kb or 131 Kb	8 Kb	1 Mb
Architecture	Striped Hyena	Striped Hyena	Striped Hyena 2
Training dataset	Prokaryotic genomes / plasmids / phages	Updated Evo1 database	Evo1 + All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

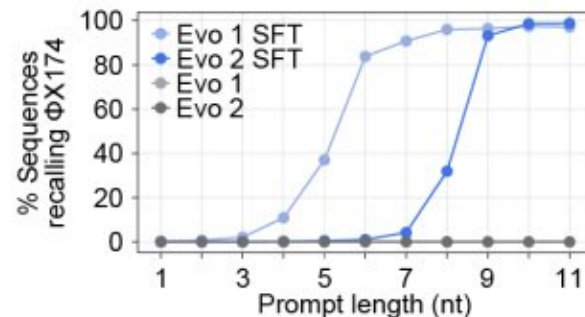
➤ The claim: first working gLM-generated genome

One of the easiest case scenario but still :

- Phage Φ X174 (Microviridae)
- Target *E. coli* cells
- Abundant in databases
- Very small: ~5Kb



- 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

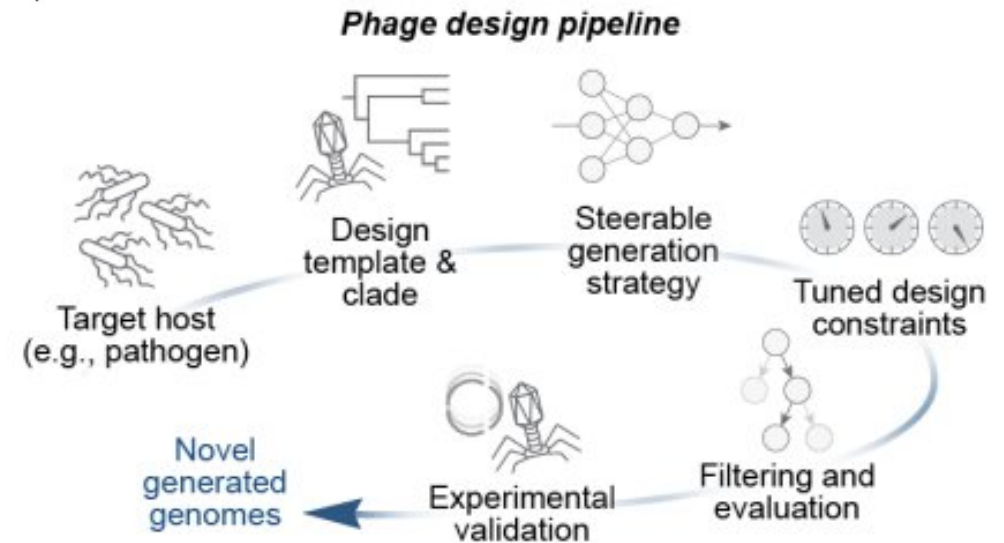


pre-print on BioRxiv (09/2025) :

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 Brix^{1,5}, Max E. Wilkinson⁶, and Brian L. Hie^{1,3,7,*}

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



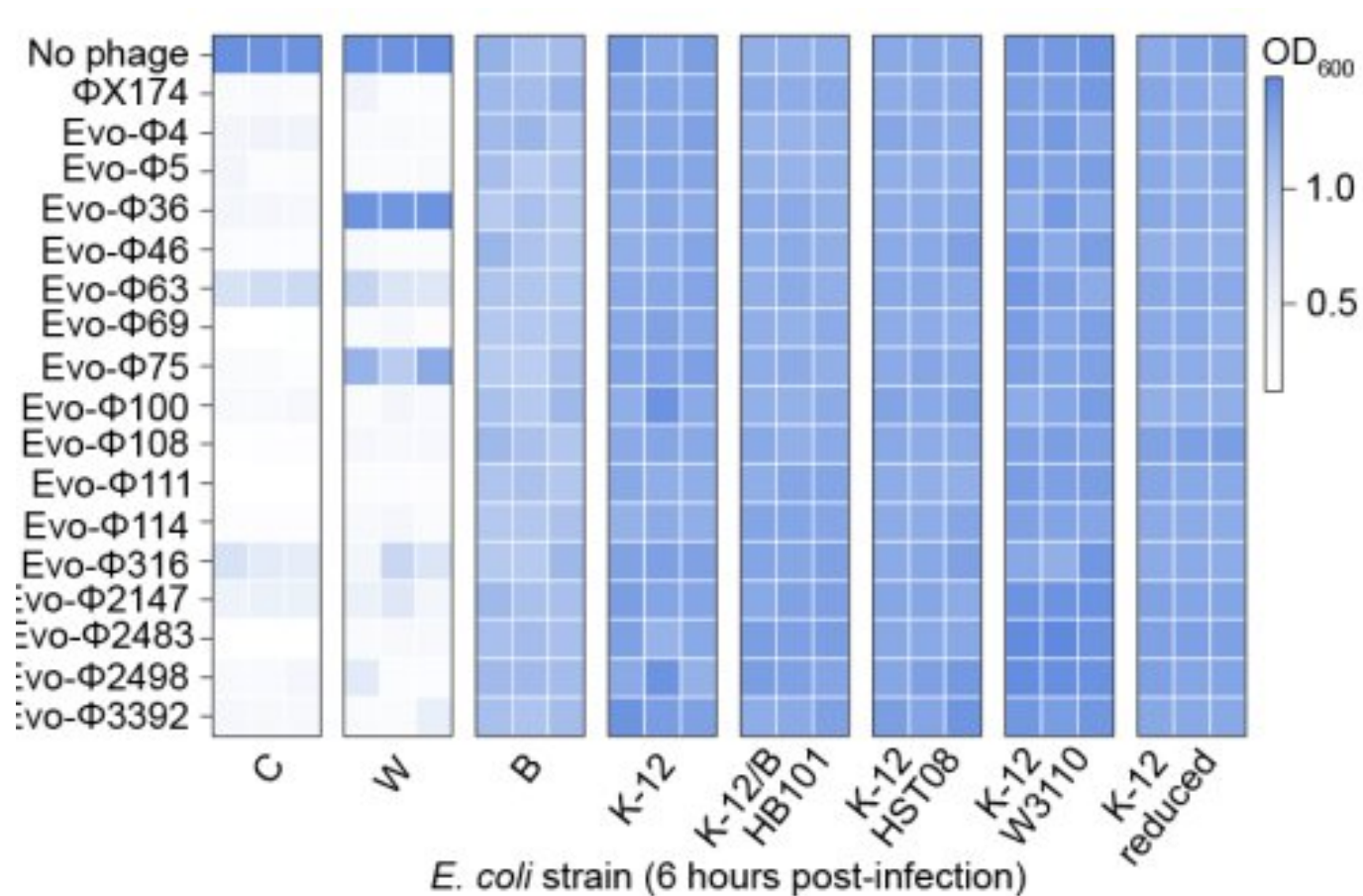
INRAE

Hands-on workshop on genomi

12/06/2026 / Guillaume GAUTREAU / StatInfOmics team / MalAGE unit

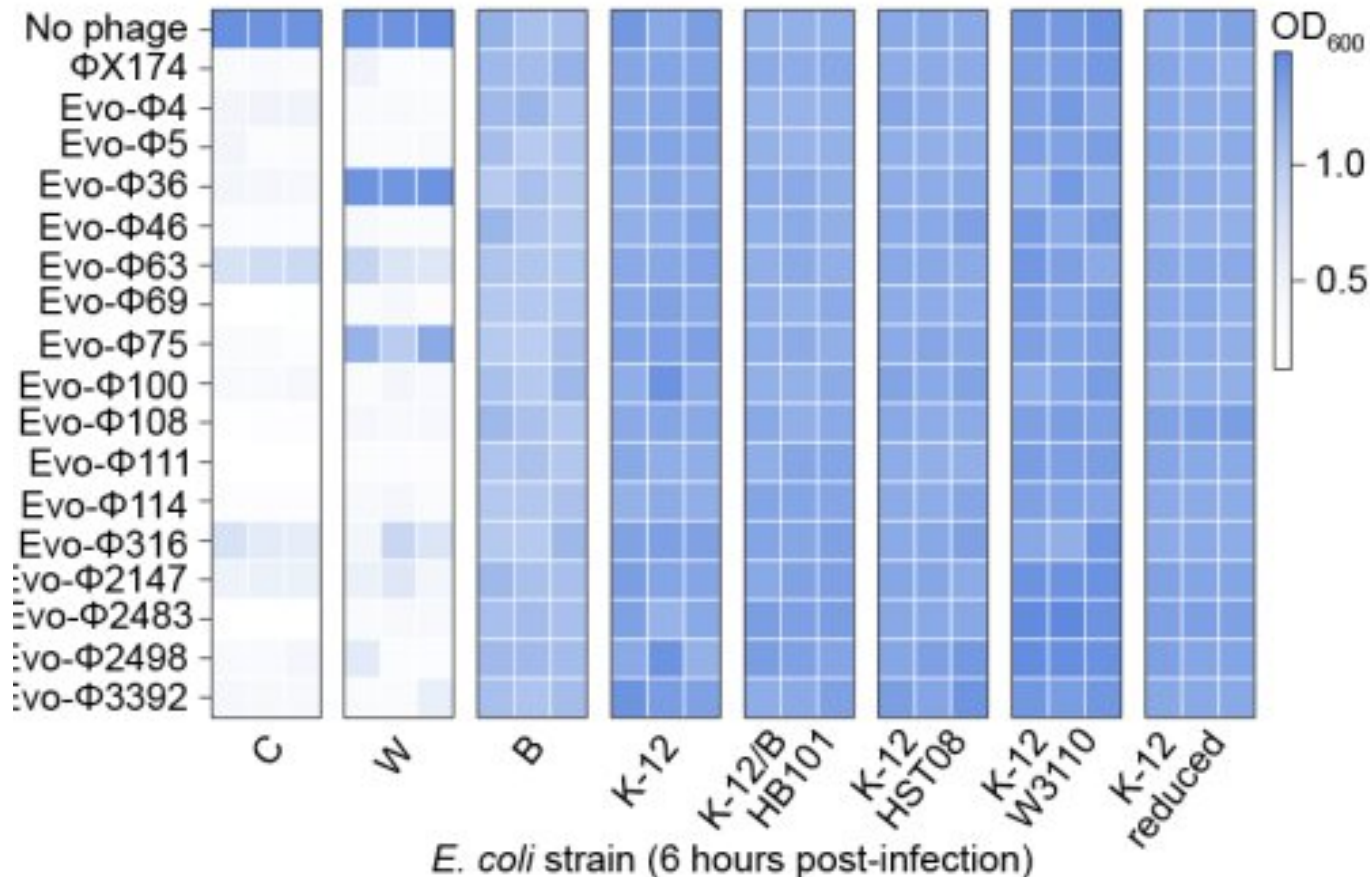
➤ The claim: first working AI-generated genome

Evo-generated phages actually lyse *E. coli* genomes!

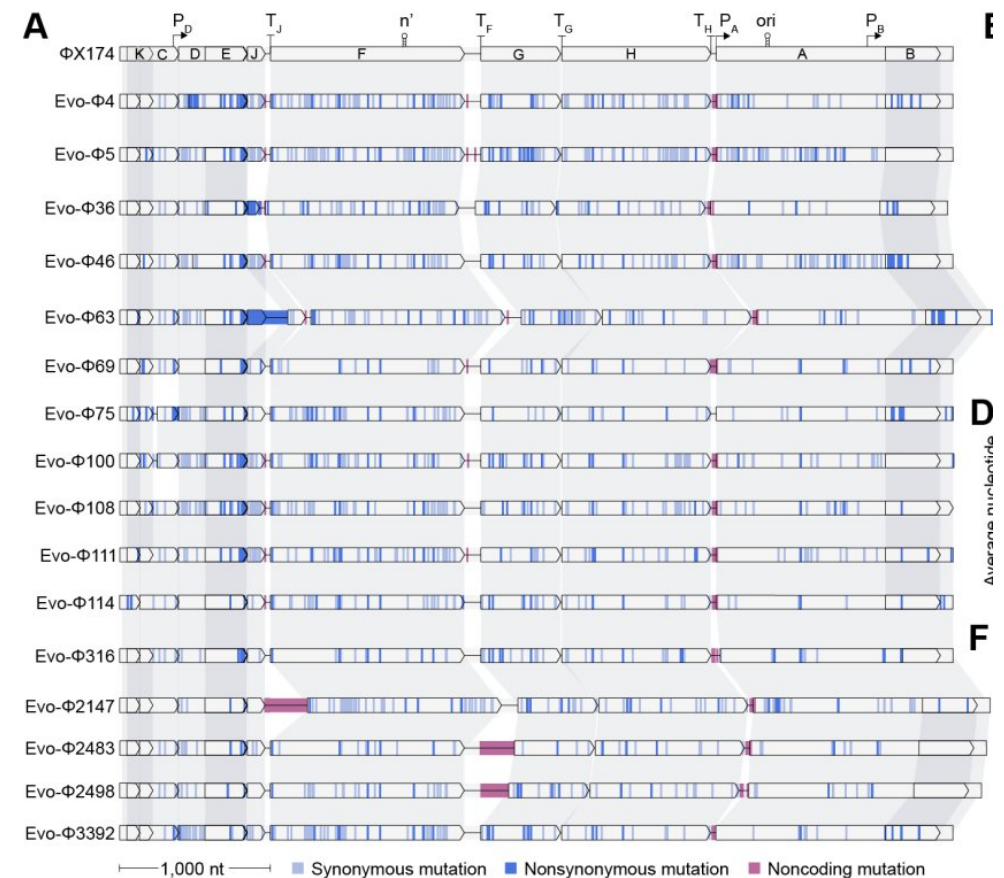


➤ The claim: first working AI-generated genome

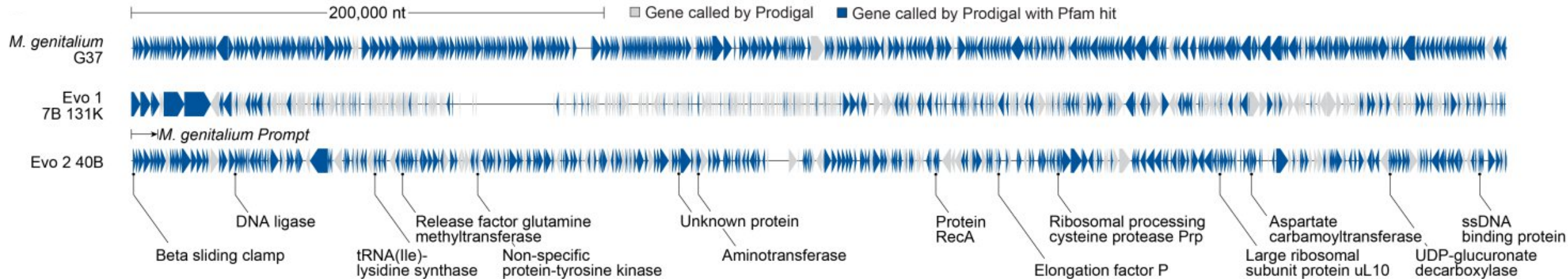
Evo-generated phages actually lyse *E. coli* genomes!



...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
- but are there actually working genomes?

- Any other idea?

➤ 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)
```

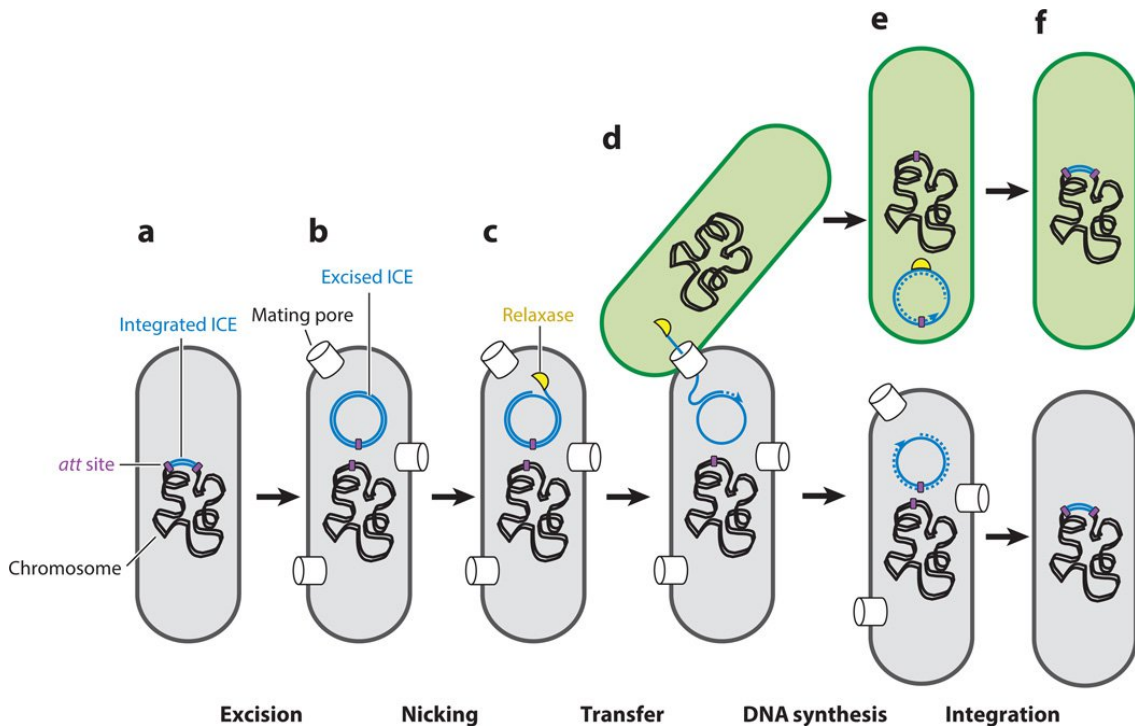
https://github.com/ArcInstitute/evo2/blob/main/notebooks/generation/generation_notebook.ipynb

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
Lactocaseibacillus 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

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

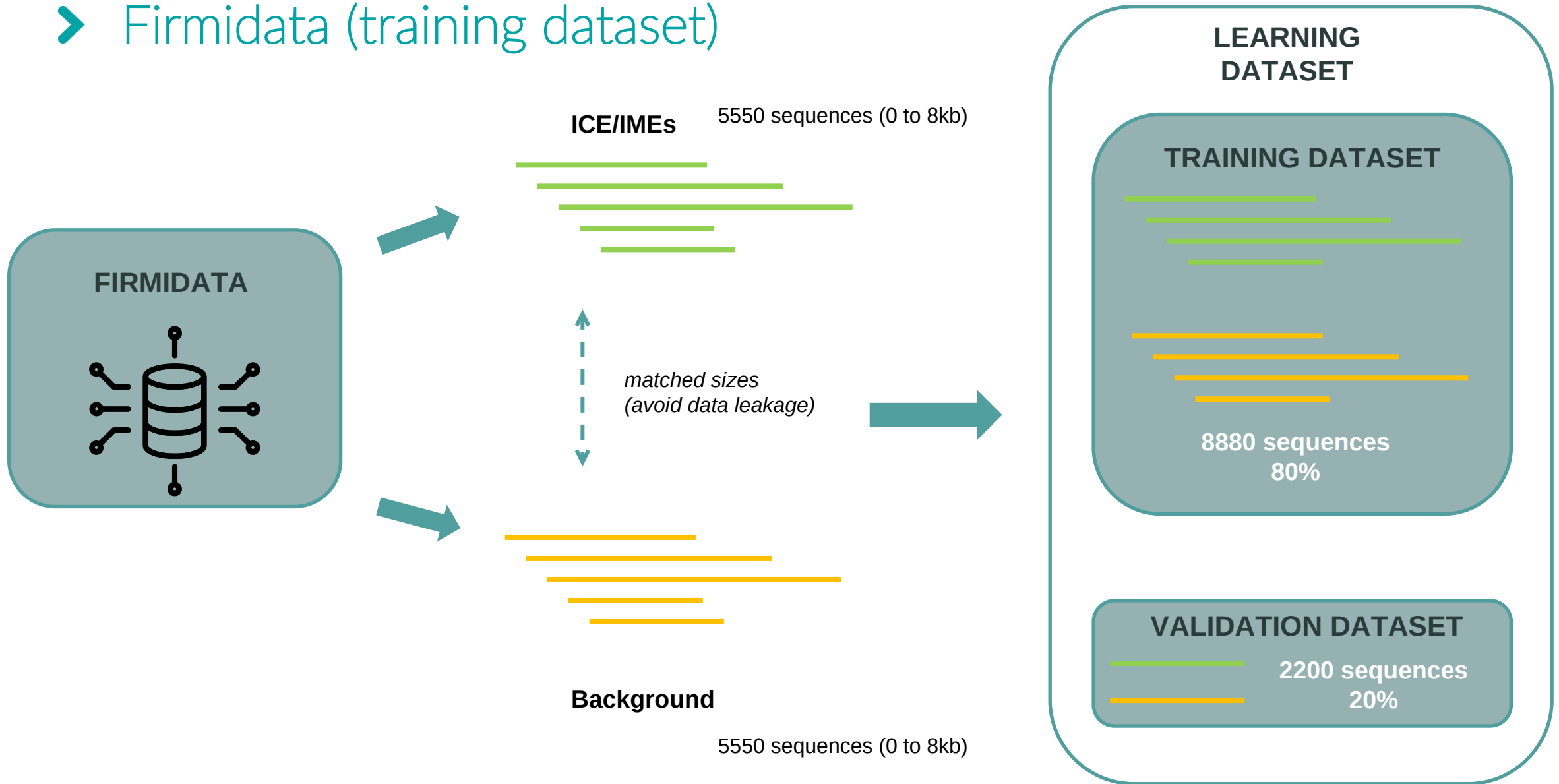


G rard Gu don¹, Julie Lao^{1,2}, Sophie Payot¹, Thomas Lacroix², H l ne Chiapello² and Nathalie Leblond-Bourget^{1*}

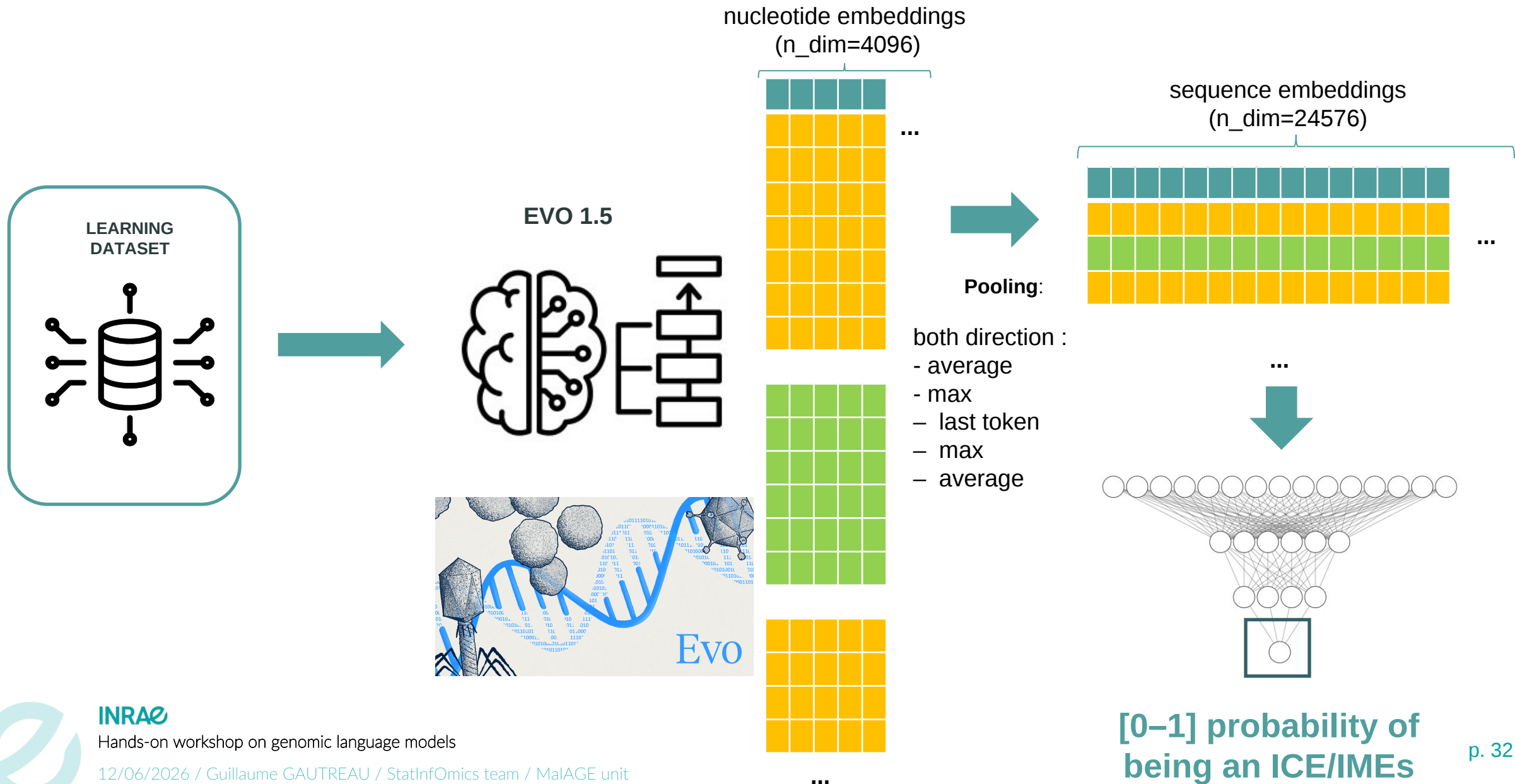
Streptococcus equi subsp. zooepidemicus MGCS10565 NC_011134.1
Streptococcus equi subsp. zooepidemicus ATCC35246 NC_017582.1
Streptococcus gallolyticus ATCCBAA-2069 NC_015215.1
Streptococcus parasanguinis ATCC15912 NC_015678.1
Streptococcus pneumoniae SPN034183 NC_021028.1
Streptococcus pneumoniae P1031 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



➤ 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]]
```

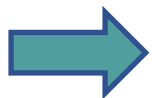
➤ Firmidata transfer learning using Evo

```
              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      2220
 weighted avg           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



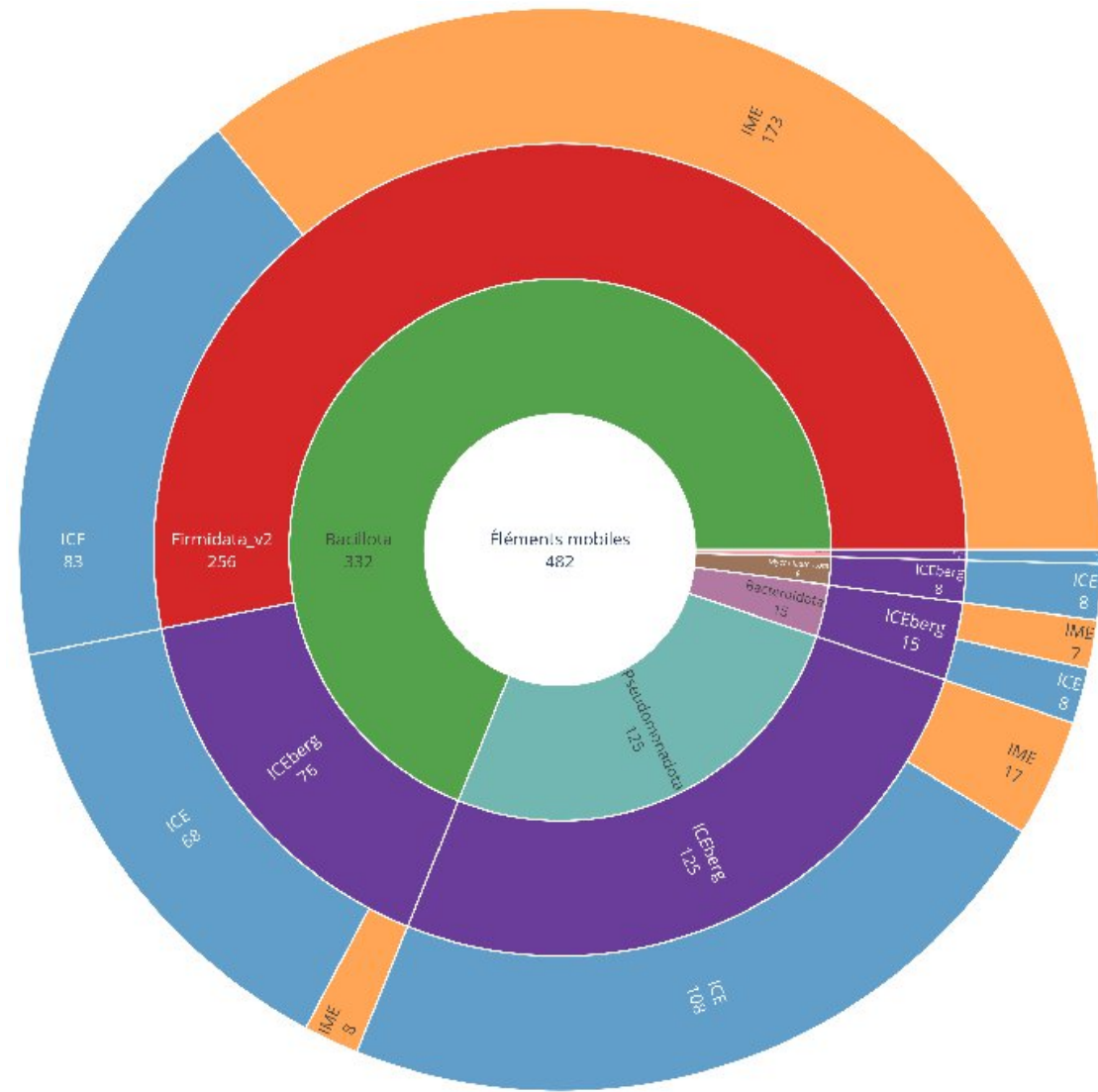
Aliou Diallo, M2 student, StatInfOmiscs team

Firmidata + ICEberg experimental: distribution of ICEs and IMEs

- Phylum**
- Bacillota
 - Pseudomonadota
 - Bacteroidota
 - Mycoplasmata
 - Campylobacterota

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

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



INRAE

Hands-on workshop on genomic language models

12/06/2026 / Guillaume GAUTREAU / StatInfOmiscs team / MalAGE unit

➤ Ongoing transfer learning experiments using a broader training dataset



Aliou Diallo,
M2 student,
StatInfOmics team

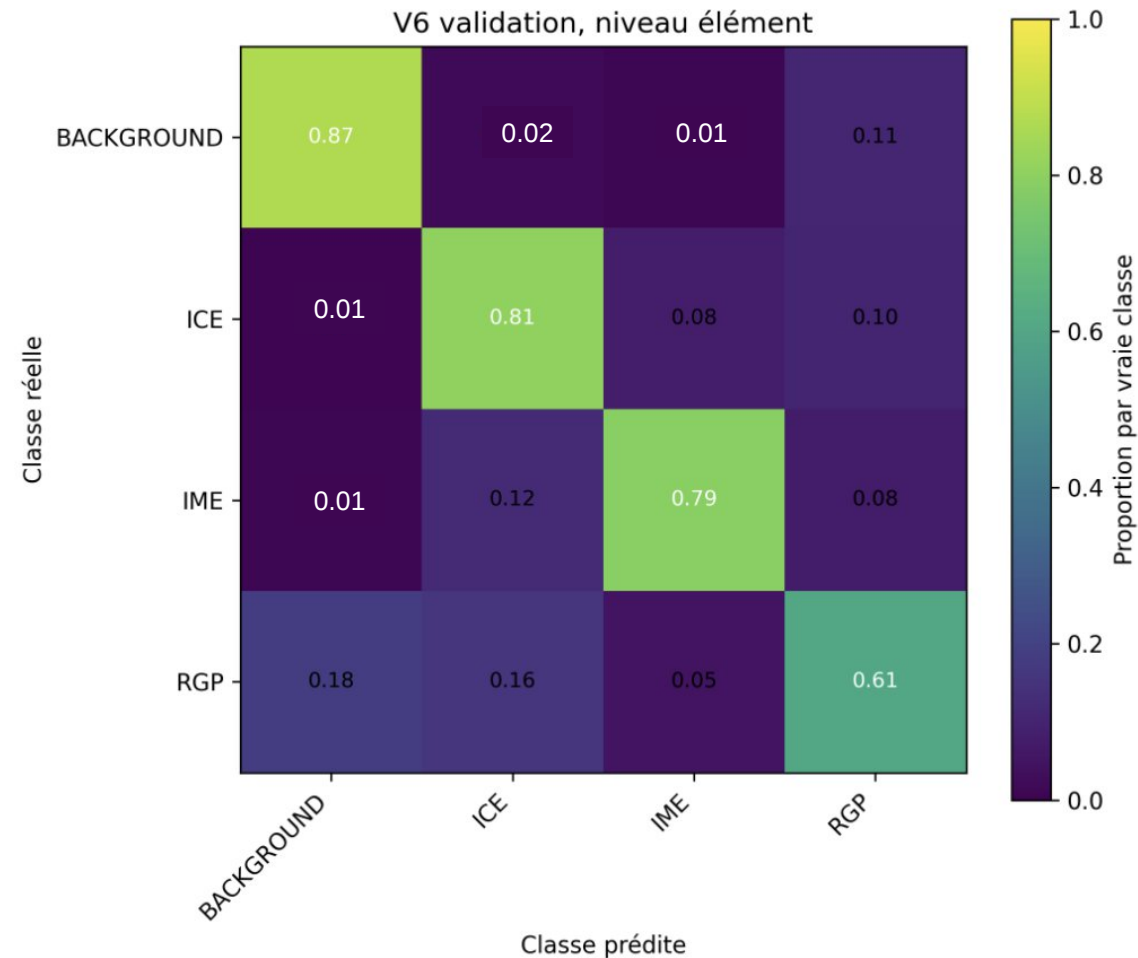
Firmidata + ICEberg experimental: distribution of ICEs and IMEs

- 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)

external validation



JOBIM 2026 Poster



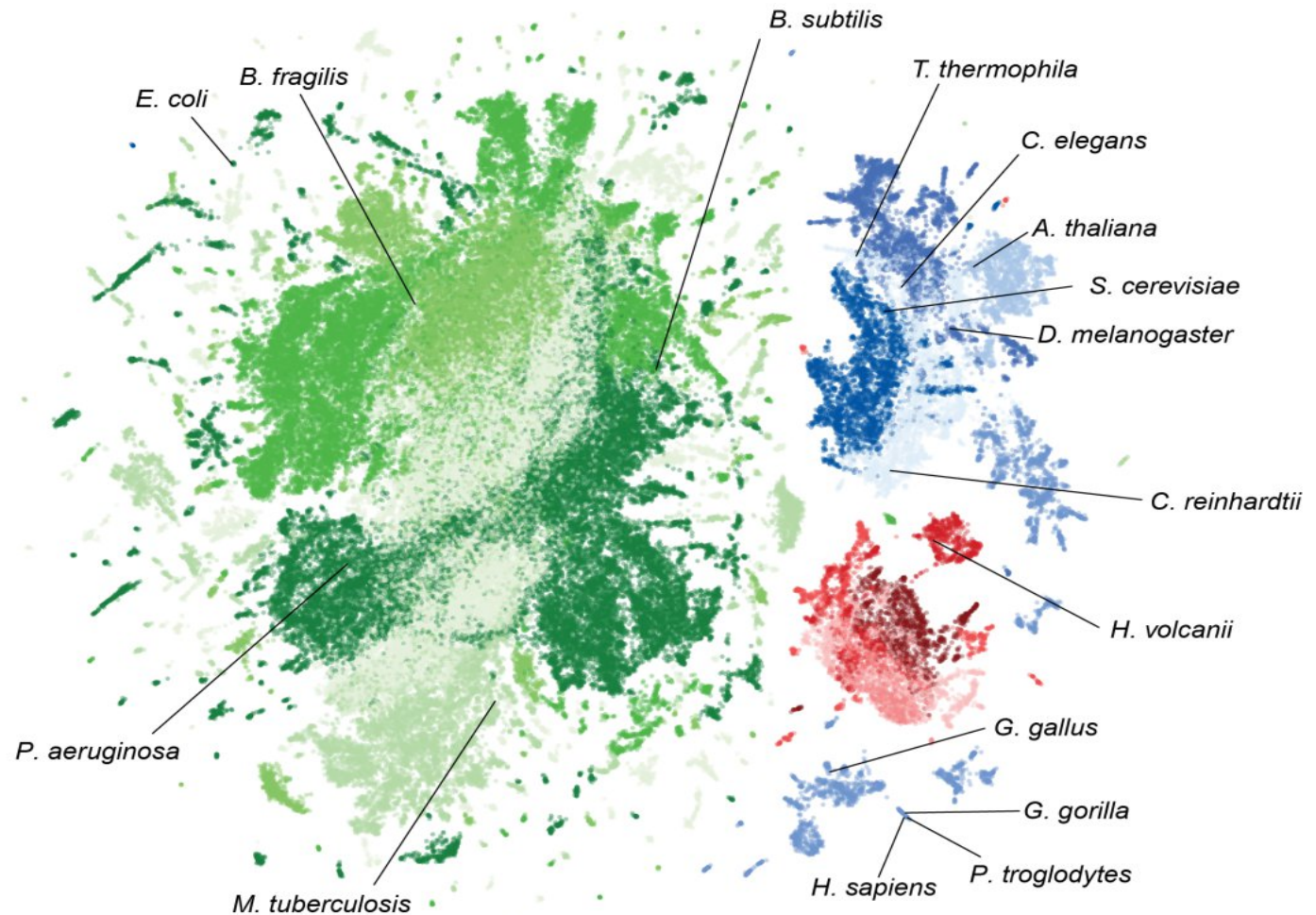
INRAE

➤ **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)



Bacterial phylum

- Pseudomonadota
- Bacillota
- Bacteroidota
- Actinomycetota
- Other/Unknown Bacteria

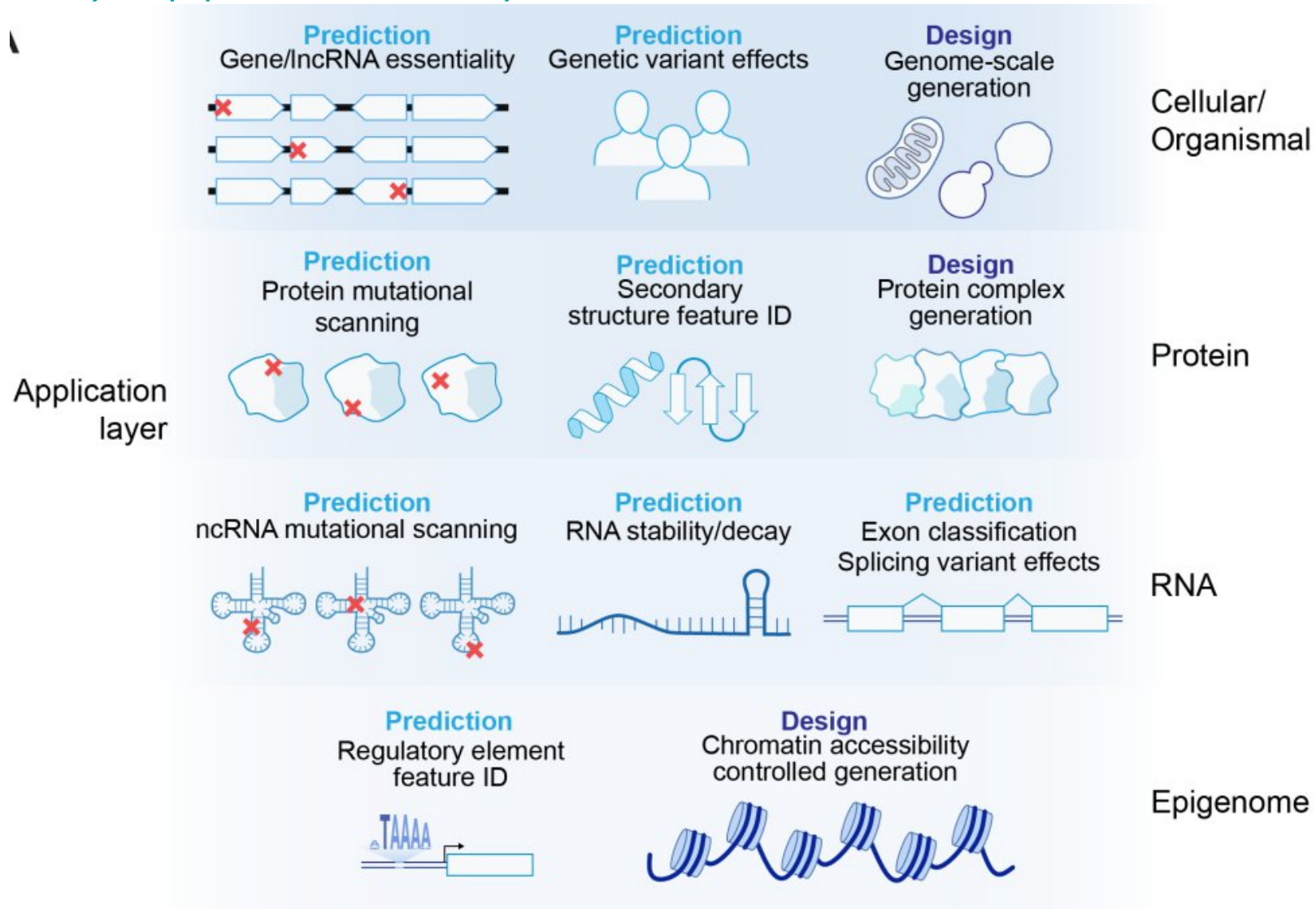
Eukaryotic phylum

- Ascomycota
- Arthropoda
- Chordata
- Streptophyta
- Other/Unknown Eukaryota

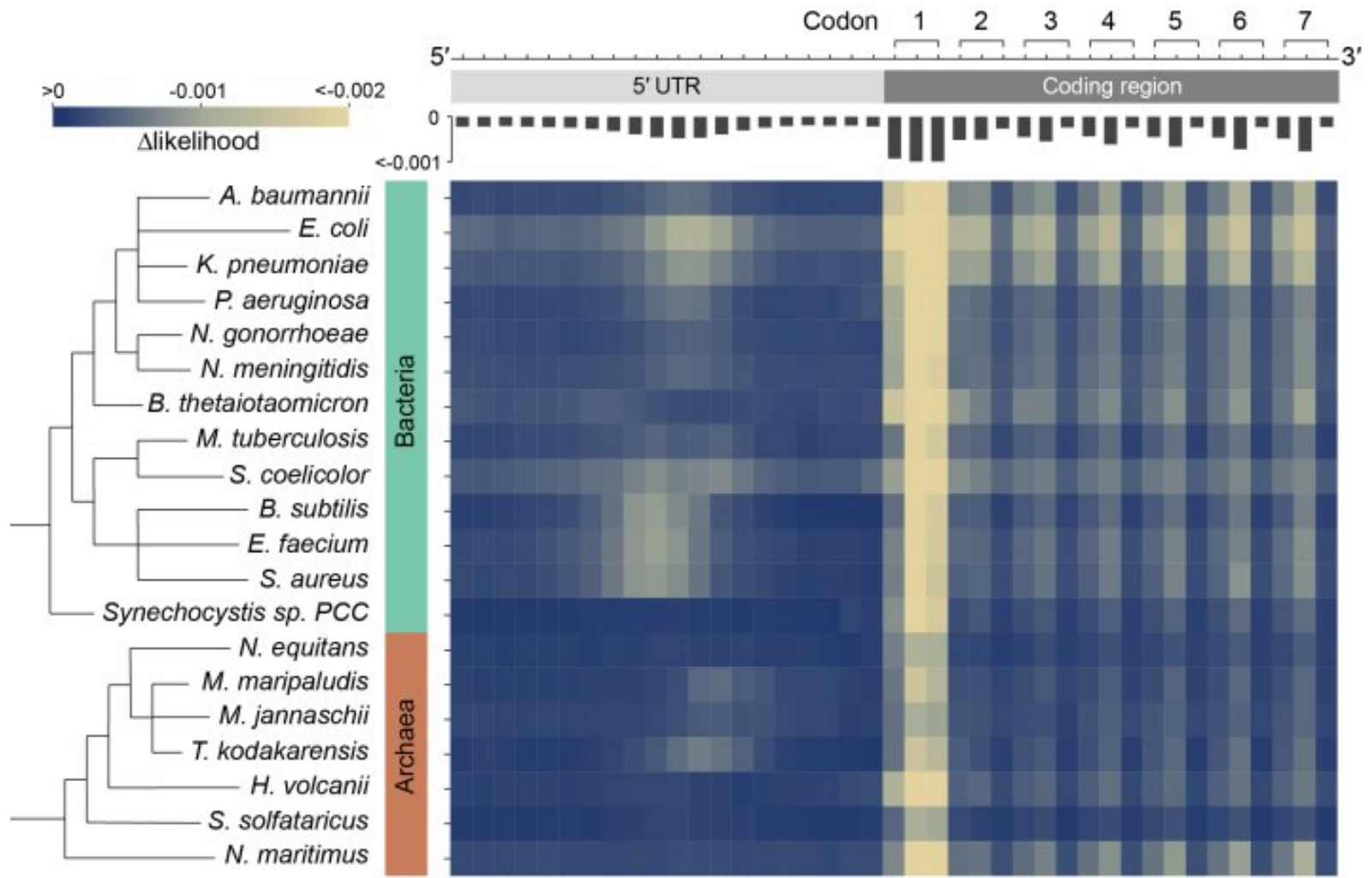
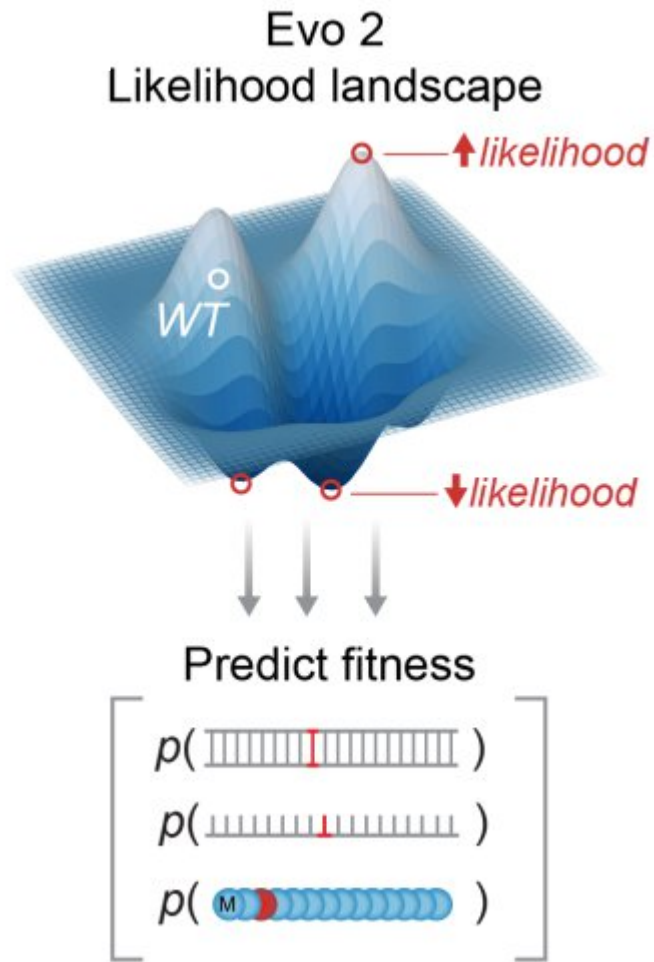
Archaeal phylum

- Thermoproteota
- Halobacteriota
- Thermoplasmata
- Nanoarchaeota
- Other/Unknown Archaea

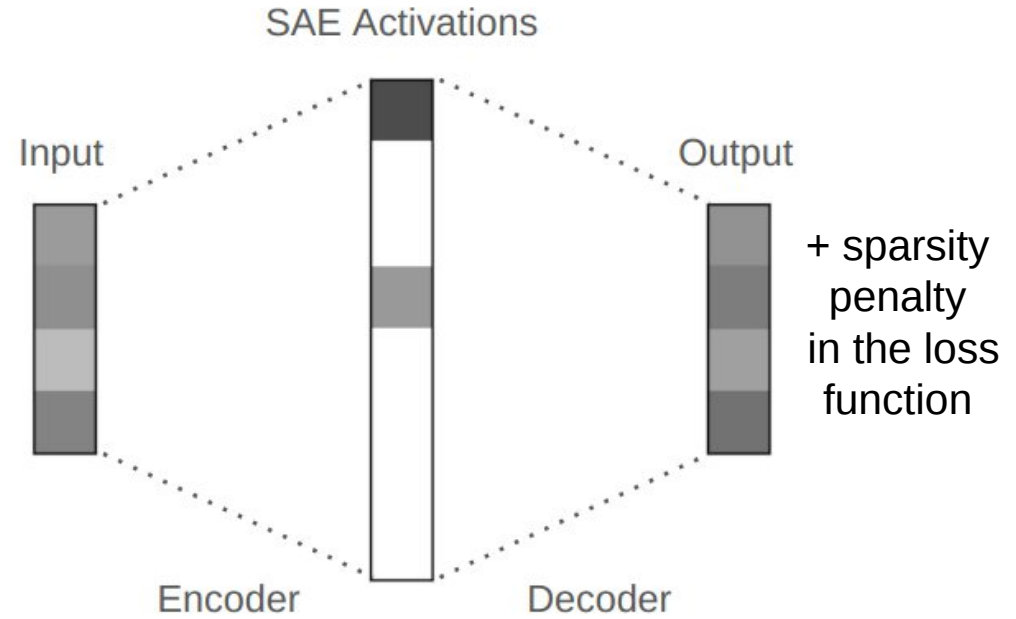
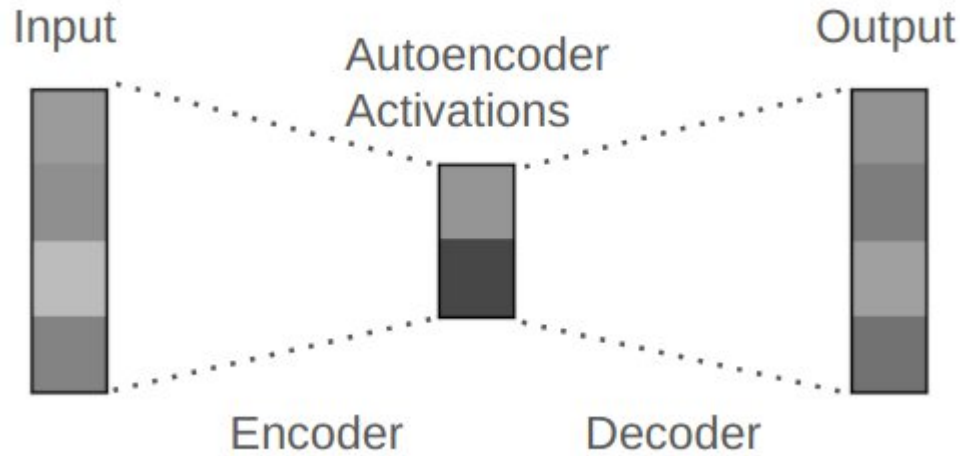
➤ So many application layers



➤ Applications: zero-shot prediction of variant effects



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



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

Feature #34M/31164353 Golden Gate Bridge feature example

The feature activates strongly on English descriptions and associated concepts

in the Presidio at the end (that's the huge park right next to the Golden Gate bridge), perfect. But not all people

repainted, roughly, every dozen years." "while across the country in san francisco, the golden gate bridge was

it is a suspension bridge and has similar coloring, it is often compared to the Golden Gate Bridge in San Francisco, US

They also activate in multiple other languages on the same concepts

ゴールデン・ゲート・ブリッジ、金門橋は、アメリカ西海岸のサンフランシスコ湾と太平洋が接続するゴールデンゲート海

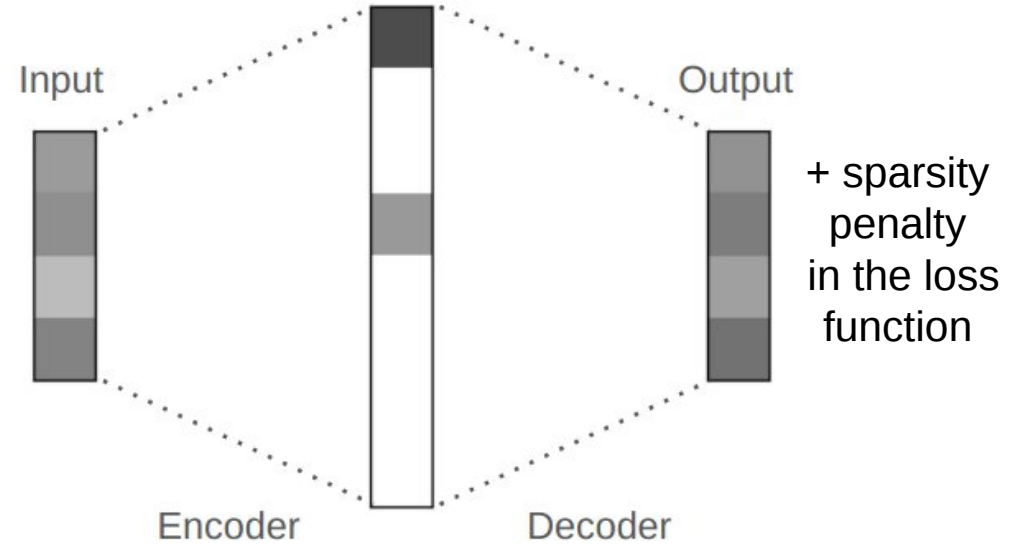
골든게이트 교 또는 금문교 는 미국 캘리포니아주 골든게이트 해협에 위치한 현수교이다. 골든게이트 교는 캘리포니아주 샌프란시

мост золотые ворота – висячий мост через пролив золотые ворота. он соединяет город сан-фран

And on relevant images as well



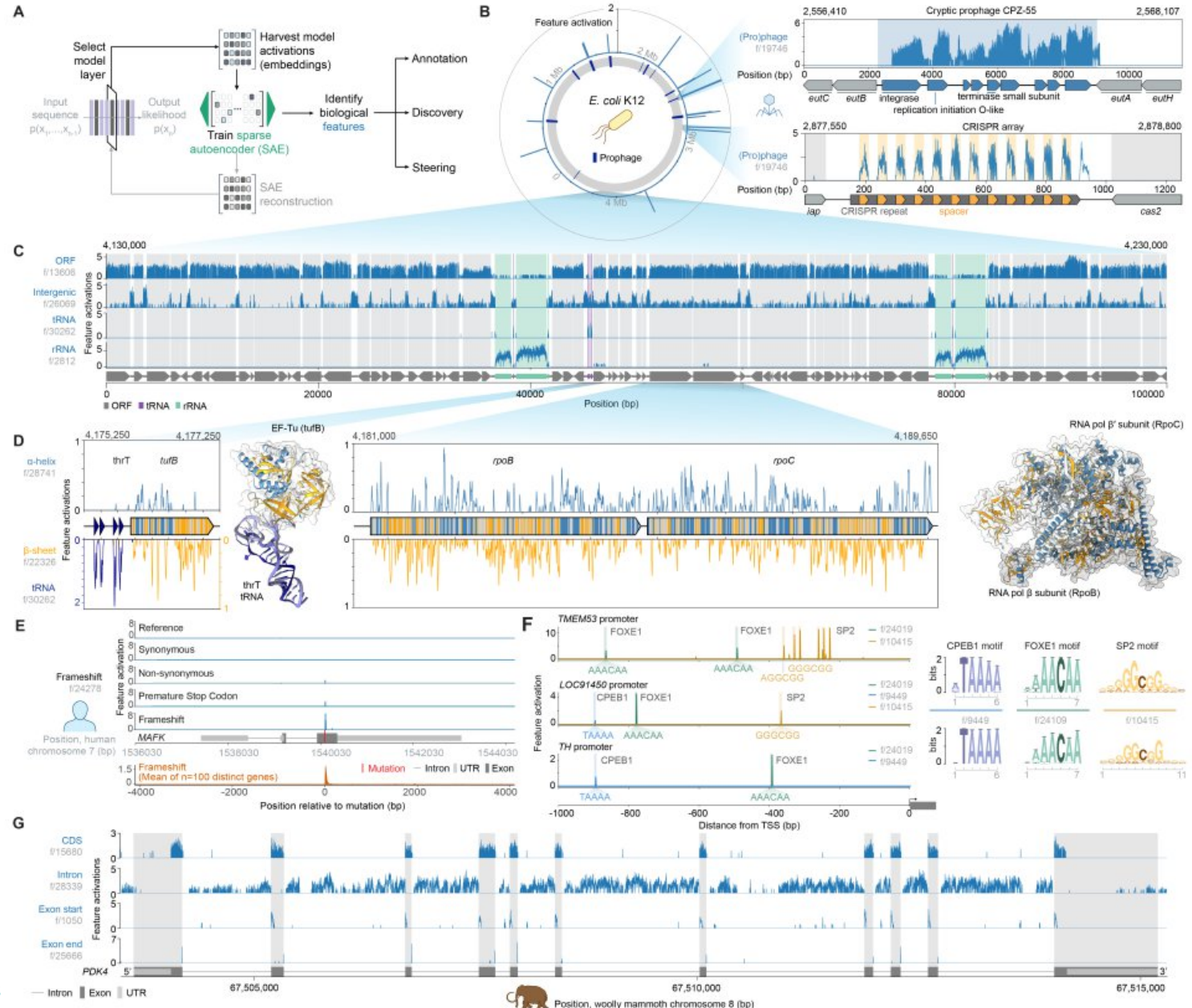
SAE Activations



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

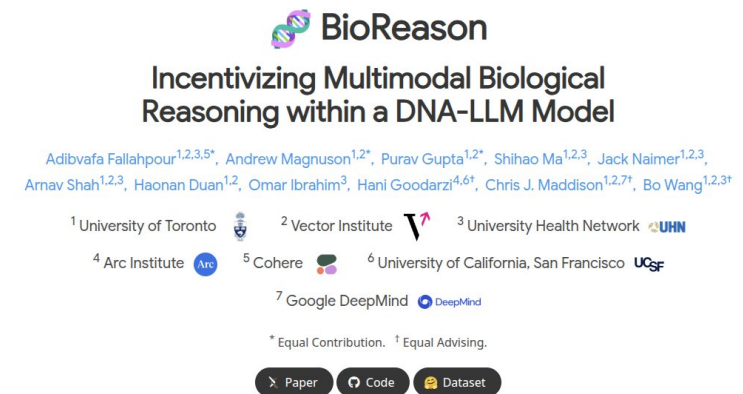
Features observed :

- 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



+ 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:



GPU providers:

– Evo 1: LabIA (Paris Saclay)



– Evo 2: Jean Zay (IDRIS)



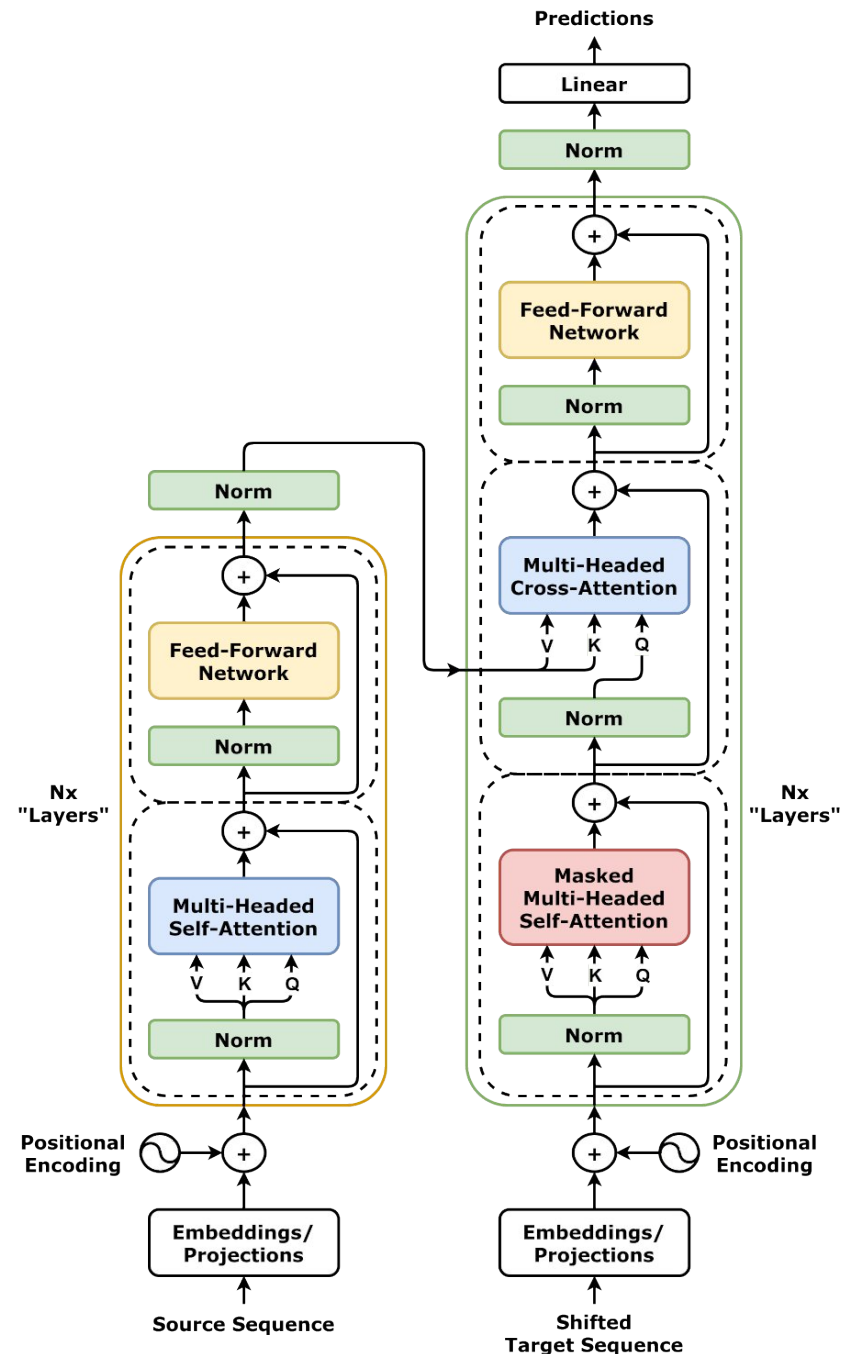
➤ Evo: not (exactly) a transformer

Transformer

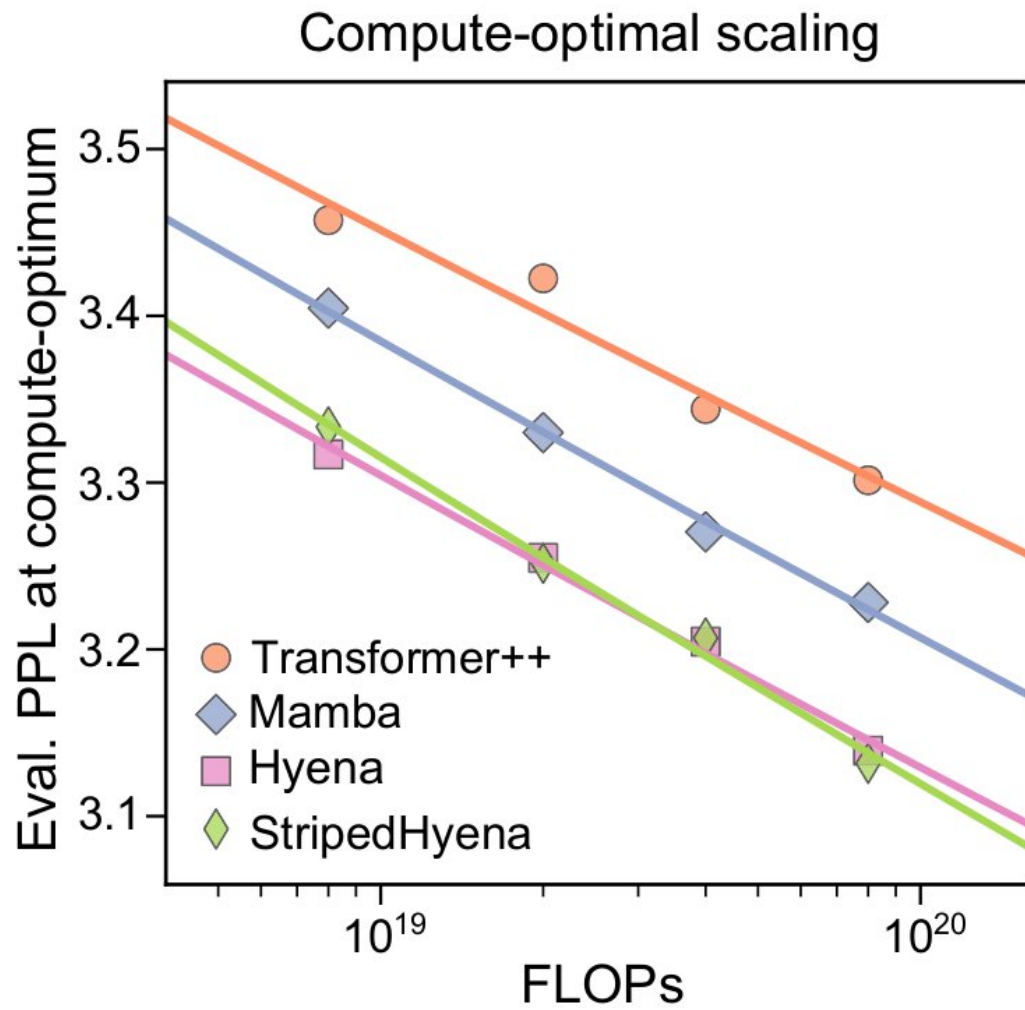
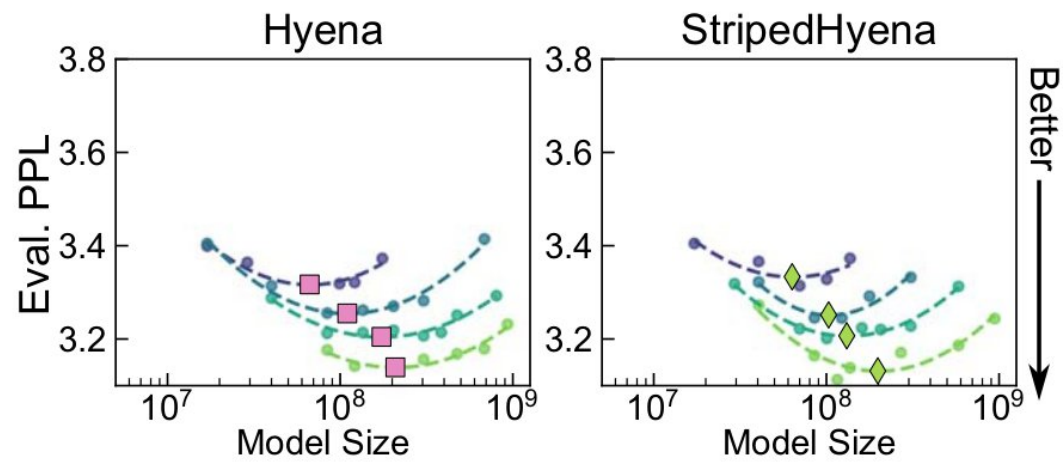
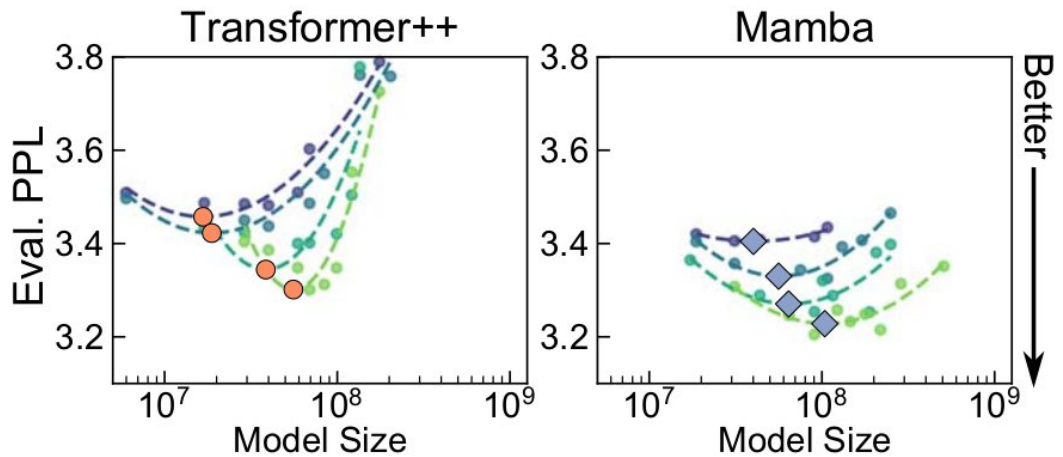
Attention Is All You Need^{*}

^{*} 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



FLOPs ● 8×10^{18} ● 2×10^{19} ● 4×10^{19} ● 8×10^{19}
 Optima ○ Transformer++ ◆ Mamba ■ Hyena ◆ StripedHyena

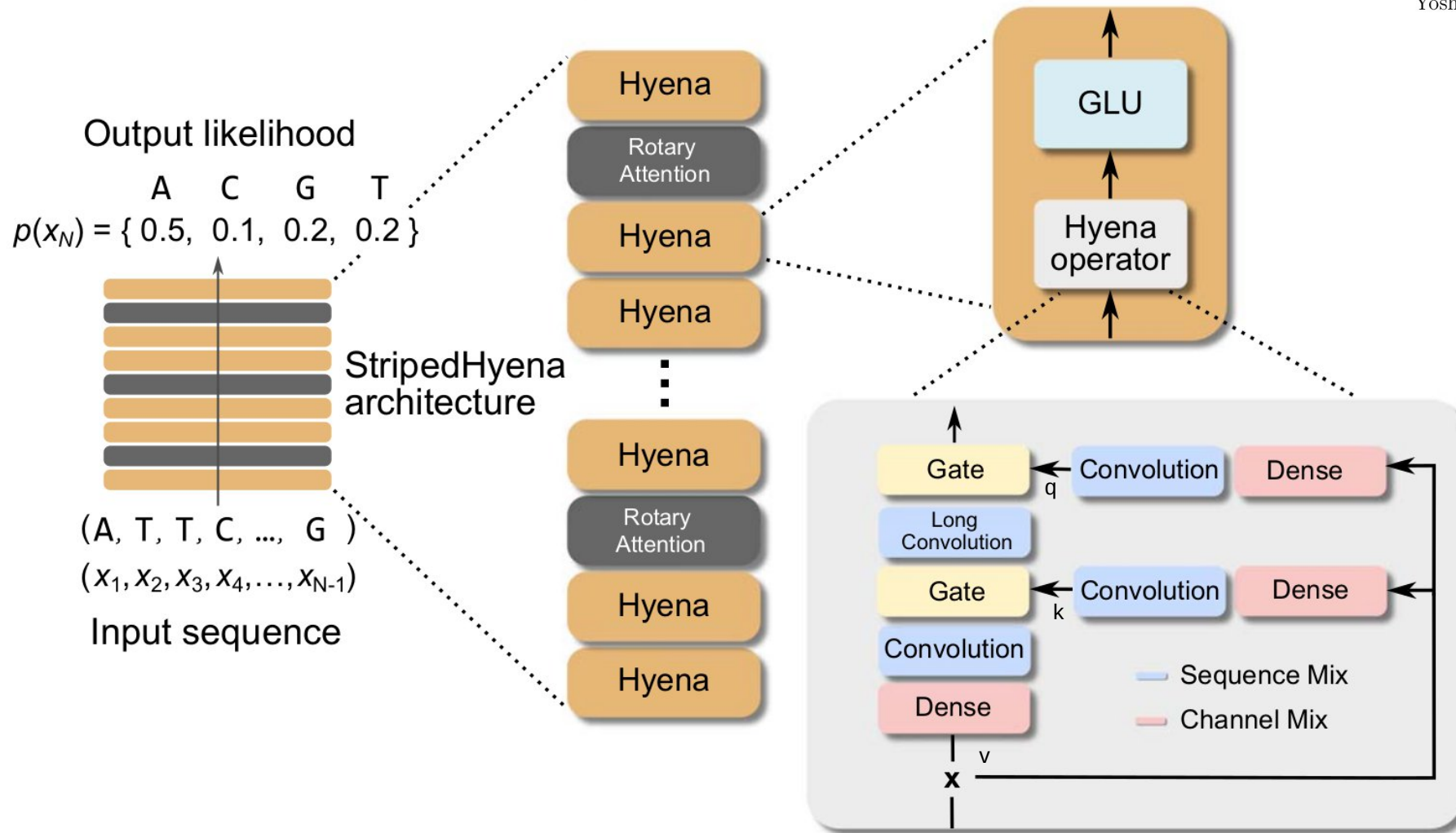


INRAE

Hands-on workshop on genomic language models

12/06/2026 / Guillaume GAUTREAU / StatInfOmics team / MalAGE unit

➤ Evo is based on Striped Hyena

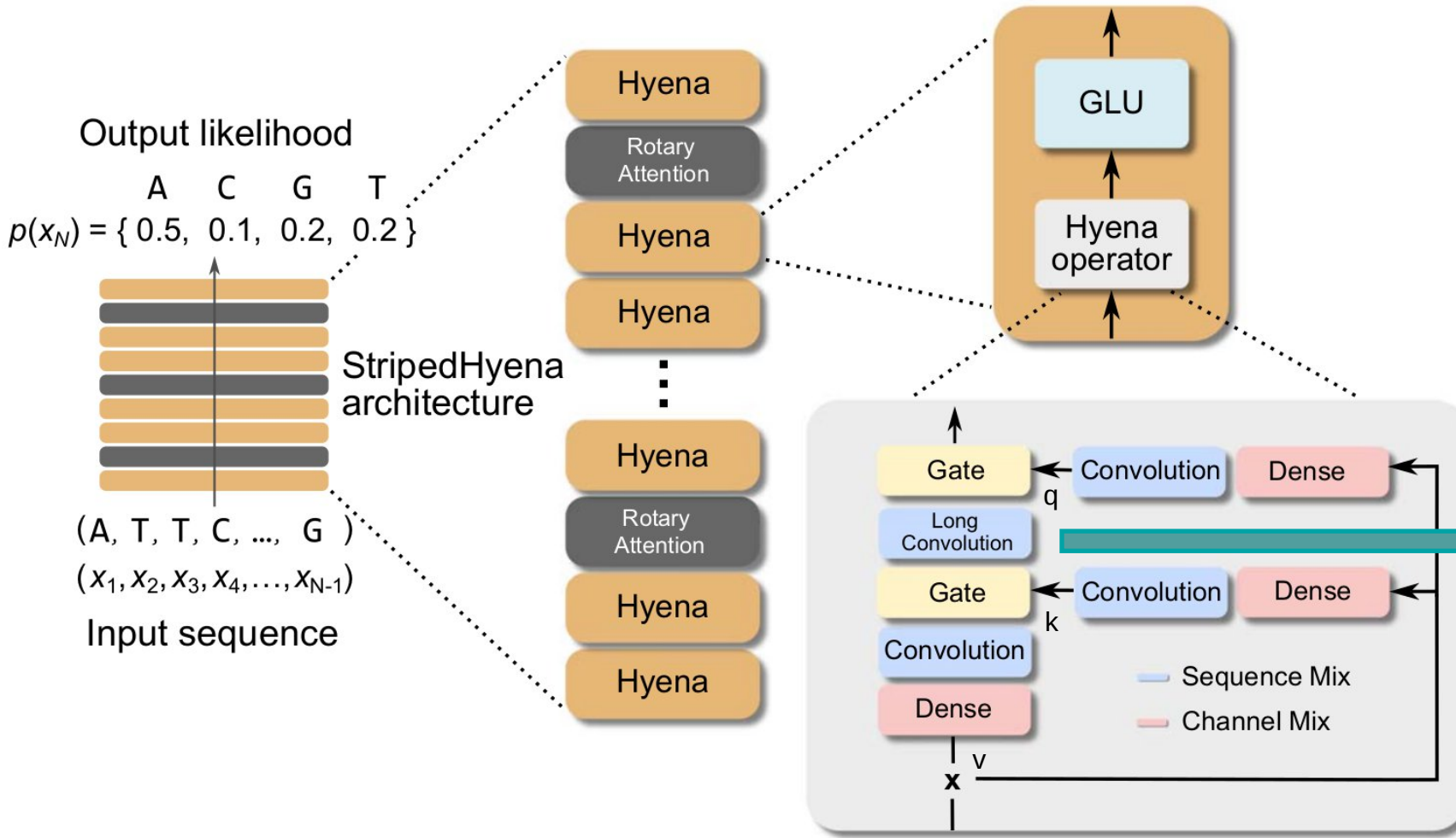


Hyena Hierarchy: Towards Larger Convolutional Language Models

Michael Poli^{*1}, Stefano Massaroli^{*2}, Eric Nguyen^{1,*},
 Daniel Y. Fu¹, Tri Dao¹, Stephen Baccus¹,
 Yoshua Bengio², Stefano Ermon^{1,†}, Christopher Ré^{1,†}



➤ Evo is based on Striped Hyena



Hyena Hierarchy: Towards Larger Convolutional Language Models

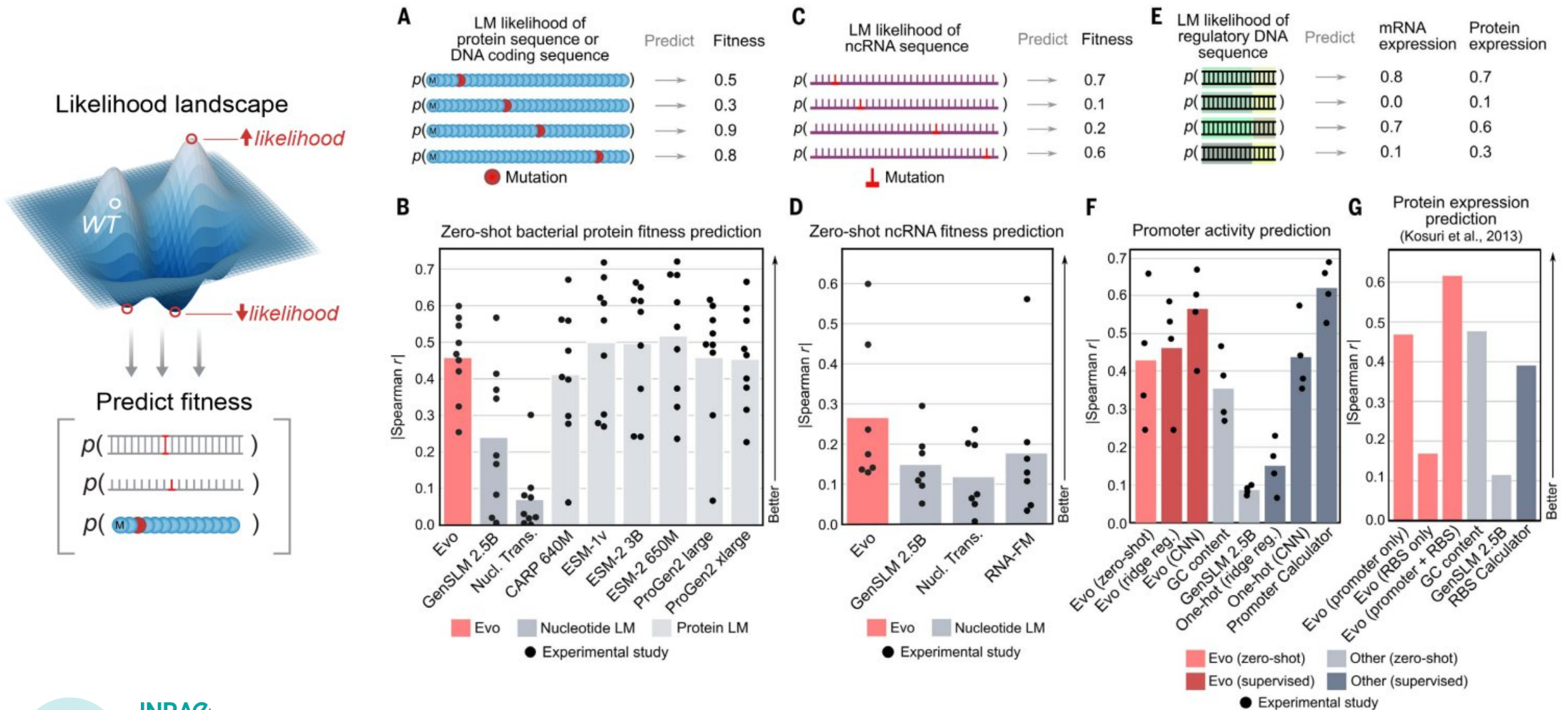
Michael Poli^{*1}, Stefano Massaroli^{*2}, Eric Nguyen^{1*},
 Daniel Y. Fu¹, Tri Dao¹, Stephen Baccus¹,
 Yoshua Bengio², Stefano Ermon^{1,†}, Christopher Ré^{1,†}



Complexity: $O(Nk)$
 N : # of tokens in window
 k : convolution kernel size
 if $k \rightarrow N \Rightarrow$ 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

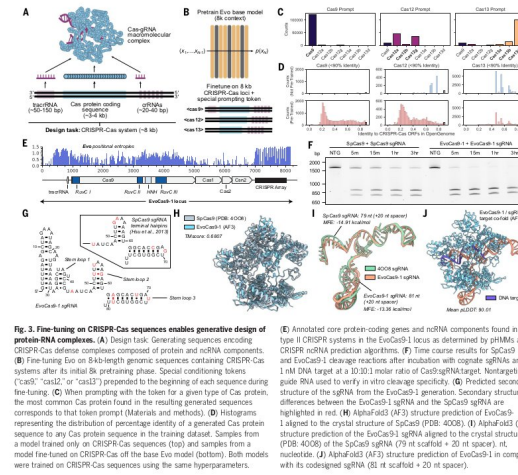


Fig. 3. Fine-tuning on CRISPR-Cas sequences enables generative design of protein-RNA complexes. (A) Design task: Generating sequences encoding CRISPR-Cas systems. (B) Fine-tuning on 84k-length genomic sequences. (C) Histogram representing the distribution of percentage identity of a generated Cas protein sequence to any Cas protein sequence in the training dataset. (D) Predicted secondary structure of the sgRNA from the EvocCas9-1 generation. (E) Annotated core protein-coding genes and ncRNA components found in type II CRISPR systems in the EvocCas9-1 locus. (F) Time course results for SgCas9 and EvocCas9-1 cleavage reactions. (G) SgCas9 sgRNA structure. (H) SgCas9 sgRNA structure. (I) AlphaFold3 structure prediction of EvocCas9-1 aligned to the crystal structure of SpCas9 (PDB: 4008). (J) AlphaFold3 structure prediction of the SpCas9 sgRNA (21 nt spacer) at 20 nt spacer. (K) AlphaFold3 structure prediction of EvocCas9-1 in complex with its cognate sgRNA (81 nt scaffold + 20 nt spacer).

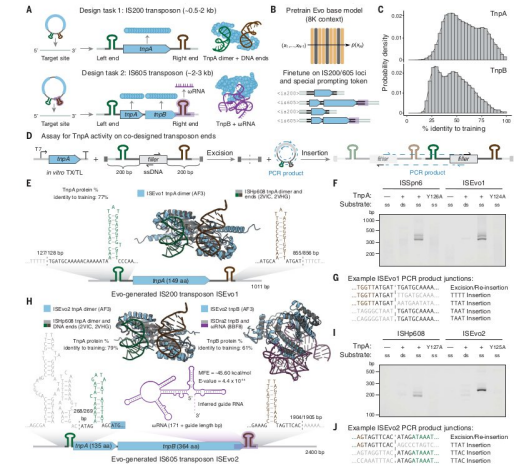


Fig. 4. Fine-tuning on IS200/IS605 sequences enables generative design of transposable biological systems. (A) IS200 and IS605 MGEs contain a TriptA transposase and are flanked by left and right end terminal repeats that interact with the TriptA to accomplish transposition. IS605 MGEs additionally encode a TriptA-*RNA* complex that performs DNA cleavage. Our design task is to produce sequences that contain these DNA, ncRNA, and protein components. (B) We fine-tuned Evo after its initial pretraining phase on natural sequences containing IS200/IS605 systems. (C) Histograms representing the distribution of the percentage identity of Evo-generated TriptA and TriptA proteins to their natural counterparts. (D) Schematic of the Evoc-generated IS200 transposon. (E) Schematic of the Evoc-generated IS605-like system. (F) Example reads from nanopore sequencing of PCR products from the Evoc-generated IS200 transposon. (G) Schematic of the Evoc-generated IS605-like system. (H) Example reads from nanopore sequencing of PCR products from the Evoc-generated IS605 transposon.

- Prediction of gene essentiality

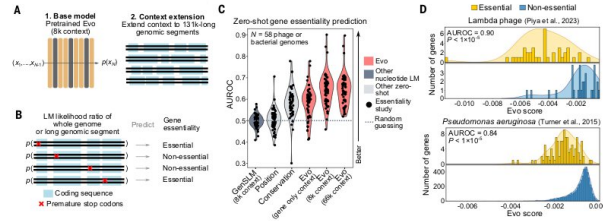


Fig. 5. Evo learns mutational effects on organismal fitness across diverse bacterial and phage genomes. (A) For genome-scale prediction and generation tasks, we first pretrained Evo on sequences with 8322 tokens and then extended its context window size in a second pretraining phase to sequences of 131,072 tokens. (B) We performed an in silico, genome-wide mutagenesis screen in which we introduced premature stop codons at each coding sequence in a genome. We computed the language model (LM) likelihood of the mutated gene sequence plus some amount of additional genomic context (up to 66 kb). We then took the ratio of this likelihood to the likelihood of the unmutated sequence. We tested whether these likelihood ratios would be predictive of gene essentiality. (C) Volin and strip plots of the distribution of the strength of gene essentiality prediction across 58 studies (each dot corresponds to a different study), in which each study conducted a genome-wide essentiality screen in a bacterial ($N = 56$) or phage ($N = 2$) species. We measured predictive performance as the AUROC in which the LM likelihood ratio is used to predict a binary label of "essential" or "nonessential." "Gene-only context" indicates that the model is provided with only the gene sequence and no additional flanking genomic context. "8k context" and "66k context" indicate that the LM is provided with the gene sequence and flanking genomic context up to a total of 8392 or 65,536 tokens, respectively. Evo has some predictive performance with gene-only context, has vastly improved performance from gene-only to 8k context, and some outlier improvements from 8k to 66k context. (D) Histograms representing the distribution of the log of the likelihood ratios ("Evo score") for the essential genes (blue) and the nonessential genes (yellow) in two genomes: lambda phage (top) and *P. aeruginosa* (bottom). These results are based on providing Evo with 66k context.

- Evo generates megabase-scale sequences with plausible genomic architecture

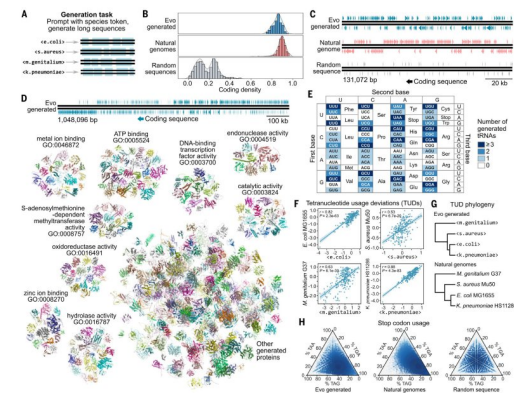
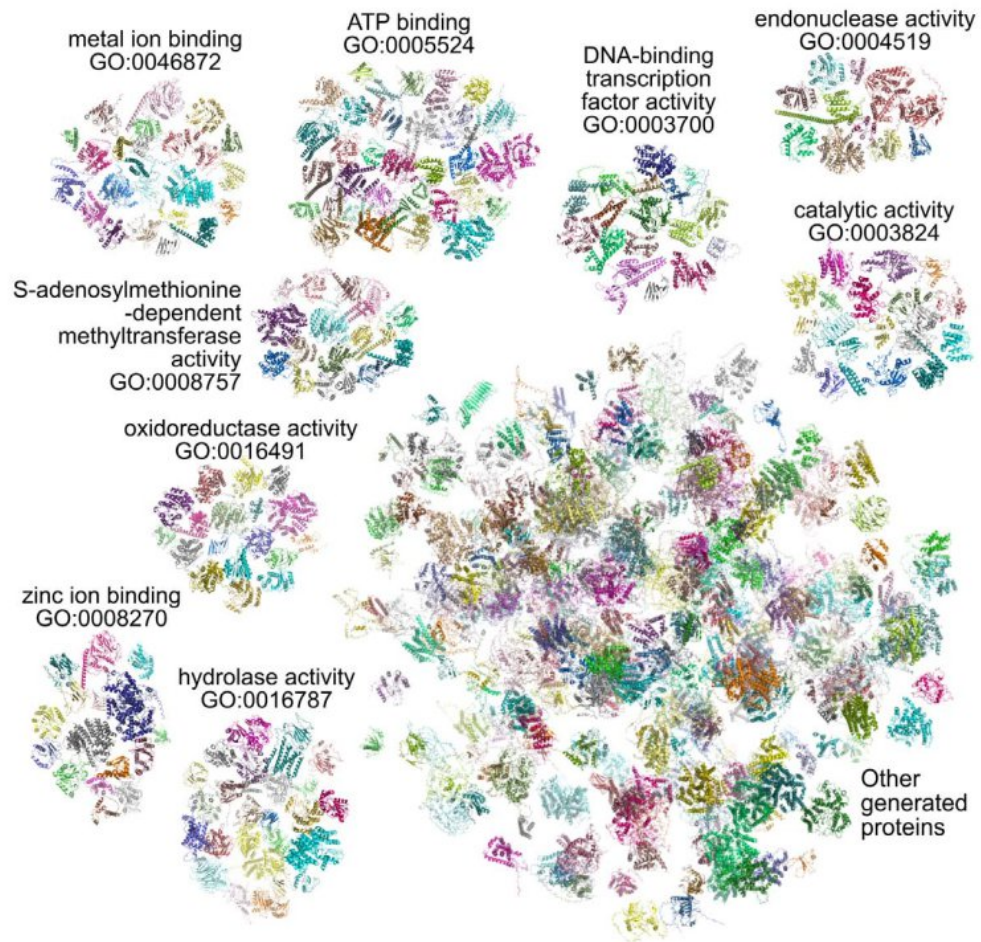


Fig. 6. Evo generates megabase-scale sequences with plausible genomic architecture. (A) We prompted Evo with species-level tokens during the second pretraining stage. We use bacterial species prompts and generate sequences of ~650 kb in length. (B) Histograms depicting the distribution of coding density scores among 131 kb crops of sequences generated by Evo ("Evo generated") sequences from natural bacteria ("natural genomes"), or sequences in which the two base pairs were sampled uniformly at random (random sequences). (C) Arrow plots depicting the organization of coding sequences on an example 131 kb sequence generated by Evo, derived from a natural genome, or sampled randomly. Coding sequences are depicted as arrows in which the horizontal length of the arrow corresponds to the genomic interval and the direction of the arrow indicates the strand. The top and bottom rows of arrows indicate the 5'-3' and 3'-5' strands, respectively, and the Evo-generated sequence was designated as the 5'-3' strand. Both Evo-generated and natural genomes exhibit sporn-like structure in which clusters of colocalized genes are on the same strand. (D) and (E) ~1 kb generated sequence is represented as an arrow plot, as in (C). Below this arrow plot are ESMFold structure predictions of all protein coding sequences from 100 through 1024 amino acids in length, as identified by Pfam. Structure predictions are aligned to natural bacteria ("natural genomes"), or sequences from natural genomes, which are then mapped to associated GO molecular function terms (Methods and Methods). The largest GO categories are displayed as clusters alongside a large cluster containing all other proteins. ATP adenose triphosphatase. (F) Log of TUD of Evo-generated versus natural genomes for each species prompt. Shaded areas are the Pearson correlation coefficient test. Shaded regions indicate a 95% confidence interval. (G) Hierarchical clustering of Evo-generated and natural genomes based on Euclidean distances of the TUD. (H) Percent usage of each stop codon in all three reading frames of Evo-generated, natural, and random ORFs.



➤ 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).

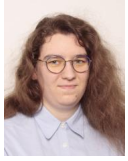
The largest GO categories are displayed as clusters, alongside a large cluster containing all other proteins. ATP, adenosine triphosphate. »

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

INRAE

➤ Feedback on Evo 1

Evaluation of the credibility of the output



Methodology

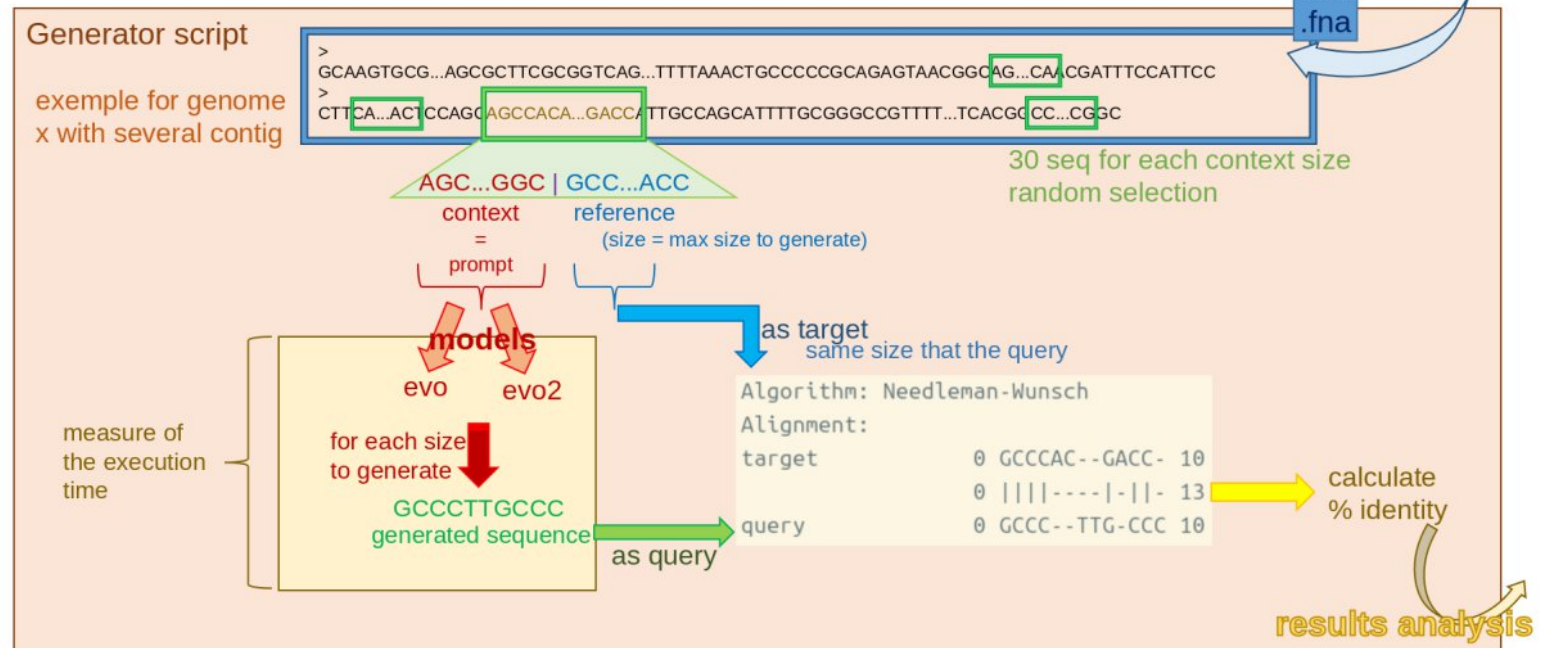
Naïa Périnelle
(M2 student)

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	Corallocooccus exiguus	19	* GCF_009909105.1
Nitrospirota	1001	Nitrospira_A sp900170025	23	* GCF_900170025.1
Patescibacteria	8106	Nanosynococcus 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

from NCBI

fasta
.fna



INRAE

Hands-on workshop on genomic language models

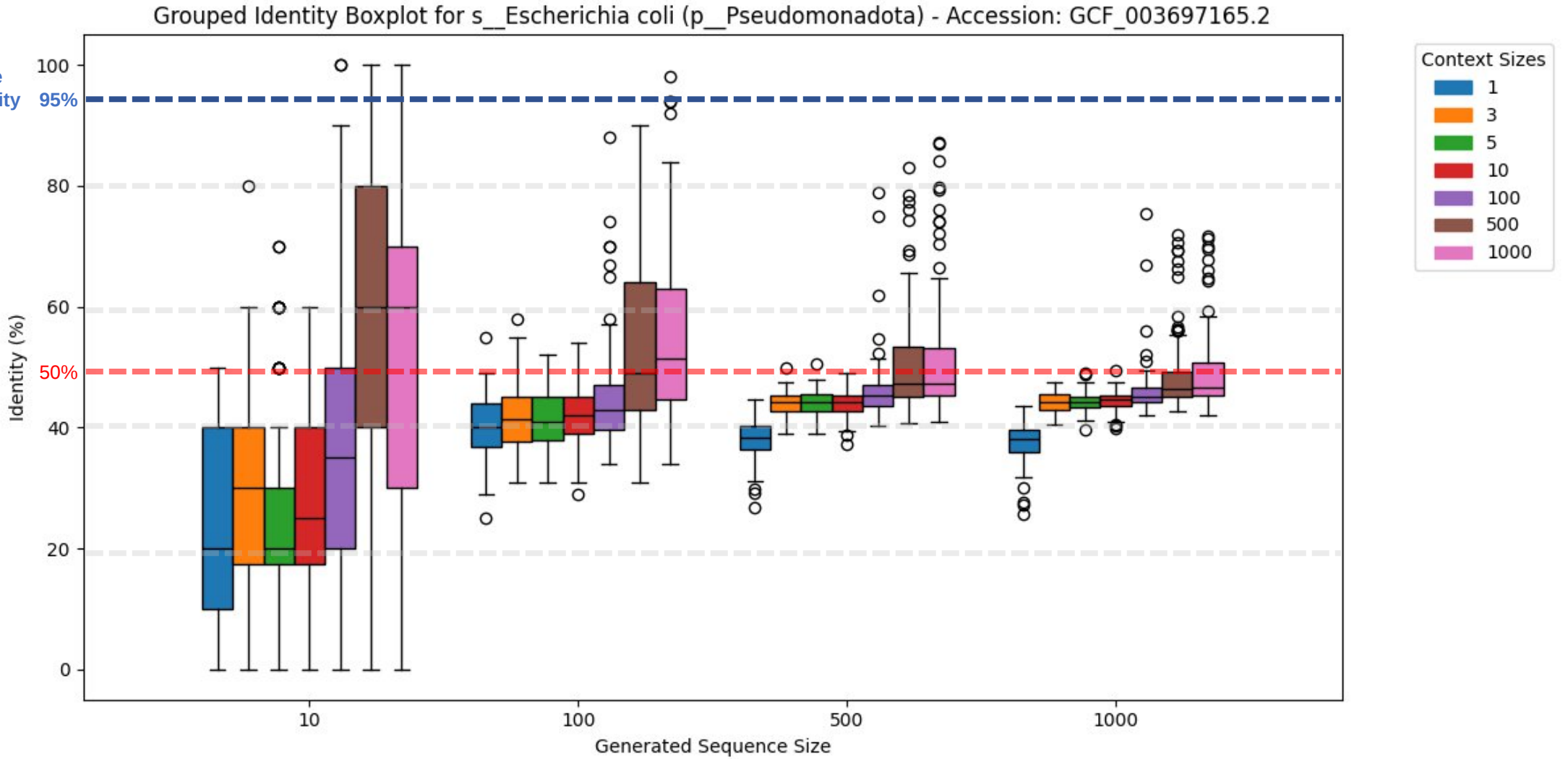
12/06/2026 / Guillaume GAUTREAU / StatInfOmics team / INRAE site



➤ Results Evo 1 131k on E. coli

Naïa Périnelle
(M2 student)

Species Average
Nucleotide Identity
in Prokaryotes



INRAE

Hands-on workshop on genomic language models

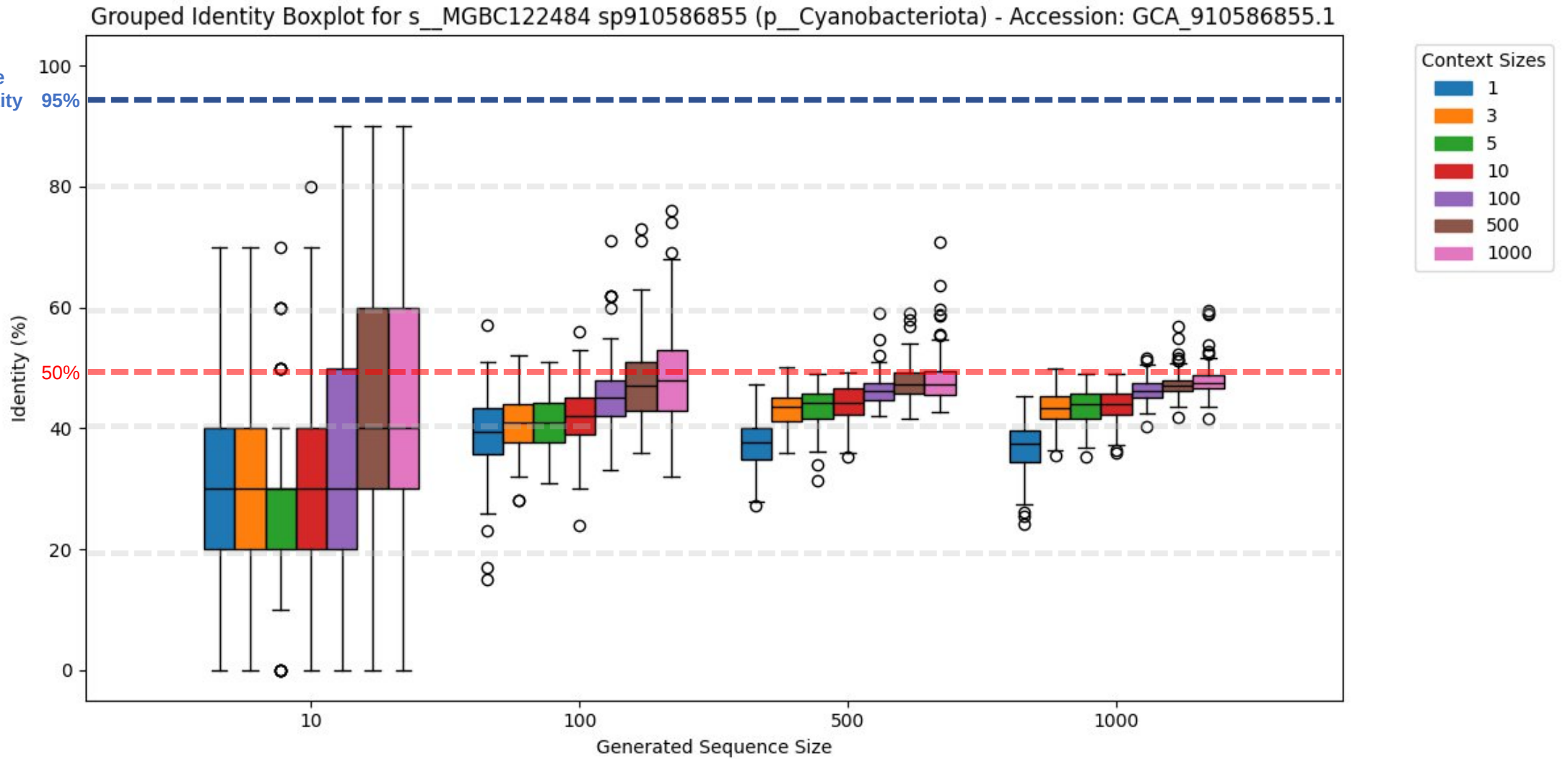
12/06/2026 / Guillaume GAUTREAU / StatInfOmics team / MalAGE unit



Naïa Périnelle
(M2 student)

Species Average
Nucleotide Identity
in Prokaryotes

➤ Results Evo 1 131k on « unnamed cyanobacteria »



INRAE

Hands-on workshop on genomic language models

12/06/2026 / Guillaume GAUTREAU / StatInfOmics team / MalAGE unit

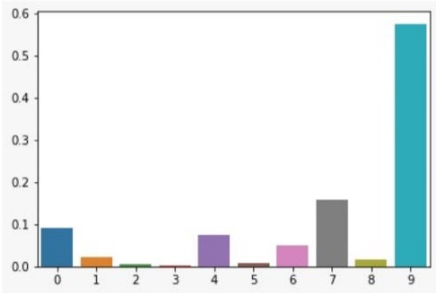


➤ Impact of temperature

Naïa Périnelle
(M2 student)

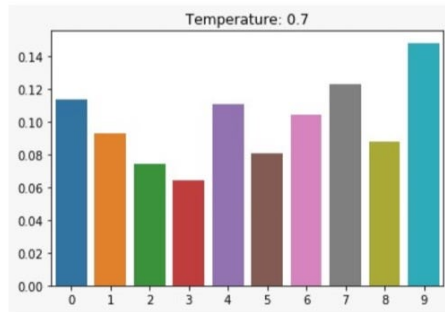
SOFTMAX WITHOUT TEMPERATURE (T=1)

$$\frac{e^{z_i}}{\sum_j e^{z_j}}$$



SOFTMAX WITH TEMPERATURE

$$\frac{e^{z_i/T}}{\sum_j e^{z_j/T}}$$

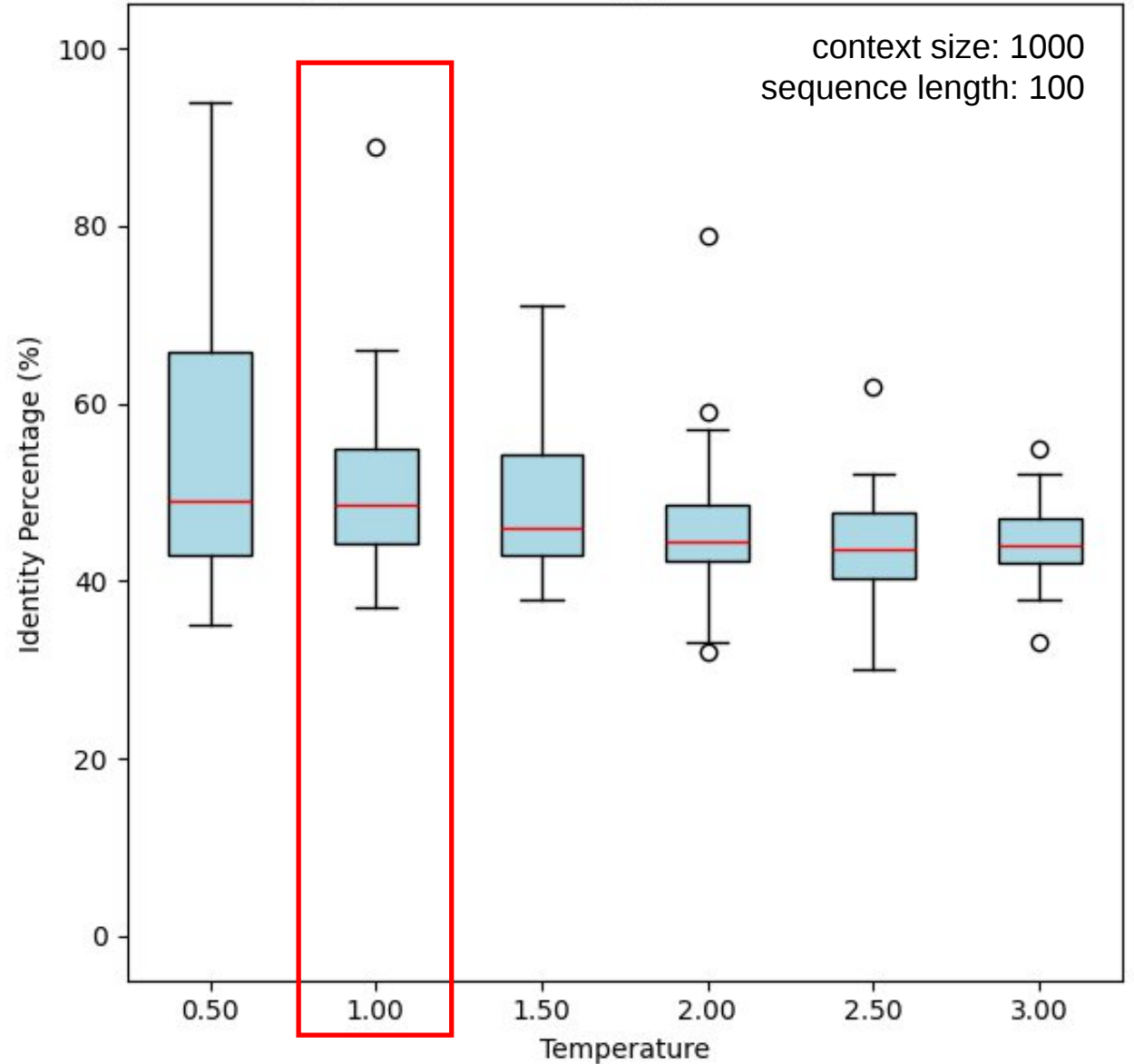


LESS ENTROPY

INCREASE IN ENTROPY
WITH INCREASE IN T

MORE ENTROPY

s_Escherichia coli (p_Pseudomonadota)



INRAE

Hands-on workshop on genomic language models

12/06/2026 / Guillaume GAUTREAU / StatInfOmics team / MalAGE unit

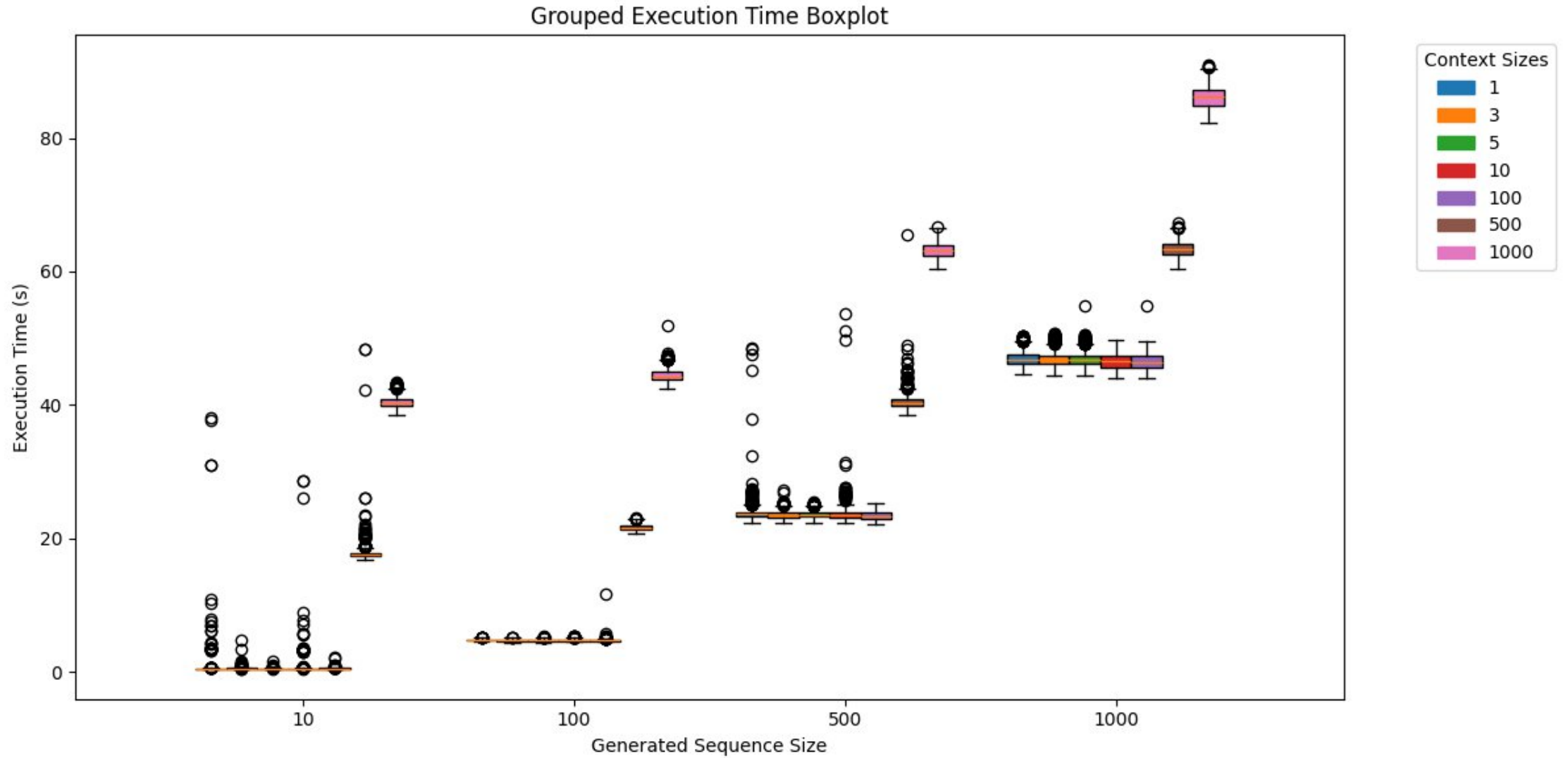
+ top_p=1, top_k=4



Execution times

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

Naïa Périnelle
(M2 student)



INRAE

Hands-on workshop on genomic language models

12/06/2026 / Guillaume GAUTREAU / StatInfOmics team / MalAGE unit

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

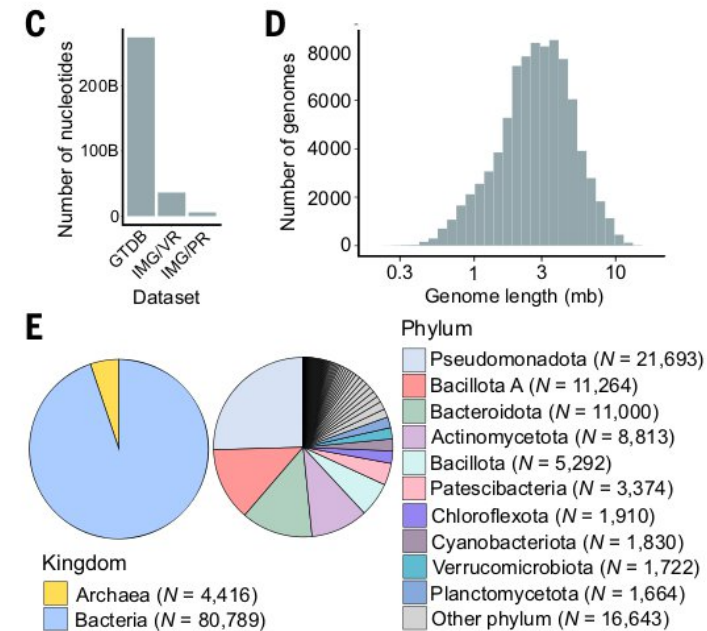
Published online: 19 November 2025

Open access

 Check for updates

Aditi T. Merchant^{1*}, Samuel H. King^{1,2}, Eric Nguyen^{1,2} & Brian L. Hie^{2,4,5}

Generative genomic models can design increasingly complex biological systems¹. 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 genes², 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.



INRAE

Hands-on workshop on genomic language models

12/06/2026 / Guillaume GAUTREAU / StatInfOmics team / MalAGE unit