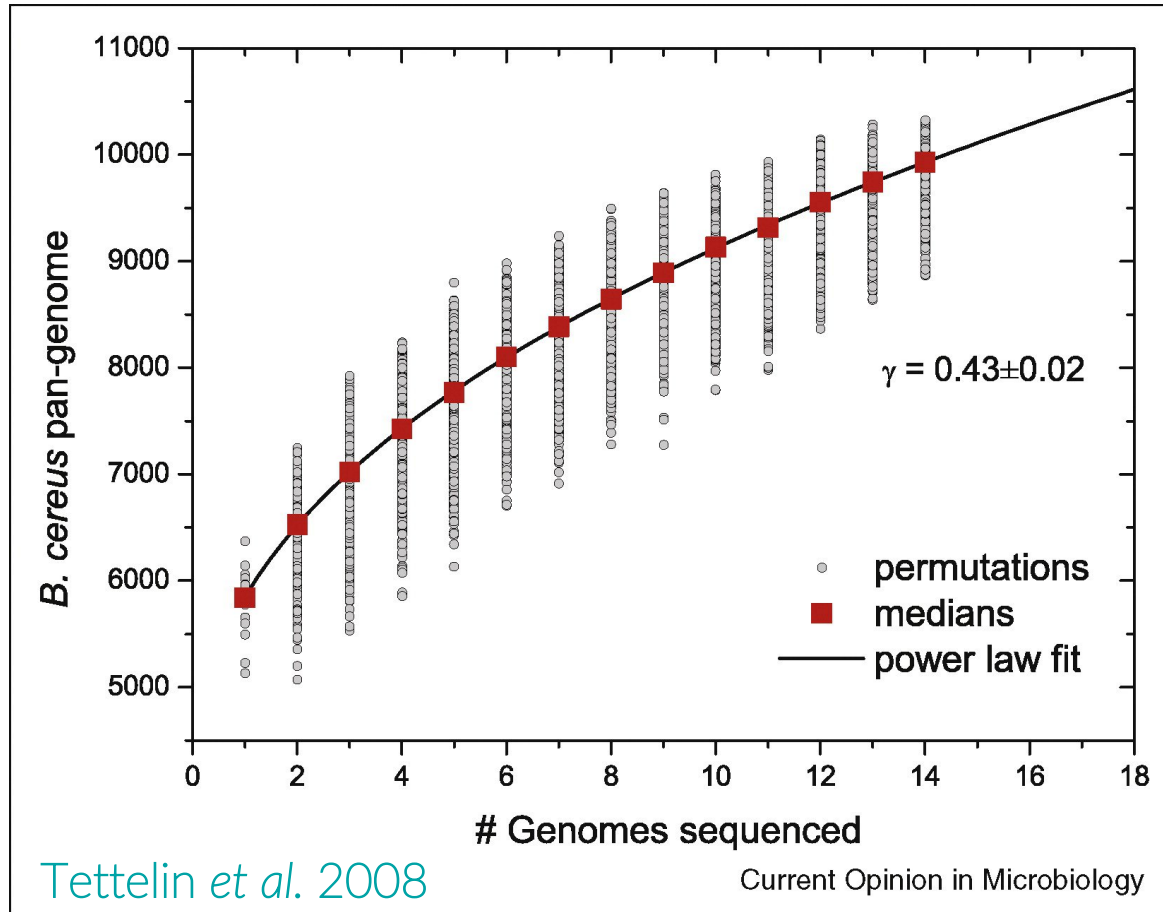


# ➤ What genomics can learn from Natural Language Processing?

Guillaume GAUTREAU, MaIAGE unit, StatInfOmics team  
13th of June, labioscripting

## ➤ Pangenomics inspired by linguistic

The pangenomic concept to measure genomic diversity and stability (core and accessory genomes)



Harold Heaps  
(1922-2016)

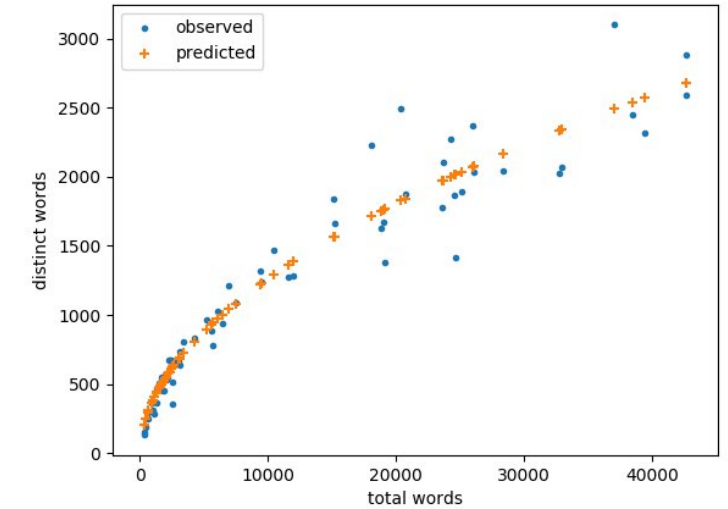
Heaps' Law is an empirical law that describes the number of distinct words in a document as an increasing function of the document length.

$$F = p(N) = \kappa N^\gamma$$

$\gamma > 1$  : supralinear tendency

$\gamma = 1$  : linear tendency

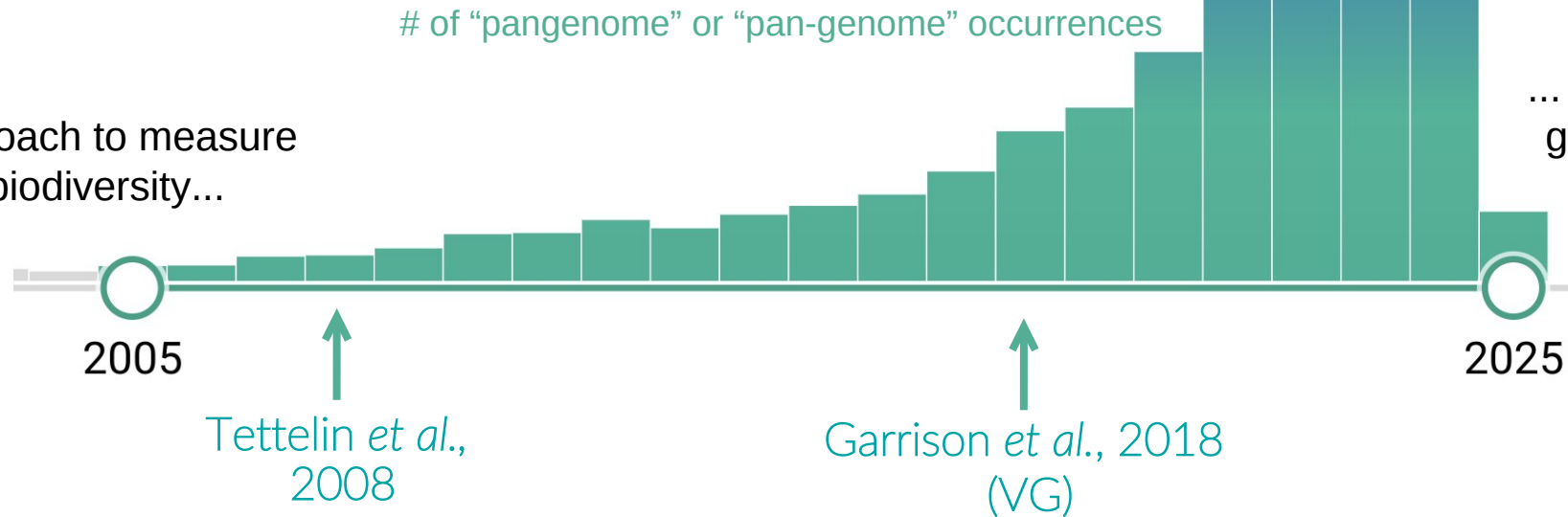
$0 < \gamma < 1$  : sublinear tendency



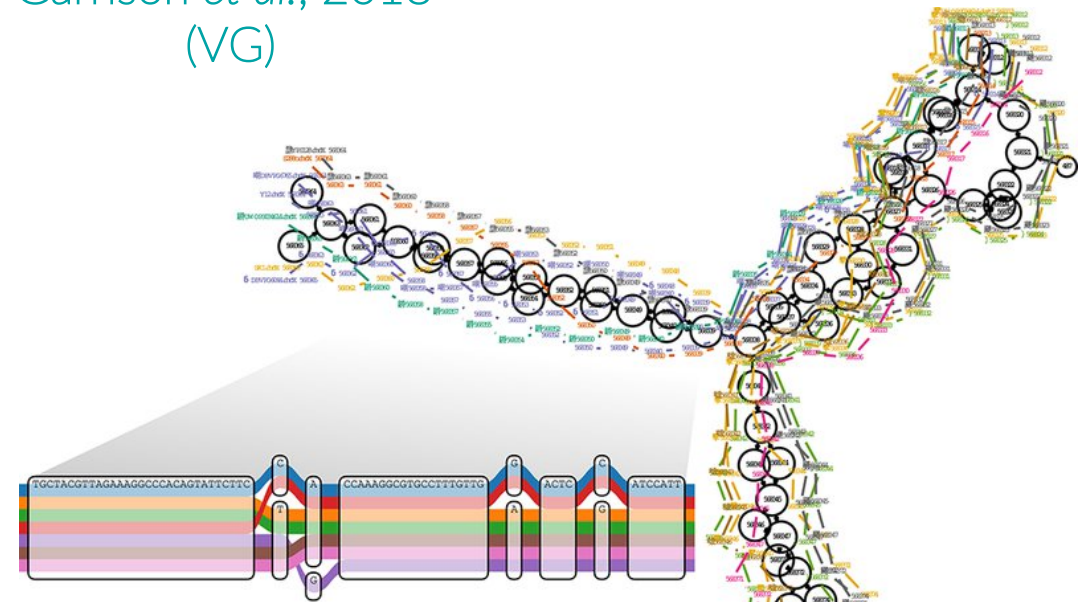
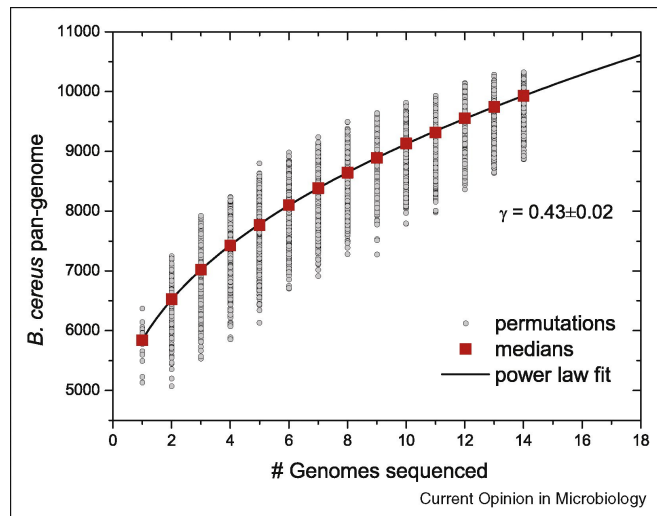
**The Bible, King James  
Version, Complete**

# ➤ What does pangenome mean?

From an approach to measure genomic biodiversity...



... to methods for managing graphs of ever-increasing genomic sequences



## ➤ Methods to handle ever-increasing genomic sequences

Pangenomics now deeply relies on pangenome graphs:

- pangenome graphs tries to compile all the possible (or at least as much as possible) genomic variations sampled in a population in a graph (DBG, variation graphs..).
- But each day, genomic information became and more abundant, so these graphs grew tremendously complex !

=> Efforts to optimize these graphs and make them computationally manageable open up a significant area of research





## ➤ Pangenome and AI

- It is uncertain whether constructing increasingly larger pangenome graphs using sophisticated data structures and algorithms is the best approach to deciphering and understanding biology
- Instead, due to the open-ended nature of genetic diversity (much like language ?) constructing pangenomes as large foundational models could help to model and capture unknown genomic relationships until now.

In other words, is pangenome the concept of a Foundation Model ?



INRAE

## ➤ IA and Large Language Models

**Back to the basics**



## ➤ AI legal definition (AI act)

### *Article 3*

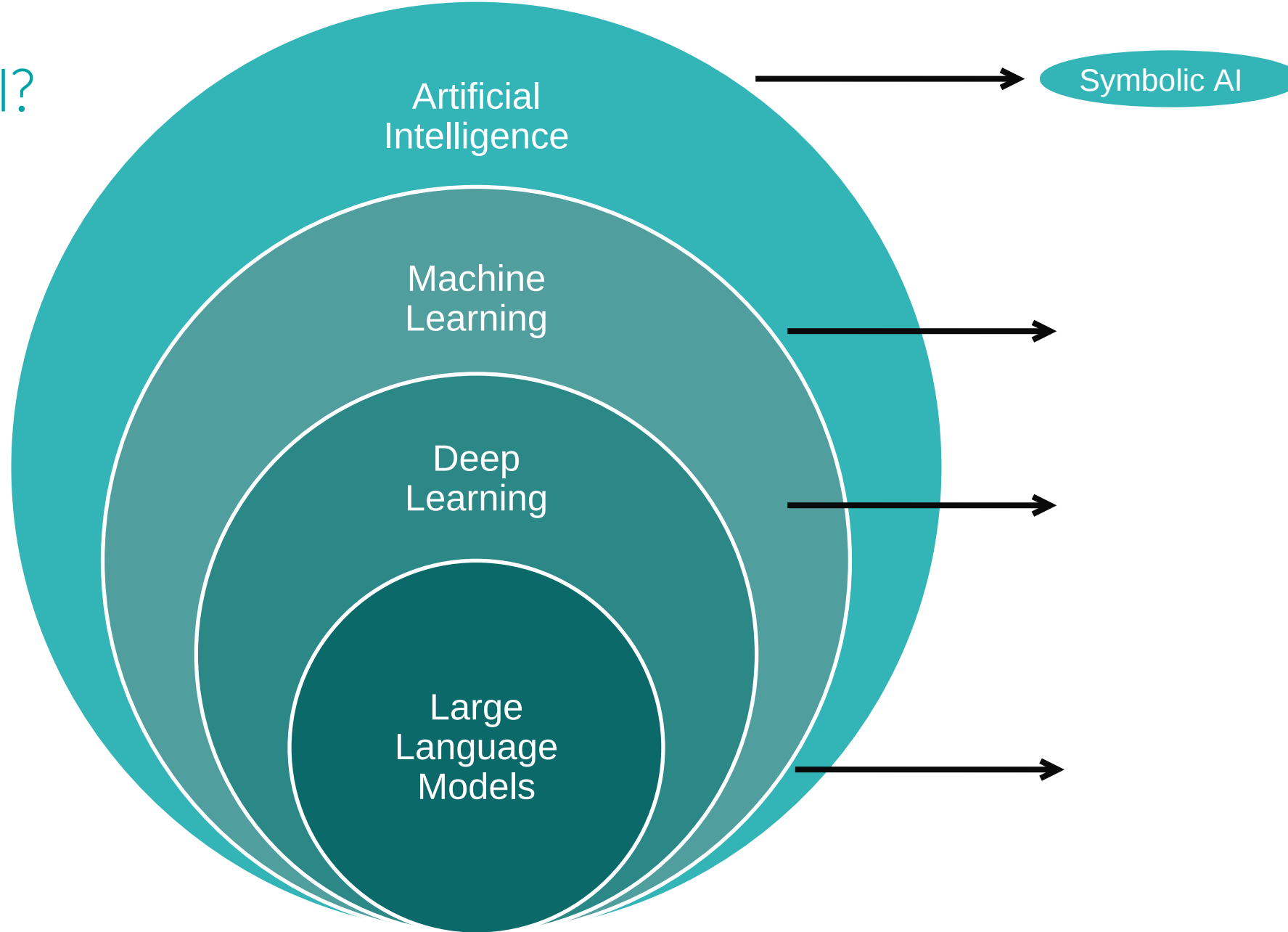
#### **Definitions**

For the purposes of this Regulation, the following definitions apply:

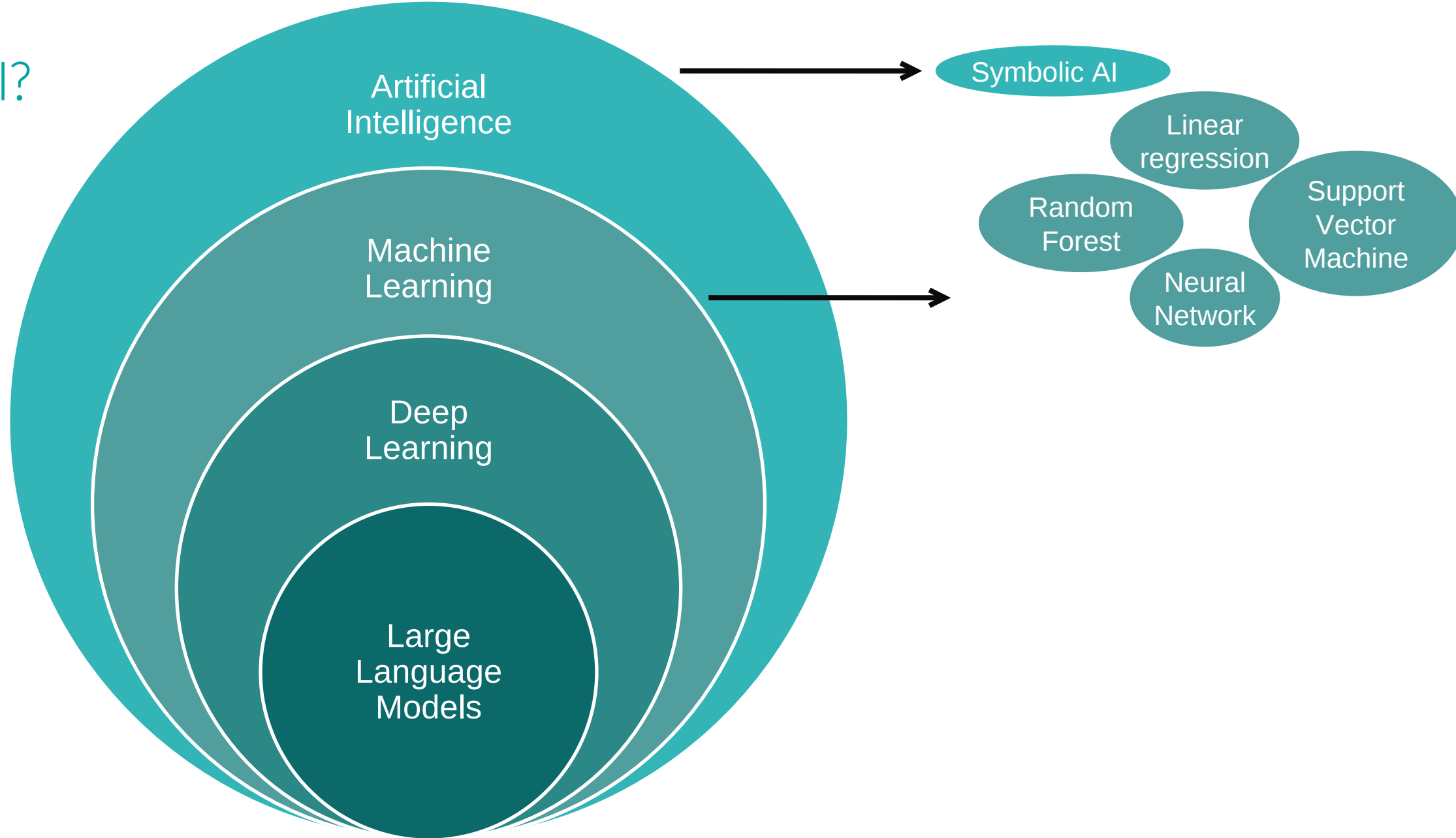
- (1) ‘AI system’ means a machine-based system that is designed to operate with varying levels of autonomy and that may exhibit adaptiveness after deployment, and that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments;



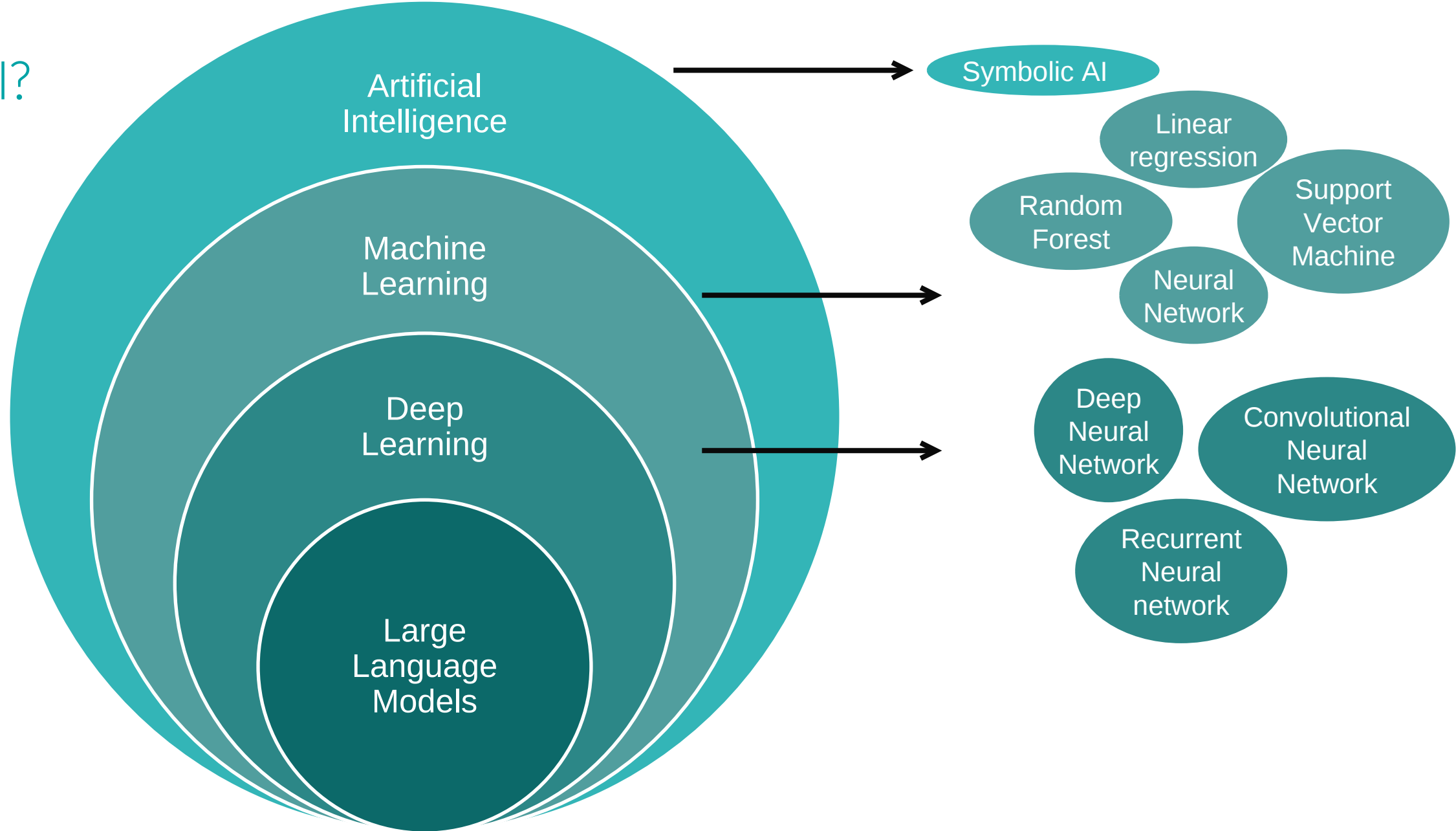
## > AI?



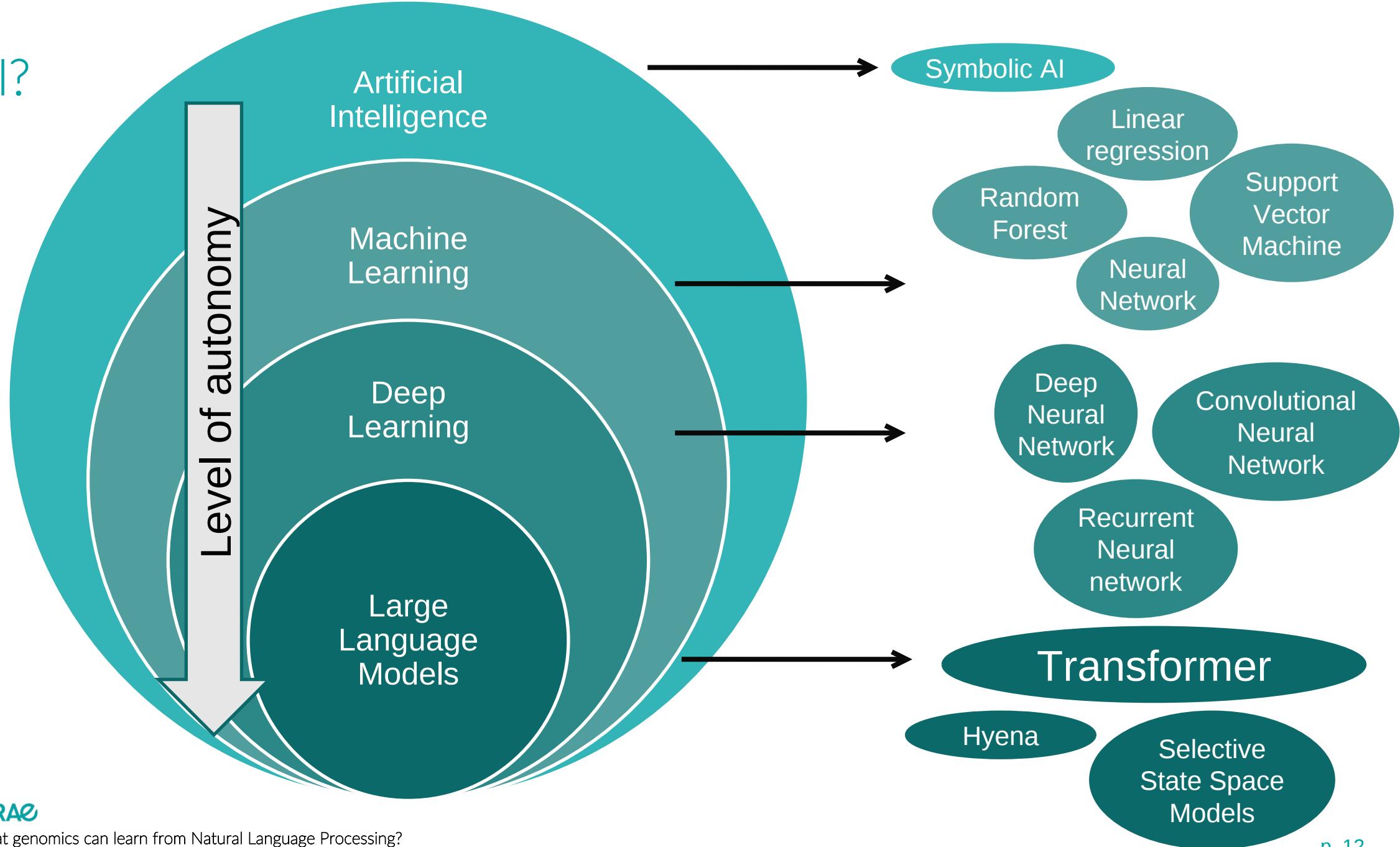
## > AI?



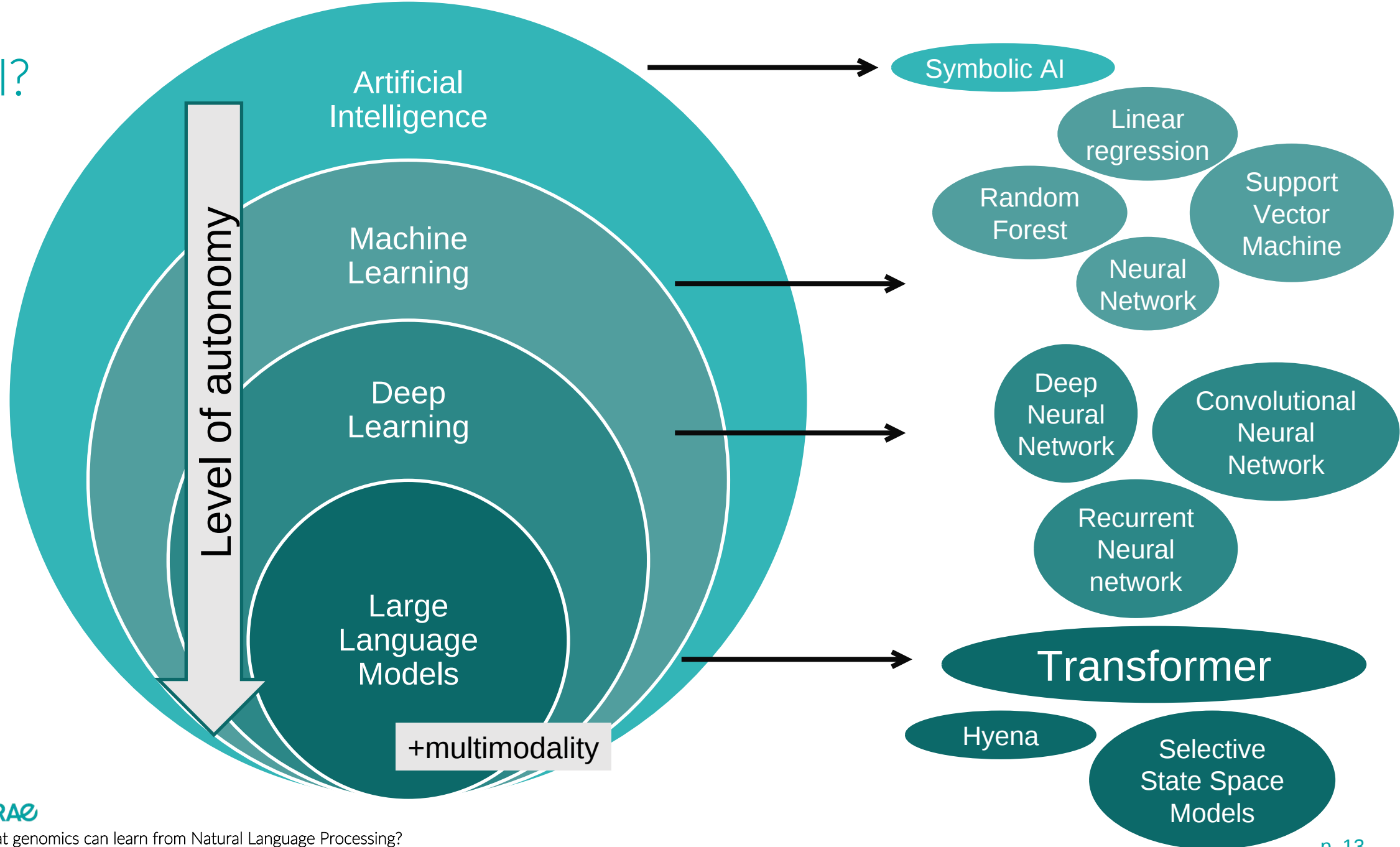
# > AI?



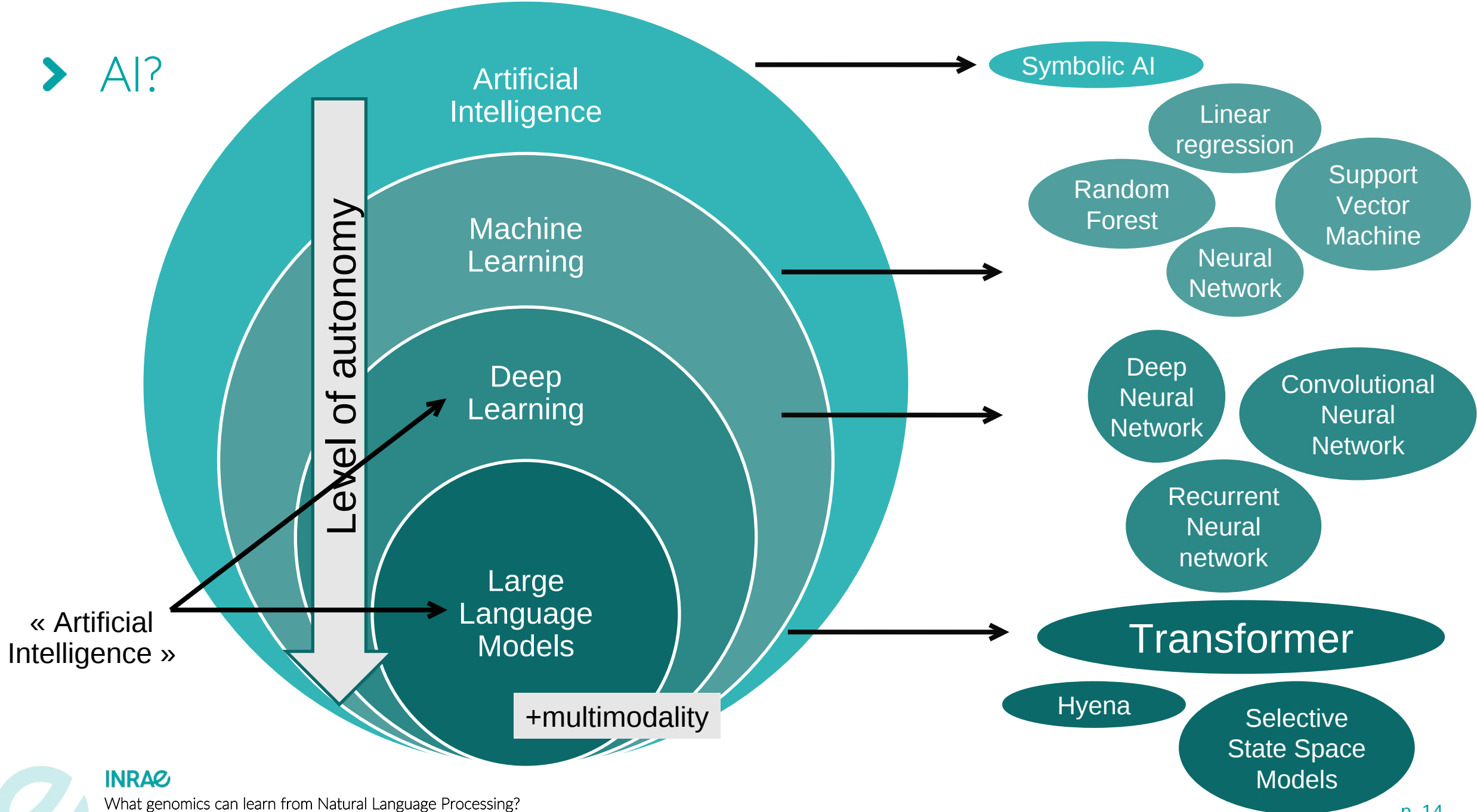
# > AI?



# > AI?



# > AI?



# ➤ 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  
lukaszkaizer@google.com

Illia Polosukhin\* ‡  
illia.polosukhin@gmail.com

### Abstract

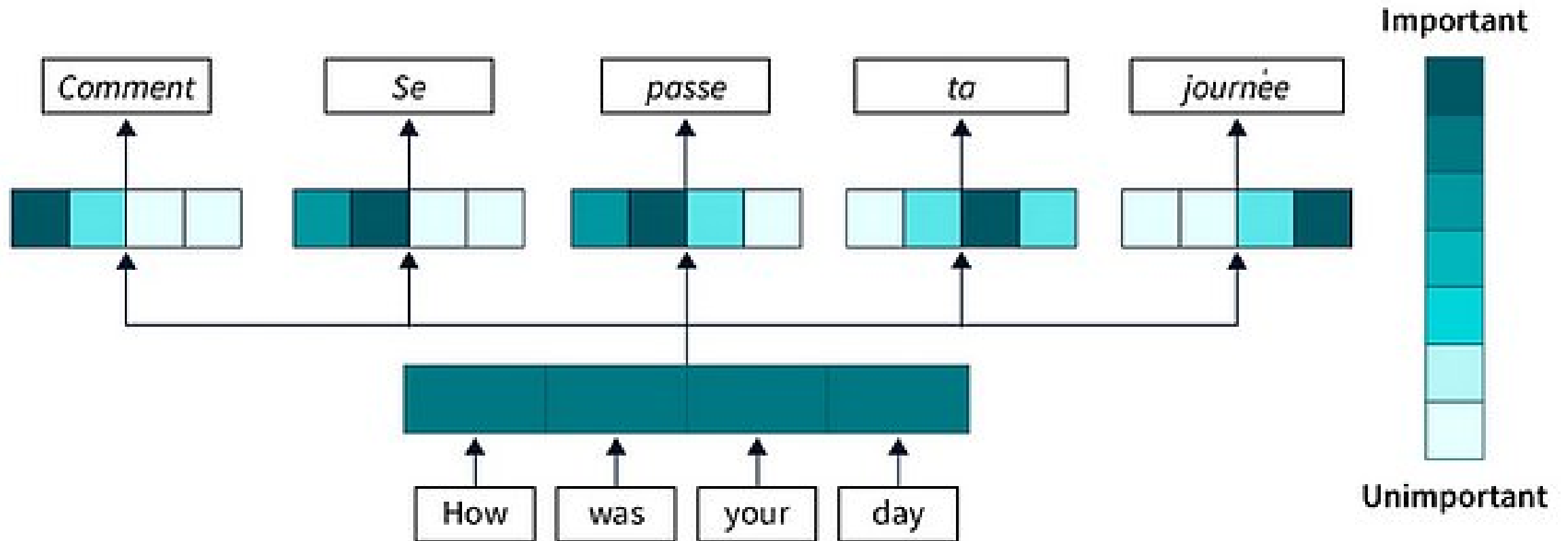
The dominant sequence transduction models are based on complex recurrent or convolutional neural networks that include an encoder and a decoder. The best performing models also connect the encoder and decoder through an attention mechanism. We propose a new simple network architecture, the Transformer, 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 (>180k)
- Design for translation purposes



## ➤ The intuition of attention: look at everything, but not equally

Instead of translating word-by-word, attention looks at which English words are important for each French word



Attention finds what is most important in context to understand the actual meaning of each token

## ➤ « 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



# » « Attention Is All You Need »

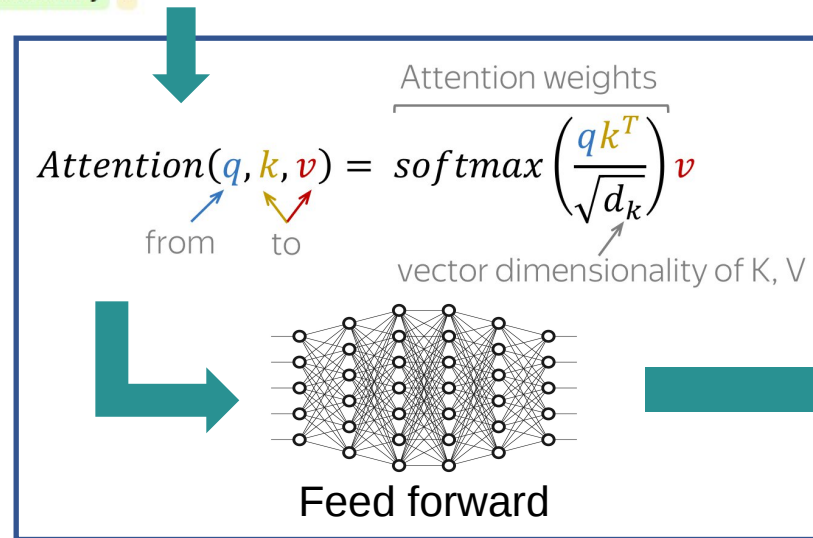
Given a prompt:

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

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

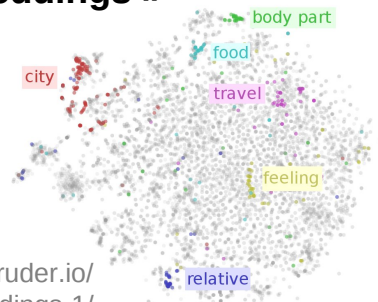
# » « Attention Is All You Need »

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.

Billions of trained weights in the model

Tokens represented in a multidimensional space « embeddings »



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

Encoder-only



Encoder-Decoder



Decoder-only



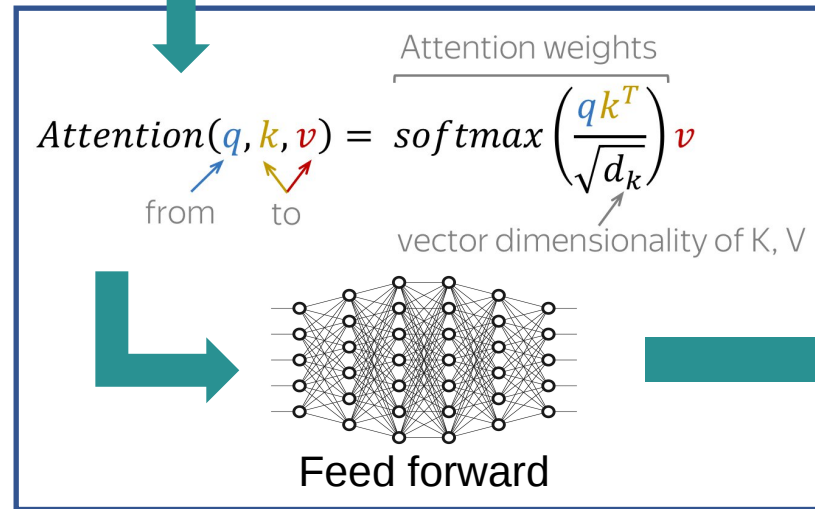
Lors de mon voyage dans la ville de Barcelone, mon parent préféré et moi avons parcouru des rues animées jusqu'à ce que nos jambes, la partie du corps la plus fatiguée, réclament du repos. Nous avons trouvé un petit café, partagé des spécialités locales et ri jusqu'à ce que le sentiment d'épuisement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

Translate into another language

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

Predicts next tokens

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

## ➤ Self-supervised learning to build foundational models



# ➤ Self-supervised learning to build foundational models

## 1. Pre-training

Computationally  
Intensive !!!



# ➤ Self-supervised learning to build foundational models

## 1. Pre-training

Computationally  
Intensive !!!



## 2. Evaluation / refinement / specialisation

Expert guidance  
highly required



INRAE

What genomics can learn from Natural Language Processing?

11/06/2025 / Guillaume GAUTREAU / StatInfOmics team / MaIAGE unit



# ➤ Self-supervised learning to build foundational models

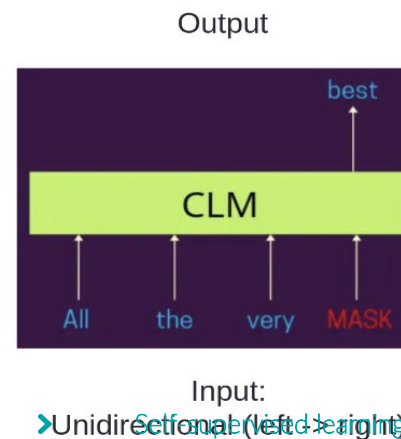
## 1. Pre-training

Computationally  
Intensive !!!

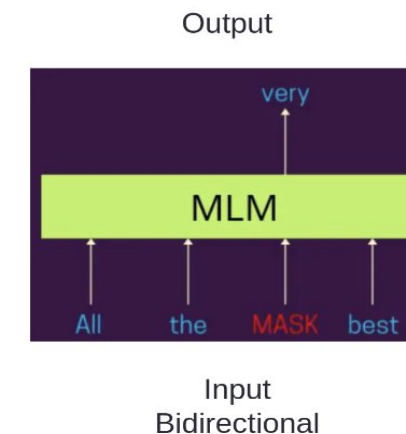


### Two self-supervised learning strategies:

Causal  
Language  
Model  
(e.g.: GPT)



Masked  
Language  
Model  
(e.g.: BERT)



## 2. Evaluation / refinement / specialisation

Expert guidance  
highly required



INRAE

What genomics can learn from Natural Language Processing?

11/06/2025 / Guillaume GAUTREAU / StatInfOmics team / MalAGE unit

# ➤ Self-supervised learning to build foundational models

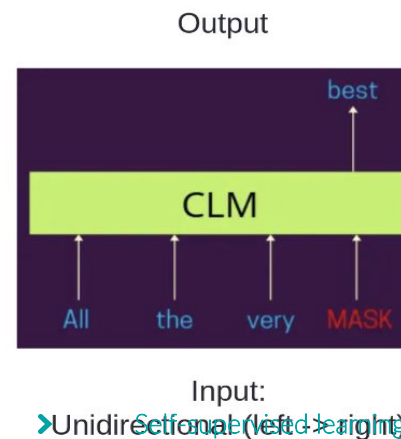
## 1. Pre-training

Computationally  
Intensive !!!

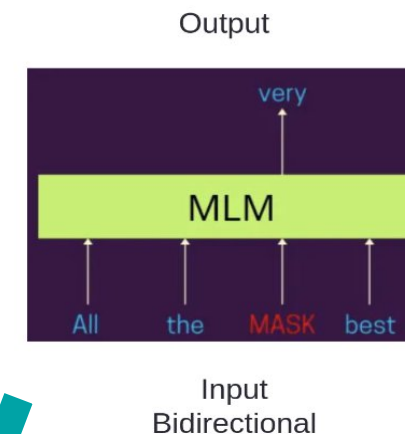


## Two self-supervised learning strategies:

Causal  
Language  
Model  
(e.g.: GPT)



Masked  
Language  
Model  
(e.g.: BERT)



**Both  
leads to:**

**Foundational models**

## 2. Evaluation / refinement / specialisation

Expert guidance  
highly required



INRAE

What genomics can learn from Natural Language Processing?

11/06/2025 / Guillaume GAUTREAU / StatInfOmics team / MaIAGE unit

# ➤ Self-supervised learning to build foundational models

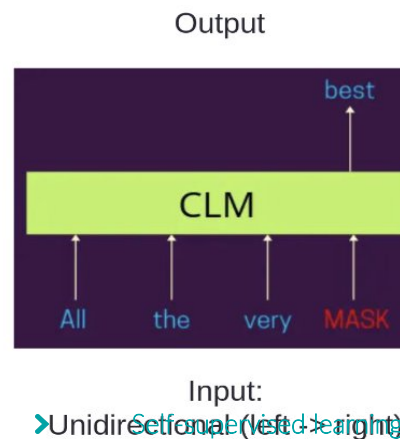
## 1. Pre-training

Computationally  
Intensive !!!

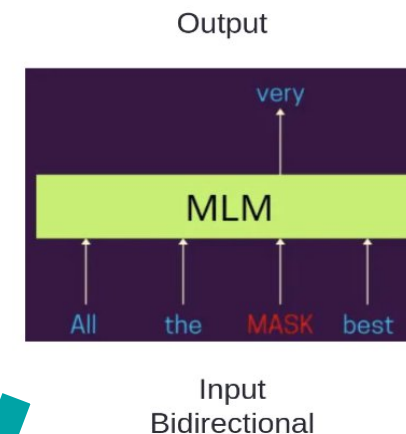


## Two self-supervised learning strategies:

Causal  
Language  
Model  
(e.g.: GPT)



Masked  
Language  
Model  
(e.g.: BERT)

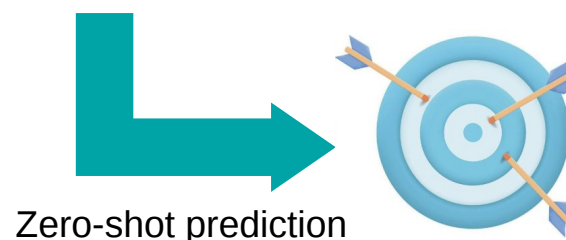


**Both  
leads to:**

**Foundational models**

## 2. Evaluation / refinement / specialisation

Expert guidance  
highly required



INRAE

What genomics can learn from Natural Language Processing?

11/06/2025 / Guillaume GAUTREAU / StatInfOmics team / MaIAGE unit

# ➤ Self-supervised learning to build foundational models

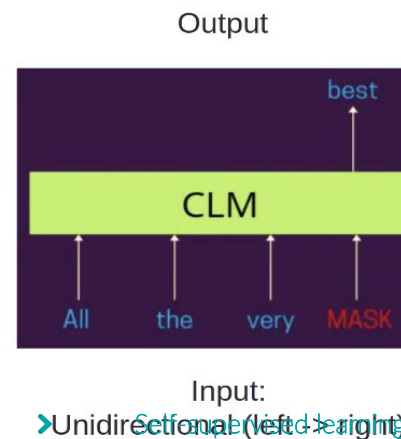
## 1. Pre-training

Computationally Intensive !!!



## Two self-supervised learning strategies:

Causal Language Model  
(e.g.: GPT)



Output



Masked Language Model  
(e.g.: BERT)

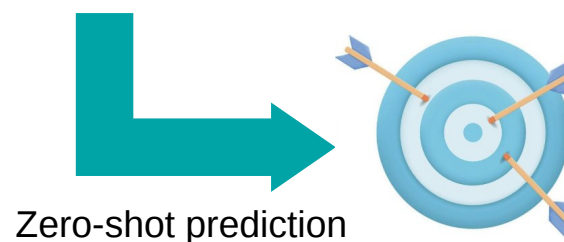
## 2. Evaluation / refinement / specialisation

Expert guidance highly required

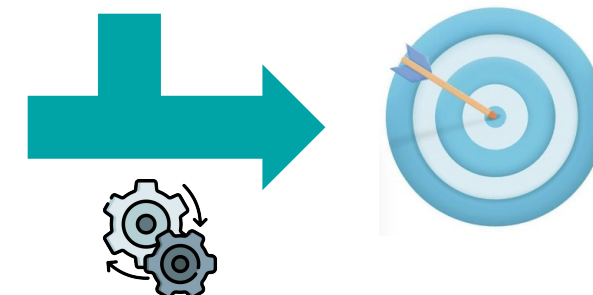


**Both leads to:**

**Foundational models**



Additional Specialized Datasets



- Reinforcement learning
- Fine-tuning on specific tasks



➤ From NLP to genomic: Evo 2

**Towards generalistic Large Language of Life Models (LLLM)**



# ➤ Large Language of Life Models (LLLM)

## 1. English



"THE QUICK BROWN FOX  
JUMPS OVER THE LAZY DOG"

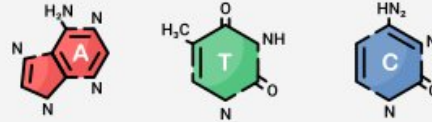
A  
Aback  
Abandon  
...

- BioMedLM
- PubMedBERT
- BioBERT

## 2. DNA



AGGACTGGACCT

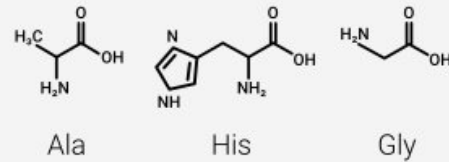


- DNABERT
- DNABERT-2
- DNABERT-S

## 3. Protein

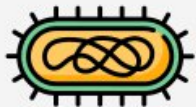


FYERATIMKHWE



- ESM
- ProteinBERT
- ProtMamba

## 4. Genome



*dnaA*  
*rpoB*  
*recA*  
...

- Nucleotide Transformer
- Evo
- Evo2

# ➤ Evo (november 2024)

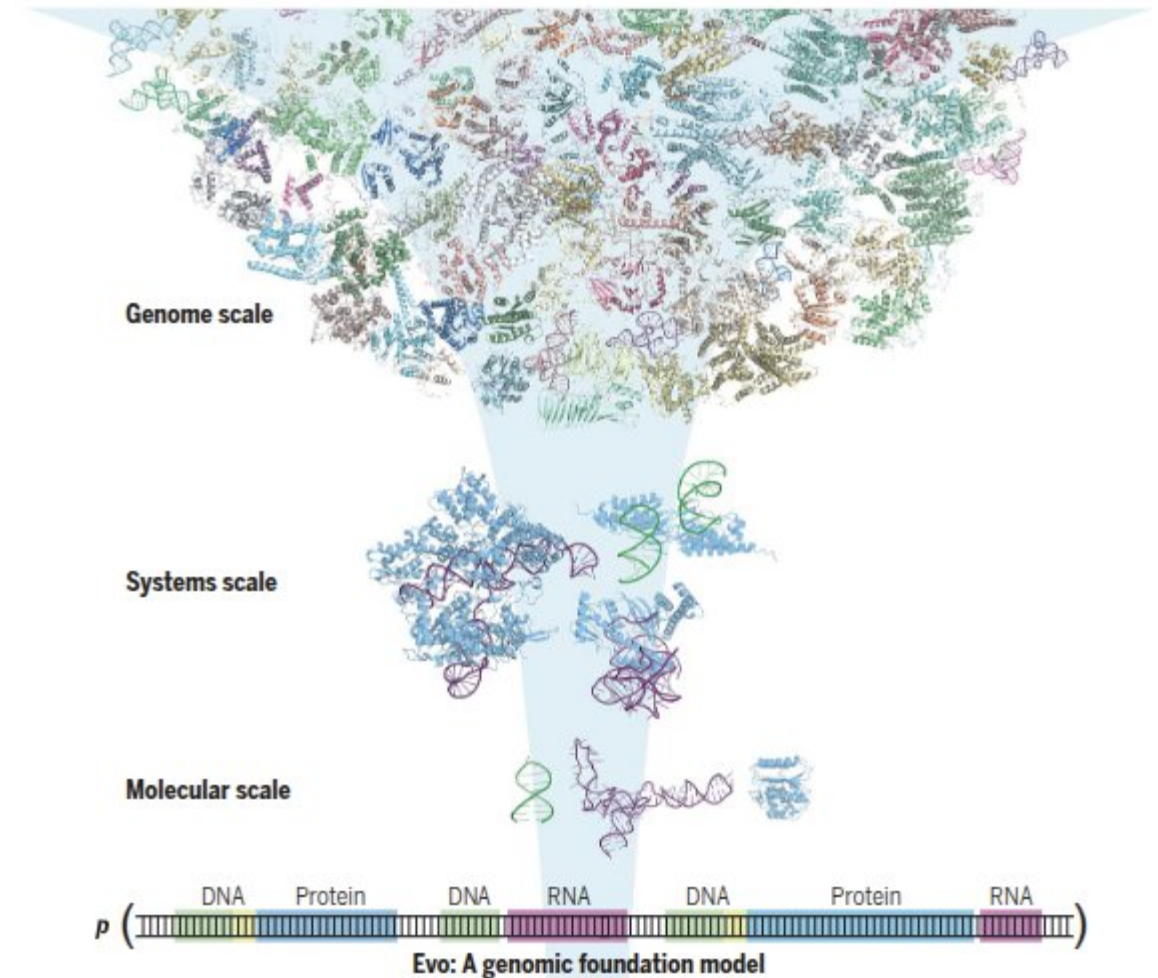
## RESEARCH ARTICLE

### GENERATIVE GENOMICS

## Sequence modeling and design from molecular to genome scale with Evo

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

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.

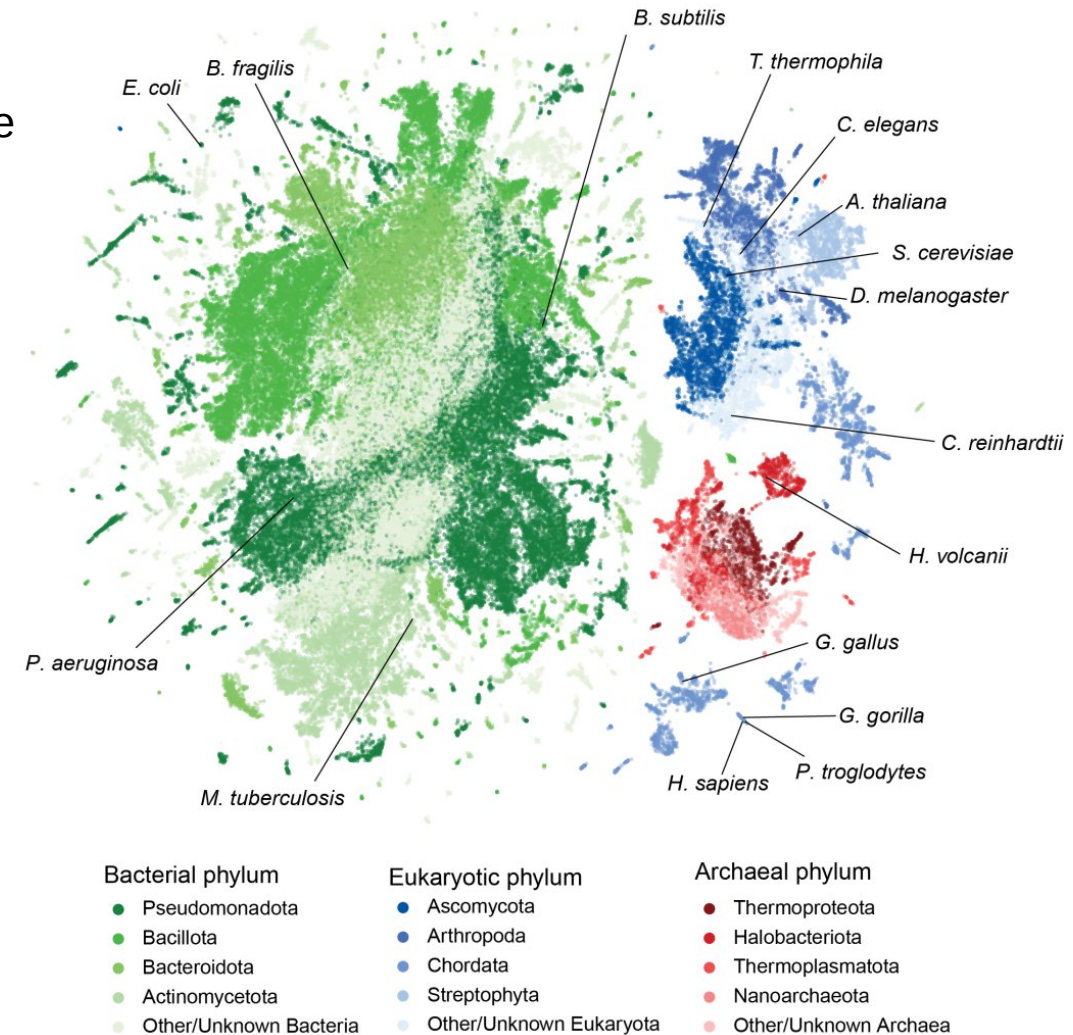
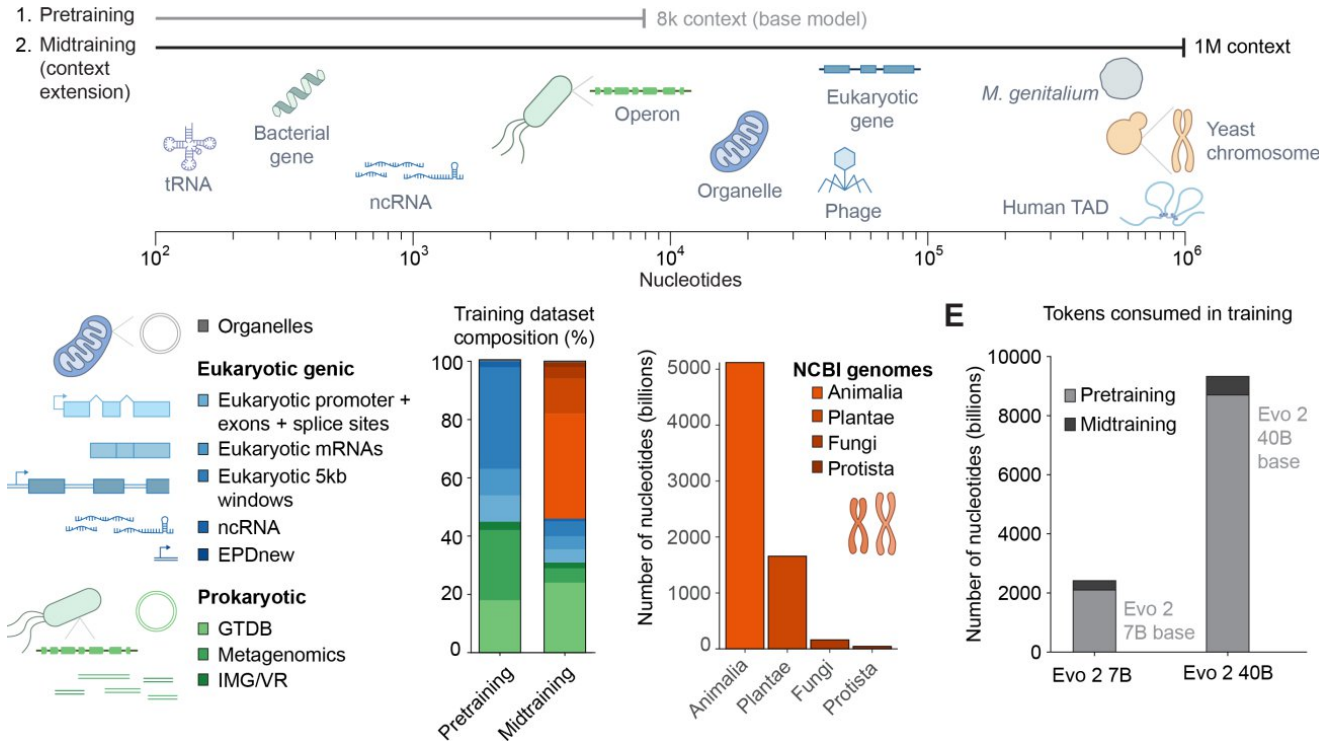


**Evo, a 7-billion-parameter genomic foundation model, learns biological complexity from individual nucleotides to whole genomes.** Trained on 2.7 million raw prokaryotic and phage genome sequences, Evo is naturally multimodal, enabling the codesign of DNA, RNA, and protein molecules that form higher-order functional systems. Evo is also inherently multiscale, enabling prediction and generation tasks at the level of molecules, systems, and genomes.

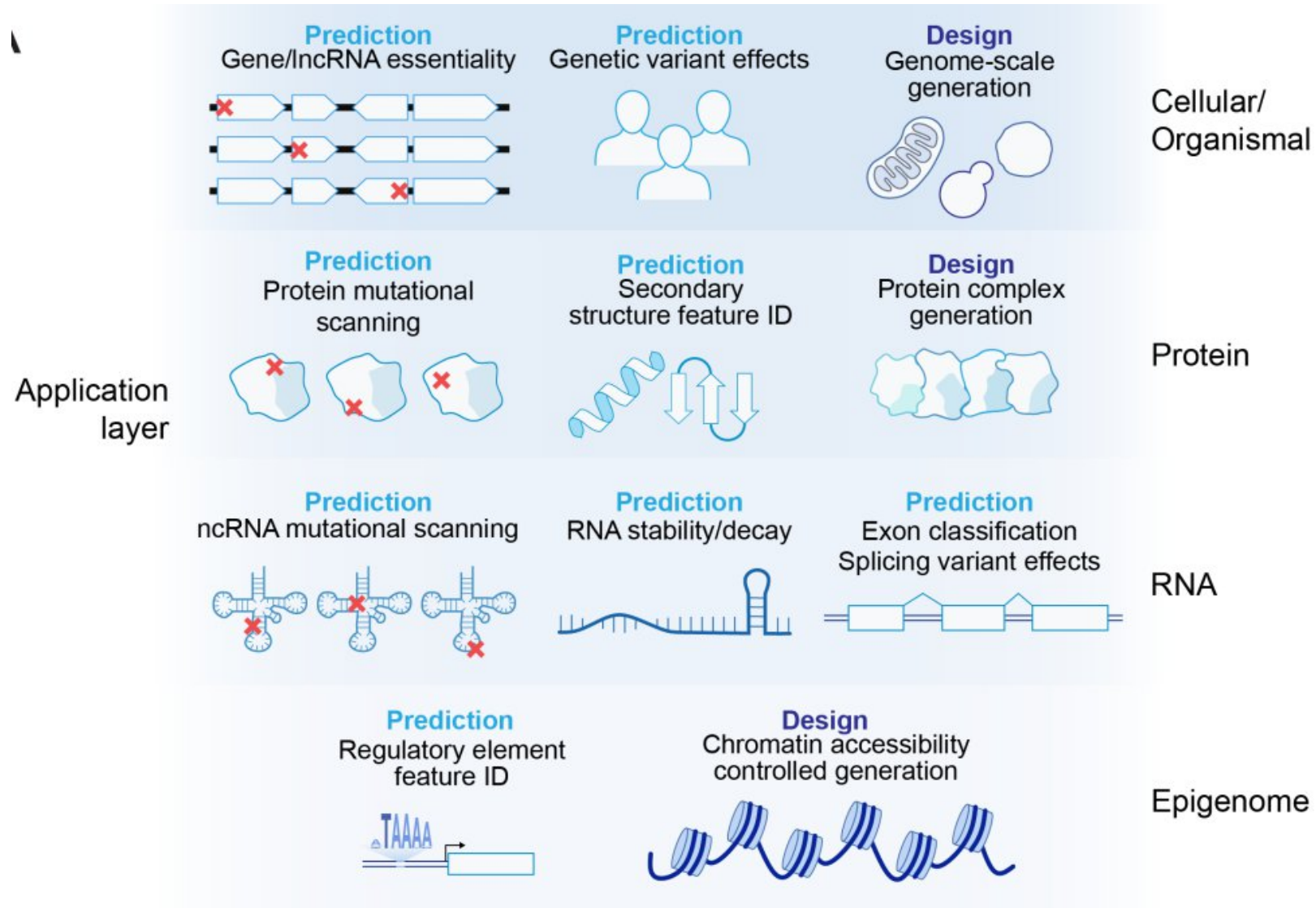


## ➤ Evo2 (preprint released in february)

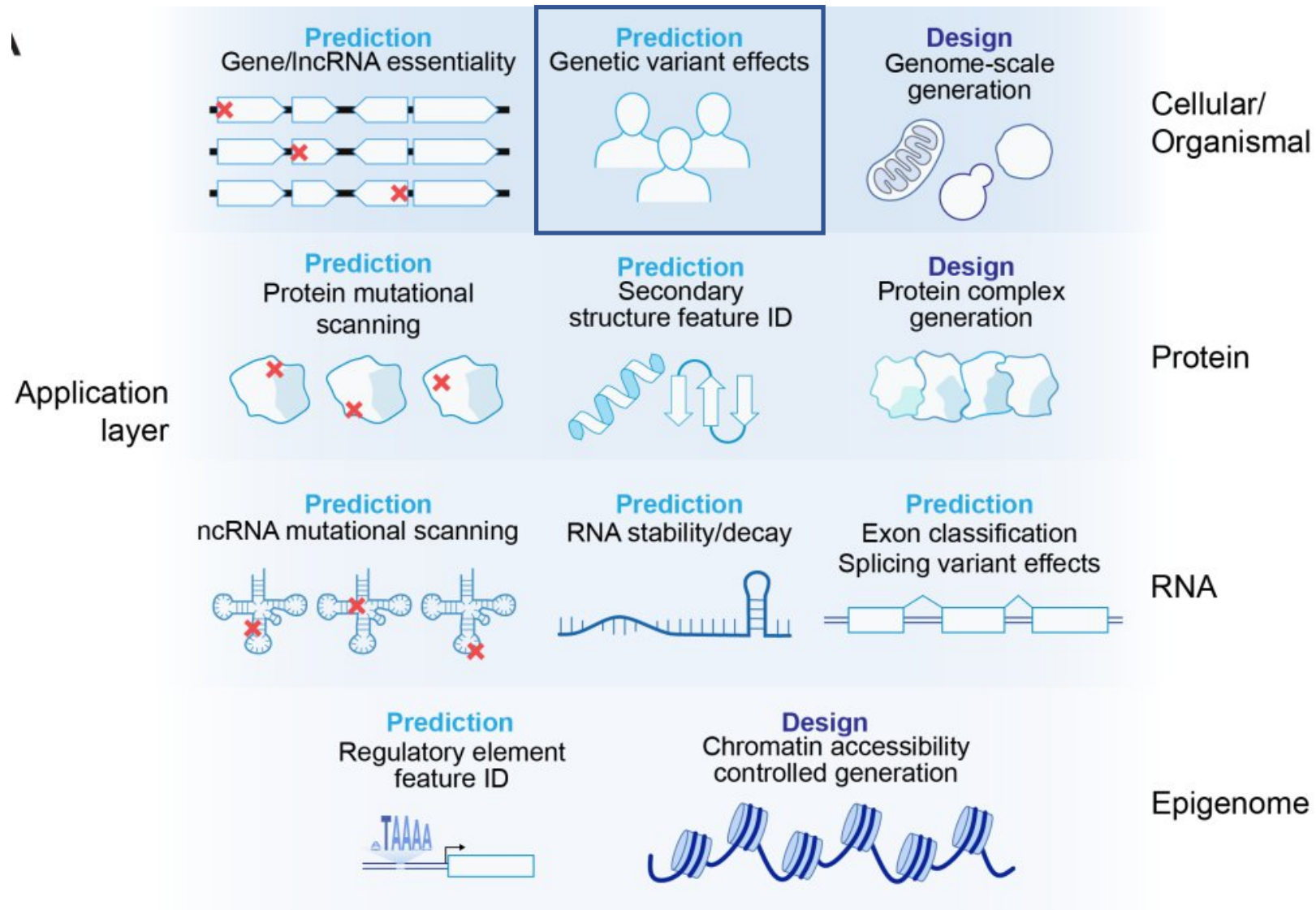
- Built to predict the next token (GPT-like architecture).
- Single-nucleotide resolution
- Uses a different architecture than Transformers to handle genome-scale context efficiently (Hyena, optimized sub-quadratic architecture)
- Training cost ~5-10 M\$ (GPU + electricity only)
- Developed in direct collaboration with NVIDIA to manage computation



# > Applications

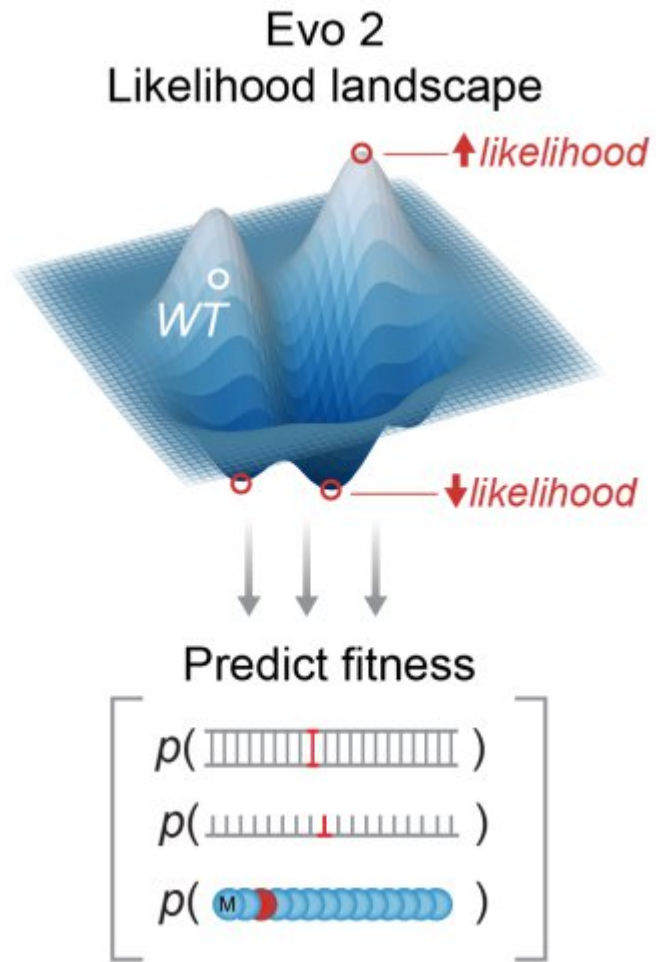


# > Applications

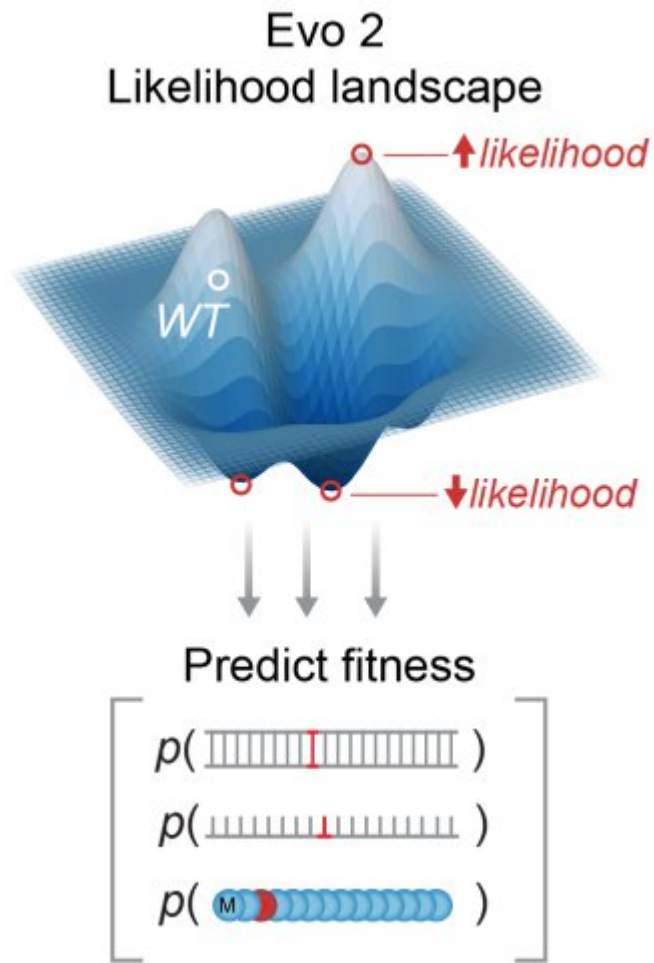




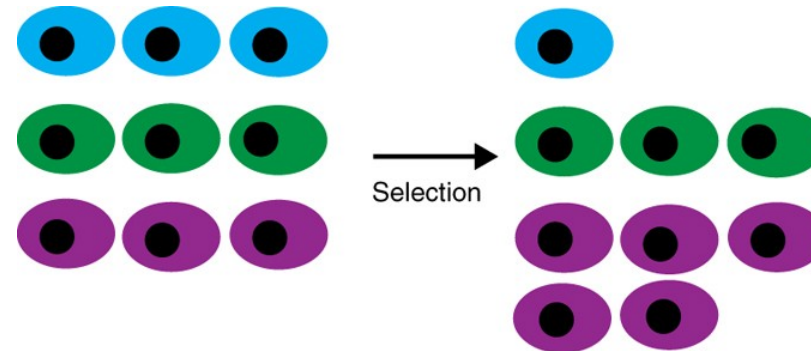
## ➤ Applications: zero-shot prediction of variant effects



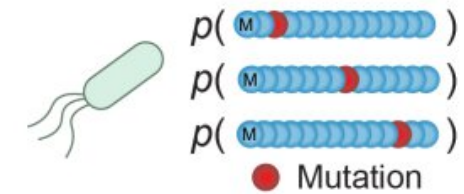
# ➤ Applications: zero-shot prediction of variant effects



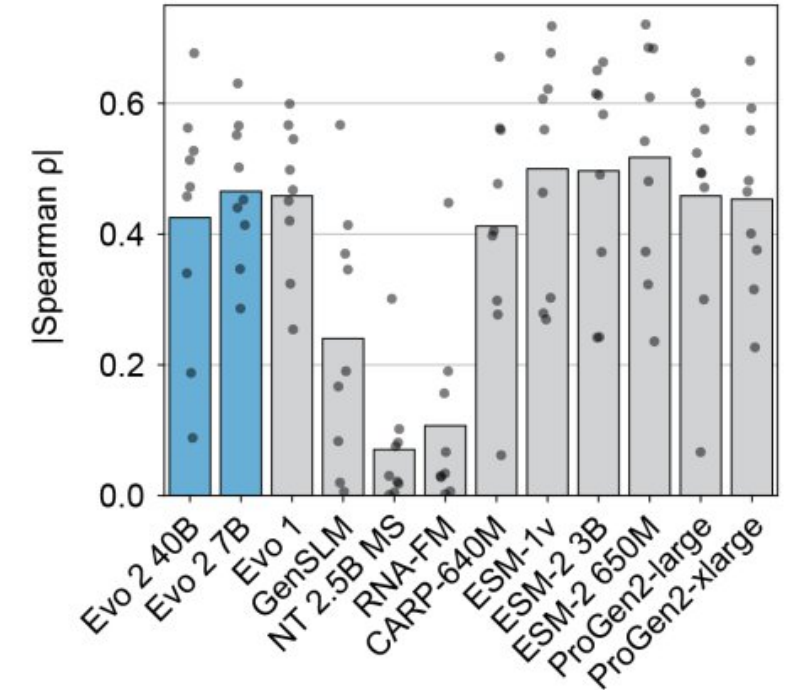
Correlation to Deep Mutation Scanning (DMS) assays  
(<https://proteingym.org/>)



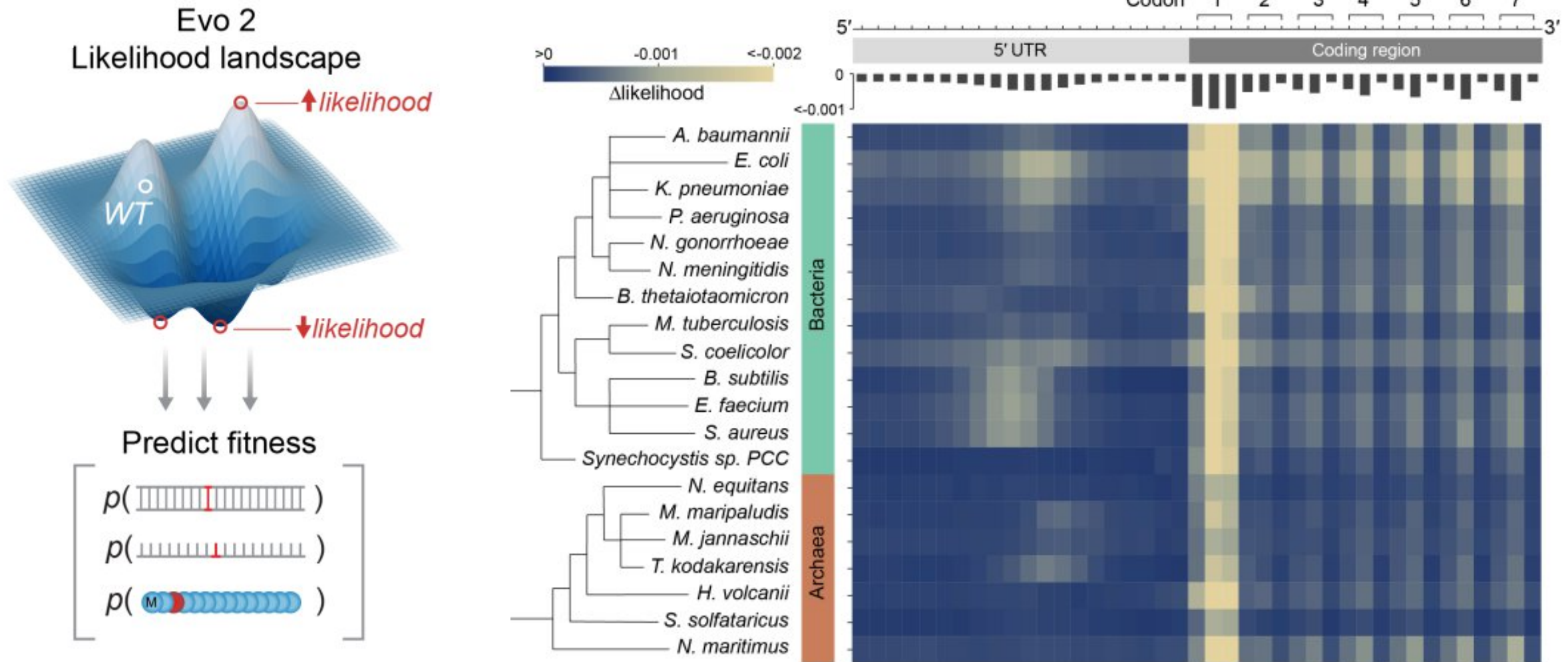
Variant	Mutation	Counts (input)	Counts (selected)	Functional score
Blue	A60P	3	1	0.33
Green	WT	3	3	1
Purple	S36T	3	5	1.67



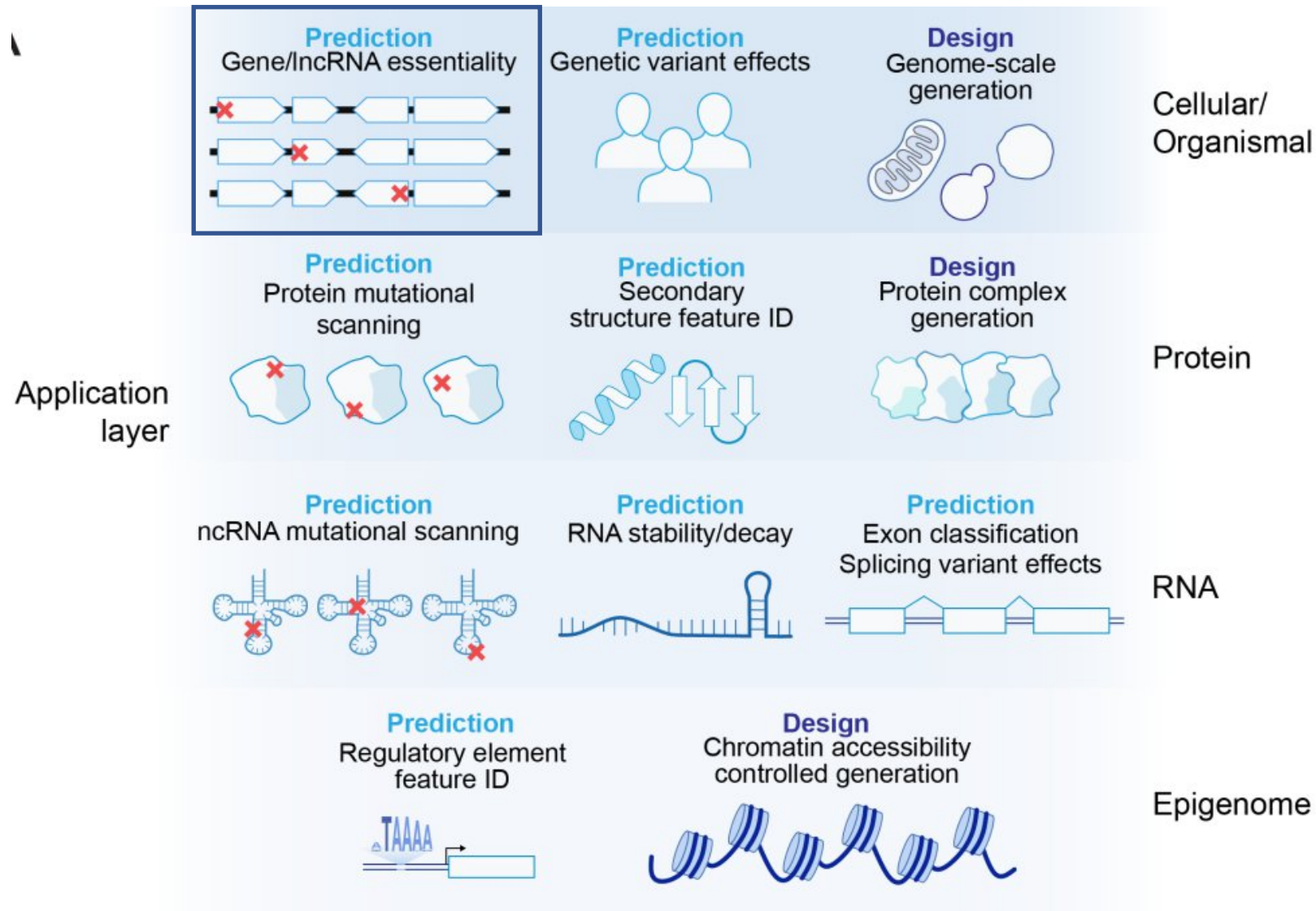
Zero-shot bacterial protein fitness prediction



## ➤ Applications: zero-shot prediction of variant effects

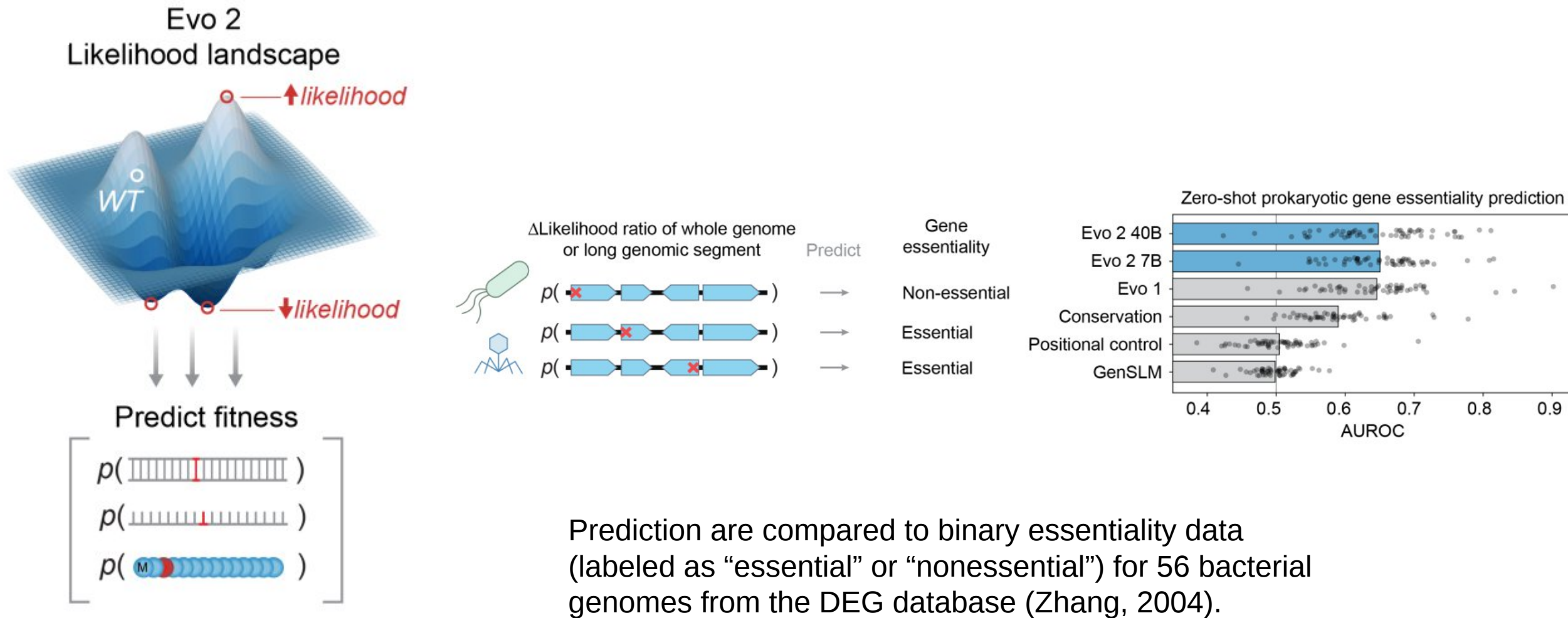


# > Applications

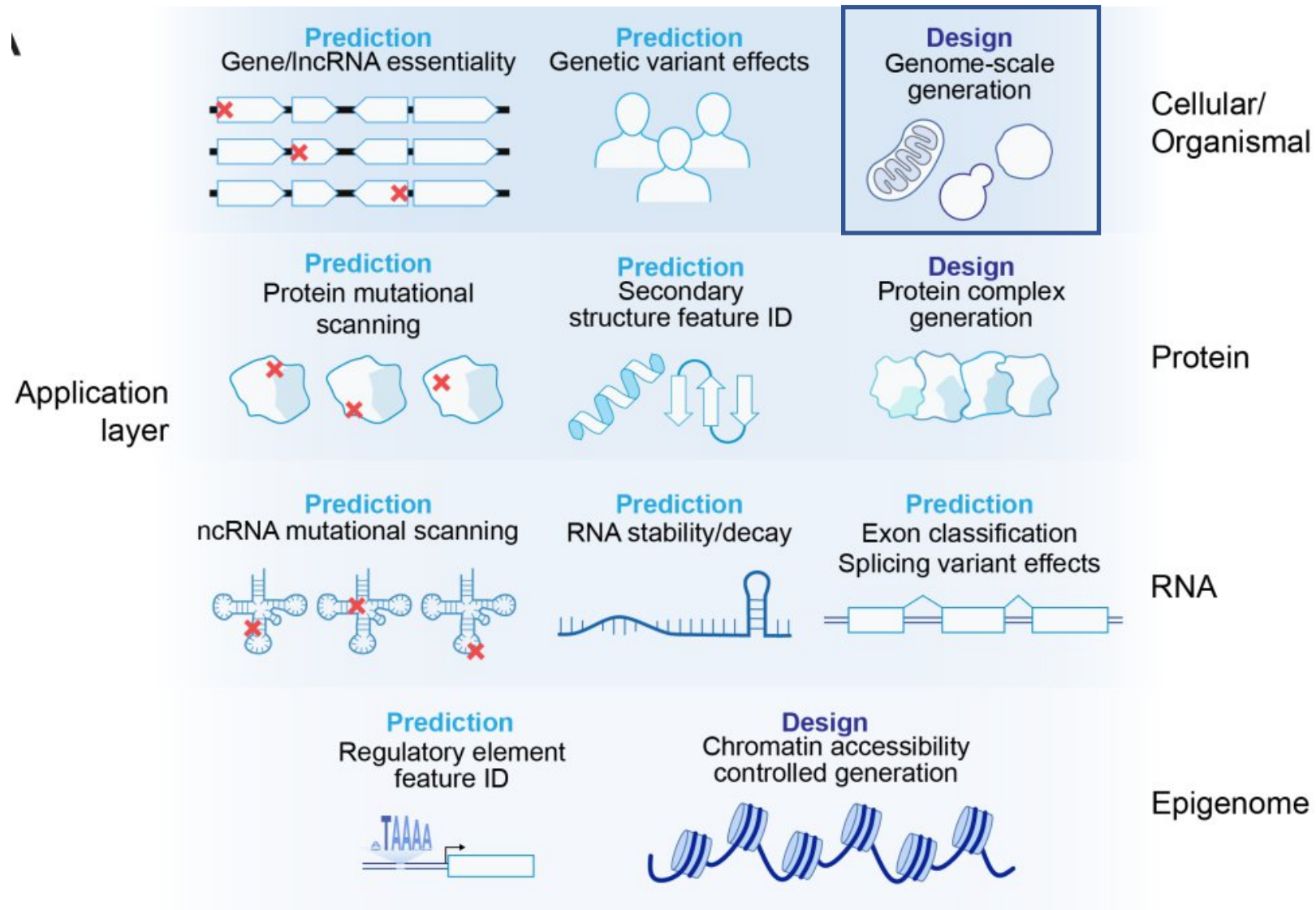




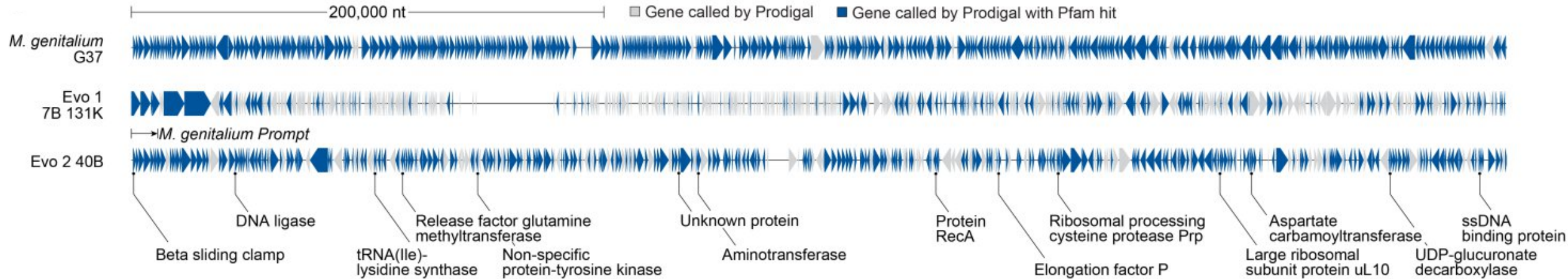
## ➤ Applications: prediction of gene essentiality



# > Applications



## ➤ Generative applications: DNA designing up to genome scale



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

## ➤ Conclusion on Evo2

- Highly versatile model. Idea : reinforcement learning to generate microbial ecosystems (synthetic ecology) ?
- Openly available, but requires an Nvidia H100 GPU (~30k€) to run locally
- User-friendly interface available: <https://arcinstitute.org/tools/evo/evo-designer>
- Ethical considerations to anticipate (biosecurity, misuses)
  - Bioterrorism (human pathogenic viruses weren't included in Evo2)
  - Fake dataset pollution, science integrity issues...
- Partially interpretable via Sparse AutoEncoders (explored in the Evo2 article)
- Personal opinion from our first tries (M2 intern): still a big margin of improvement (at least on Evo1)
- Next step: LLM x LLLM : « Talk to your genome »
  - ChatNT (published last week in Nature Machine Intelligence)
  - BioReason (Evo2 x Qwen3), preprint on Arxiv
- Retrieval Augmented Generation ?



	Question	Ground Truth (KEGG)	BioReason's Output
KEGG Data Point	<p>Question</p> <p>&lt;[im_start]&gt;user</p> <p>Chromosome Number: 17</p> <p>Network Definition of the Pathway:</p> <p>Actin(monomeric) // PFN1* // Actin(filamentous)</p> <p>Genes in the Pathway:</p> <p>ACTB; actin beta   ACTG1; actin gamma 1   PFN1; profilin 1</p> <p>Given this context, what is the biological effect of this PFN1 allele, specifically what disease does this contribute to?</p> <p>&lt;[im_end]&gt;&lt;[im_start]&gt;assistant</p>	<p>Step 1: The variant KEGG_800 represents a C&gt;G substitution at position 4945969 on chromosome 17, occurring in the PFN1 gene which encodes profilin-1 protein.</p> <p>Step 2: This nucleotide change ... potentially affecting a functionally important domain.</p> <p>Step 10: The progressive loss of motor neurons results in muscle weakness, atrophy, and eventually paralysis - the hallmark clinical features of ALS.</p> <p>The C&gt;G mutation in the PFN1 gene ... This cytoskeletal dysfunction contributes to axonal transport defects and motor neuron degeneration, ultimately leading to familial Amyotrophic Lateral Sclerosis (ALS).</p>	<p>&lt;think&gt;</p> <p>... Step 3: This mutation likely alters the protein structure or function of profilin-1, potentially affecting its ability to bind actin monomers or regulate actin polymerization.</p> <p>Step 10: This specific mutation contributes to ALS through a cytoskeletal/proteostasis pathway...</p> <p>This C&gt;G mutation in the PFN1 gene likely disrupts profilin-1 function, impairing its ability to regulate actin dynamics by binding monomeric actin and promoting polymerization. The resulting cytoskeletal dysfunction affects motor neuron axonal transport and cellular homeostasis, contributing to the development of familial amyotrophic lateral sclerosis (ALS).</p> <p>&lt;/think&gt;</p> <p>Answer: amyotrophic lateral sclerosis (ALS)&lt;[im_end]&gt;</p>



## ➤ Conclusion on Evo2

- Highly versatile model. Idea : reinforcement learning to generate microbial ecosystems (synthetic ecology) ?
- Openly available, but requires an H100 GPU (~30k€) to run locally
- User-friendly interface available: <https://arcinstitute.org/tools/evo/evo-designer>
- Ethical considerations to anticipate (biosecurity, misuses)
  - Bioterrorism (human pathogenic viruses weren't included in Evo2)



KEGG Data Point

### Question

#### Question

<|im\_start|>user

Chromosome Number: 17

#### Network Definition of the Pathway:

Actin(monomeric) // PFN1\* // Actin(filamentous)

#### Genes in the Pathway:

ACTB; actin beta | ACTG1; actin gamma 1 | PFN1; profilin  
1

Given this context, what is the **biological effect** of this **PFN1 allele**, specifically what **disease** does this contribute to?

<|im\_end|><|im\_start|>assistant

### Ground Truth (KEGG)

**Step 1:** The variant KEGG\_800 represents a C>G substitution at position 4945969 on chromosome 17, occurring in the **PFN1** gene which encodes **profilin-1 protein**.

**Step 2:** This nucleotide change ... potentially affecting a **functionally important domain**.

⋮

**Step 10:** The progressive loss of motor neurons results in **muscle weakness, atrophy**, and eventually **paralysis** - the hallmark clinical features of **ALS**.

The **C>G mutation** in the **PFN1** gene ... This cytoskeletal dysfunction contributes to **axonal transport defects** and **motor neuron degeneration**, ultimately leading to familial **Amyotrophic Lateral Sclerosis (ALS)**.

### BioReason's Output

<think>

... **Step 3:** This mutation likely **alters the protein structure or function** of **profilin-1**, potentially affecting its ability to bind actin monomers or regulate actin polymerization.

⋮

**Step 10:** This specific mutation contributes to ALS through a cytoskeletal/proteostasis pathway...

⋮

This **C>G mutation** in the **PFN1** gene likely **disrupts profilin-1 function, impairing its ability to regulate actin dynamics** by binding monomeric actin and promoting polymerization. The resulting cytoskeletal dysfunction affects motor neuron axonal transport and cellular homeostasis, contributing to the development of familial **amyotrophic lateral sclerosis (ALS)**.

</think>

**Answer: amyotrophic lateral sclerosis (ALS)**<|im\_end|>