TECHNISCHE UNIVERSITÄT DRESDEN

DRESDEN concept

DNA language models in biomedical research

Anna Poetsch

Research Group „Biomedical Genomics“, Biotechnology Center TU Dresden, NCT Dresden, and CSBD

@apoetsch

Outline

- What are Large Language Models?
- A short glimpse into a transformer foundation model
- DNA language models
- The inner workings of GROVER
- applications in biomedical research

Pre-training and fine-tuning

Foundation models train language on large corpora of data

Parameters:	1.5B	175B	400B+	562B	1.7 T(?)
Year:	2019	2020	soon	2023	2023

BIG

expensive

rigid

Fine-tuning

- assistant
- image generation
- translation
- speech recognition
- ...

"DallE2, please give me an illustration of damaged DNA in comic style":

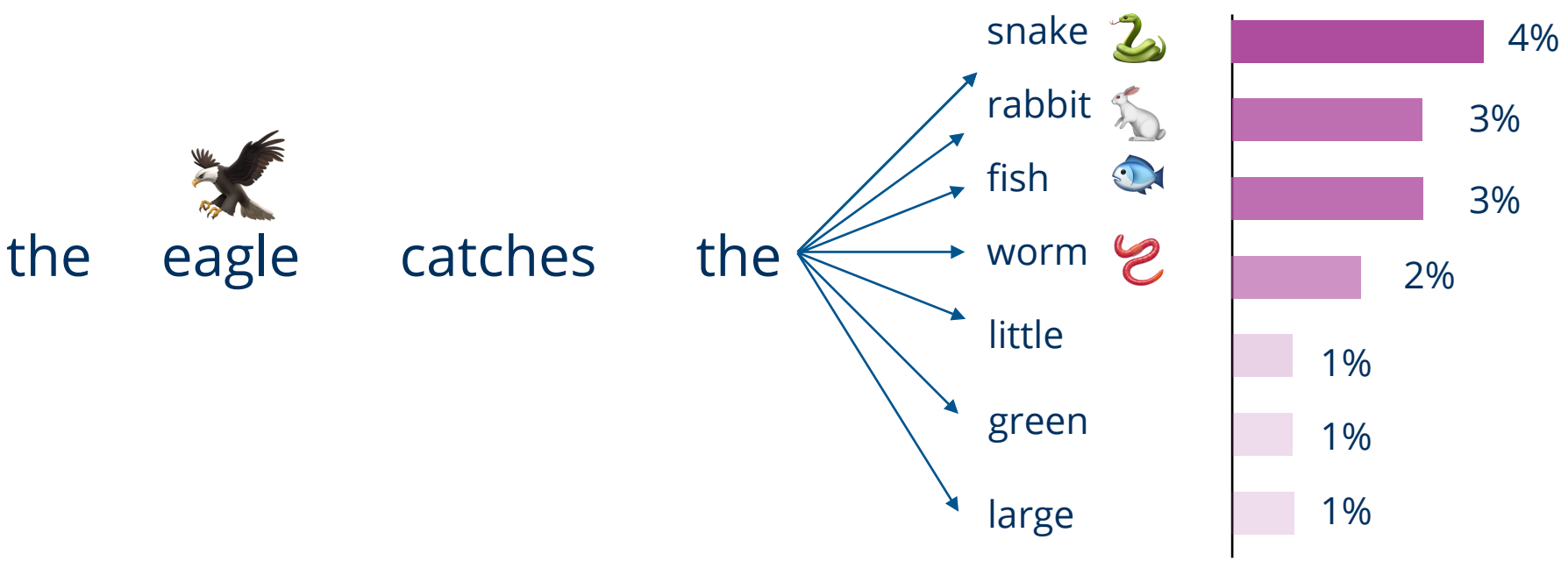
efficient

flexible

cheap

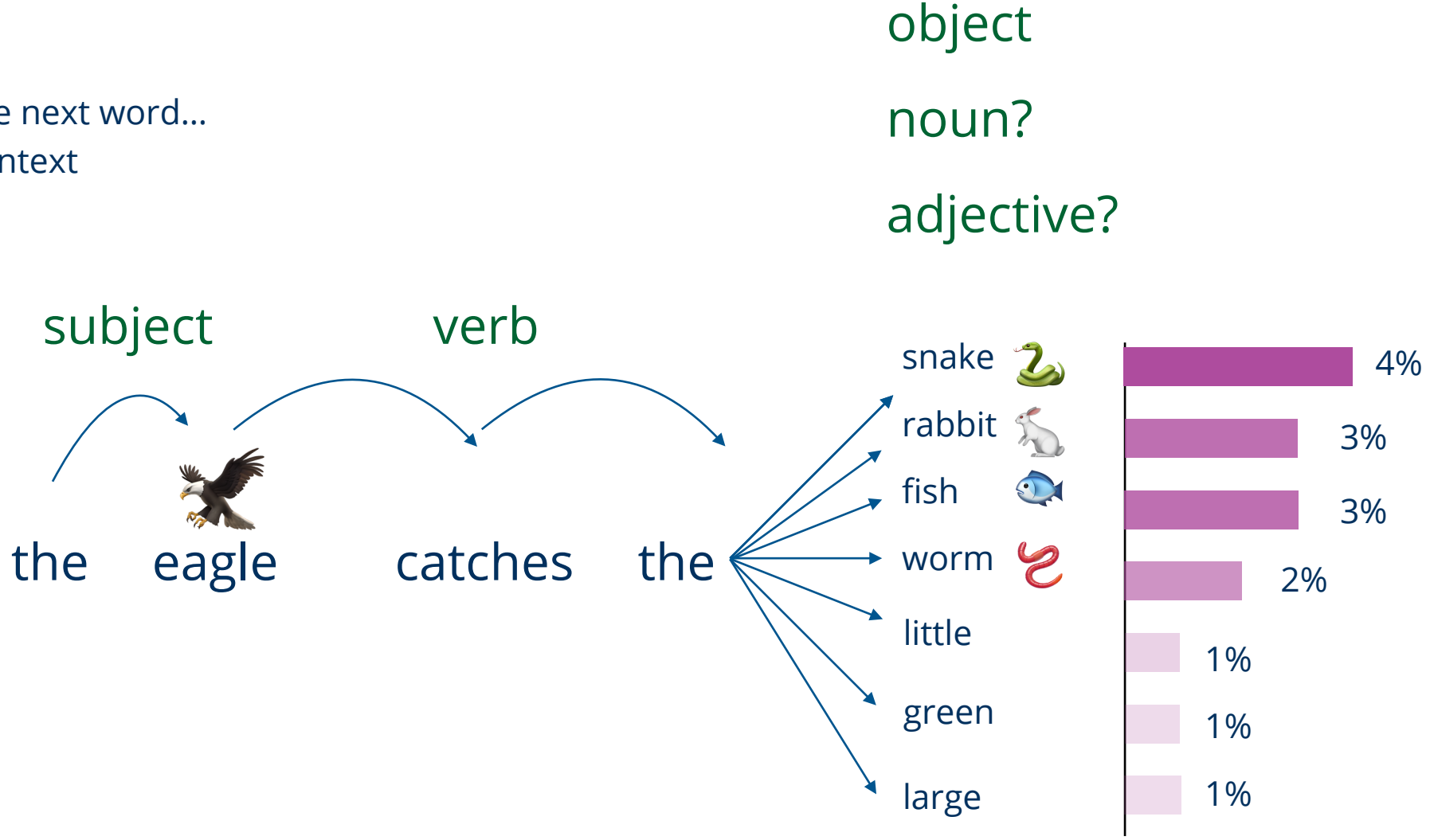
Large Language Models are trained on next-token prediction

Predicting the next word...



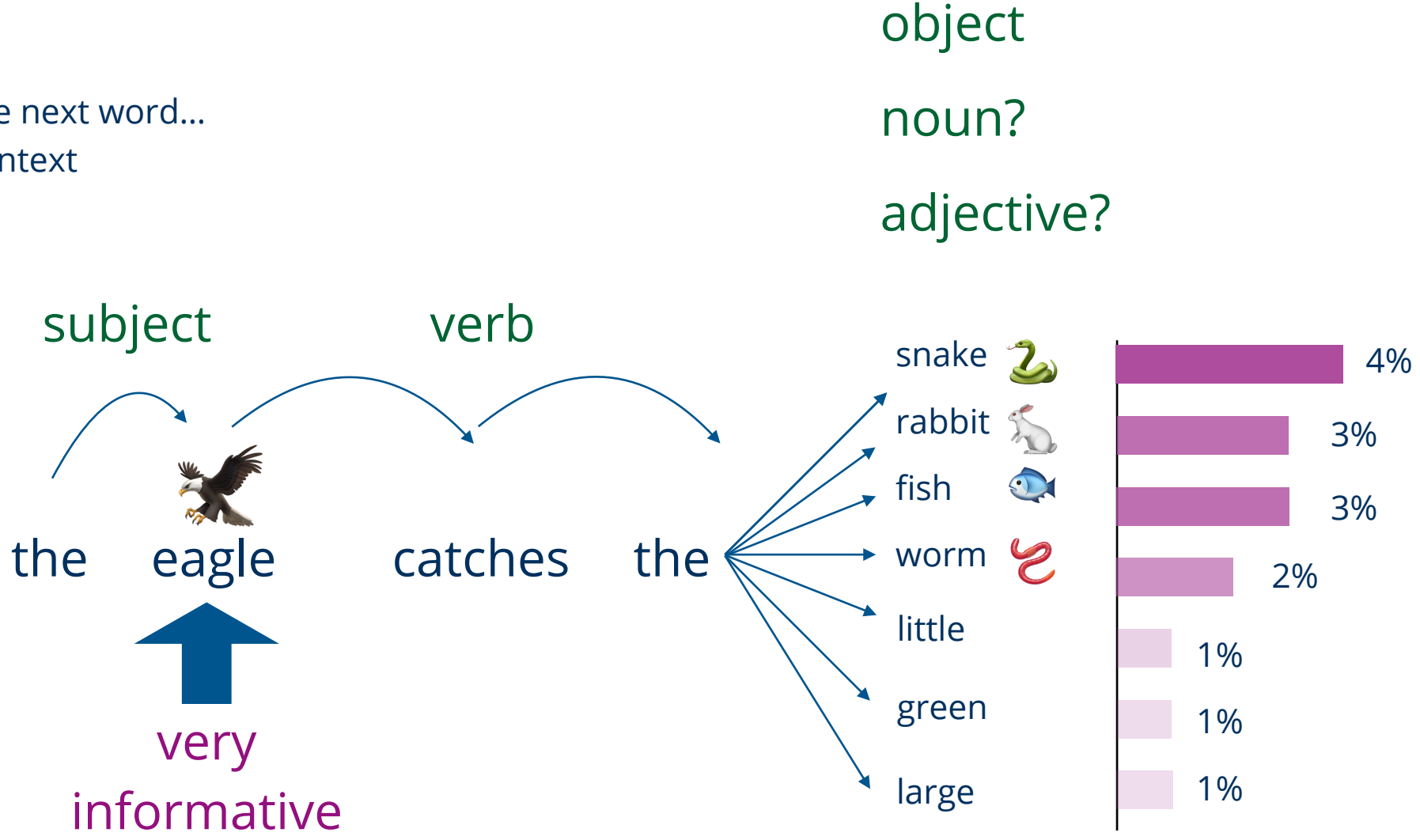
Large Language Models

Predicting the next word...
...requires context



Large Language Models

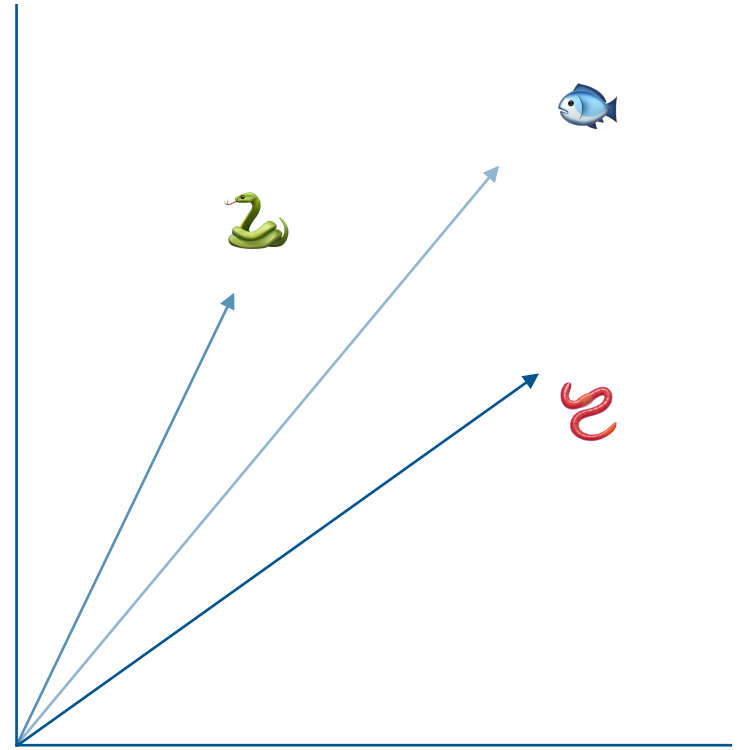
Predicting the next word...
...requires context



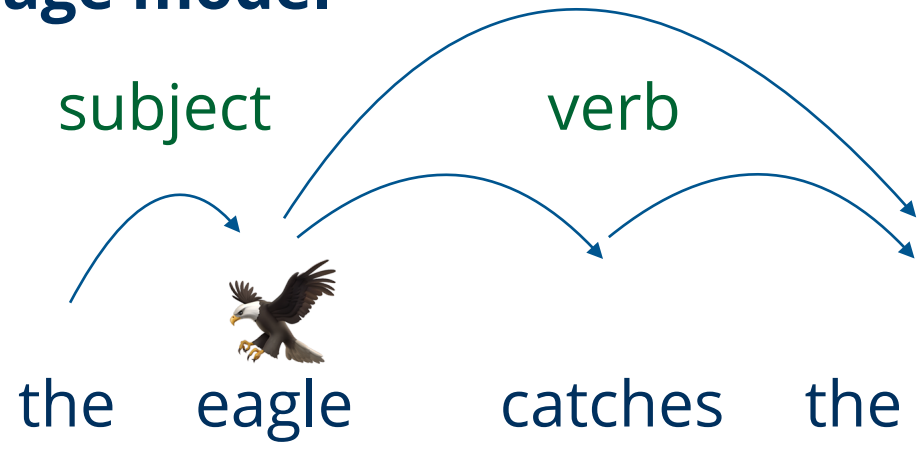
The embedding of words/ tokens

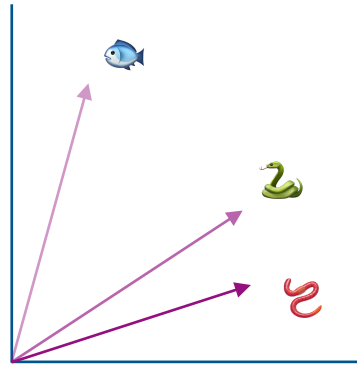
Tokens are assigned a vector, which places them into a multi-dimensional space

the	eagle	catches	the	???
5.4	3.2	0.2	5.4	
2.3	1.4	0.3	2.3	
1.2	0.4	5.6	1.2	
.	.	.	.	
.	.	.	.	
.	.	.	.	
6.4	1.2	3.2	6.4	
0.4	8.6	4.1	0.4	



Training a large language model





Input embedding

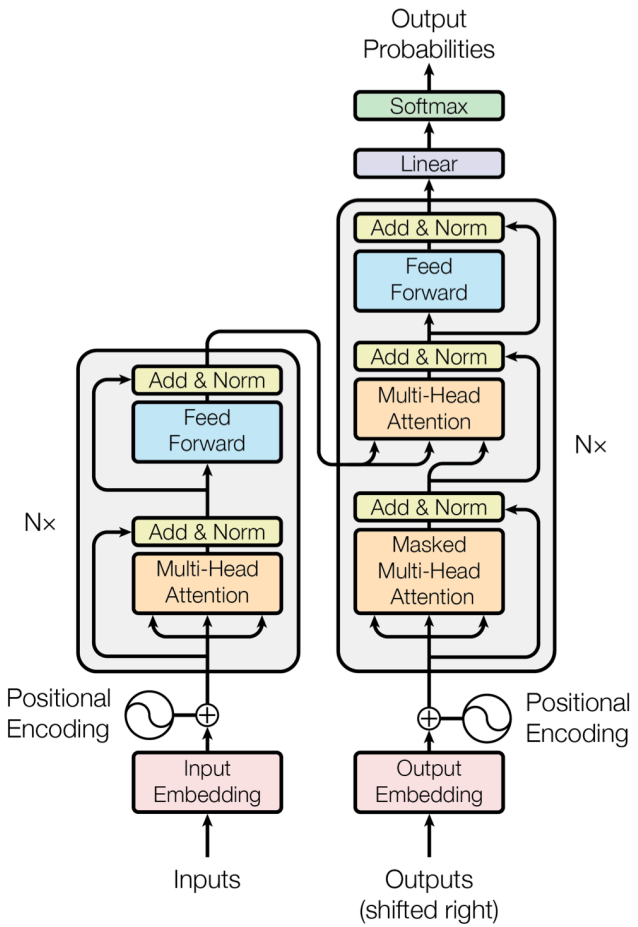
5.4	3.2	0.2	5.4
2.3	1.4	0.3	2.3
1.2	0.4	5.6	1.2
.	.	.	.
.	.	.	.
.	.	.	.
6.4	1.2	3.2	6.4
0.4	8.6	4.1	0.4

context

Trained embedding

3.2	6.3	0.6	3.7
4.6	0.4	1.7	5.3
5.2	8.2	4.8	7.2
.	.	.	
.	.	.	
.	.	.	
5.9	3.2	3.9	2.1
0.3	0.4	6.2	0.9

Context is grasped through transformers: “Attention is all you need”



GPT-3

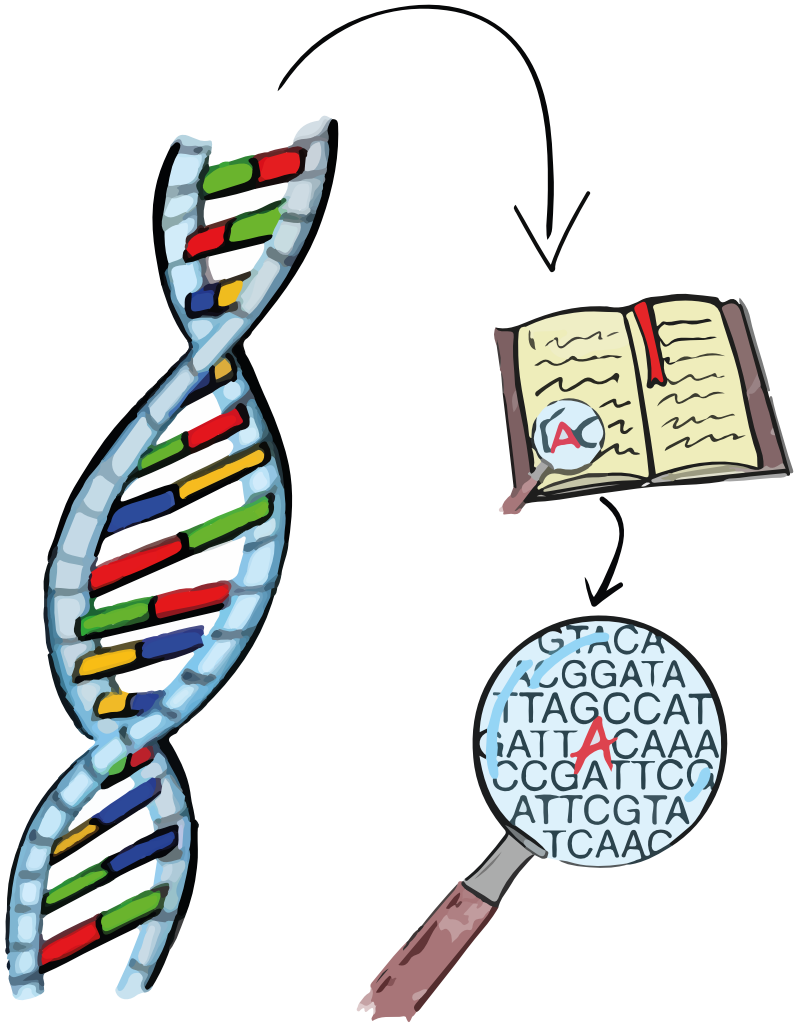
Total parameters:
175,181,291,520
27,938 matrices

	Embedding	12,288 d embed	* 50,257 n vocab			=	617,558,016
attention	Key	128 d query	* 12,288 d embed	* 96 n heads	* 96 n layers	=	14,495,514,624
attention	Query	128 d query	* 12,288 d embed	* 96 n heads	* 96 n layers	=	14,495,514,624
attention	Value	128 d query	* 12,288 d embed	* 96 n heads	* 96 n layers	=	14,495,514,624
attention	Output	12,288 d embed	* 128 d value	* 96 n heads	* 96 n layers	=	14,495,514,624
	Up-projection	49,152 n neurons	* 12,288 d embed	* 96 n layers		=	57,982,058,496
	Down-projection	12,288 d embed	* 49,152 n neurons	* 96 n layers		=	57,982,058,496
	Unembedding	50,257 n vocab	* 12,288 d embed			=	617,558,016

$$\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V$$

https://arxiv.org/abs/1706.03762

Can we treat DNA as if it were text?



DNABERT-2

GROVER

HyenaDNA

Nucleotide Transformer

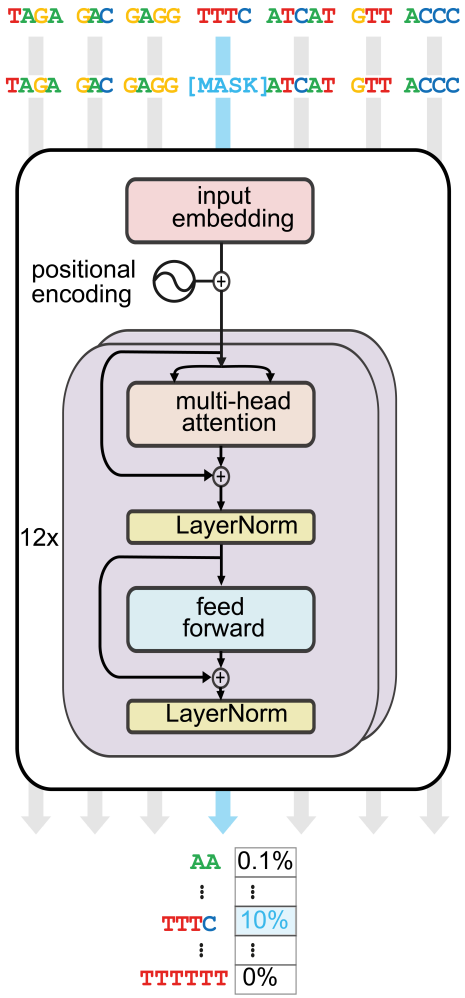
Proformer

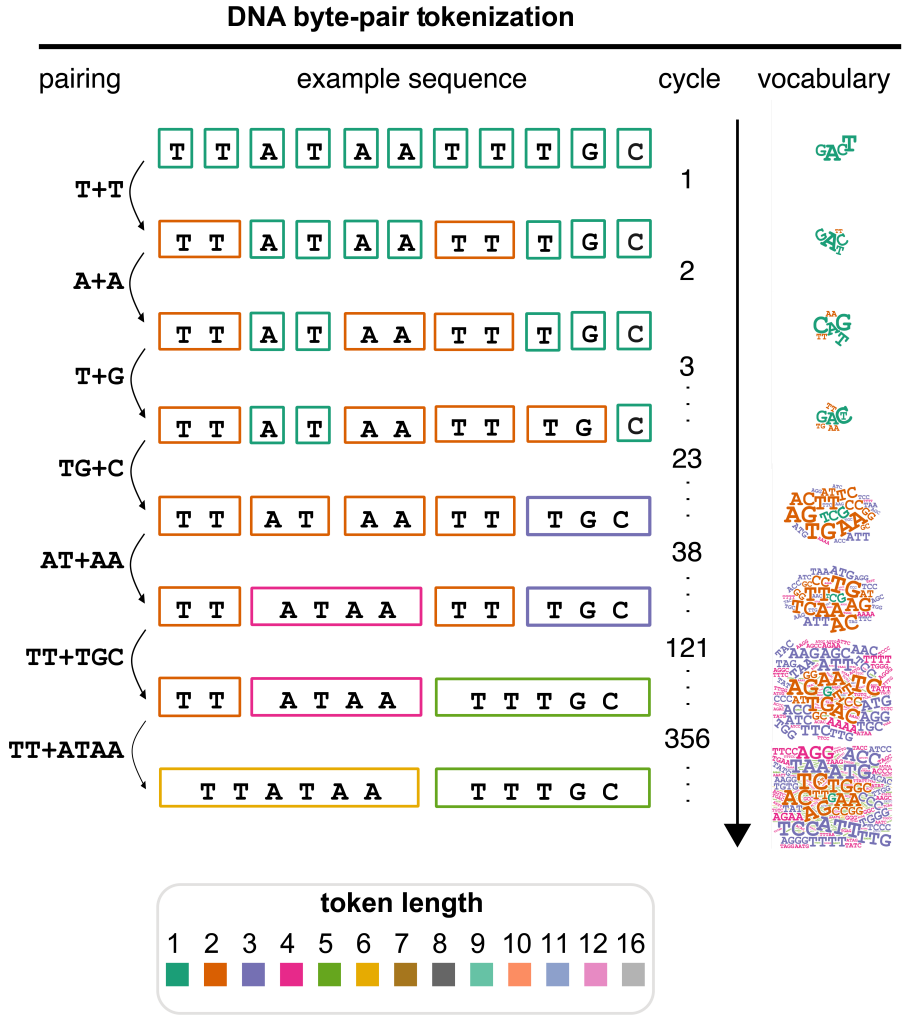
Enformer

⋮

GROVER (Genome Rules Obtained Via Extracted Representations)

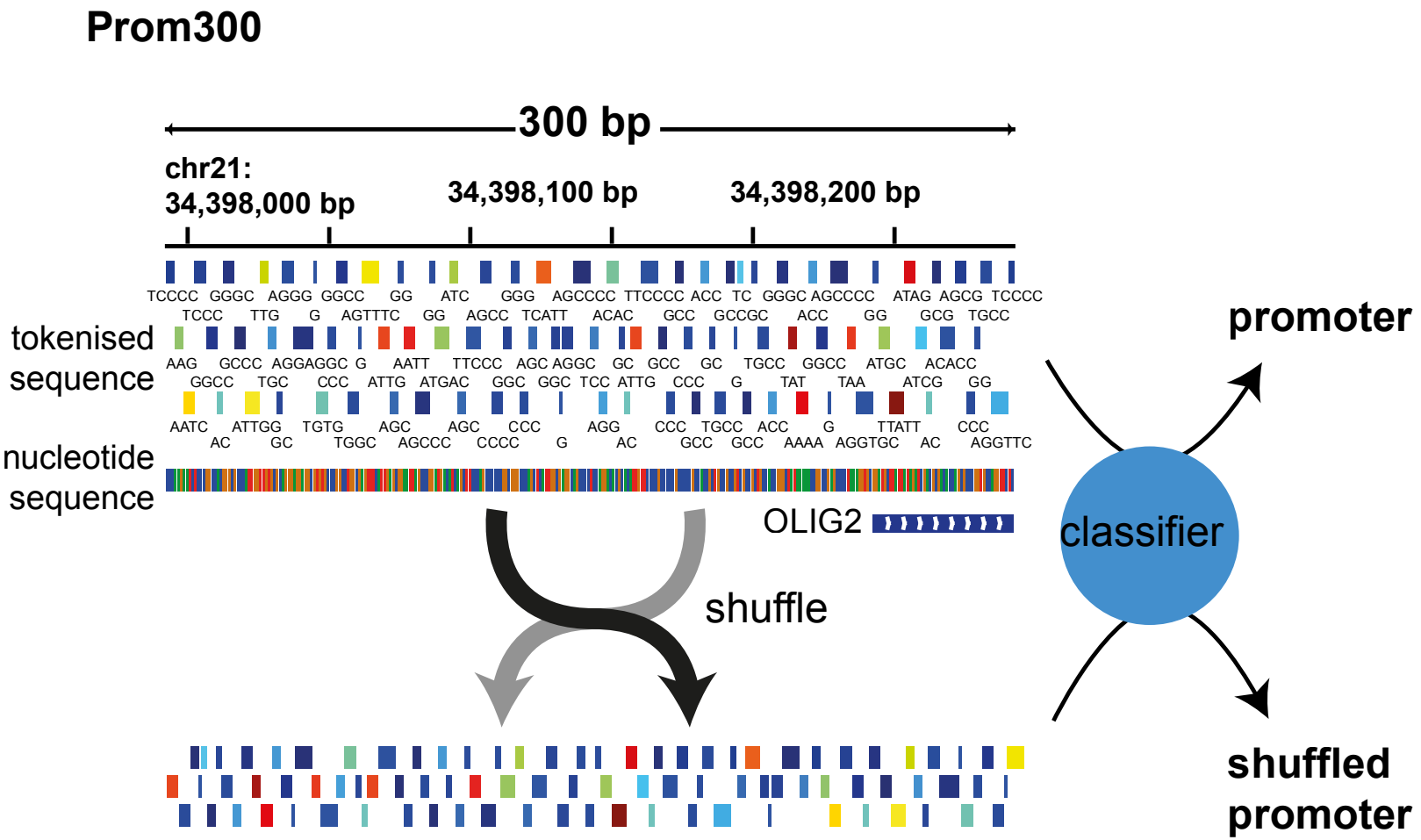
trains on masked token prediction

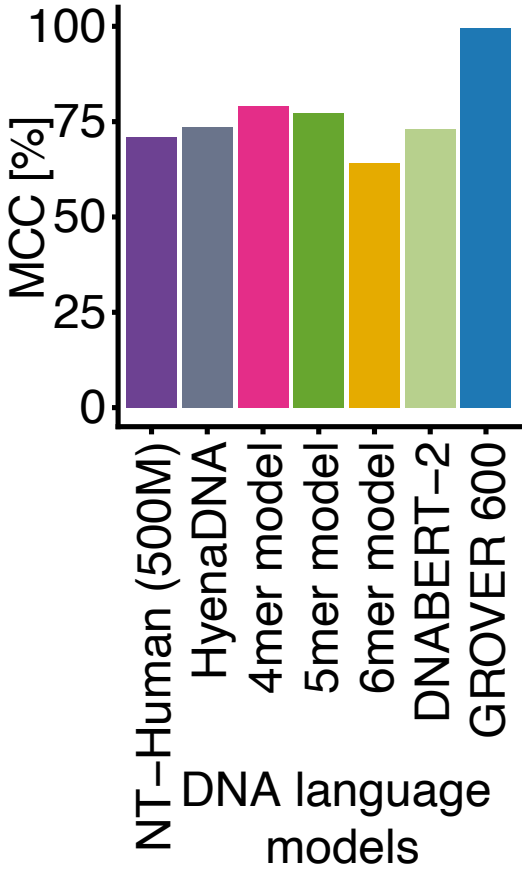




Sanabria *et al* Nature Machine Intelligence (accepted) & Sanabria *et al* BMC Bioinformatics (accepted)

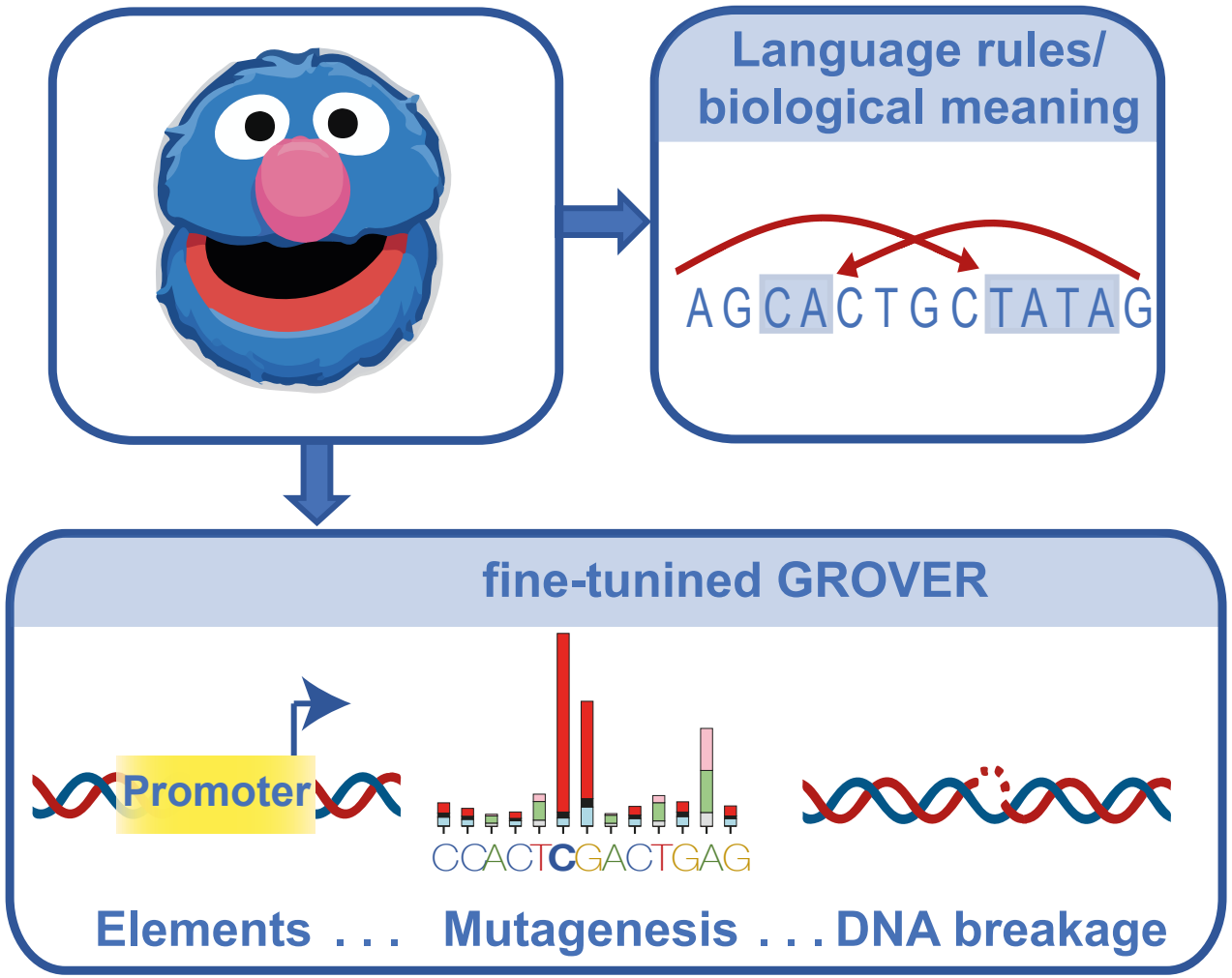
Fine-tuning tasks to understand genome biology





Sanabria *et al* Nature Machine Intelligence (accepted)

What can we use GROVER for?

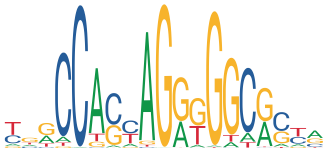


The model:

https://huggingface.co/PoetschLab/GROVER

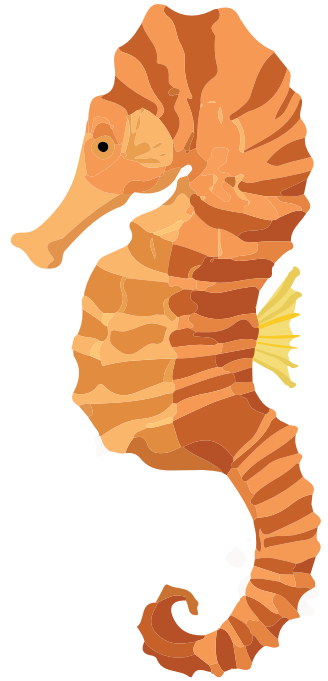
Tutorial for GROVER-based CTCF binding prediction:

https://doi.org/10.5281/zenodo.8373158

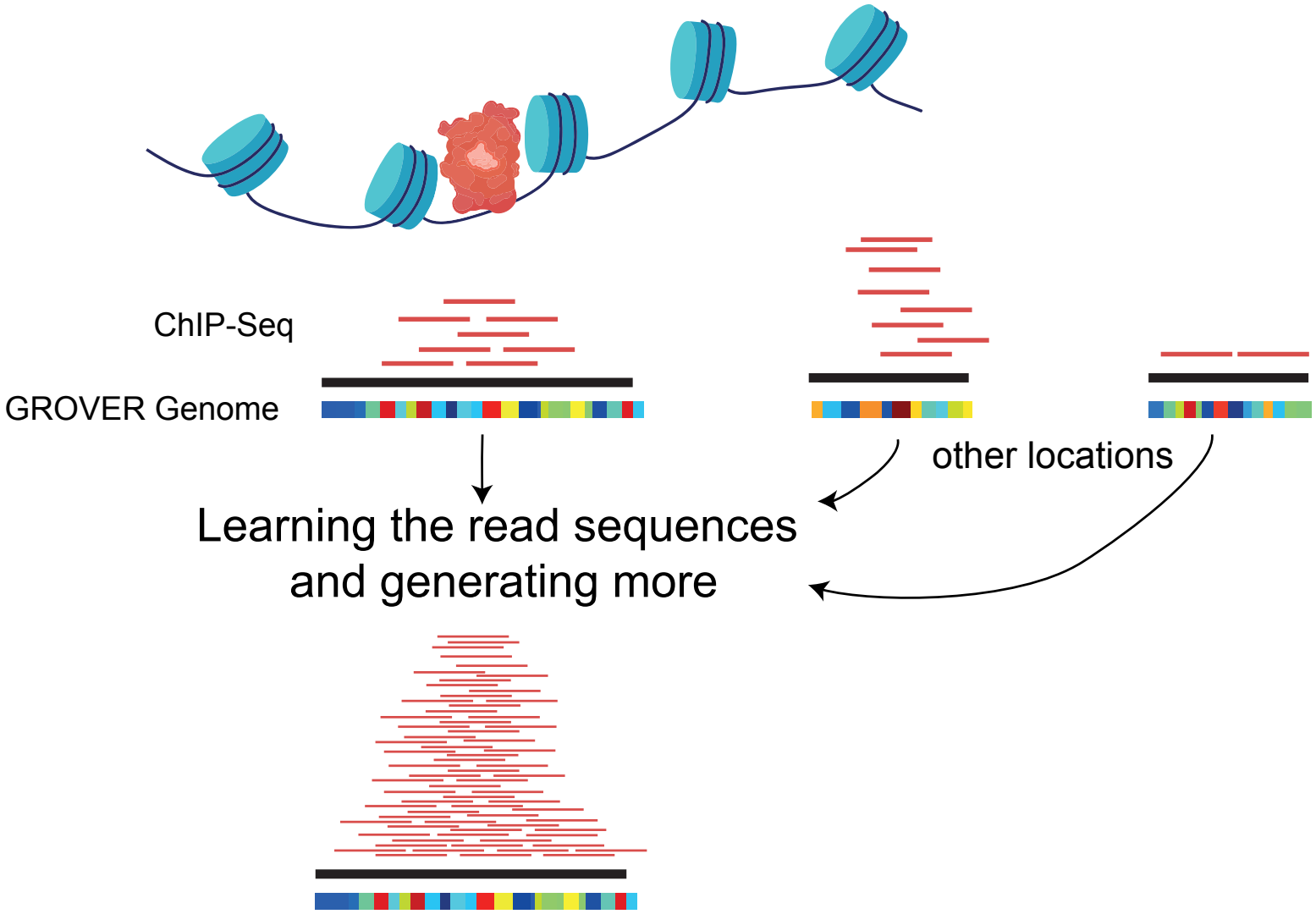


Translation AI - translation of genomics data across species

annotation
regulatory genomics data
⋮



Generative AI - Boosting of genomics data



TextToImage - GenotypeToPhenotype

“DallE2, please give me an illustration of damaged DNA in comic style”:

“...ACGCGTAAAATCGATTTAGCGATTTGCAA...”:

https://www.ebi.ac.uk/training/online/courses/human-genetic-variation-introduction/what-is-genetic-variation/genotype-or-phenotype/

Summary, take-home messages, and further information

- Large Language models are models trained on next- or masked-token prediction
- They learn a sense of grammar and syntax, as well as language context
- DNA resembles “language” and can be trained analogously
- DNA language models can be used to understand genomes
- DNA language models can be fine-tuned for a myriad of tasks

Attention is all you need: https://arxiv.org/abs/1706.03762

How Transformers Work: A Detailed Exploration of Transformer Architecture: https://www.datacamp.com/tutorial/how-transformers-work

But what is a GPT? Visual intro to transformers: https://www.youtube.com/watch?v=wjZofJX0v4M

Attention in transformers, visually explained: https://www.youtube.com/watch?v=eMlx5fFNoYc