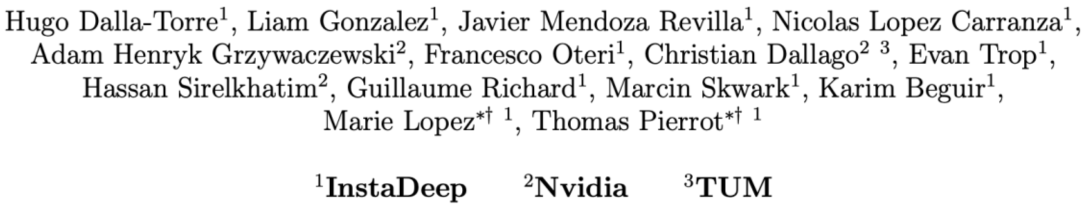
The Nucleotide Transformer: Building and Evaluating Robust Foundation Models for Human Genomics



Hugo Dalla-Torre[1], Liam Gonzalez[1], Javier Mendoza Revilla[1], Nicolas Lopez Carranza[1],
Adam Henryk Grzywaczewski[2], Francesco Oteri[1], Christian Dallago[2 3], Evan Trop[1],
Hassan Sirelkhatim[2], Guillaume Richard[1], Marcin Skwark[1], Karim Beguir[1],
Marie Lopez*†[1], Thomas Pierrot*†[1]

[1]**InstaDeep** [2]**Nvidia** [3]**TUM**

Shentong Mo

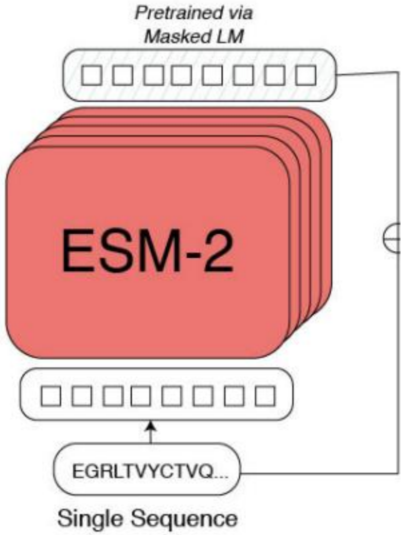
Aug 24, 2023

Presentation Line

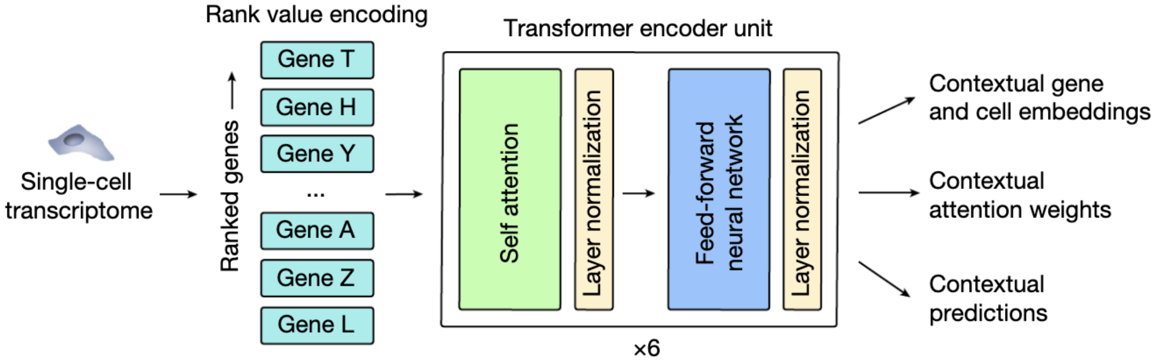
- Background & Motivation
- NT architecture
- NT methods
- NT training data
- Main experimental results
- Discussions

Background

- Foundation models in natural language processing:
  - BERT, GPT-3, GPT-4
- Protein foundation models
  - ESM-2 (15B, from Alexander Rives' group)
  - xTrimoPGLM (100B, from Le Song's group)
- Single-cell foundation models
  - Geneformer (from Patrick T. Ellinor's group)
  - scFoundation (100M, from Le Song's group)

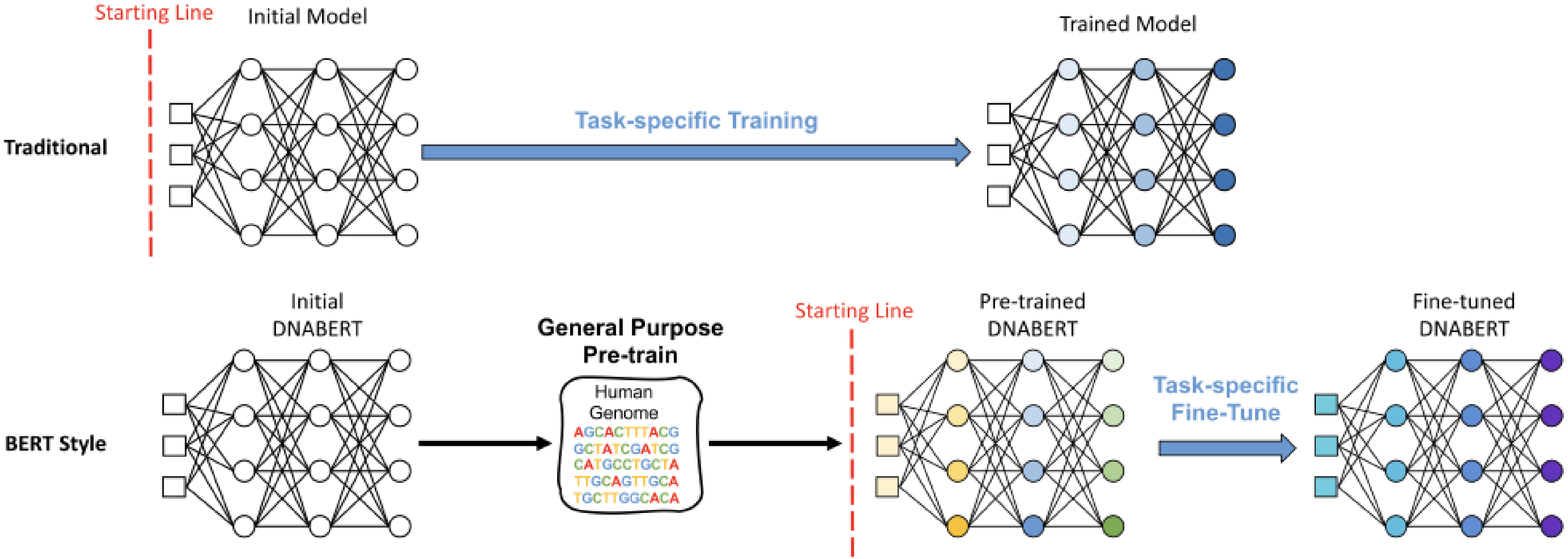


(ESM-2, Alexander Rives, Science 2023)



(Geneformer, Patrick T. Ellinor, Nature 2023)

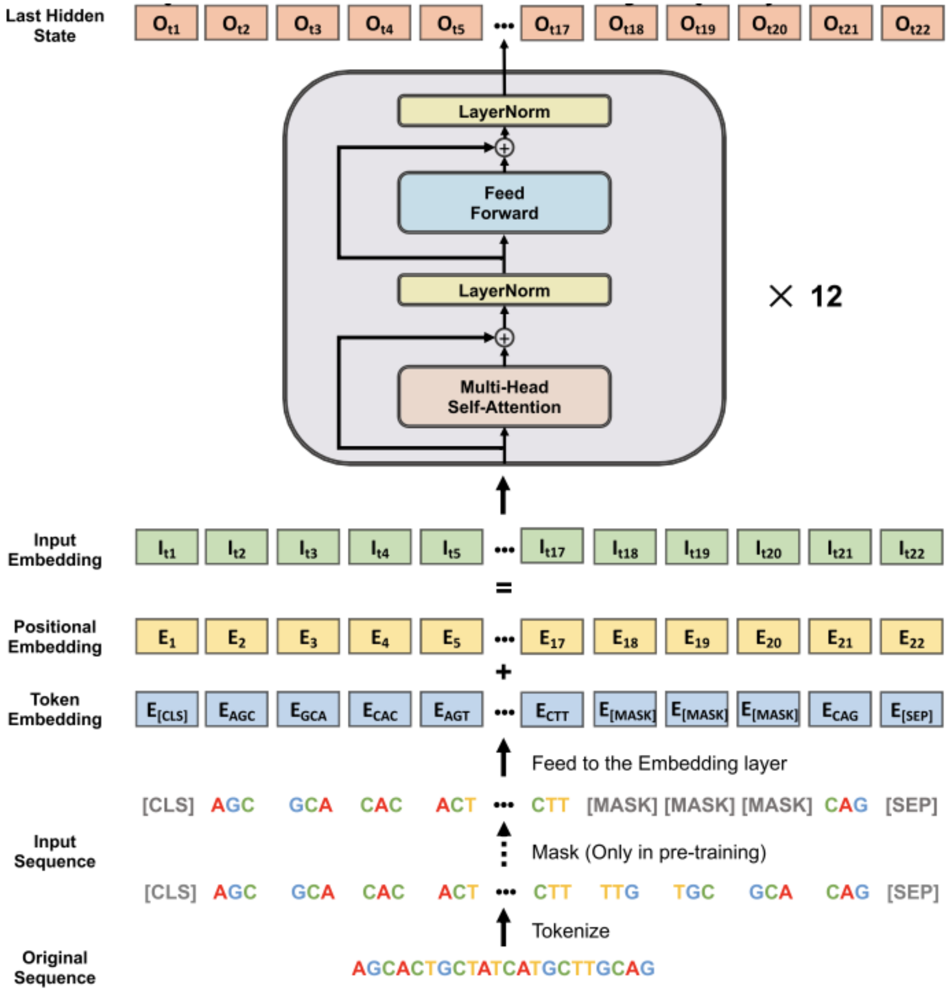
Motivation



- **Abundant genomic data** able to uncover the intricate patterns of natural variability across species and populations is vastly available.
- Powerful deep learning methods, that can operate **at large scale**, are required to perform accurate signal extraction from this large amount of unlabelled genomic data.

Architecture

- An **encoder-only transformer** architecture.
- Input Embeddings:
  - **Tokens sequence** embeddings (k-mer).
  - **Learnable positional** encodings.
- Each transformer layer transforms its input through a layer normalization layer followed by a multi-head self-attention layer.
- A language model **head** (MLP) in the final layer to predict a **probability distribution** over the existing tokens at each position in the sequence.


Last Hidden State
O_{t1} O_{t2} O_{t3} O_{t4} O_{t5} ••• O_{t17} O_{t18} O_{t19} O_{t20} O_{t21} O_{t22}
LayerNorm
Feed Forward
LayerNorm
Multi-Head Self-Attention
× 12
Input Embedding
I_{t1} I_{t2} I_{t3} I_{t4} I_{t5} ••• I_{t17} I_{t18} I_{t19} I_{t20} I_{t21} I_{t22}
=
Positional Embedding
E_1 E_2 E_3 E_4 E_5 ••• E_{17} E_{18} E_{19} E_{20} E_{21} E_{22}
+
Token Embedding
$E_{[CLS]}$ E_{AGC} E_{GCA} E_{CAC} E_{AGT} ••• E_{CTT} $E_{[MASK]}$ $E_{[MASK]}$ $E_{[MASK]}$ E_{CAG} $E_{[SEP]}$
Feed to the Embedding layer
Input Sequence
[CLS] AGC GCA CAC ACT ••• CTT [MASK] [MASK] [MASK] CAG [SEP]
Mask (Only in pre-training)
[CLS] AGC GCA CAC ACT ••• CTT TTG TGC GCA CAG [SEP]
Tokenize
Original Sequence
AGCACTGCTATCATGCTTGCAG

From DNA-BERT

Masked Self-supervised Training

- GPU & Batch Size: **128 A100 GPUs**, 14 and 2 sequences for 500M and 2.5B.
- Within a sequence, of a subset of 15% of tokens, 80% are replaced by a special mask [MASK] token, others are replaced by randomly selected standard tokens.
- Loss function: Sum of the **cross-entropy losses**, between the predicted probabilities over tokens and the ground truth tokens, at each selected position
- Gradients were accumulated to reach an effective batch size of **1M tokens** per batch.
- The **500M** parameters models were trained on **a single node for a day**, while the **2.5B** models models required the whole cluster for **28 days** to be trained.

Parameter-Efficient Fine-tuning

- The weights of transformer layers and embedding layers are **frozen** and **new learnable weights** are introduced (0.1%). 9.5×18 = **171 GBs**, whereas parameter-efficient fine tuning required only **171 MB**.
- For **each transformer layer**, we introduced **three learned vectors** $l_k \in \mathbb{R}^{d_k}$, $l_v \in \mathbb{R}^{d_v}$ and $l_{ff} \in \mathbb{R}^{d_{ff}}$, which were introduced in the self-attention mechanism as:

$$\operatorname{softmax}\left(\frac{Q(l_k) \odot K^T}{\sqrt{d_k}}\right)(l_v \odot V)$$

- in the position-wise feed-forward networks as $(l_{ff}\ \gamma(W_1 x))W_2$, where γ is the feed-forward network nonlinearity, and $\odot$ represents the element-wise multiplication.
- These weights will weigh the transformer layers to improve the final representation of the model on a downstream task, so that the **classification/regression head** can more accurately solve the problem.

Training Data

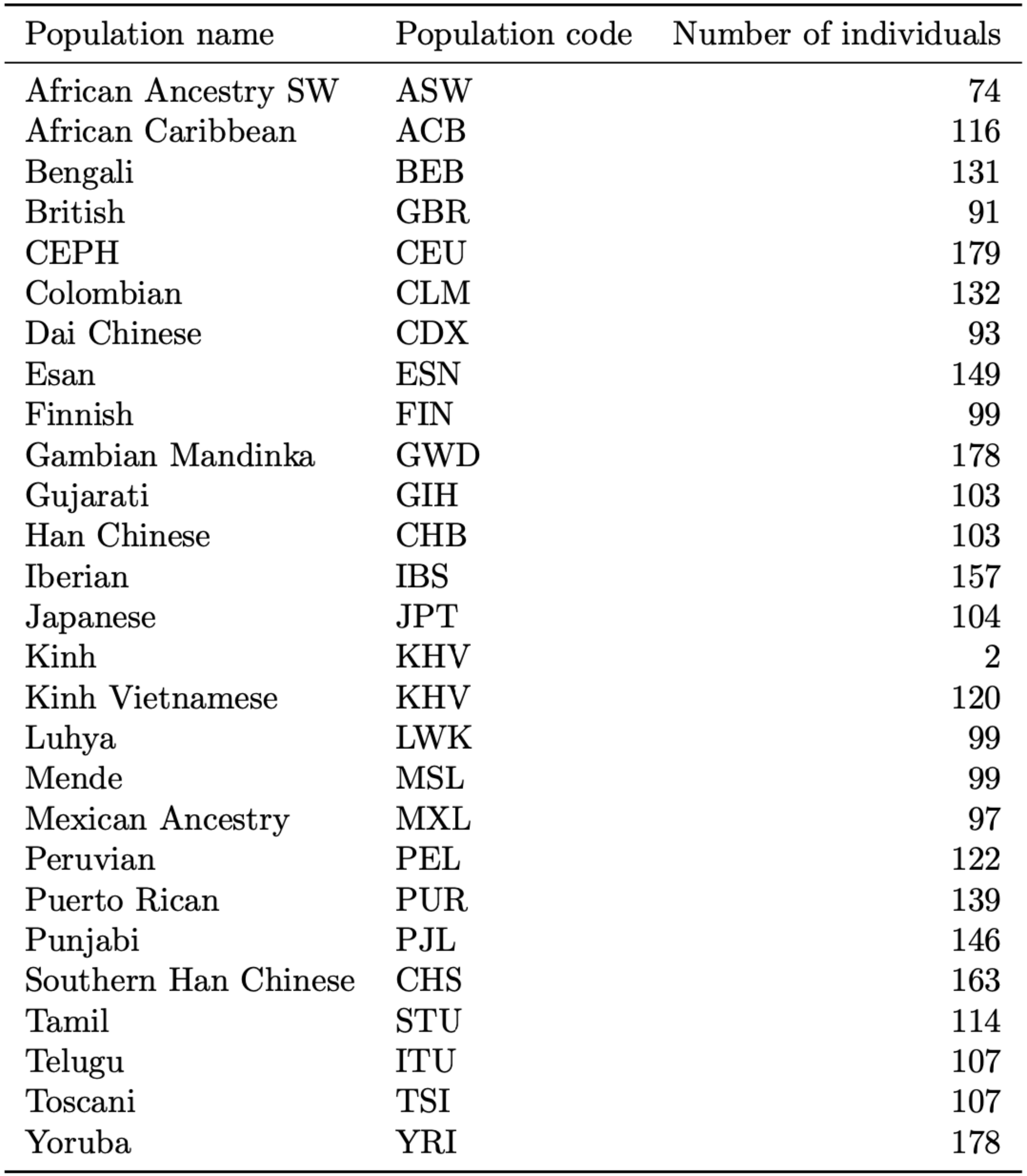
- **The Human reference genome dataset**
  - all autosomal and sex chromosomes sequences from reference assembly GRCh38/hg38
  - 3.2 billion nucleotides
- **The 1000G dataset**
  - 3202 high-coverage human genomes, originating from 27 geographically structured populations
  - 20.5 trillion nucleotides
- **The Multispecies dataset**
  - Plant and virus genomes were not taken into account, as their regulatory elements differ
  - A total of 850 species, 174 billion nucleotides

The Human reference genome dataset

- 3.2 billion nucleotides
- Each sequence is 6,200 or 12,200 base pairs (nucleotides) long
- Each data instance:
    - sequence: a string containing a DNA sequence from the human reference genome
    - chromosome: a string indicating the chromosome (1,2,...,21,X,Y)
    - start_pos: an integer indicating the index of the sequence's first nucleotide
    - end_pos: an integer indicating the index of the sequence's last nucleotide
- Train/Val/Test (6kbp): 498,444/7,784/8,469

The 1000G dataset

- **3202** high-coverage individual human genomes, originating from 27 geographically structured populations
- 20.5 trillion nucleotides
- Such diversity allowed the dataset to encode a better representation of human genetic variation.

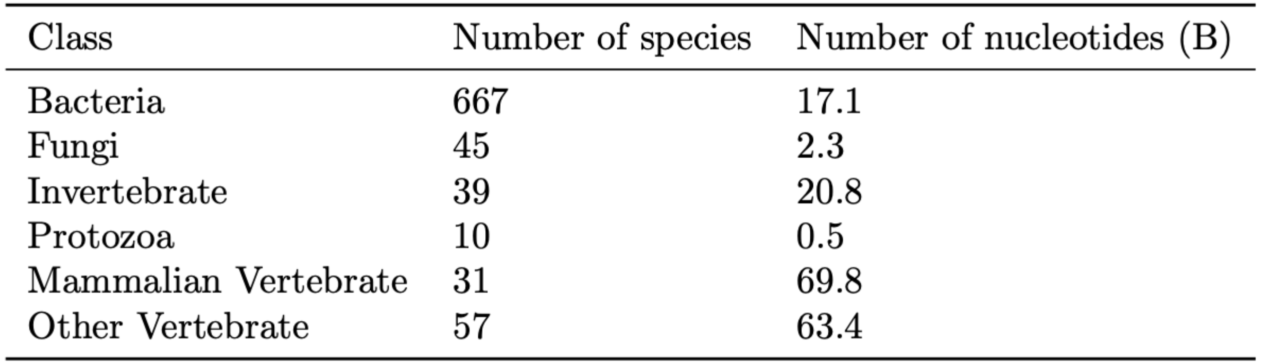


Population name	Population code	Number of individuals
African Ancestry SW	ASW	74
African Caribbean	ACB	116
Bengali	BEB	131
British	GBR	91
CEPH	CEU	179
Colombian	CLM	132
Dai Chinese	CDX	93
Esan	ESN	149
Finnish	FIN	99
Gambian Mandinka	GWD	178
Gujarati	GIH	103
Han Chinese	CHB	103
Iberian	IBS	157
Japanese	JPT	104
Kinh	KHV	2
Kinh Vietnamese	KHV	120
Luhya	LWK	99
Mende	MSL	99
Mexican Ancestry	MXL	97
Peruvian	PEL	122
Puerto Rican	PUR	139
Punjabi	PJL	146
Southern Han Chinese	CHS	163
Tamil	STU	114
Telugu	ITU	107
Toscani	TSI	107
Yoruba	YRI	178

Table 2: Number of individuals per population in the 1000G dataset.

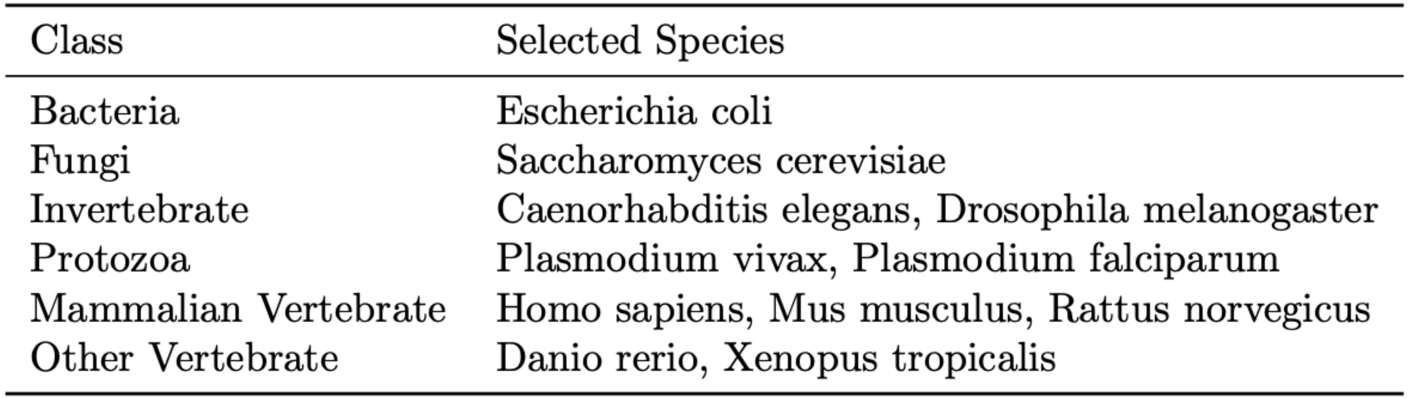
The Multispecies dataset

- Plant and virus genomes were not taken into account, as their regulatory elements differ
- A total of 850 species, 174 billion nucleotides



Class	Number of species	Number of nucleotides (B)
Bacteria	667	17.1
Fungi	45	2.3
Invertebrate	39	20.8
Protozoa	10	0.5
Mammalian Vertebrate	31	69.8
Other Vertebrate	57	63.4

Table 6: Distribution of the genomes present in the multispecies dataset.



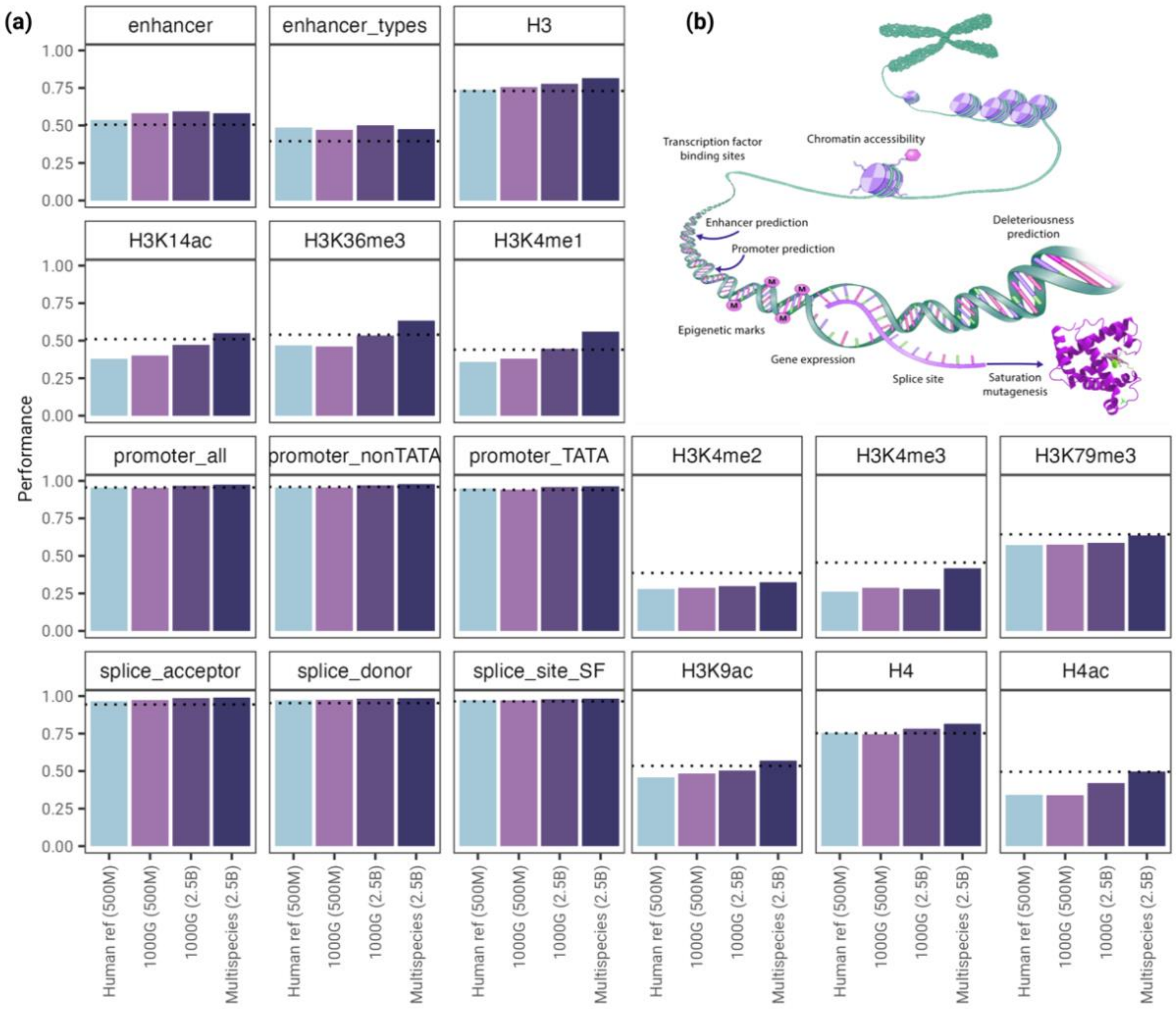
Class	Selected Species
Bacteria	Escherichia coli
Fungi	Saccharomyces cerevisiae
Invertebrate	Caenorhabditis elegans, Drosophila melanogaster
Protozoa	Plasmodium vivax, Plasmodium falciparum
Mammalian Vertebrate	Homo sapiens, Mus musculus, Rattus norvegicus
Other Vertebrate	Danio rerio, Xenopus tropicalis

Table 7: Model organisms selected to be included in the multispecies dataset.

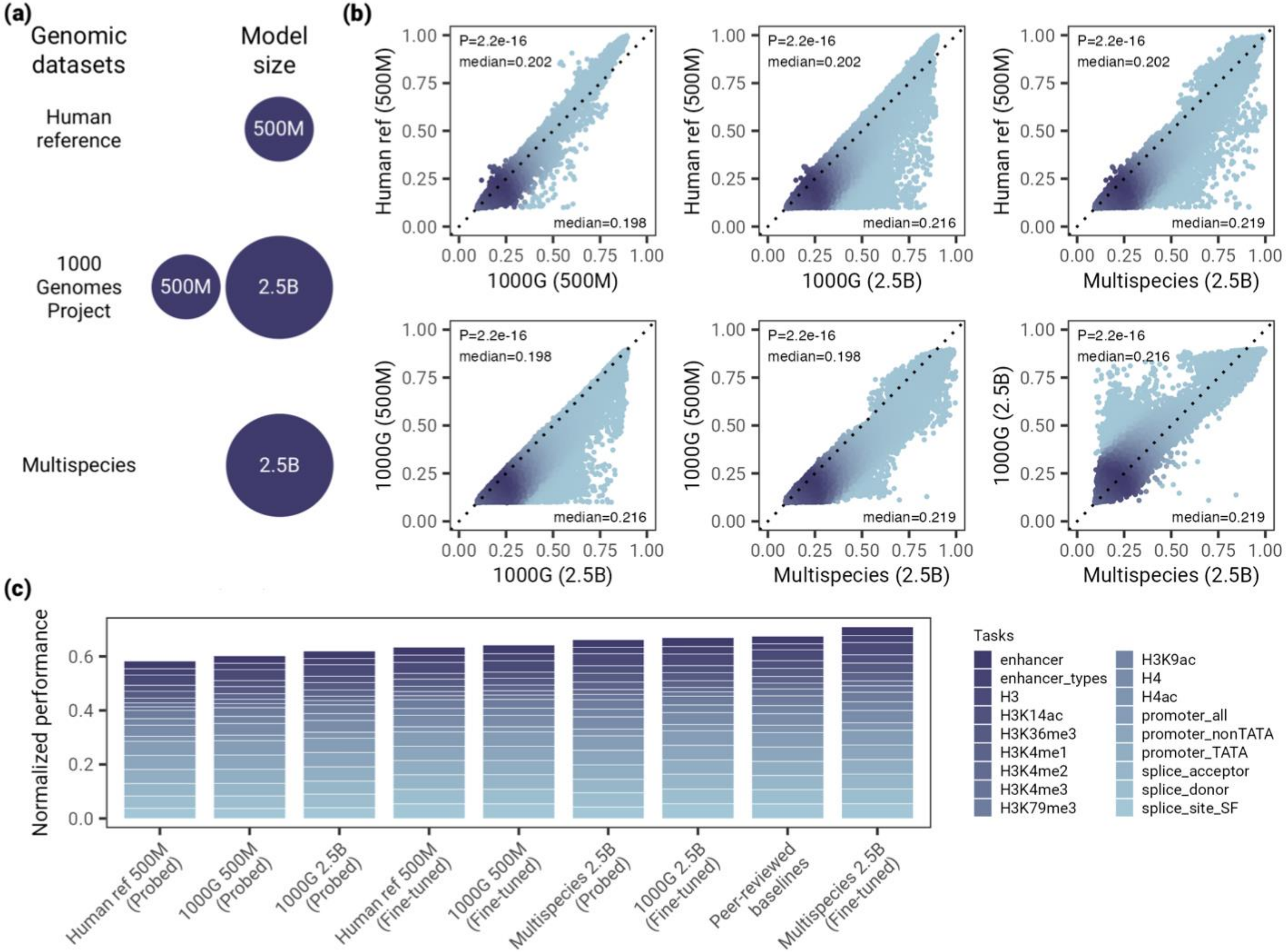
Downstream Tasks

- **Epigenetic marks prediction**
  - histone marks: H3, H4, H3K9ac, H3K14ac, H4ac, H3K4me1, H3K4me2, H3K4me3, H3K36me3 and H3K79me3.
- **Promoter sequence prediction**
  - 29,597 promoter regions, 3,065 of which were TATA-box promoters
- **Enhancer sequence prediction**
  - 742 strong enhancers, 742 weak enhancers and 1484 non-enhancers
- **Splice site prediction**
  - donor, acceptor, and non-splice sites, containing sequences detected in human genes

NT outperform or matched 15 18 genomic baselines after fine-tuning

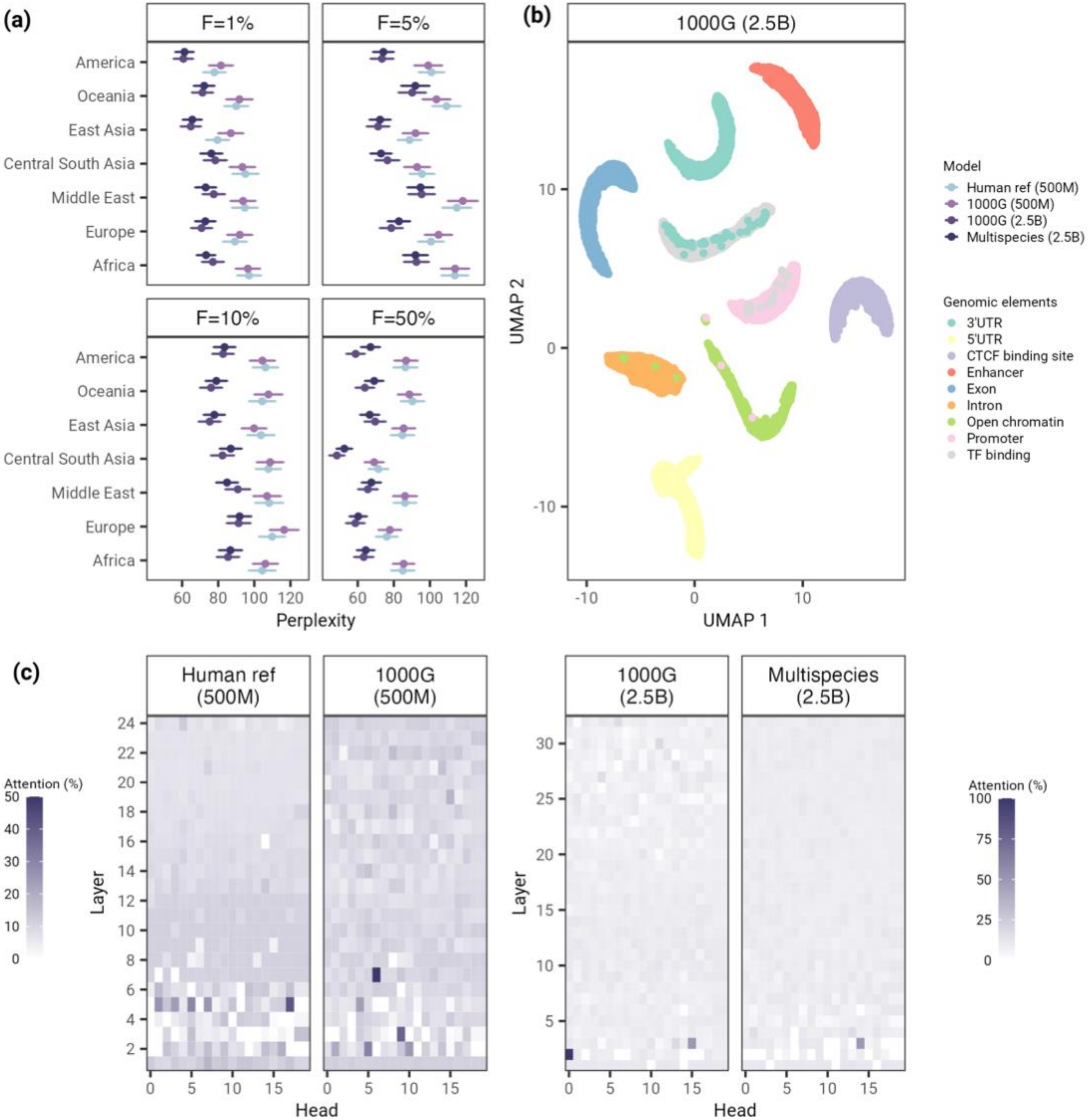


Parameter scaling and increasing data diversity improve performance



NT models learnt to detect known genomic elements and human genetic variation

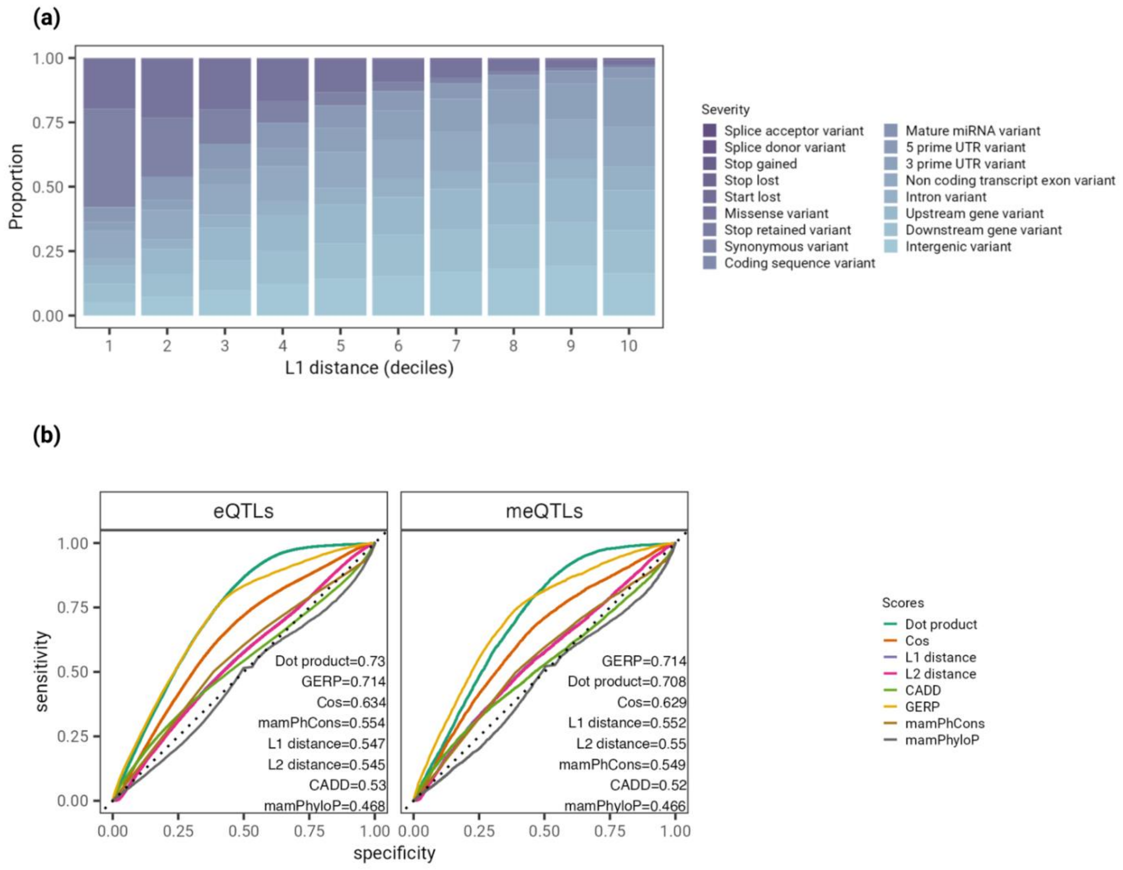
the effectiveness of sequence reconstruction by recording the model perplexity scores over **single nucleotide polymorphisms** (SNPs) occurring at 1%, 5%, 10% and 50% frequencies.



NT embeddings predict the impact of mutations

(a) all segregating 1000 Genomes Project SNPs from chromosome 22 and the distribution of these embedding-based divergence scores across 17 variant categories (e.g. stop gained, missense, intergenic), where embedding- based distances alone encodes information about variant functional categories.

(b) assessed the ability of the divergence scores to classify genetic variants exerting regulatory effects on gene expression (i.e., expression quantitative trait loci [eQTLs]) and genetic variants associated with DNA methylation variation (i.e., methylation quantitative trait loci [meQTLs])



Evolution informs representations: models trained on genomes from different species top charts

- **intra-** (i.e. when training on multiple genomes of a single species) and **inter-species** (i.e., on genomes across different species) variability play an important factor driving accuracy across tasks
- the models trained on genomes coming from **different specie**s perform well on categorical human genomics downstream tasks, as well as on human variant prediction, even when compared to models trained exclusively on the human genome
- genome LMs capture **a signal of evolution** so fundamental across species that it better generalizes to shared functions

Model scale drives performance

- The Nucleotide Transformer models trained ranged from 500 million up to 2.5 billion parameters, which is **five times** larger than DNA-BERT and **ten times** larger than the Enformer.
- As has been the case in NLP, results in the genomic space confirmed that increasing model size yields better performance.
- Training the largest parameter model required a total of 128 GPUs across 16 compute nodes for 28 days.
- Significant investments were made to **engineer efficient training routines** that took full advantage of the infrastructure, highlighting the need for both specialized infrastructure and dedicated software solutions.

Diverse datasets, large models and fine-tuning: best results require it all

- Best probing performance for intermediate transformer layers, aligning with recent work in computational biology and common practice in NLP, suggesting that even for existing biological-related LMs new highs could be reached.
- We additionally explored a recent downstream fine-tuning technique that introduces a small number of trainable weights in the transformer for **weight-efficient fine-tuning**.
- Compared to the extensive probing exercise, this technique yielded better results using **fewer compute resources** and rose the number of tasks outperforming or matching baselines to 15 out of 18, confirming that downstream model engineering can yield performance improvements.

Genomics insights are captured during training despite no supervision

- The Nucleotide Transformer models learned insights about key **regulatory genomic elements** through attention, as demonstrated through the analysis of attention maps, embedding spaces, and probability distributions.
- Elements such as **enhancers and promoters** were detected by all models, and at several heads and layers. We also observed that each model contained at least one layer that produced embeddings which clearly separated five of the genomic elements analyzed.
- Building transformer models (currently, limited to 6kbp) with the ability to model language that can work with long inputs, **up to 200kbp** or more, is a promising direction for the field, such as sparse attention.

Thanks