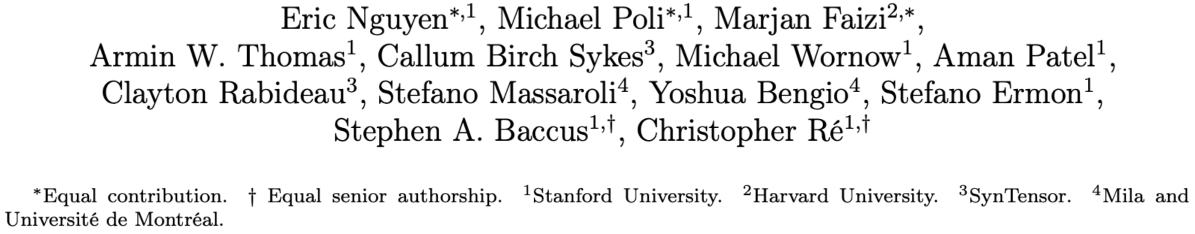
HyenaDNA: Long-Range Genomic Sequence Modeling at Single Nucleotide Resolution



Eric Nguyen[*,1], Michael Poli[*,1], Marjan Faizi[2,*],
Armin W. Thomas[1], Callum Birch Sykes[3], Michael Wornow[1], Aman Patel[1],
Clayton Rabideau[3], Stefano Massaroli[4], Yoshua Bengio[4], Stefano Ermon[1],
Stephen A. Baccus[1,†], Christopher Ré[1,†]

[*]Equal contribution. † Equal senior authorship. [1]Stanford University. [2]Harvard University. [3]SynTensor. [4]Mila and Université de Montréal.

Shentong Mo

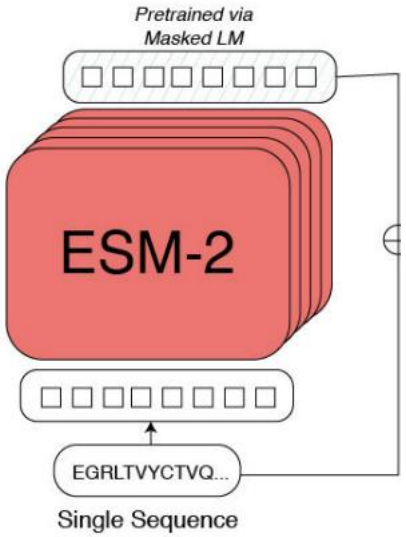
Aug 31, 2023

Presentation Line

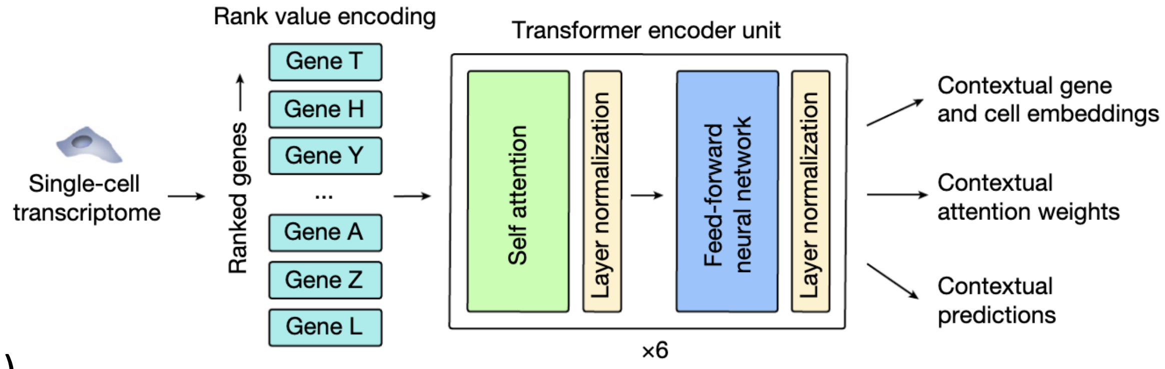
- Background & Motivation
- HyenaDNA architecture
- HyenaDNA methods
- HyenaDNA training data
- Main experimental results
- Discussions

Background

- 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)
- DNA foundation models
  - DNABERT (100M, from Ramana V. Davuluri's group)
  - Nucleotide Transformer (2.5B, from Nvidia & TUM)



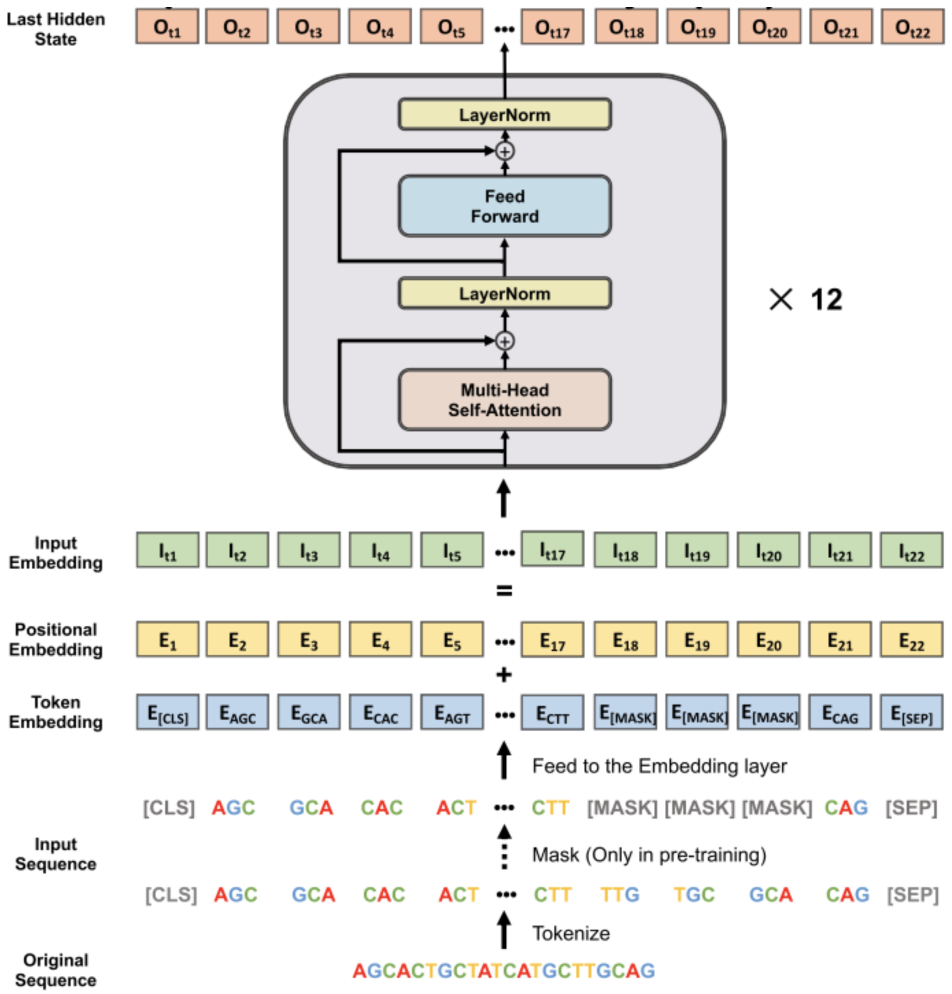
(ESM-2, Alexander Rives, Science 2023)



(Geneformer, Patrick T. Ellinor, Nature 2023)

Motivation

- **Long-range context**: The attention mechanism scales quadratically in sequence length, with current genomic FMs pretraining on only 512 to 4,096 tokens as context, <0.001% of the human genome (**3.2B nucleotides**, interactions that span over **100k+ nucleotides**).
- **Single Nucleotide Resolution**: the reliance on **fixed k-mers**, akin to DNA "words", and tokenizers to aggregate meaningful DNA units. However, single nucleotide alterations represent physical analogs where, for example, **single nucleotide polymorphisms (SNPs)** and mutations can have a profound impact on biological properties including regulatory activity.


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
[CLS] AGC GCA CAC ACT ... CTT [MASK] [MASK] [MASK] CAG [SEP]
Input Sequence
Mask (Only in pre-training)
[CLS] AGC GCA CAC ACT ... CTT TTG TGC GCA CAG [SEP]
Tokenize
Original Sequence
AGCACTGCTATCATGCTTGCAG

From DNA-BERT

Motivation (cont.)

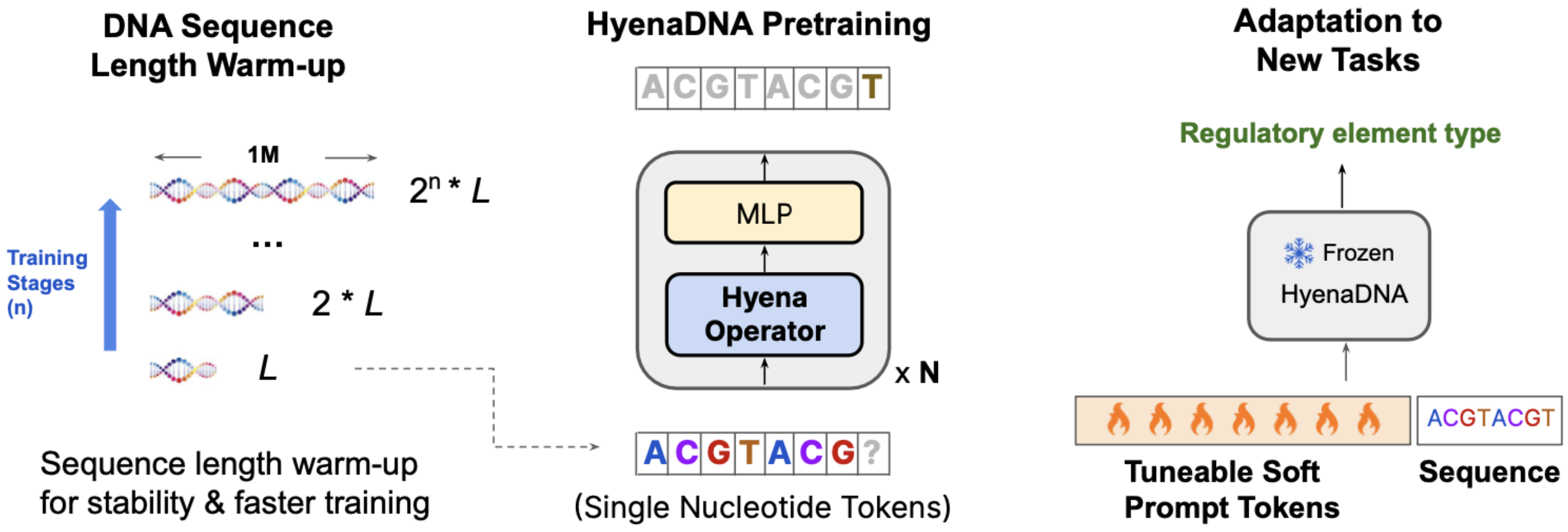
- **Toward longer context models**: **Hyena**, a large language model based on implicit convolutions, was shown to match attention in quality while reducing computational time complexity, thereby allowing a longer context to be processed.
- **Hyena**: a parameter efficient global convolutional filter along with a data-controlled gating mechanism, which enables a context-specific operation over every token. A shallow 2 layer model could effectively process context lengths at **131k tokens**. We hypothesize that Hyena's core operations can unlock the potential to capture both the long-range and single nucleotide resolution of real genomic sequences over attention-based approaches.

(i.) Can a convolutional long-context model be used effectively at single nucleotide resolution?

(ii.) What new capabilities could long-context genomic foundations models enable?

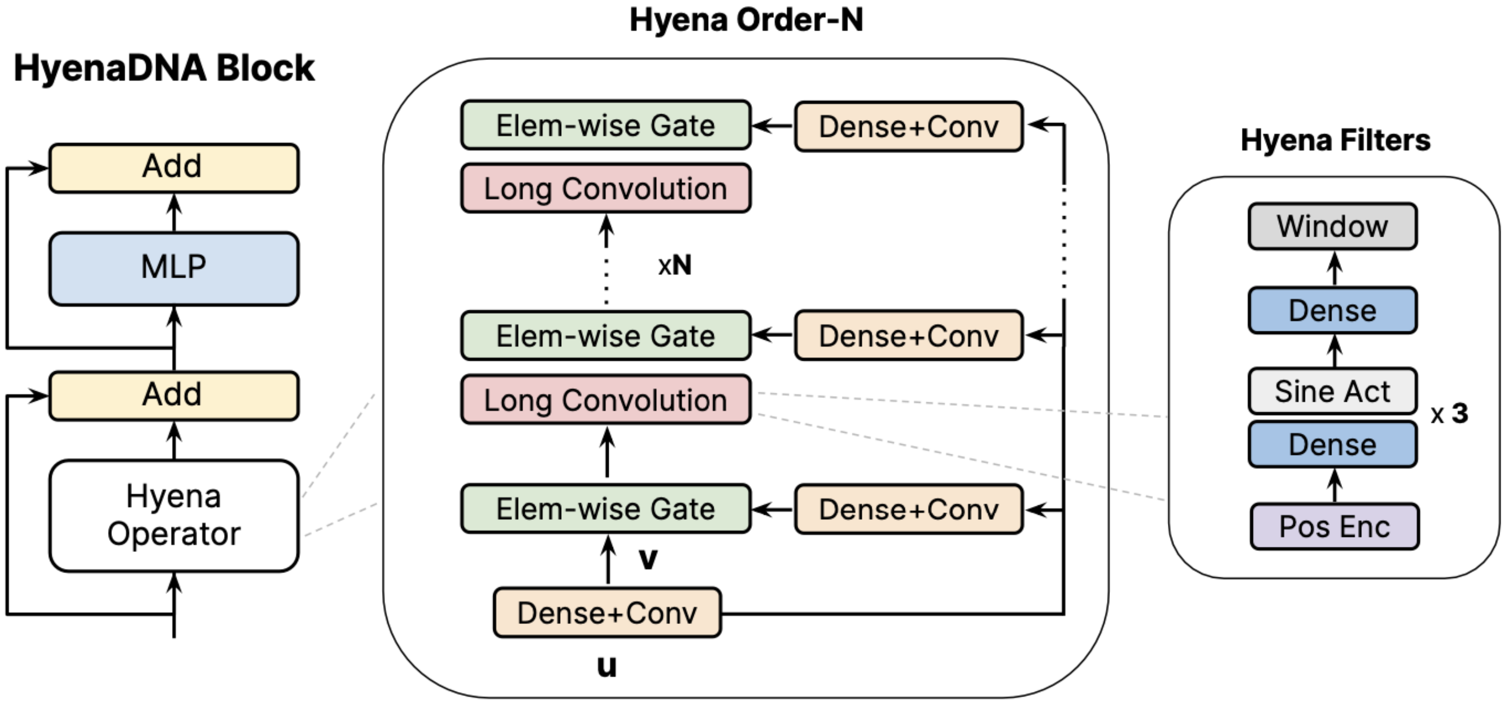
M. Poli, S. Massaroli, E. Nguyen, D. Y. Fu, T. Dao, S. Baccus, Y. Bengio, S. Ermon, and C. Ré. Hyena Hierarchy: Towards larger convolutional language models. arXiv preprint arXiv:2302.10866, 2023.

Architecture



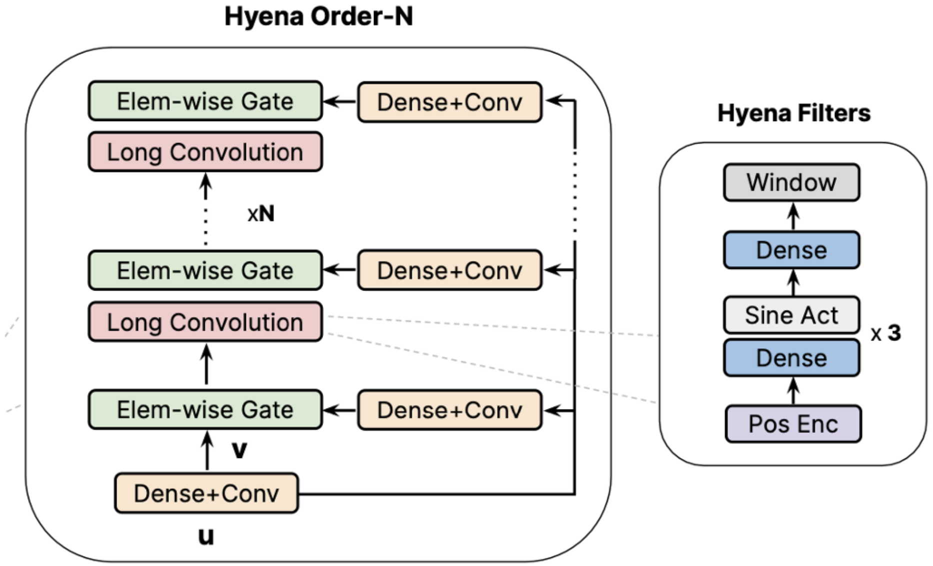
- A genomic FM pretrained on the human reference genome at context lengths up to **1 million** tokens at single nucleotide resolution - an up to **500x increase** over existing genomic FMs using dense-attention.
- HyenaDNA scales sub-quadratically in sequence length (training up to **160x faster** than attention at sequence length 1M), uses single nucleotide tokens, and has a global receptive field at each layer.

The HyenaDNA Model



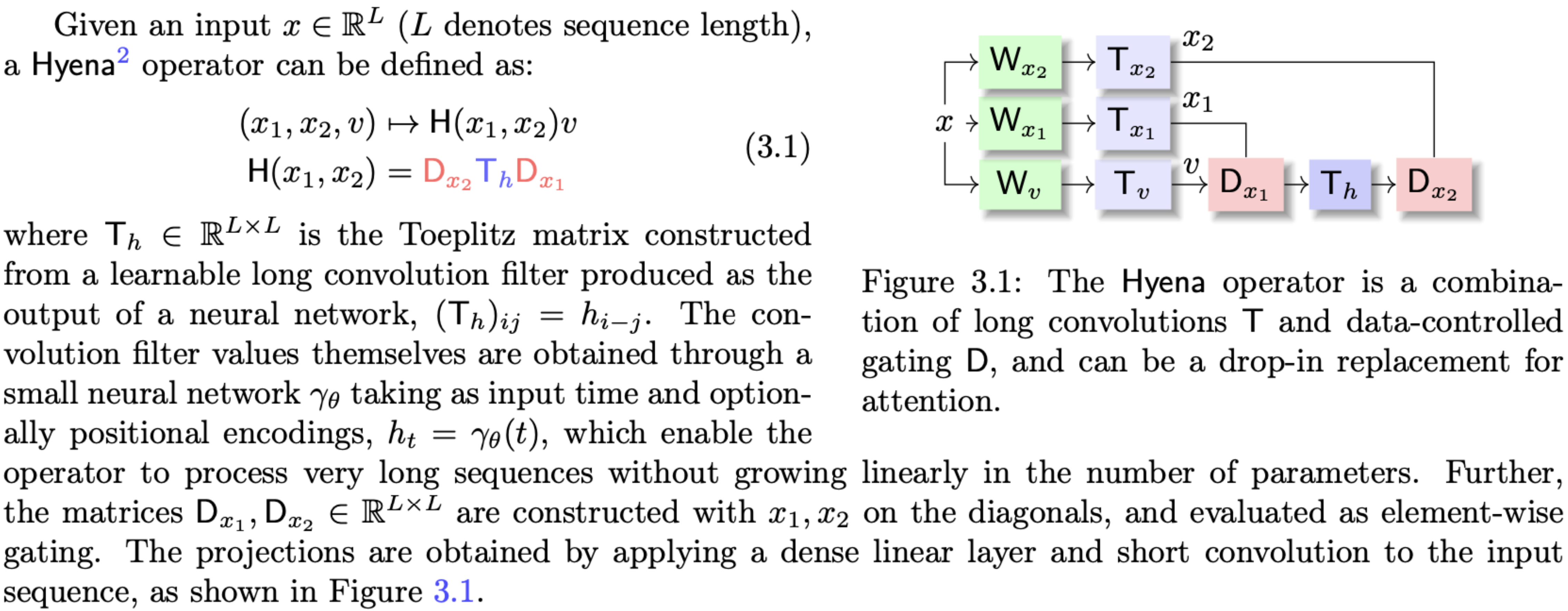
- The HyenaDNA model is a decoder-only, sequence-to-sequence architecture defined by a stack of blocks consisting of a **Hyena operator**, followed by normalization and a feed-forward neural network.
- A Hyena operator is composed of **long convolutions** and **element-wise gate layers**.

The Hyena Operator



- A Hyena operator is composed of **long convolutions** and **element-wise gate layers**.
- The gates are composed of projections of the input using dense layers and short convolutions.
- The long convolutions are parameterized implicitly via an **MLP** that produces the **convolutional filters**.
- The convolution is evaluated using a **Fast Fourier Transform convolution** with time complexity $O(Llog_2L)$.

The Hyena Operator (cont.)



Given an input $x \in \mathbb{R}^L$ (L denotes sequence length), a **Hyena**[2] operator can be defined as:

$$\begin{aligned} (x_1, x_2, v) &\mapsto \mathsf{H}(x_1, x_2)v \\ \mathsf{H}(x_1, x_2) &= \mathsf{D}_{x_2}\mathsf{T}_h\mathsf{D}_{x_1} \end{aligned} \tag{3.1}$$

where $\mathsf{T}_h \in \mathbb{R}^{L \times L}$ is the Toeplitz matrix constructed from a learnable long convolution filter produced as the output of a neural network, $(\mathsf{T}_h)_{ij} = h_{i-j}$. The convolution filter values themselves are obtained through a small neural network γ_θ taking as input time and optionally positional encodings, $h_t = \gamma_\theta(t)$, which enable the operator to process very long sequences without growing linearly in the number of parameters. Further, the matrices $\mathsf{D}_{x_1}, \mathsf{D}_{x_2} \in \mathbb{R}^{L \times L}$ are constructed with x_1, x_2 on the diagonals, and evaluated as element-wise gating. The projections are obtained by applying a dense linear layer and short convolution to the input sequence, as shown in Figure 3.1.

Figure 3.1: The **Hyena** operator is a combination of long convolutions T and data-controlled gating D, and can be a drop-in replacement for attention.

M. Poli, S. Massaroli, E. Nguyen, D. Y. Fu, T. Dao, S. Baccus, Y. Bengio, S. Ermon, and C. Ré. Hyena Hierarchy: Towards larger convolutional language models. arXiv preprint arXiv:2302.10866, 2023.

A Hyena operator can be evaluated in $O(L\,log_2 L)$ time

Given an input $x \in \mathbb{R}^L$ (L denotes sequence length), a Hyena[2] operator can be defined as:

$$\begin{aligned}(x_1, x_2, v) &\mapsto \mathsf{H}(x_1, x_2)v \\ \mathsf{H}(x_1, x_2) &= \mathsf{D}_{x_2}\mathsf{T}_h\mathsf{D}_{x_1}\end{aligned} \tag{3.1}$$

where $\mathsf{T}_h \in \mathbb{R}^{L\times L}$ is the Toeplitz matrix constructed from a learnable long convolution filter produced as the output of a neural network, $(\mathsf{T}_h)_{ij} = h_{i-j}$.

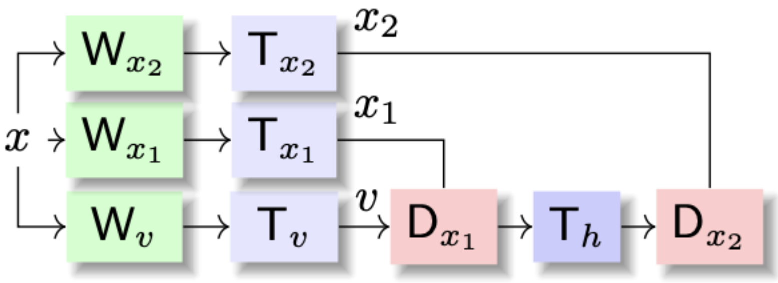
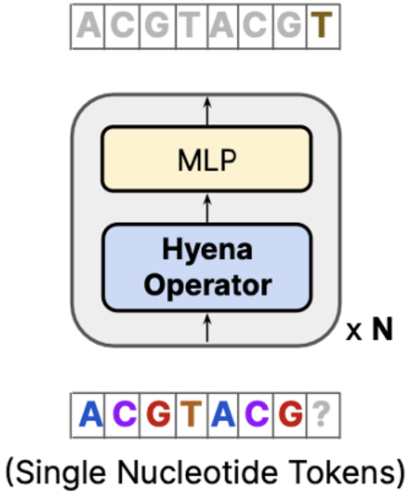


Figure 3.1: The Hyena operator is a combination of long convolutions T and data-controlled gating D, and can be a drop-in replacement for attention.

Efficient evaluation is crucial on settings involving extremely long sequences such as genomics. In the general case where the embedding dimension $D > 1$ and $u \in \mathbb{R}^{L\times D}$, the linear projections $\mathsf{W}_{x_1}, \mathsf{W}_{x_2}, \mathsf{W}_v \in \mathbb{R}^{D\times D}$ are right multiplied to x, and D independent Hyena operators are then applied to each dimension.

M. Poli, S. Massaroli, E. Nguyen, D. Y. Fu, T. Dao, S. Baccus, Y. Bengio, S. Ermon, and C. Ré. Hyena Hierarchy: Towards larger convolutional language models. arXiv preprint arXiv:2302.10866, 2023.

Training Long Sequence Models



- **Tokenization**
  - Use the natural DNA vocabulary and refer to each nucleotide as a token.
  - This vocabulary lookup table includes the nucleotides "A", "G", "C", "T", and "N" representing a non-specific nucleotide, and use **next token prediction** for pre-training.
  - Special character tokens are also added for padding, separation, and unknown characters.
- **Sequence length warm-up for ultralong sequences**
  - Directly training on long sequences can affect training stability as the variance in gradient increases.
  - For ultralong sequences (200k+), we develop a new warm-up schedule that gradually increases the sequence length in stages to improve both stability and decrease training time.

Training Long Sequence Models (cont.)

- **Sequence length warm-up for ultralong sequences**
  - Our sequence length schedule starts at $L_1 = 64$, then **doubles the window at each stage** while keeping the global batch size constant.
  - By doing so, iterations at each consecutive stage will include more tokens, ensuring the scheduler can also act as a form of **batch size warm-up**.
  - sequence length scheduling to be particularly important at sequence lengths **greater than 450k**.

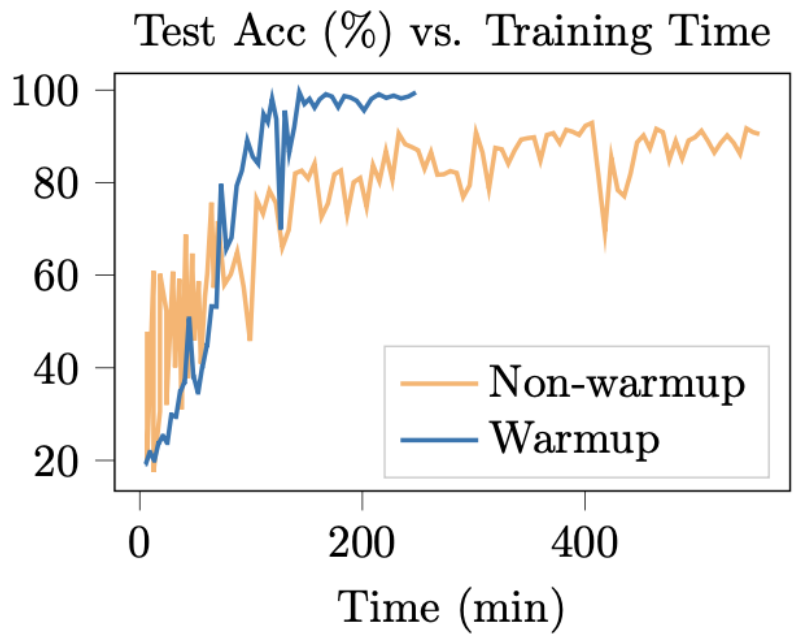
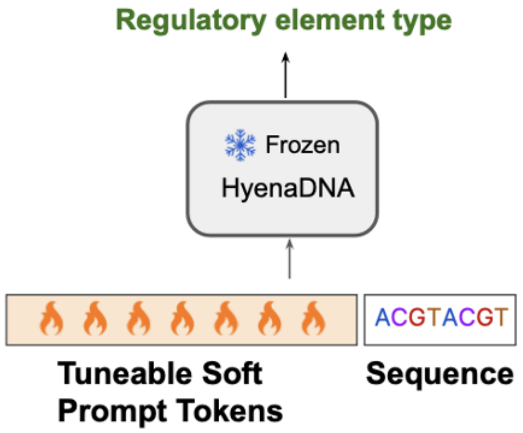
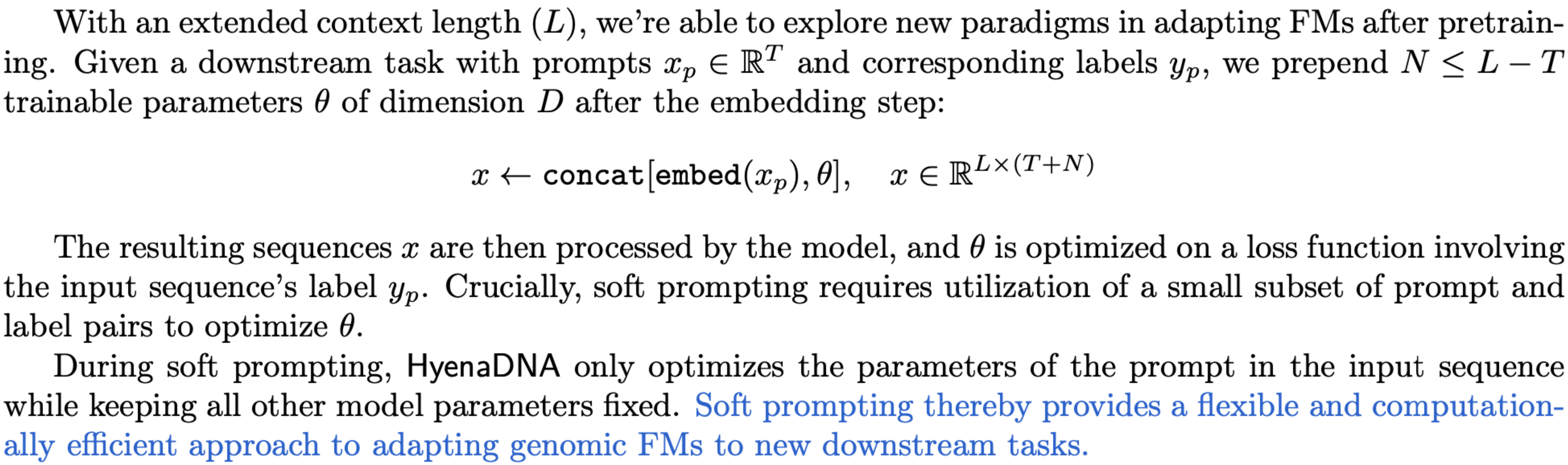


Figure 3.2: Sequence length warm-up reduces the training time of HyenaDNA at sequence length 450k by 40% and boosts accuracy by 7.5 points on species classification.

Downstream Adaptation

- **Tuneable prompting for long-context models**





With an extended context length (L), we're able to explore new paradigms in adapting FMs after pretraining. Given a downstream task with prompts $x_p \in \mathbb{R}^T$ and corresponding labels y_p, we prepend $N \leq L - T$ trainable parameters θ of dimension D after the embedding step:

$$x \leftarrow \texttt{concat}[\texttt{embed}(x_p), \theta], \quad x \in \mathbb{R}^{L \times (T+N)}$$

The resulting sequences x are then processed by the model, and θ is optimized on a loss function involving the input sequence's label y_p. Crucially, soft prompting requires utilization of a small subset of prompt and label pairs to optimize θ.

During soft prompting, HyenaDNA only optimizes the parameters of the prompt in the input sequence while keeping all other model parameters fixed. Soft prompting thereby provides a flexible and computationally efficient approach to adapting genomic FMs to new downstream tasks.

Training Data

- **The Human reference genome dataset**
  - all autosomal and sex chromosomes sequences from reference assembly GRCh38/hg38
  - 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/end_pos: an integer indicating the index of the sequence's first/last nucleotide
- Train/Val/Test (6kbp): 498,444/7,784/8,469

Downstream Tasks

- **Regulatory element prediction:** 8 regulatory element prediction datasets (7 binary, 1 3-way) with sequence lengths of 200-500, and one up to 4,776.
- **Epigenetic marks prediction:** histone marks, including 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

Pretraining on the Human Genome

- As context length increases, perplexity improves during pretraining.

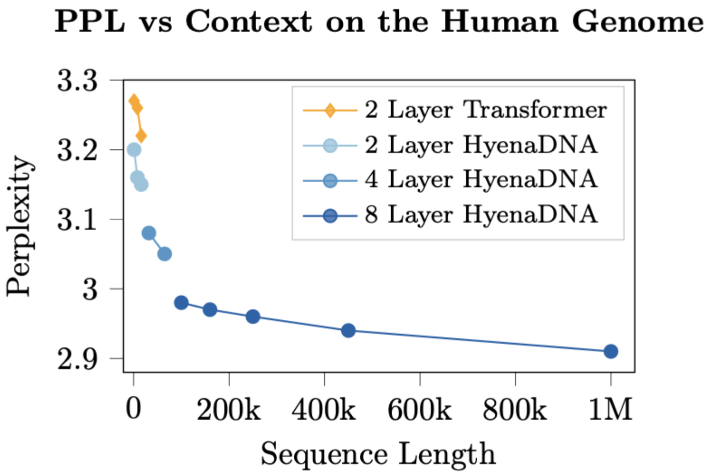


Figure 1.2: Pretraining on the human reference genome using longer sequences leads to better perplexity (improved prediction of next token).

- At sequence length 1M, HyenaDNA is 160x faster than its Transformer counterpart.

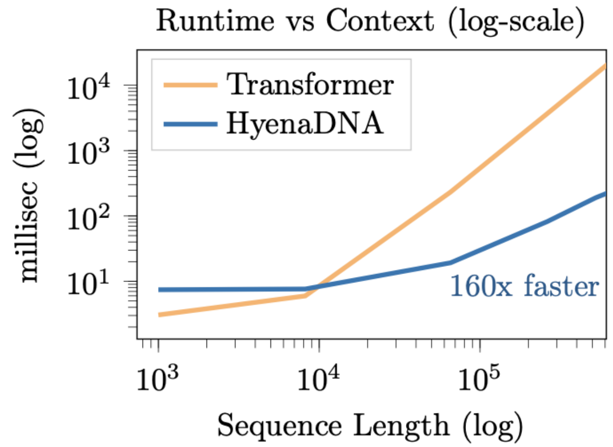


Figure 4.1: Runtime (forward & backward pass) for Transformer and HyenaDNA: 2 layers, width=128, gradient checkpointing, batch size=1, A100 80GB. At 1M tokens HyenaDNA is **160x faster** than Transformer.

- In optimizing for faster training time, **shorter context** enable lower perplexity to be reached faster.
- In optimizing for best overall perplexity, **longer context** allows for **lower perplexity** at the cost of training on more tokens.

Single Nucleotide Resolution

- Use 8 regulatory element prediction datasets (7 binary, 1 3-way) with sequence lengths of 200-500, and one up to 4,776.
- The original baseline model uses a CNN with a 1D kernel of length 16, and a baseline Transformer with Flash Attention.
- HyenaDNA sets a new SotA on all datasets by wide margins and up to 20% points on the human enhancer identification task.

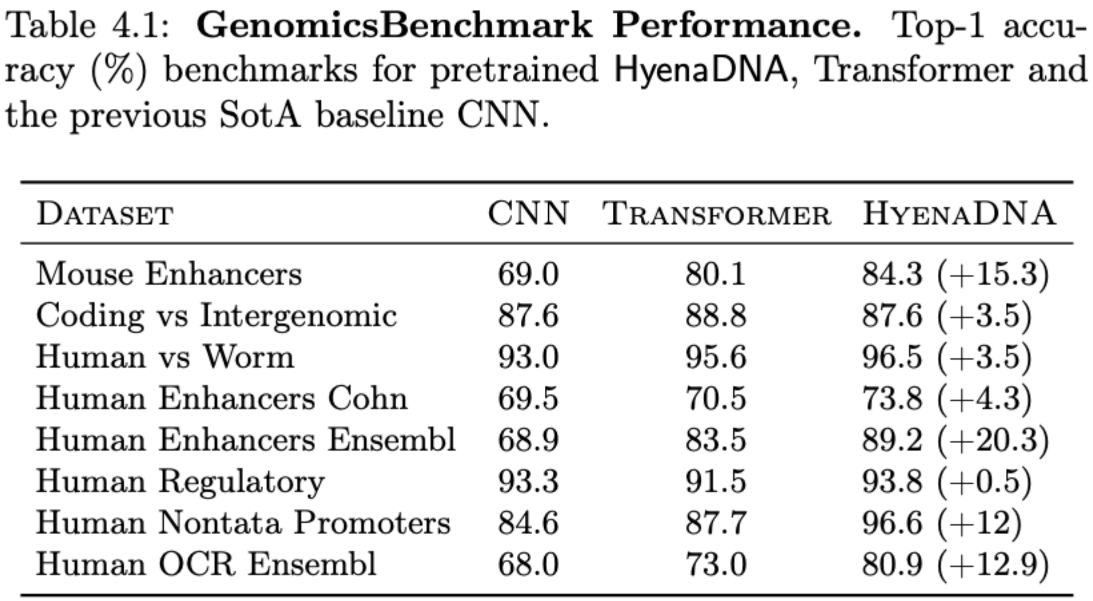


Table 4.1: **GenomicsBenchmark Performance.** Top-1 accuracy (%) benchmarks for pretrained HyenaDNA, Transformer and the previous SotA baseline CNN.

Dataset	CNN	Transformer	HyenaDNA
Mouse Enhancers	69.0	80.1	84.3 (+15.3)
Coding vs Intergenomic	87.6	88.8	87.6 (+3.5)
Human vs Worm	93.0	95.6	96.5 (+3.5)
Human Enhancers Cohn	69.5	70.5	73.8 (+4.3)
Human Enhancers Ensembl	68.9	83.5	89.2 (+20.3)
Human Regulatory	93.3	91.5	93.8 (+0.5)
Human Nontata Promoters	84.6	87.7	96.6 (+12)
Human OCR Ensembl	68.0	73.0	80.9 (+12.9)

Single Nucleotide Resolution (cont.)

- The NT models ranged from **500M to 2.5B** parameters, and pretrained on up to 3202 genomes. All NT models use **6-mer** sequences of 1000 tokens long.
- HyenaDNA uses a model with 1500x fewer parameters (1.6M) and 3200x less pretraining data (1 human reference genome).
- For HyenaDNA, we attach a linear decoder head and fine-tune a pretrained model, surpassing SotA on 12 of 17 datasets using a model with orders of magnitude less parameters and pretraining data.

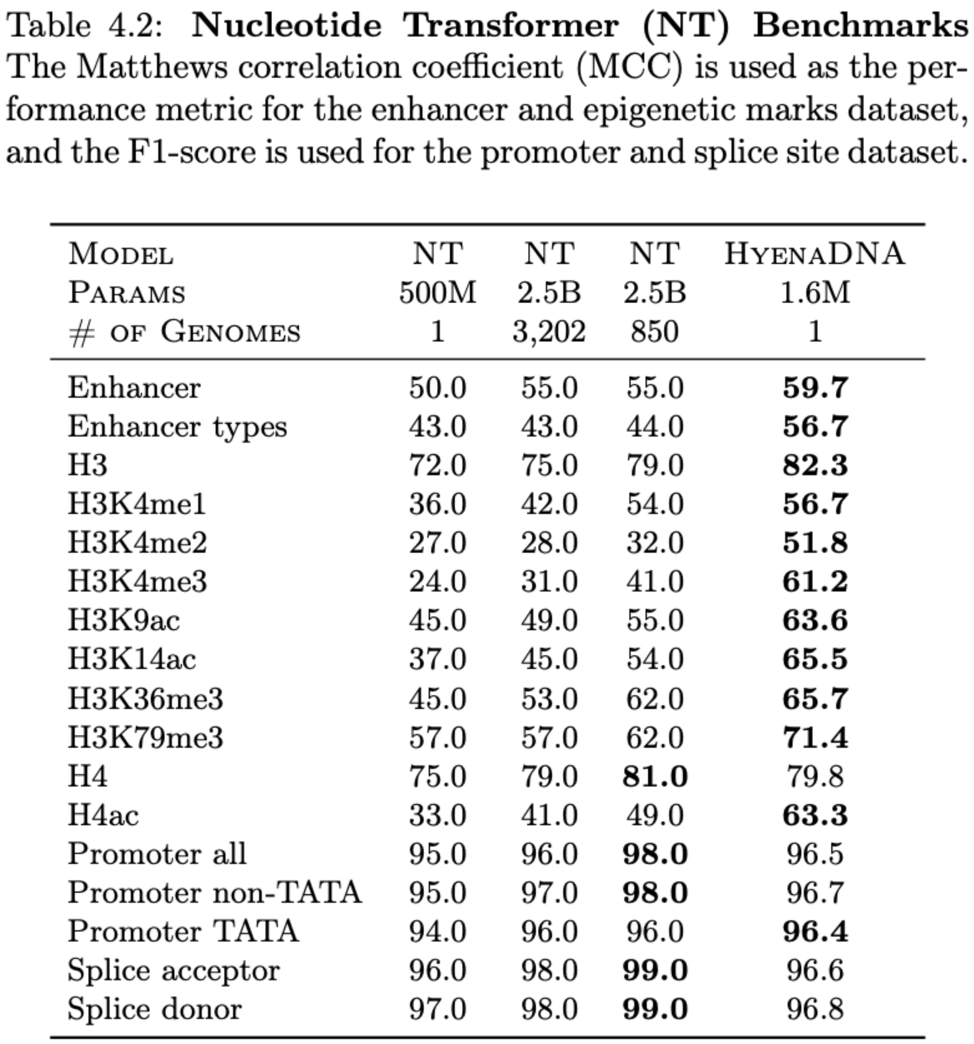


Table 4.2: **Nucleotide Transformer (NT) Benchmarks** The Matthews correlation coefficient (MCC) is used as the performance metric for the enhancer and epigenetic marks dataset, and the F1-score is used for the promoter and splice site dataset.

Model	NT	NT	NT	HyenaDNA
Params	500M	2.5B	2.5B	1.6M
# of Genomes	1	3,202	850	1
Enhancer	50.0	55.0	55.0	**59.7**
Enhancer types	43.0	43.0	44.0	**56.7**
H3	72.0	75.0	79.0	**82.3**
H3K4me1	36.0	42.0	54.0	**56.7**
H3K4me2	27.0	28.0	32.0	**51.8**
H3K4me3	24.0	31.0	41.0	**61.2**
H3K9ac	45.0	49.0	55.0	**63.6**
H3K14ac	37.0	45.0	54.0	**65.5**
H3K36me3	45.0	53.0	62.0	**65.7**
H3K79me3	57.0	57.0	62.0	**71.4**
H4	75.0	79.0	**81.0**	79.8
H4ac	33.0	41.0	49.0	**63.3**
Promoter all	95.0	96.0	**98.0**	96.5
Promoter non-TATA	95.0	97.0	**98.0**	96.7
Promoter TATA	94.0	96.0	96.0	**96.4**
Splice acceptor	96.0	98.0	**99.0**	96.6
Splice donor	97.0	98.0	**99.0**	96.8

In-context Learning for Genomic Sequences

- **soft prompting** $\{T_e, X, SEP\}$
    - prepending T_e, a sequence of **soft tunable tokens** (2 to 32k) directly in the input sequences.
    - include a brief tuning phase (< 20 epochs), updating the soft tokens only, to provide HyenaDNA with the ability to indicate the target classes.
    - for binary classification, for example, indicate the two classes with the letters "A" and "N".
- **instruction fine-tuning** $X: \quad \{X_1, \text{SEP}, Y_1, \text{SEP}, X_2, \text{SEP}, Y_2, \text{SEP}, X, \text{SEP}\},$
    - few-shot learning approach to in-context learning by prepending, consecutively, k (2 to 32) demonstrations of each class and its sequence into the prompt.
    - encode class labels by the use of **individual letters** of HyenaDNA's existing vocabulary.
    - perform a brief instruction-tuning period for each dataset to familiarize HyenaDNA with this task structure by tuning the pretrained model on **a small subset** of the dataset.

In-context Learning for Genomic Sequences (cont.)

- HyenaDNA’s performance on novel tasks improves as more tunable tokens are added into the input sequences, and saturates close to baseline performance.

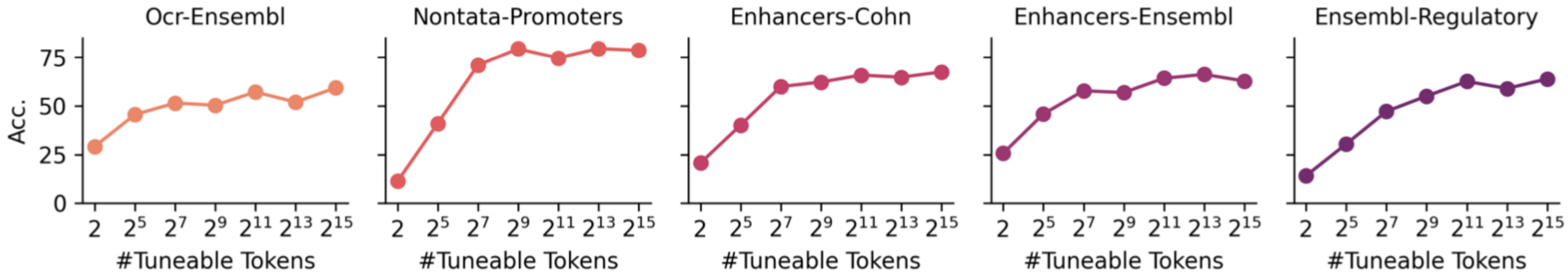
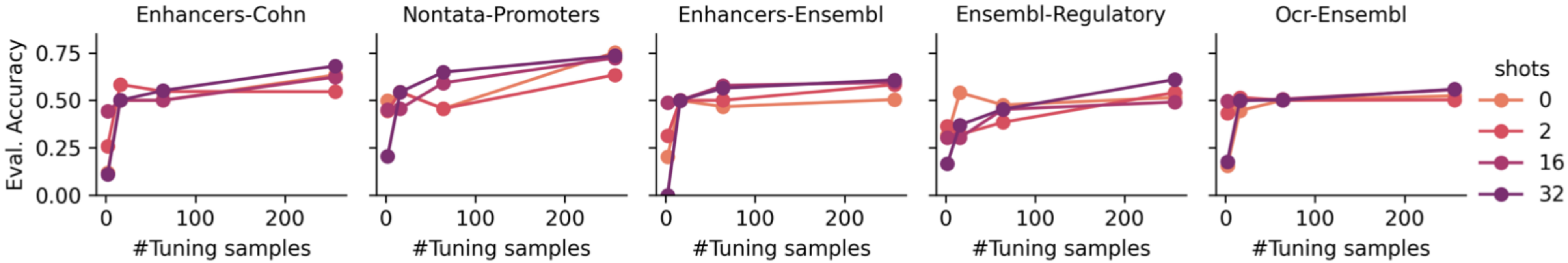


Figure 4.2: **Filling long-context with soft tuneable tokens.** HyenaDNA is able to learn new tasks in-context when adding a sequence of tuneable tokens to the input sequences. Longer sequences of tuneable tokens lead to better performance.

In-context Learning for Genomic Sequences (cont.)

- Increasing k-shot demonstrations to the input **<u>does not necessarily</u>** improve performance. A higher number of tuning samples is needed before k-shot demonstrations start to boost accuracy.



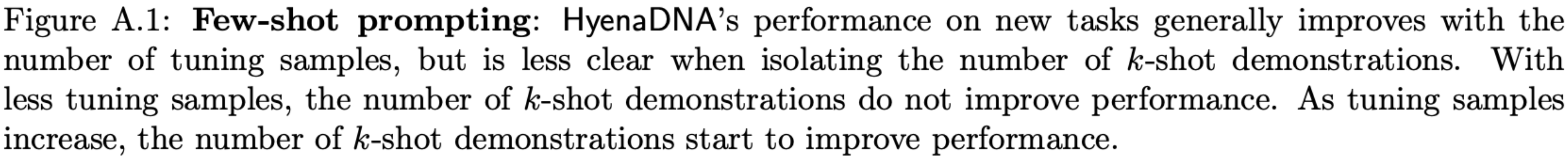


Figure A.1: **Few-shot prompting**: HyenaDNA's performance on new tasks generally improves with the number of tuning samples, but is less clear when isolating the number of k-shot demonstrations. With less tuning samples, the number of k-shot demonstrations do not improve performance. As tuning samples increase, the number of k-shot demonstrations start to improve performance.

Ultralong-Range Genomics

- **Chromatin Profile Prediction**
  - The prediction of chromatin profiles and epigenetic markers from DNA sequences is an important and challenging task to quantify the functional effects of non-coding variants.
- **The DeepSEA dataset**
  - 919 chromatin features including transcription factor (TF) binding profiles, DNase I-hypersensitive sites (DHS) and histone mark (HM) profiles.
  - The task is to jointly predict **919 labels** corresponding to the chromatin profile (similar to peak detection) of a central region of the sequence, indicating the presence of such functional effects.
  - The input also includes flanking regions that provide broader contextual information needed to incorporate **long-range interactions**.

Chromatin Profile Prediction

- DeepSEA, a **convolutional** sequence model, and BigBird, a **sparse attention** based language model.
- The authors of BigBird fine-tune on the DeepSEA dataset with input sequences extended to **8000 bp** (asymmetrically about the center-point by -5000 and +3000 bp).
- HyenaDNA perform competitively against a DeepSea CNN and the SotA sparse attention BigBird baselines using 5-30× fewer parameters.

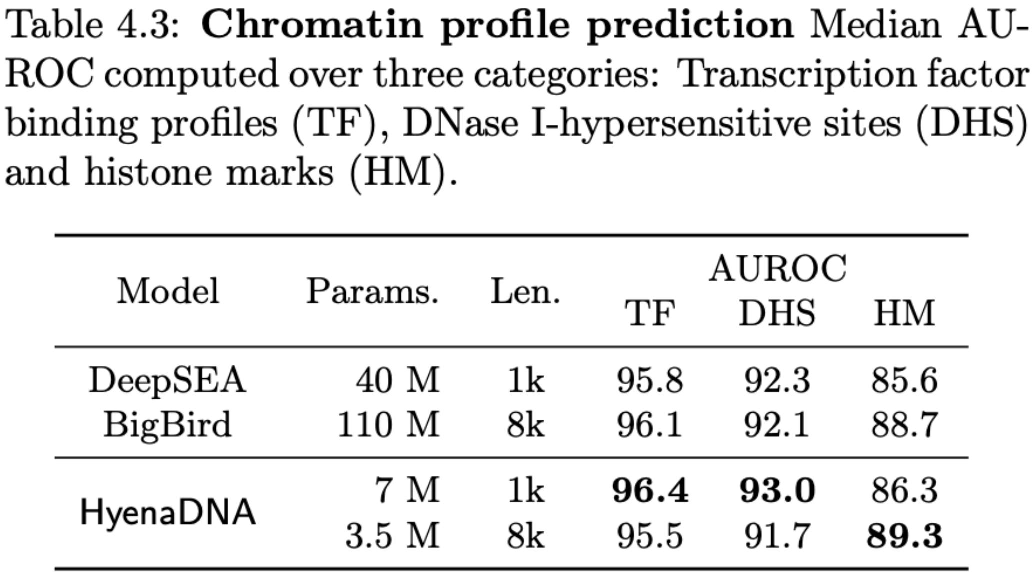


Table 4.3: **Chromatin profile prediction** Median AUROC computed over three categories: Transcription factor binding profiles (TF), DNase I-hypersensitive sites (DHS) and histone marks (HM).

Model	Params.	Len.	AUROC		
			TF	DHS	HM
DeepSEA	40 M	1k	95.8	92.3	85.6
BigBird	110 M	8k	96.1	92.1	88.7
HyenaDNA	7 M	1k	**96.4**	**93.0**	86.3
	3.5 M	8k	95.5	91.7	**89.3**

Biotype Embeddings

- Fit the embeddings using an XGBoost classifier on the 10 most frequent biotypes in the Ensembl dataset, and apply for visualization, **<u>distinct clusterings emerge visually</u>**.

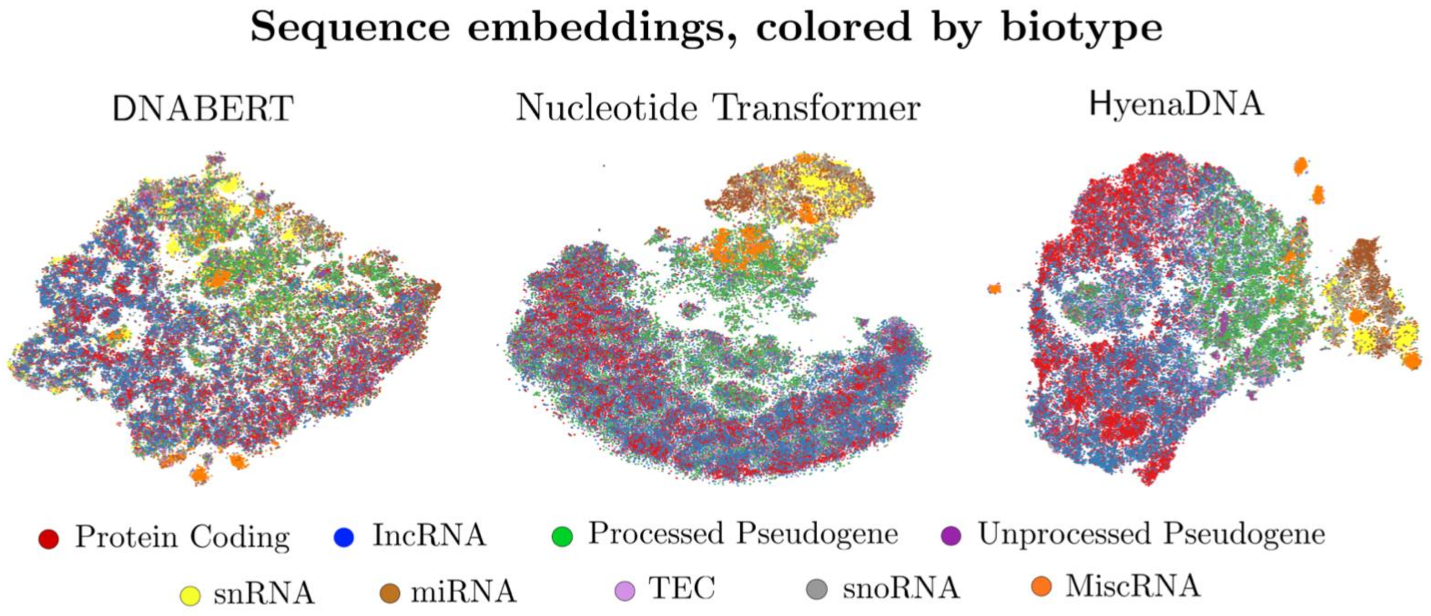


Figure 4.3: **Embedding visualisation.** t-SNE of the embeddings generated by DNABERT, Nucleotide Transformer and HyenaDNA coloured by Ensembl biotype annotations.

Biotype Embeddings (cont.)

- Quantitatively, HyenaDNA produces the highest F1 score in biotype classification (with a much smaller model), indicating that during pretraining, HyenaDNA learns informative features related to biological function.

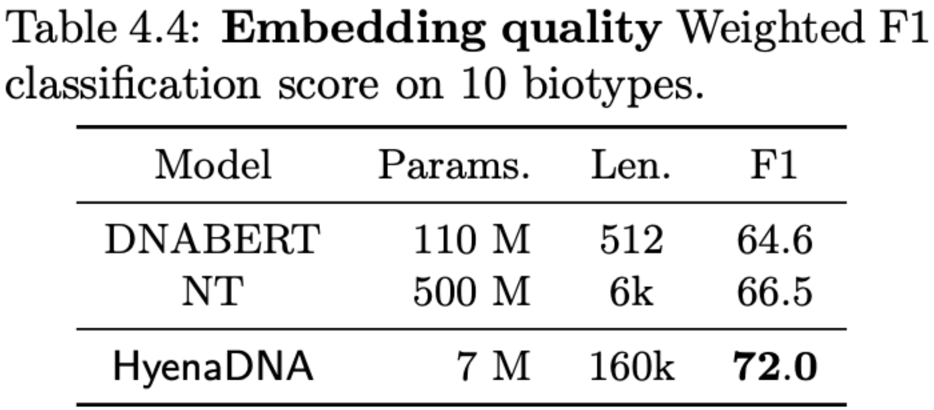


Table 4.4: **Embedding quality** Weighted F1 classification score on 10 biotypes.

Model	Params.	Len.	F1
DNABERT	110 M	512	64.6
NT	500 M	6k	66.5
HyenaDNA	7 M	160k	**72.0**

Species Classification

- Task: to design an **ultralong-range sequence modeling** task to test whether a model can determine the source species of a random genetic sequence
- Randomly sample DNA sequences from **5 different species**: human (homo sapien), lemur (lemur catta), mouse (mus mus-culus), pig (sus scrofa), and hippo (hippopotamus amphibius)
- Hold out **four chromosomes** from each species (chromosome numbers 1, 3, 12, and 13) for evaluation, and use the rest of each species' chromosomes for training
- Compare HyenaDNA against a **baseline Transformer**, which uses Flash Attention in the mixing layer instead of a Hyena operator.
- Either pool across all tokens (1k and 32k models) or use the last token for classification (250k - 1M models).

Species Classification (cont.)

- Both models struggle on shorter sequences of length 1024, but performance improves with longer contexts as the distinct mutational profile of each species becomes more evident.
- HyenaDNA effectively solves the task by using a context length of 450k to 1 million, where Transformer cannot due to infeasible training time limitations.

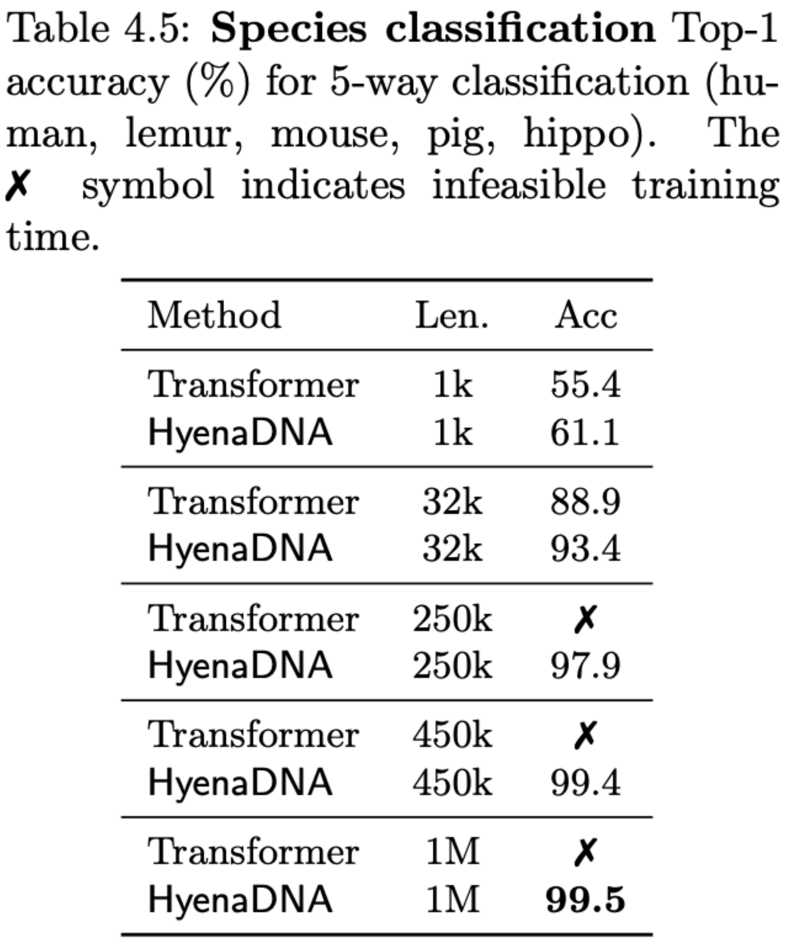


Table 4.5: **Species classification** Top-1 accuracy (%) for 5-way classification (human, lemur, mouse, pig, hippo). The ✗ symbol indicates infeasible training time.

Method	Len.	Acc
Transformer	1k	55.4
HyenaDNA	1k	61.1
Transformer	32k	88.9
HyenaDNA	32k	93.4
Transformer	250k	✗
HyenaDNA	250k	97.9
Transformer	450k	✗
HyenaDNA	450k	99.4
Transformer	1M	✗
HyenaDNA	1M	**99.5**

Thanks