DNAChunker:

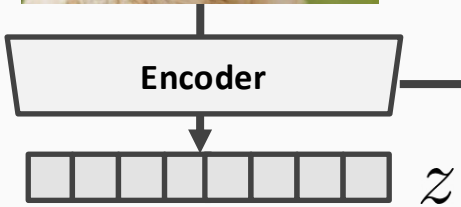
Learnable Tokenization for DNA Language Models

Taewon Kim, Jihwan Shin, Hyomin Kim,
Youngmok Jung, Jonghoon Lee, Won-Chul Lee, Sungsoo Ahn*, Insu Han*

Background

GLMs are essentially representation encoders

Genetic Domain



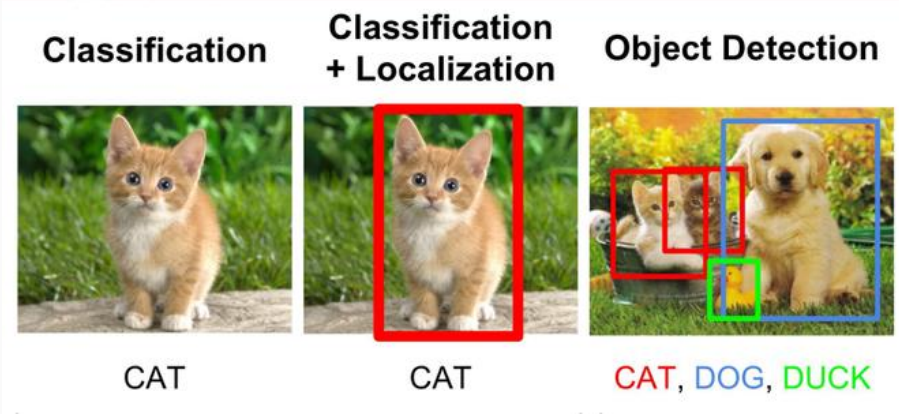
OpenAI CLIP

Meta DINO v2, v3

BEiT, Florence

Diverse SSL approaches have enabled

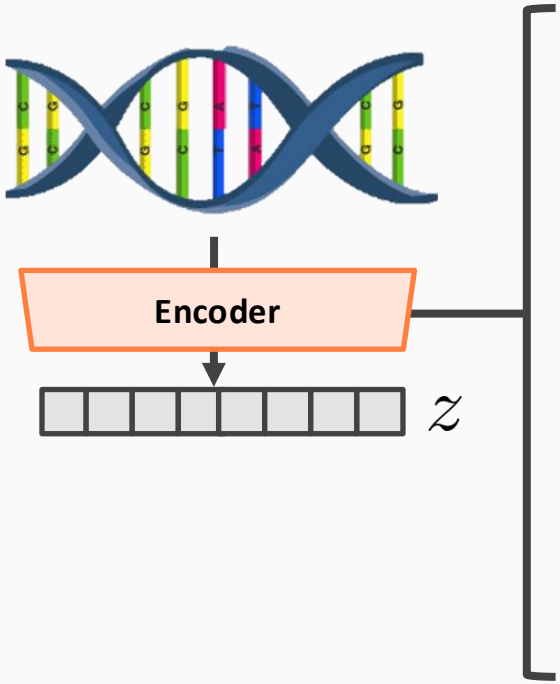
Compression of images to vectors, which can solve:



Background

GLMs are essentially representation encoders

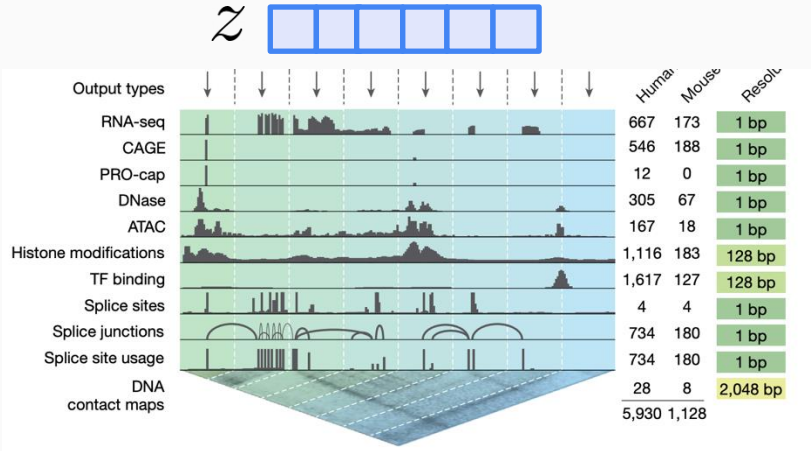
Genetic Domain



e.g. AlphaGenome

Input: Genomic Sequence -> Output: Representation

Either trained with SSL (autoregressive, MLM) or SL (label guidance)



Can solve DNA-specific tasks such as DNA contact map prediction, etc.

Motivation

In this work, we focus on tokenization

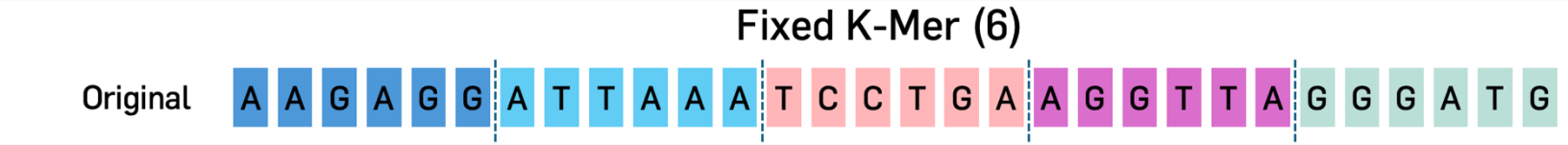
Natural Language : I / love / genomic / language / models

Many meaningful separation exists – ex. **spaces already divide input in a semantically-aware manner**

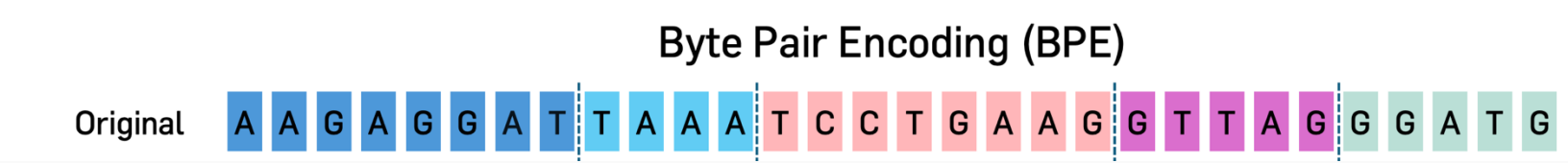
Genomic Lanugage: ATGCATTACATGCATTACATGCATTAC…

No natural separation exists

Thus, people have naively used NLP-style tokenizations



Divide input sequence into 6 bp-length chunks, each with unique numbering.

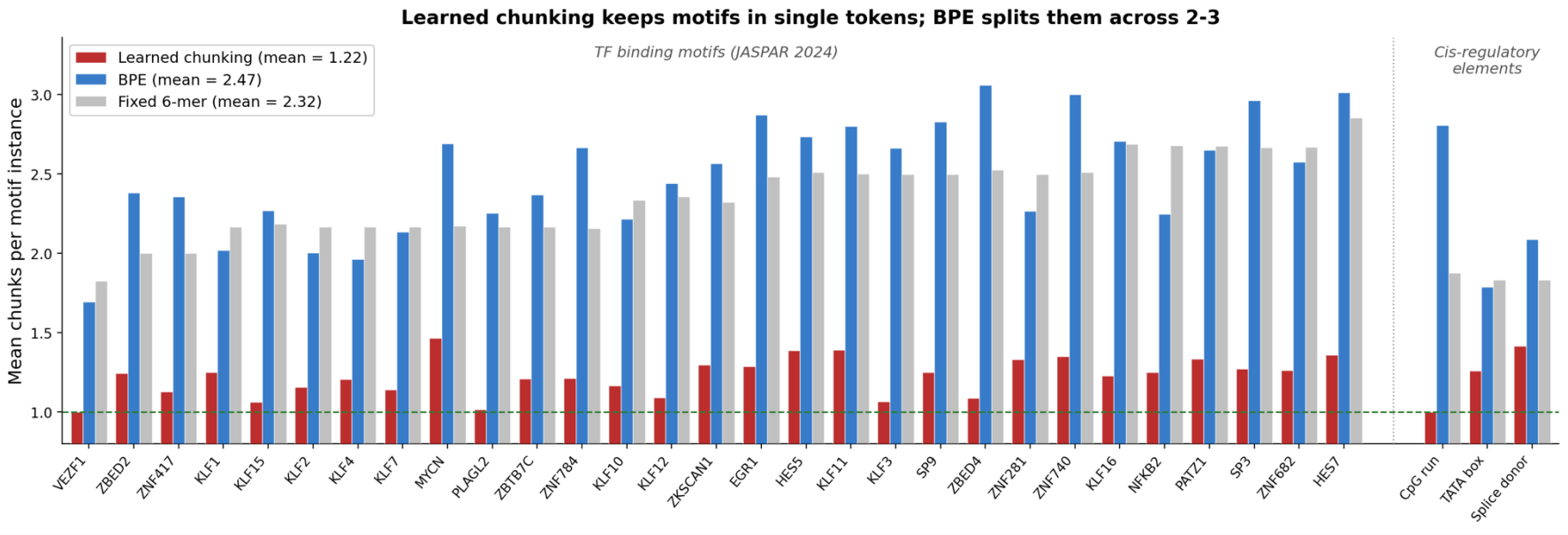


Predefine a set of chunks that are statistically – merging frequent pairs of characters.

Motivation

Prior approaches are ‘semantically blind’ of DNA

Question: Does prior tokenization preserve the known “Grammar (motif)” of DNA?



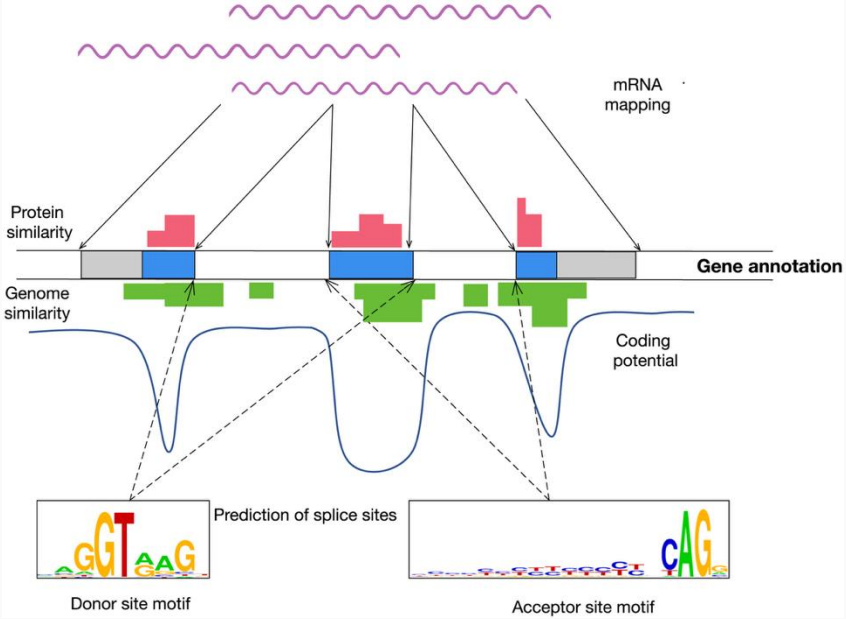
Answer: NO! in fact, it ACTIVELY fragments it!

Single Motif -> BPE divides it into 2.47, K-Mer into 2.32 tokens on average

Motivation

Why Not Use Pre-annotated Gene Elements as Guidance?

TL;DR: Lack of annotation across genomes

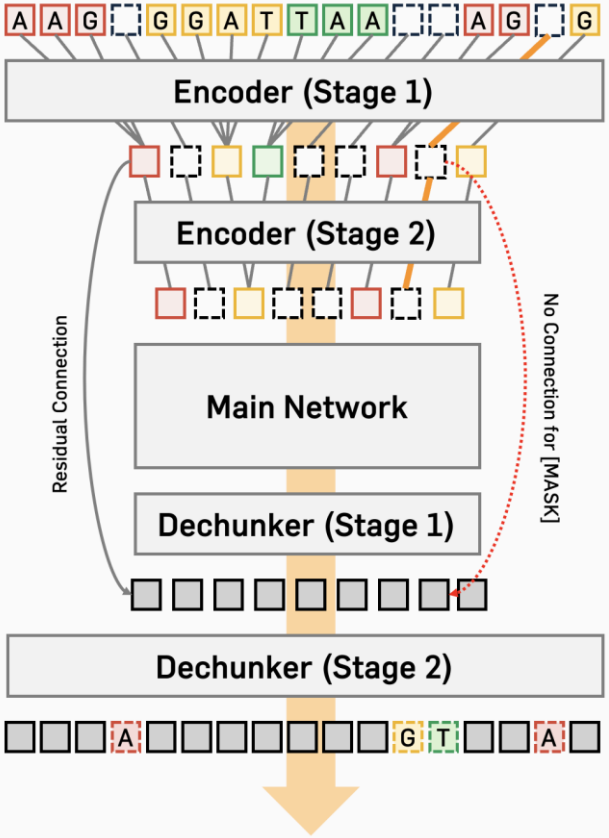


1. Conservation-based reasoning
   - ✓ Highly conservative == Important
   - ✓ **Lack of lineage-specific functionals**
2. Association-based reasoning
   - ✓ Statistically associated with disease
   - ✓ **Correlation != Causal**
3. Experimental Assays …
   - ✓ **Hard to obtain…**

Furthermore, we do not know
WHICH are best for tokenization

Method

DNAChunker: Unsupervised, Data-driven tokenization strategy



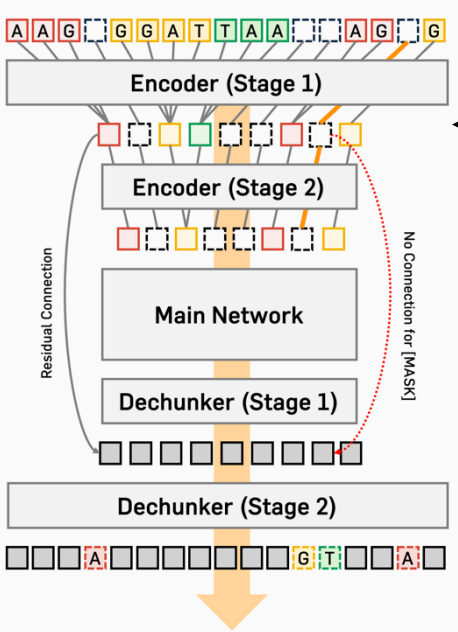
TL;DR: Learn the tokenization strategy from data!

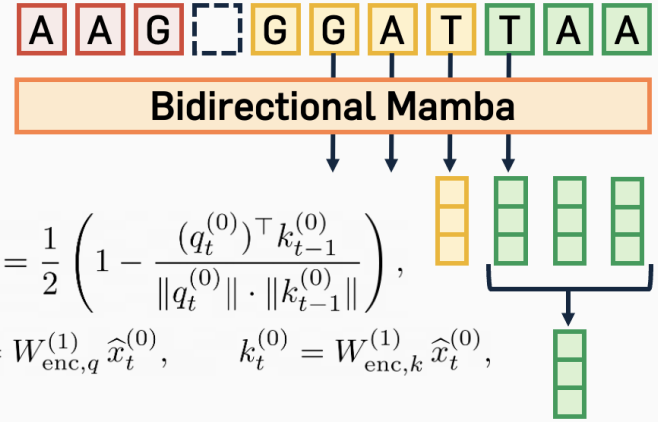
- ✓ First approach to use **learnable tokenization in gLMs**
- ✓ Enforce bidirectional nature of DNA sequences
- ✓ Almost **NO inductive bias**
- ✓ **SOTA performance** on Benchmark Tasks

Method

Grouping similar tokens to a singular chunk

TL;DR: Merge adjacent tokens using cosine similarity





$$p_t^{(0)} = \frac{1}{2}\left(1 - \frac{(q_t^{(0)})^\top k_{t-1}^{(0)}}{\|q_t^{(0)}\| \cdot \|k_{t-1}^{(0)}\|}\right),$$

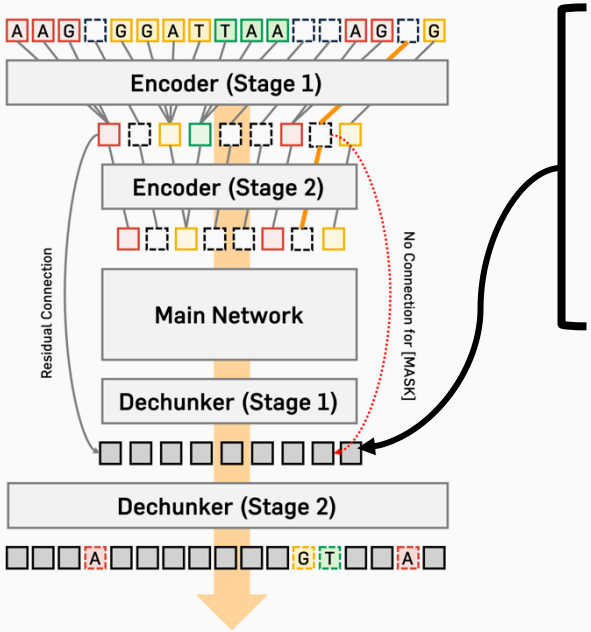
$$q_t^{(0)} = W_{\text{enc},q}^{(1)}\, \widehat{x}_t^{(0)}, \qquad k_t^{(0)} = W_{\text{enc},k}^{(1)}\, \widehat{x}_t^{(0)},$$

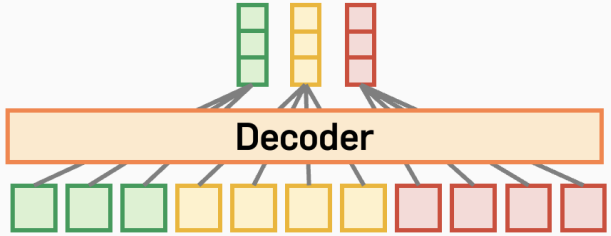
1. Calculate cosine similarity among adjacent embeddings
2. If similarity > 0.5; *assign to different chunk*
3. Else; *assign to same chunk*

Method

Decoding chunks back to original resolution

TL;DR: Tricks need be in place for gradient propagation





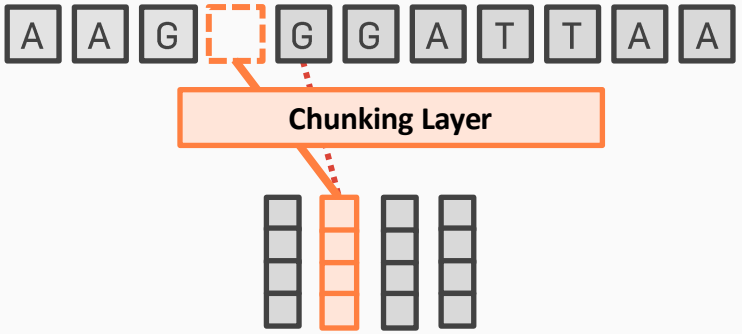
$$z_t^{(s+1)} = \tfrac{1}{2}\left(\textsc{Scan}_{\rightarrow}(\tilde{z}^{(s+1)}, p)_t + \textsc{Scan}_{\leftarrow}(\tilde{z}^{(s+1)}, p)_t\right),$$

1. Embeddings are upsampled to original resolution
2. **Bidirectional EMA scan** is applied
   Trick to ensure gradient propagation to the chunking module (via mixing between boundaries)
3. + Residual connections (like U-Net) to original embedding

Method

[MASK] are different from other letters

- ✓ [MASK] is *NOT seem in test data*. -> Train vs. Test gap exits
- ✓ Learning a chunker WITH *[MASK] makes the tokenization mask-dependent*



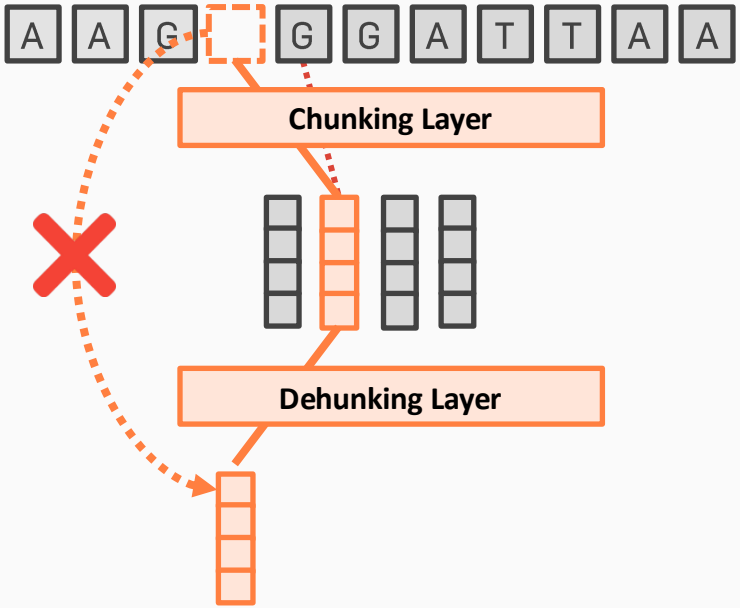
Trick#1: [MASK] Protection Mechanism

- [MASK] should not be merged with other tokens (A,T,G,C,N).

Method

[MASK] are different from other letters

- ✓ [MASK] is *NOT seem in test data*. -> Train vs. Test gap exits
- ✓ Learning a chunker WITH *[MASK] makes the tokenization mask-dependent*



Trick#1: [MASK] Protection Mechanism

- [MASK] should not be merged with other tokens (A,T,G,C,N).

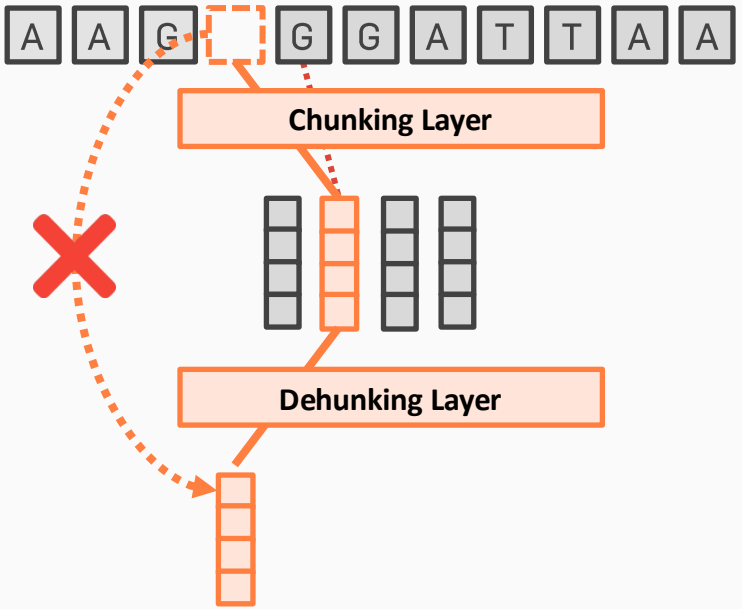
Trick#2: Residual Gating of [MASK]

- If residual connection is ALSO given to the mask, encoder leaks information to decoder

Method

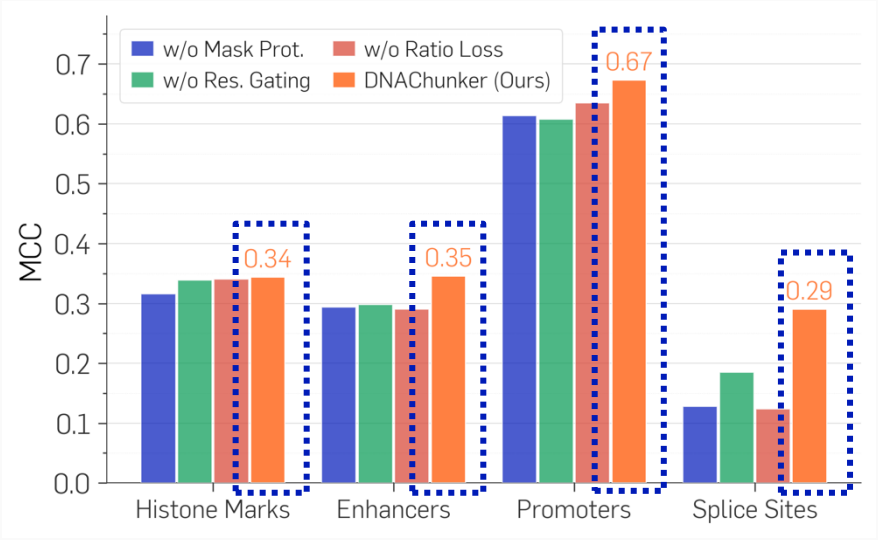
[MASK] are different from other letters

- ✓ [MASK] is *NOT seem in test data*. -> Train vs. Test gap exits
- ✓ Learning a chunker WITH *[MASK] makes the tokenization mask-dependent*



[MASK] handeling is crucial for performance

Linear probing upon NT-revised Benchmark in MCC



Method

Loss Function

Standard MLM loss + Repeat region down-weighting + Ratio Loss

$$\mathcal{L}_{\text{MLM}} = \sum_{t \in M} w_t \mathcal{L}_{\text{CE}}(t) \quad w_t = \begin{cases} 0.1 & \text{if position } t \text{ is in a repetitive region} \\ 1.0 & \text{otherwise} \end{cases}$$

- Following formulation of Evo2[1], down-weight repeat regions
- Repeat regions lack biological significance

[1] Brixi et al. Genome modeling and design across all domains of life with Evo 2

Method

Loss Function

Standard MLM loss + Repeat region down-weighting + Ratio Loss [1]

$$\mathcal{L}_{\text{ratio}}^{(s)} = \frac{\bar{b}^{(s)}\bar{p}^{(s)}}{\alpha^{(s)}} + \frac{(1-\bar{b}^{(s)})(1-\bar{p}^{(s)})}{1-\alpha^{(s)}}, \quad \bar{b}^{(s)} = \frac{1}{T}\sum_{t=1}^{T} b_t^{(s)}, \quad \bar{p}^{(s)} = \frac{1}{T}\sum_{t=1}^{T} p_t^{(s)},$$

- To avoid collapse to trivial solutions (i.e. assigning each character to a separate chunk)
- Loss attains minimum when the fraction of selected boundaries($\bar{b}^{(s)}$) and boundary probability ($\bar{p}^{(s)}$) becomes the target compression ratio ($\alpha^{(s)}$)

[1] Hwang et al. Dynamic Chunking for End-to-End Hierarchical Sequence Modeling

Results

State-of-the-art upon Benchmarks

Table 1. **Nucleotide Transformer Benchmark.** The reported values represent the Matthews Correlation Coefficient (MCC; mean ± standard error) averaged over 10-fold cross-validation. Best results are **bold**; second best are underlined.

	Enformer (252M)	DNABERT-2 (117M)	HyenaDNA (55M)	NT-multi (2.5B)	NT-v2 (500M)	Caduceus-Ph (8M)	Caduceus-PS (8M)	GROVER (87M)	GENErator (1.2B)	DNACHUNKER (172M)
Histone Markers										
H3	0.724 ± 0.018	0.785 ± 0.012	0.781 ± 0.015	0.793 ± 0.013	0.788 ± 0.010	0.794 ± 0.012	0.772 ± 0.022	0.768 ± 0.008	0.806 ± 0.005	**0.817** ± 0.011
H3K14ac	0.284 ± 0.024	0.515 ± 0.009	0.608 ± 0.020	0.538 ± 0.009	0.538 ± 0.015	0.564 ± 0.033	0.596 ± 0.038	0.548 ± 0.020	0.605 ± 0.008	**0.711** ± 0.021
H3K36me3	0.345 ± 0.019	0.591 ± 0.005	0.614 ± 0.014	0.618 ± 0.011	0.618 ± 0.015	0.590 ± 0.018	0.611 ± 0.048	0.563 ± 0.017	0.657 ± 0.007	**0.677** ± 0.003
H3K4me1	0.291 ± 0.016	0.512 ± 0.008	0.512 ± 0.008	0.541 ± 0.005	0.544 ± 0.009	0.468 ± 0.015	0.487 ± 0.029	0.461 ± 0.018	0.553 ± 0.009	**0.631** ± 0.009
H3K4me2	0.207 ± 0.021	0.333 ± 0.013	0.455 ± 0.028	0.324 ± 0.014	0.302 ± 0.020	0.332 ± 0.034	0.431 ± 0.016	0.403 ± 0.042	0.424 ± 0.013	**0.599** ± 0.011
H3K4me3	0.156 ± 0.022	0.353 ± 0.021	0.550 ± 0.015	0.408 ± 0.011	0.437 ± 0.028	0.490 ± 0.042	0.528 ± 0.033	0.458 ± 0.022	0.512 ± 0.009	**0.660** ± 0.045
H3K79me3	0.498 ± 0.013	0.615 ± 0.010	0.669 ± 0.014	0.623 ± 0.010	0.621 ± 0.012	0.641 ± 0.028	0.682 ± 0.018	0.626 ± 0.026	0.670 ± 0.011	**0.731** ± 0.012
H3K9ac	0.415 ± 0.020	0.545 ± 0.009	0.586 ± 0.021	0.547 ± 0.011	0.567 ± 0.020	0.575 ± 0.024	0.564 ± 0.018	0.581 ± 0.015	0.612 ± 0.006	**0.678** ± 0.007
H4	0.735 ± 0.023	0.797 ± 0.008	0.763 ± 0.012	0.808 ± 0.007	0.795 ± 0.008	0.788 ± 0.010	0.799 ± 0.010	0.769 ± 0.017	**0.815** ± 0.008	0.813 ± 0.012
H4ac	0.275 ± 0.022	0.465 ± 0.013	0.564 ± 0.011	0.492 ± 0.014	0.502 ± 0.025	0.548 ± 0.027	0.585 ± 0.018	0.530 ± 0.017	0.592 ± 0.015	**0.687** ± 0.027
Average MCC (↑)	0.393	0.551	0.610	0.569	0.571	0.579	0.606	0.571	0.625	**0.701**
Regulatory Annotation										
Enhancer	0.454 ± 0.029	0.525 ± 0.026	0.520 ± 0.031	0.545 ± 0.028	0.561 ± 0.029	0.522 ± 0.024	0.511 ± 0.026	0.516 ± 0.018	**0.580** ± 0.015	0.558 ± 0.011
Enhancer Type	0.312 ± 0.043	0.423 ± 0.018	0.403 ± 0.056	0.444 ± 0.022	0.444 ± 0.036	0.403 ± 0.028	0.410 ± 0.026	0.433 ± 0.029	0.477 ± 0.017	**0.519** ± 0.005
Promoter All	0.910 ± 0.004	0.945 ± 0.003	0.919 ± 0.003	0.951 ± 0.004	0.952 ± 0.002	0.937 ± 0.002	0.941 ± 0.003	0.926 ± 0.004	0.962 ± 0.002	**0.967** ± 0.013
Promoter NonTATA	0.910 ± 0.006	0.944 ± 0.003	0.919 ± 0.004	0.969 ± 0.003	0.952 ± 0.003	0.935 ± 0.007	0.940 ± 0.002	0.925 ± 0.006	0.962 ± 0.001	**0.971** ± 0.007
Promoter TATA	0.920 ± 0.012	0.911 ± 0.011	0.881 ± 0.020	0.919 ± 0.008	0.933 ± 0.009	0.895 ± 0.010	0.903 ± 0.010	0.891 ± 0.009	0.948 ± 0.008	**0.961** ± 0.015
Average MCC (↑)	0.701	0.750	0.728	0.766	0.768	0.738	0.741	0.738	0.786	**0.796**
Splice Site Annotation										
Splice Acceptor	0.772 ± 0.007	0.909 ± 0.004	0.935 ± 0.005	0.973 ± 0.002	0.973 ± 0.004	0.918 ± 0.017	0.907 ± 0.015	0.912 ± 0.010	**0.981** ± 0.002	0.969 ± 0.013
Splice Site All	0.831 ± 0.012	0.950 ± 0.003	0.917 ± 0.006	0.974 ± 0.004	0.975 ± 0.002	0.935 ± 0.011	0.953 ± 0.005	0.919 ± 0.005	**0.976** ± 0.011	0.968 ± 0.030
Splice Donor	0.813 ± 0.015	0.927 ± 0.003	0.894 ± 0.013	0.974 ± 0.002	0.977 ± 0.007	0.912 ± 0.009	0.930 ± 0.010	0.888 ± 0.012	**0.978** ± 0.001	0.960 ± 0.007
Average MCC (↑)	0.805	0.929	0.915	0.974	0.975	0.922	0.930	0.906	**0.979**	0.965
Total Average MCC (↑)	0.547	0.669	0.694	0.690	0.693	0.680	0.697	0.673	0.728	**0.772**
Total Average Rank (↓)	9.67	6.72	6.00	4.83	4.56	6.33	5.61	7.22	2.06	**1.67**

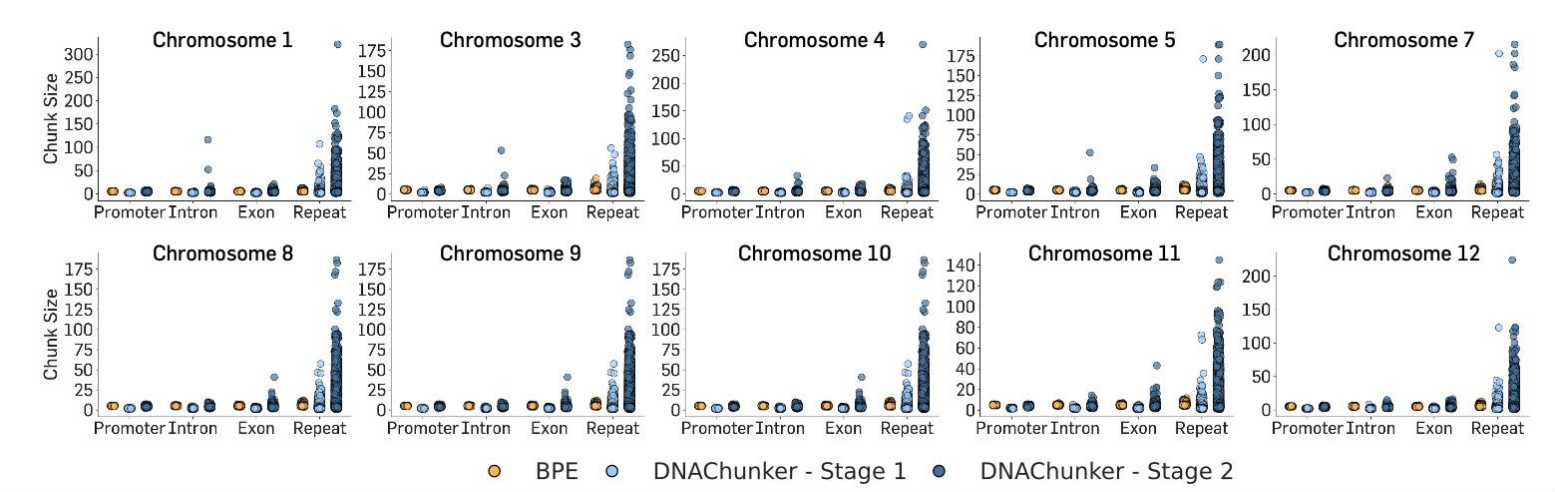
Table 2. **Genomics Benchmark.** The reported values represent the Top-1 Accuracy (Acc; mean ± standard error) averaged over 10-fold cross-validation. Best results are **bold**; second best are underlined.

	DNABERT-2 (117M)	HyenaDNA (55M)	NT-v2 (500M)	Caduceus-Ph (8M)	Caduceus-PS (8M)	GROVER (87M)	GENErator (1.2B)	GENErator-All (1.2B)	DNACHUNKER (172M)
Coding vs. Intergenomic	0.951 ± 0.002	0.902 ± 0.004	0.955 ± 0.001	0.933 ± 0.001	0.944 ± 0.002	0.919 ± 0.002	**0.963** ± 0.000	0.959 ± 0.001	0.955 ± 0.012
Drosophila Enhancers Stark	0.774 ± 0.011	0.770 ± 0.016	0.797 ± 0.009	**0.827** ± 0.010	0.816 ± 0.015	0.761 ± 0.011	0.821 ± 0.005	0.768 ± 0.015	0.779 ± 0.021
Human Enhancers Cohn	0.758 ± 0.005	0.725 ± 0.009	0.756 ± 0.006	0.747 ± 0.003	0.749 ± 0.003	0.738 ± 0.003	**0.763** ± 0.002	0.754 ± 0.006	0.761 ± 0.011
Human Enhancers Ensembl	0.918 ± 0.003	0.901 ± 0.003	0.921 ± 0.004	**0.924** ± 0.002	0.923 ± 0.002	0.911 ± 0.004	0.917 ± 0.002	0.912 ± 0.002	0.922 ± 0.007
Human Ensembl Regulatory	0.874 ± 0.007	0.932 ± 0.001	**0.941** ± 0.001	0.938 ± 0.004	**0.941** ± 0.002	0.897 ± 0.001	0.928 ± 0.001	0.926 ± 0.001	0.935 ± 0.005
Human non-TATA Promoters	0.957 ± 0.008	0.894 ± 0.023	0.932 ± 0.006	0.961 ± 0.003	0.961 ± 0.002	0.950 ± 0.005	0.958 ± 0.001	0.955 ± 0.005	**0.962** ± 0.001
Human OCR Ensembl	0.806 ± 0.003	0.774 ± 0.004	0.813 ± 0.001	0.825 ± 0.004	**0.826** ± 0.003	0.789 ± 0.002	0.823 ± 0.002	0.812 ± 0.003	0.810 ± 0.007
Human vs. Worm	0.977 ± 0.001	0.958 ± 0.004	0.976 ± 0.001	0.975 ± 0.001	0.976 ± 0.001	0.966 ± 0.001	**0.980** ± 0.000	0.978 ± 0.001	0.969 ± 0.001
Mouse Enhancers Ensembl	0.865 ± 0.014	0.756 ± 0.030	0.855 ± 0.018	0.788 ± 0.028	0.826 ± 0.021	0.742 ± 0.025	0.871 ± 0.015	0.784 ± 0.027	**0.874** ± 0.020
Average Acc (↑)	0.876	0.846	0.883	0.880	0.885	0.853	**0.892**	0.872	0.885
Average Rank (↓)	5.11	8.22	4.17	3.89	3.33	8.11	**2.89**	5.44	3.29

- On average, **BEST performance**
- Evaluated upon:
  - NT-Benchmark (Best avg. Perf. & Rank)
  - NT-revised Benchmark (Best avg. Rank)
  - Genomics Benchmark (2nd avg. Rank)
  - BEND Benchmark (Best avg. Rank)
  - DNALongBench (Best avg. Perf. & Rank)

Analysis

DNAChunker learns biological grammar

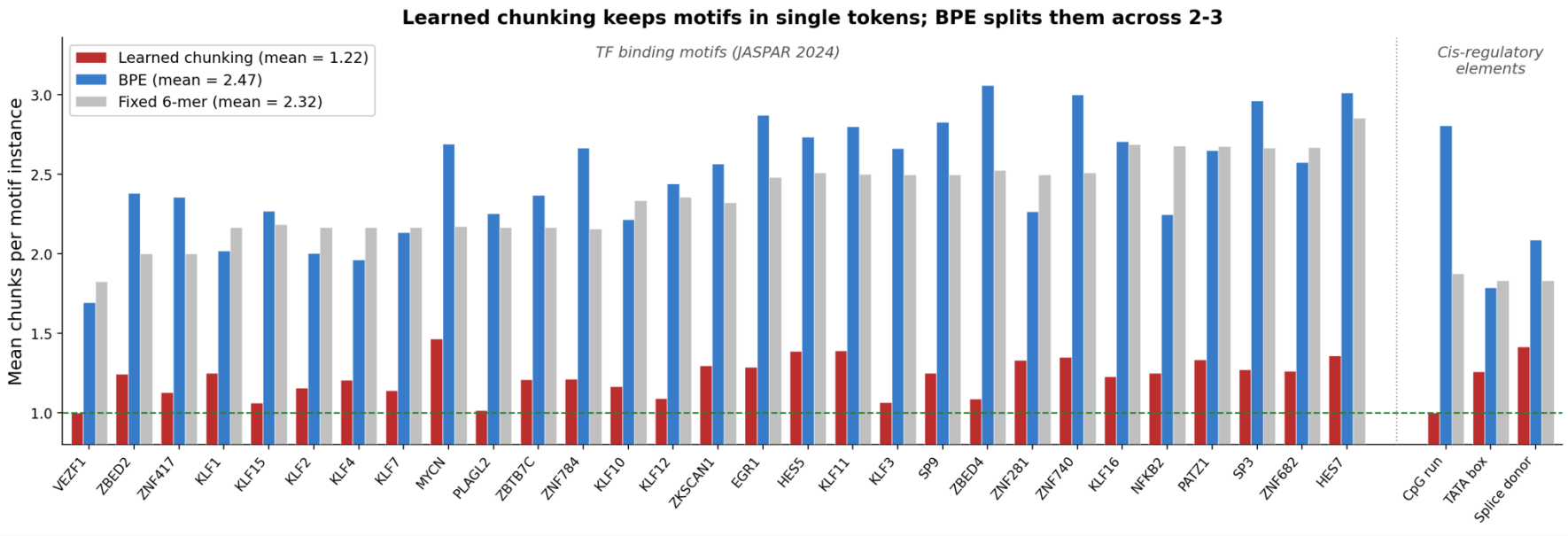


- Scatter Plot upon ***chunk size distribution*** in chromosomes, ***divided by annotation.***
- Large chunk sizes assigned to repeats – where chunk sizes are larger than 150+ bp

Analysis

Revisiting ‘semantic blindness’ to biological grammar

Question: DNAChunker preserve the known “Grammar (motif)” of DNA?

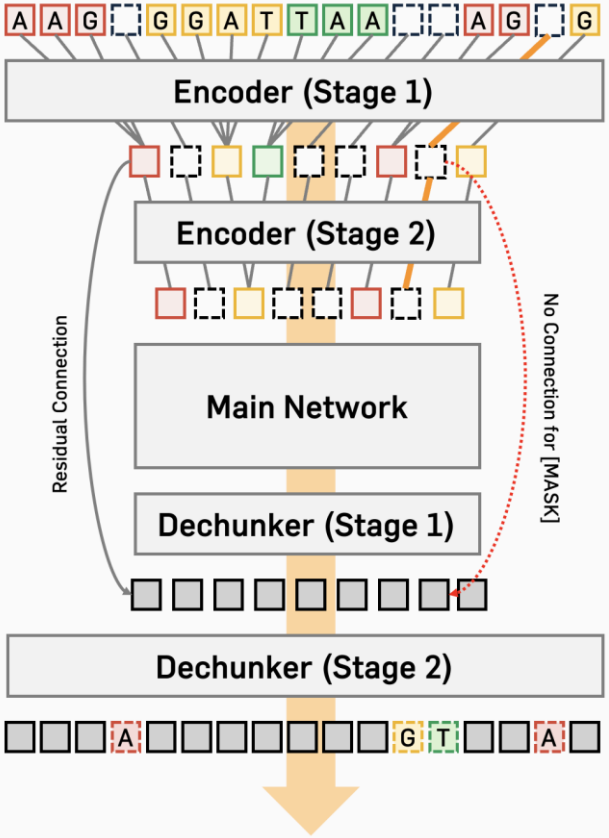


Answer: Yes! in fact, it assigns in average, 1.22 tokens to motifs, ACTIVELY preserving it.

Single Motif -> BPE divides it into 2.47, K-Mer into 2.32 tokens on average

Summary

Data-driven tokenization is beneficial for DNAs!



TL;DR: Learn the tokenization strategy from data!

- ✓ First approach to use **learnable tokenization in gLMs**
- ✓ Enforce bidirectional nature of DNA sequences
- ✓ Almost **NO inductive bias**
- ✓ **Learned tokenization respects biological grammar**
- ✓ Currently working on **cancer-subtype classification with DNAChunker** on **patient data**

Personal Link

Paper Link