

NucEL: Single-Nucleotide ELECTRA-Style Genomic Pre-training for Efficient and Interpretable Representations

Ke Ding^{1,2,3}, Brian Parker^{3,4,5*}, Jiayu Wen^{1,2,3*}

¹John Curtin School of Medical Research, Australian National University

²The Shine-Dalgarno Centre for RNA Innovation, Australian National University

³ARC Centre of Excellence for the Mathematical Analysis of Cellular Systems

⁴School of Computing, Australian National University

⁵Biological Data Science Institute, Australian National University

Ke.Ding@anu.edu.au, Brian.Parker@anu.edu.au, Jiayu.Wen@anu.edu.au

Abstract

Pre-training large language models on genomic sequences has become a powerful approach for learning biologically meaningful representations. While masked language modeling (MLM)-based approaches, such as DNABERT and Nucleotide Transformer (NT), achieve strong performance, they are hindered by inefficiencies due to partial token supervision, pre-training/fine-tuning mismatches, and high computational costs. We introduce NucEL, the first ELECTRA-style pre-training framework for genomic foundation models, which overcomes these challenges. Through a discriminator network identifying tokens modified by a generator, NucEL achieves comprehensive token-level supervision across all sequence positions, thereby markedly improving training efficiency relative to the partial supervision of masked positions inherent in MLM frameworks. By integrating ModernBERT’s architectural advancements, including hybrid local-global attention and flash attention mechanisms, NucEL establishes an optimized BERT architecture for genomic sequence modeling. Unlike traditional methods that tokenize genomic sequences into 6-mers, NucEL implements single-nucleotide tokenization, enabling fine-grained resolution and improving both efficiency and interpretability. Pre-trained on the human genome only, NucEL achieves state-of-the-art performance on benchmark datasets across diverse downstream tasks in both human and non-human species, including regulatory element identification (e.g., promoters, enhancers), transcription factor binding prediction in human and mouse, open chromatin region classification, and histone modification profiles, surpassing MLM-based models of similar size and rivaling models 25 times larger, such as NT. Ablation studies provide critical insights into tokenization and masking strategies, optimizing ELECTRA-style pretraining for DNA sequences. Attention analyses reveal NucEL’s superior ability to capture biologically relevant sequence motifs compared to NT, offering valuable insights into its hierarchical learning process and regulatory element modeling capabilities. This work highlights the potential of ELECTRA-style pretraining as an efficient and effective strategy for advancing genomic representation learning with broad implications for future genomic research.

Code — <https://github.com/FreakingPotato/NucEL>

*Corresponding authors.

Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

Introduction

Understanding genomic sequences is a fundamental challenge in biology, as DNA encodes the instructions for cellular function, development, and evolution. Large language models (LLMs), which have transformed natural language processing (NLP), are now being adapted for biological sequences, showing remarkable success across genomics, transcriptomics, proteomics, metagenomics, and epigenomics (Dalla-Torre et al. 2025; Sasse, Chikina, and Mostafavi 2024; Jumper et al. 2021; Abramson et al. 2024). Genomic language models (gLMs) are particularly promising for decoding the functional logic of genomes, revealing how DNA sequences are organized and interact to produce the complexity of biological systems.

Traditional deep learning methods for genomics, such as convolutional neural networks (CNNs) (LeCun, Bengio, and Hinton 2015) and recurrent neural networks (RNNs) (Hochreiter and Schmidhuber 1997), are limited: CNNs struggle with long-range dependencies in DNA sequences, while RNNs process sequences inefficiently. Transformer-based architectures overcome these shortcomings by leveraging pre-training on vast unlabeled genomic datasets to learn meaningful representations of functional elements such as promoters and enhancers, facilitating effective transfer learning for downstream tasks (Devlin et al. 2019). However, existing transformer-based models, such as DNABERT (Ji et al. 2021) and Nucleotide Transformer (Dalla-Torre et al. 2025), rely on masked language modeling (MLM), which is computationally inefficient. MLM learns from only a small subset of the input—typically 15% of masked tokens—underutilizing the majority of the sequence during training (Clark et al. 2020). Additionally, the pre-training (with masking) and fine-tuning (without masking) mismatch can impair transfer learning efficacy.

To address these limitations, we introduce NucEL, the first genomic foundation model to adopt the ELECTRA framework’s replaced token detection (RTD) objective (Clark et al. 2020). Unlike MLM, RTD employs a generator-discriminator architecture, enabling dense token-level supervision across all input positions and significantly boosting training efficiency. NucEL integrates ModernBERT’s architectural innovations (Warner et al. 2024), including hybrid local-global attention (Beltagy, Peters, and Cohan 2020)

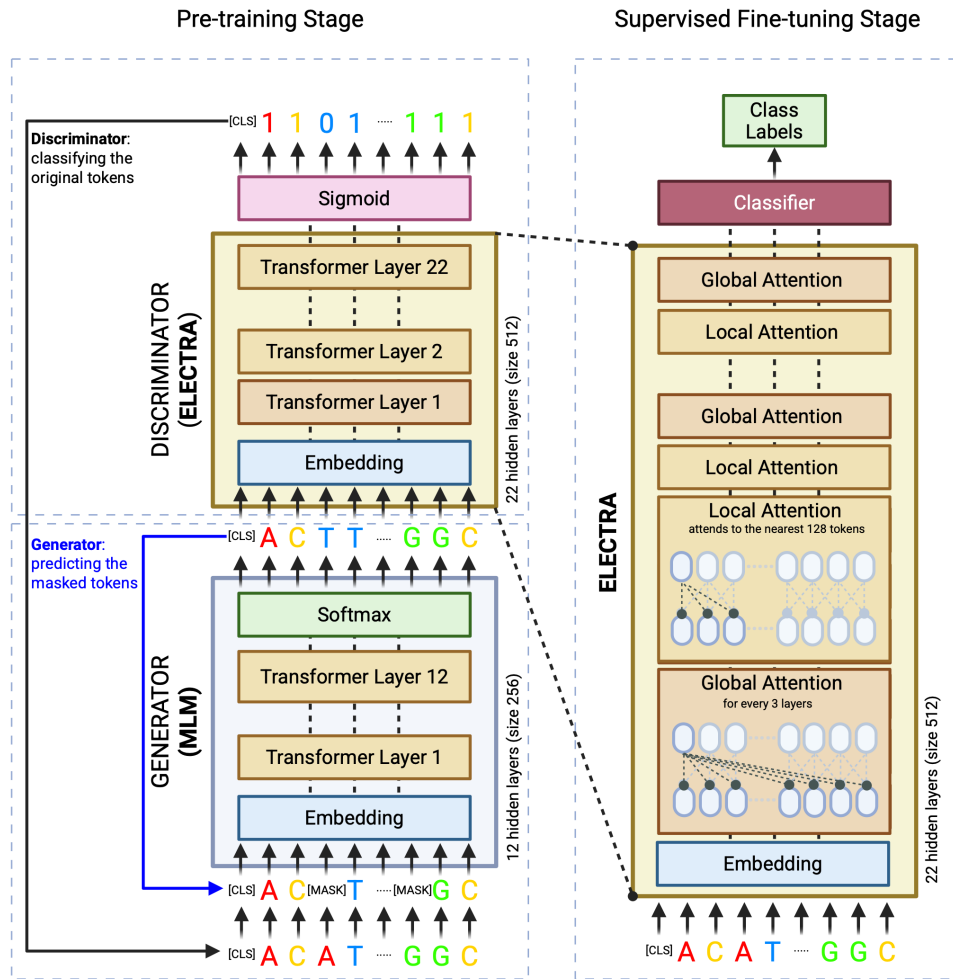


Figure 1: Overview of NucEL the Electra-style Genomic Pre-training Framework. Illustrates the two-stage ELECTRA-style pre-training framework for NucEL, depicting the generator-discriminator architecture during the pre-training stage and the supervised fine-tuning stage for downstream genomic tasks.

and flash attention mechanisms (Dao 2023), to capture both short and long-range genomic dependencies effectively. Unlike k-mer (Dalla-Torre et al. 2025; Ji et al. 2021) or BPE tokenization (Zhou et al. 2023; Sanabria et al. 2024), NucEL uses single-nucleotide tokenization, which may yield more tokens for long sequences but ensures fine-grained resolution and interpretability (Dotan et al. 2024). Combining single-nucleotide resolution with RTD optimizes efficiency and precision: it preserves base-level detail for tasks like mutation analysis, while dense supervision and hybrid attention mitigate the computational cost of longer sequences and enhance long-range interaction modeling. We evaluate NucEL across a range of benchmark tasks—including promoter classification, transcription factor binding prediction, and epigenetic mark profiling—and demonstrate that it achieves state-of-the-art performance while using significantly fewer parameters than larger MLM-based models. Our contributions include:

- Introduced the first ELECTRA-style genomic foundation model, leveraging Replaced Token Detection (RTD) to enhance training efficiency over MLM-based approaches by utilizing all input tokens.
- Achieved state-of-the-art performance on genome foundation model benchmarks with significantly fewer parameters than MLM-based models.
- Utilized single-nucleotide resolution and fine-grained attention analyses to enable precise genomic modeling while enhancing interpretability by revealing complex biological representations.
- Integrated ModernBERT’s hybrid local-global attention and flash attention mechanisms to efficiently capture genomic dependencies while optimizing computational performance.

Our results demonstrate that NucEL provides a practical and efficient approach to learning biologically meaningful genomic representations, advancing genomic lan-

guage modeling while reducing computational demands, thus making advanced genomic models more accessible to the broader research community.

Related Work

Genomic sequence modeling has evolved rapidly with deep learning, particularly through transformer-based models adapted from natural language processing. Here, we review key developments in genomic foundation models, pre-training objectives and tokenization strategies, positioning NucEL within this landscape.

Genomic Language Models

Transformer-based models have become a cornerstone of genomic representation learning. DNABERT (Ji et al. 2021) pioneered this approach by adapting BERT’s masked language modeling (MLM) to DNA sequences, using k -mer tokenization to predict masked tokens for tasks like promoter prediction. DNABERT-2 (Zhou et al. 2023) refined this with Byte Pair Encoding (BPE) tokenization and architectural improvements, enhancing performance on genomic benchmarks. The Nucleotide Transformer (NT) series (Dalla-Torre et al. 2025) scaled this paradigm to billions of parameters, pretrained on multi-species datasets, demonstrating strong transfer learning capabilities. Additionally, ModernBERT (Warner et al. 2024) introduced architectural innovations such as hybrid local-global attention and flash attention mechanisms, improving transformer efficiency for sequence modeling. However, MLM-based models are limited by partial token supervision (only $\sim 15\%$ of tokens) and a pretraining-finetuning mismatch, reducing efficiency.

Alternatives to transformers include state-space models (SSMs) like HyenaDNA (Nguyen et al. 2023) and Caduceus (Schiff et al. 2024), which use single-nucleotide resolution and reverse-complement awareness for efficient long-sequence modeling. Convolutional neural networks (CNNs), such as DeepBind (Alipanahi et al. 2015) and Enformer (Avsec et al. 2021), remain relevant for capturing local motifs but struggle with long-range dependencies compared to transformers.

Pretraining Objectives

Pretraining objectives shape the efficiency and effectiveness of genomic models. MLM, used in DNABERT (Ji et al. 2021) and NT (Dalla-Torre et al. 2025), masks a subset of tokens for prediction, but its sparse supervision leaves most of the forward pass unutilized. Next-Sentence Prediction (NSP), used alongside with MLM in BERT (Devlin et al. 2019), is rarely adopted in genomics due to DNA’s lack of paragraph-like structure, offering limited benefits (Poli et al. 2023). In contrast, Replaced Token Detection (RTD), introduced by ELECTRA (Clark et al. 2020), uses a generator-discriminator framework to provide dense supervision across all tokens, significantly improving efficiency. RTD has proven effective in NLP (e.g., DeBERTa-v3 (He, Gao, and Chen 2021)), but its potential in genomics remains untapped, motivating our NucEL’s approach.

Tokenization Strategies

Tokenization critically shapes a model’s ability to represent genomic sequences. K -mer tokenization, as used in DNABERT (Ji et al. 2021), compresses sequences by grouping nucleotides into k -length substrings but fractures single-nucleotide context, exponentially inflating vocabulary size and diluting base-level detail. Byte Pair Encoding (BPE), adopted by DNABERT-2 (Zhou et al. 2023) and GROVER (Sanabria et al. 2024), learns variable-length subwords to balance efficiency and motif capture but can obscure single-nucleotide variations and introduce inconsistencies for similar sequences. Single-nucleotide tokenization, as in HyenaDNA (Nguyen et al. 2023) and Caduceus (Schiff et al. 2024), preserves base-level resolution and avoids vocabulary explosion, making it ideal for tasks requiring fine-grained precision, such as modeling point mutations and identifying regulatory motifs such as transcription factor binding motifs.

Positioning NucEL

While MLM-based models perform well on downstream tasks, their inefficiencies in supervision and computational demands drive the need for alternatives. NucEL introduces the first ELECTRA-style framework for genomic modeling, leveraging RTD to achieve dense supervision and superior sample efficiency compared to MLM-based approaches. By integrating ModernBERT’s hybrid attention and flash attention mechanisms (Warner et al. 2024) and adopting single-nucleotide tokenization, NucEL maximizes interpretability and efficiency. This enables state-of-the-art performance with fewer parameters, positioning NucEL as a scalable and effective solution for genomic representation learning.

Methods

This section describes NucEL’s training strategies, model architecture, pre-training data and tokenization, optimization, and fine-tuning approach.

Training Strategies

NucEL uses a RTD objective, training two transformer encoders in a generator-discriminator framework. The generator, a smaller transformer, performs MLM to predict masked tokens, while the discriminator, a larger transformer, identifies whether tokens are original or replaced by the generator. This provides dense supervision across all tokens, enhancing sample efficiency. Quantitatively, RTD supervises all N tokens in a sequence of length N , unlike MLM’s sparse supervision of $r \cdot N$ tokens (e.g., $r = 0.15$), yielding a supervision ratio of approximately 6.67. The total loss combines the generator’s MLM loss and the discriminator’s RTD loss, weighted by $\lambda = 50.0$:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{generator}} + \lambda \cdot \mathcal{L}_{\text{discriminator}}, \quad (1)$$

where $\mathcal{L}_{\text{generator}}$ is the cross-entropy over masked tokens, and $\mathcal{L}_{\text{discriminator}}$ is the binary cross-entropy over all tokens. For a sequence $\mathcal{X} = \{x_1, x_2, \dots, x_N\}$ with masked positions

$\mathcal{M} \subseteq \{1, 2, \dots, N\}$, the generator loss is:

$$\mathcal{L}_{\text{generator}} = -\frac{1}{|\mathcal{M}|} \sum_{i \in \mathcal{M}} \log P_G(x_i | \mathcal{X}_{\text{masked}}), \quad (2)$$

and the discriminator loss is:

$$\mathcal{L}_{\text{discriminator}} = -\frac{1}{N} \sum_{i=1}^N \left[y_i \log P_D(y_i = 1 | \mathcal{X}') + (1 - y_i) \log(1 - P_D(y_i = 1 | \mathcal{X}')) \right] \quad (3)$$

where \mathcal{X}' is the sequence with masked tokens replaced by the generator’s predictions, and $y_i = 1$ if $x'_i = x_i$, else 0.

Model Architecture

NucEL’s generator-discriminator framework is shown in Figure 1. The generator comprises 11 transformer layers, a hidden size of 256, and 8 attention heads. The discriminator, NucEL’s primary model, has 22 layers, a hidden size of 512, and 16 attention heads. Both use a hybrid attention mechanism, combining local attention windows (128 tokens) with global attention every third layer, balancing efficiency and long-range dependency modeling (see details in Appendix).

Pre-training Data and Tokenization

NucEL was pre-trained on the human genome (GRCh38/Hg38) using single-nucleotide tokenization, where each base (A, C, G, T) is a distinct token. Sequences were extracted using an overlapping sliding window (1224 bp, 100 bp overlap), with 1024 bp segments randomly sampled during training. The vocabulary consists of 27 tokens: 4 nucleotides, 7 special tokens ([PAD], [UNK], [SEP], [CLS], [MASK], [BOS], [EOS]), and 16 reserved for future extensions (see Appendix). Ambiguous bases (N) are mapped to [UNK] or omitted, though rare.

Single-nucleotide tokenization preserves base-level precision but increases sequence length compared to alternatives like k-mer ($k=6$) and Byte Pair Encoding (BPE). For a sequence of length L , the tokenized length for single-nucleotide tokenization is $N_{1\text{-mer}} = L$, while for k-mer tokenization with overlapping windows, it is $N_{k\text{-mer}} = L - k + 1$ (e.g., for $k = 6$), and for BPE, it is approximately $N_{\text{BPE}} \approx \alpha L$, where $0.1 < \alpha < 0.5$ depending on the learned vocabulary. The vocabulary size for single-nucleotide tokenization is $|\mathcal{V}_{1\text{-mer}}| = 27$, significantly smaller than $|\mathcal{V}_{k\text{-mer}}| = 4^k$ (e.g., $4^6 = 4096$ for $k = 6$) for k-mer tokenization, and $|\mathcal{V}_{\text{BPE}}| \approx 5000$ to 8000 for BPE. While single-nucleotide tokenization increases sequence length, this is mitigated by NucEL’s local attention mechanism, which reduces computational complexity while preserving base-level precision essential for fine-grained genomic tasks.

Pre-training and Fine-tuning Details

NucEL was trained using the AdamW optimizer (learning rate 1×10^{-4} , $\beta_1 = 0.9$, $\beta_2 = 0.999$) for 50 epochs, with a global batch size of 192 and mixed precision (FP16) on 8 NVIDIA A100 GPUs. Training stability was ensured by a 1000-step warmup schedule and a maximum gradient norm

of 1.0 (see Appendix). For downstream tasks, a linear layer was added atop the discriminator’s [CLS] token output, and the entire model was fine-tuned end-to-end.

Experiments

We evaluated NucEL on three human genome-focused benchmark suites: Genome Understanding Evaluation (GUE) Zhou et al. (2023), Genomic Benchmarks (GB) Grešová et al. (2023), and Nucleotide Transformer (NT) Dalla-Torre et al. (2025) tasks. These cover diverse genomic tasks, including gene regulation, epigenomics, and sequence classification (see Appendix for details). Performance was assessed using the Matthews Correlation Coefficient (MCC) for all benchmarks ensuring robust evaluation across imbalanced datasets. In addition, accuracy was reported for GB tasks.

Evaluating Regulatory Sequence Recognition on GUE Benchmark

Model	Human Tasks			
	TF-H	PD	CPD	SSP
DNABERT (3-mer)	64.43	84.63	72.96	84.14
DNABERT (4-mer)	64.41	82.99	71.10	84.05
DNABERT (5-mer)	50.46	84.04	72.03	84.02
DNABERT (6-mer)	64.17	81.70	71.81	84.07
NT-500M-human	50.82	85.51	66.54	79.71
NT-500M-1000g	58.92	86.58	69.13	80.97
NT-2500M-1000g	61.99	86.61	68.17	85.78
NT-2500M-multi	63.32	88.14	71.62	<u>89.36</u>
DNABERT-2 (117M)	70.10	84.21	70.52	84.99
NucEL (93M)	<u>67.64</u>	<u>87.10</u>	75.13	90.30
Model	Other Species			
	TF-M (Mouse)	EMP (Yeast)	CVC (Virus)	Average Performance
DNABERT (3-mer)	57.73	49.54	62.23	67.95
DNABERT (4-mer)	59.58	48.59	59.87	67.23
DNABERT (5-mer)	54.85	48.62	63.64	65.38
DNABERT (6-mer)	56.43	49.10	55.50	66.11
NT-500M-human	45.24	45.35	57.13	61.47
NT-500M-1000g	49.31	47.68	52.06	63.52
NT-2500M-1000g	56.82	50.86	66.73	68.14
NT-2500M-multi	67.01	58.06	73.04	72.94
DNABERT-2 (117M)	<u>67.99</u>	55.98	<u>71.02</u>	72.11
NucEL (93M)	70.62	65.01	70.29	75.16

Table 1: Performance comparison on GUE (MCC for all tasks except CVC which uses F1; Averaged over 3 Seeds; **Best**, Second-Best)

We first evaluate NucEL on the human subset of the GUE benchmark, including transcription factor binding (TF-H), promoter detection (PD), core promoter detection (CPD), and splice site prediction (SSP). Following DNABERT2’s protocol (Zhou et al. 2023), we report average MCC across three random seeds, selecting the model with the lowest validation loss. Compared models include DNABERT (DB1, 89M parameters), DNABERT-2 (DB2, 117M), and Nucleotide Transformer variants (500M and 2.5B parameters, trained on human or multi-species data)(Zhou et al. 2023). As shown in Table1, NucEL (93M parameters)

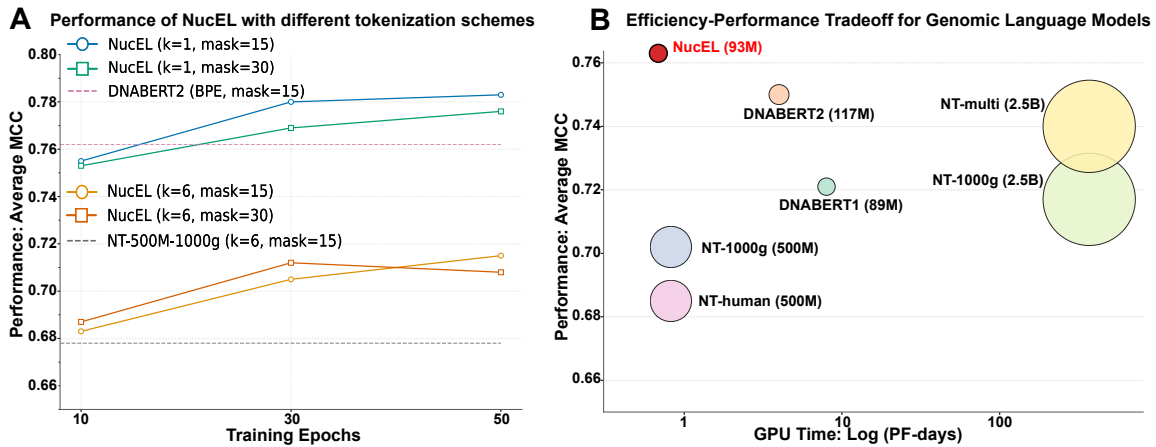


Figure 2: (A) Performance Across Tokenization Schemes. Performance of NucEL with different tokenization schemes (single-nucleotide, k-mer with $k=6$, and BPE) across training epochs on human GUE datasets, highlighting the superior representation learning of single-nucleotide tokenization. (B) Efficiency-Performance Tradeoff for Genomic Language Models. Efficiency-performance tradeoff on human GUE tasks, with GPU time (petaflop-days, log scale) on the x-axis, average MCC on the y-axis, and bubble size representing model parameter count. NucEL-93M (in red) achieves a strong balance, outperforming larger models like NT-multi-2.5B.

demonstrates competitive performance across human genomic tasks, achieving top-1 results on CPD and SSP tasks.

We further evaluate NucEL’s generalization ability across different species to assess its cross-domain transferability. Despite being trained exclusively on human genomic data, NucEL demonstrates remarkable performance on diverse genomic tasks across mouse (Transcription Factor binding, TF-M), yeast (Epigenetic Marks Prediction, EMP), and viral genomes (Covid Variant Classification, CVC), as shown in Table 1. Notably, NucEL outperforms models trained on multi-species data while NT-multi contains 25 times more parameters than NucEL. The model achieves the highest overall average performance of 75.16 across all seven genomic tasks and sets top-2 results on 6 of 7 cross-species tasks, demonstrating that fundamental genomic patterns learned from human data can effectively generalize to other species. This cross-species transferability, combined with its parameter efficiency, highlights NucEL’s ability to capture universal genomic representations that transcend species boundaries.

Assessing Regulatory DNA Classification on GB Benchmark

We assessed NucEL on seven human-centric regulatory sequence classification tasks from the Genomic Benchmarks dataset (Grešová et al. 2023), evaluating its ability to distinguish functional DNA classes across diverse sequence lengths. Adopting Caduceus’s protocol (Schiff et al. 2024), we report average accuracy and standard deviation across five random seeds, selecting the model with the lowest validation loss. Compared models include HyenaDNA, Caduceus variants (Ph and PS), and Nucleotide Transformer 2 (NT2-100M) (Schiff et al. 2024). Table 2 shows NucEL achieves the highest average accuracy of 89.9%, outperforming NT2-100M (89.0%) and Caduceus-Ph (88.2%),

with state-of-the-art results on four tasks and second-best on two, underscoring its robustness in regulatory DNA classification.

Profiling Genomic Element Prediction on NT Benchmark

We evaluated NucEL on the revised Nucleotide Transformer benchmark (Dalla-Torre et al. 2025), encompassing histone marker prediction, regulatory element identification, and splice site detection. Following Dalla-Torre et al. (2025), we report average MCC across ten random seeds. Table 3 shows NucEL (93M parameters) outperforms similarly sized models like DNABERT2 (117M) and NT2-Multi (100M), achieving state-of-the-art on 11 of 18 tasks. Its average MCC of 0.664 matches NT2-Multi (500M parameters) and slightly exceeds NT-Multi (2.5B parameters, 0.661), despite using only human genome data for pre-training, unlike the multi-species data used by DNABERT2 and NT2. NucEL’s competitive performance on chromatin profiles (e.g., histone markers) is detailed in the Appendix. With five times fewer parameters than NT2-500M and 27 times fewer than NT-Multi 2.5B, NucEL demonstrates exceptional efficiency and performance.

Impact of Tokenization Strategies

Tokenization significantly impacts genomic language model performance and efficiency. Figure 2(A) compares NucEL’s performance across tokenization schemes—single-nucleotide ($k = 1$), k-mer ($k = 6$), and Byte Pair Encoding (BPE)—over training epochs on GUE datasets. Single-nucleotide tokenization outperforms k-mer and BPE, preserving base-level detail critical for genomic tasks. K-mer tokenization, as used in NT-500M-1000g, reduces sequence length but dilutes single-nucleotide context, compromising performance. BPE, employed by DNABERT2,

Genomic Benchmark Tasks	HyenaDNA	Caduceus-Ph	Caduceus-PS	NT2-100m	NucEL
Coding vs. Intergenic	0.904 \pm 0.005	0.915 \pm 0.003	0.910 \pm 0.003	0.950 \pm 0.002	0.941 \pm 0.001
Human vs. Worm	0.964 \pm 0.002	0.973 \pm 0.001	0.968 \pm 0.002	0.972 \pm 0.001	0.975 \pm 0.001
Human Enhancers Cohn	0.729 \pm 0.014	0.747 \pm 0.004	0.745 \pm 0.007	0.736 \pm 0.004	0.735 \pm 0.008
Human Enhancer Ensembl	0.849 \pm 0.006	0.893 \pm 0.008	0.900 \pm 0.006	0.935 \pm 0.001	0.940 \pm 0.001
Human Regulatory	0.869 \pm 0.012	0.872 \pm 0.011	0.873 \pm 0.007	0.941 \pm 0.001	0.932 \pm 0.002
Human OCR Ensembl	0.783 \pm 0.007	0.828 \pm 0.006	0.818 \pm 0.006	0.776 \pm 0.002	0.794 \pm 0.002
Human NonTATA Promoters	0.944 \pm 0.002	0.946 \pm 0.007	0.945 \pm 0.010	0.920 \pm 0.003	0.973 \pm 0.001
Average	0.863	0.882	0.880	0.890	0.899

Table 2: Performance comparison on GB (Accuracy; Averaged over 5 Seeds with Standard Deviation; **Best**, Second-Best)

improves computational efficiency but obscures single-nucleotide variations and risks inconsistent tokenization for similar sequences due to mutation-induced boundary shifts. Paired with ELECTRA’s RTD objective, single-nucleotide tokenization enables efficient learning across all sequence positions, yielding superior representations.

Task	DNABERT2 (117M)	NT (500M)	NT2 (100M)	NucEL (93M)
H2AFZ	0.49	0.465	0.492	0.508
H3K27ac	0.491	0.457	0.487	0.470
H3K27me3	0.599	0.589	0.595	0.611
H3K36me3	0.637	0.594	0.617	0.584
H3K4me1	0.49	0.468	0.485	0.480
H3K4me2	0.558	0.527	0.551	0.546
H3K4me3	0.646	0.622	0.633	0.621
H3K9ac	0.564	0.524	0.538	0.550
H3K9me3	0.443	0.433	0.445	0.474
H4K20me1	0.655	0.634	0.648	0.640
Enhancer	0.517	0.515	0.507	0.578
Enhancer types	0.476	0.477	0.465	0.536
Promoter all	0.754	0.734	0.753	0.762
Promoter non-TATA	0.762	0.738	0.766	0.769
Promoter TATA	0.784	0.831	0.826	0.922
Splice acceptor	0.837	0.941	0.947	0.960
Splice site all	0.835	0.939	0.960	0.962
Splice donor	0.861	0.952	0.947	0.970
Average	0.628	0.634	0.651	0.664

Table 3: Performance comparison on NT (MCC; Averaged over 10 Seeds ; **Best**, Second-Best)

Efficiency-Performance Tradeoff Analysis

We analyzed NucEL’s efficiency-performance tradeoff against BERT-based genomic models, focusing on training time, performance (MCC), and model size. Following Zhou et al. (2023), we use GPU time as a proxy for the total number of operations performed during pretraining across different hardware. Figure 2(B) illustrates this tradeoff on GUE tasks, with GPU time (petaflop-days, log scale) on the x-axis, average MCC on the y-axis, and bubble size representing parameter count. NucEL-93M (red) achieves a superior balance, outperforming larger models like NT-multi-2.5B and NT-1000g-2.5B, which are approximately 25 times larger in parameter count and require over 100 times more computational resources. Compared to models

with similar computational demands (e.g., NT-1000g-500M, NT-human-500M), NucEL delivers over 10% higher MCC. This highlights the ELECTRA-style framework’s ability to provide robust genomic understanding with significantly reduced computational and parameter requirements.

MODEL	MACRO F1	MICRO F1	WEIGHTED F1
HyenaDNA	0.6882	0.7519	0.7377
NT2-100m	0.6325	0.7579	0.7470
DNABERT2	0.6398	0.7487	0.7358
NucEL	0.7306	0.7603	0.7490

Table 4: Embedding quality Weighted F1 classification score on biotypes.

Embedding Analysis by Gene Biotype

To assess NucEL’s ability to capture biological signals, we extracted sequences from the most prevalent Ensembl biotypes and generated embeddings using NucEL, HyenaDNA (small 32K), DNABERT2 (117M), and NT2-100M. We visualized embeddings using t-SNE (Figure 3) and trained an XGBoost classifier to predict biotypes, evaluating performance with macro, micro, and weighted F1 scores (Table 4). NucEL achieved the highest F1 scores across all metrics, demonstrating superior separation of biotype classes, particularly for small RNAs (red circle in Figure 3). Notably, lincRNA and protein-coding genes, which are often hard to distinguish due to sequence similarities and shared exon-intron structures, were effectively clustered together in NucEL’s embeddings, highlighting its robust representation of complex biological relationships compared to competing models.

Visualizing Attention and Model Interpretability

Predictive models for genomic sequences aim to uncover novel biological insights through interpretable analyses, surpassing traditional probabilistic enrichment studies by capturing complex interactions between regulatory elements. To evaluate model interpretability, we designed a synthetic “motif-order” classification task featuring two randomly placed motifs, A and B, where sequences with Motif A before B are positive and B before A are negative. This task mimics biological scenarios, such as RNA-binding proteins binding in a specific order to activate cis-regulatory elements. The motifs share identical base composition and are

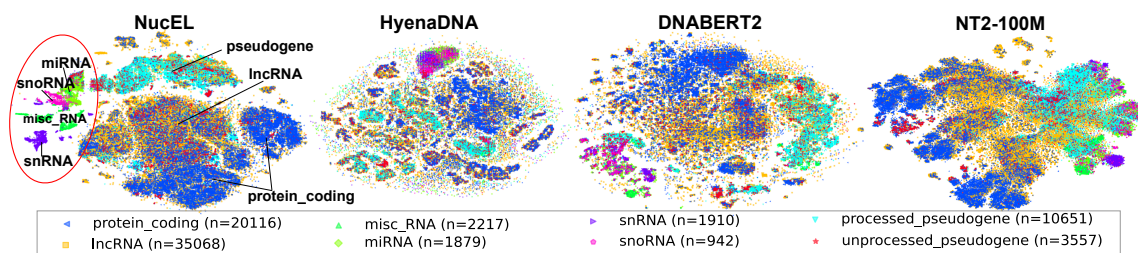


Figure 3: Visualization of Gene Biotype Embeddings. t-SNE visualizations of embeddings generated by NucEL, HyenaDNA, DNABERT2, and NT2-100M for the prevalent biotypes, demonstrating NucEL’s superior separation.

separated by a random gap, forcing the model to learn long-range interactions rather than relying on simple motif detection. See Appendix for detailed methods and results.

We compared NucEL and NT2-100M on this motif-order task, with both models fine-tuned to near-perfect accuracy. Despite similar predictive performance, their interpretability differed substantially. We extracted attention weights from the global attention layers (layers 0, 3, 6, 9, 12, 15, 18, and 21) and computed averaged attention maps across 100 test sequences aligned at the motif positions. Figure 4 shows the signal-to-noise ratio (SNR) at true motif locations, quantifying how strongly attention signals rise above background noise. NucEL consistently exhibits higher SNRs across global layers, indicating more effective localization of informative features. Attention heatmaps (see Appendix) further reveal that NucEL focuses sharply on the embedded motif regions, while NT2-100M displays more diffuse attention with elevated background activity. Quantitatively, NucEL achieves substantial improvements in maximum SNR—65% higher for Motif A and 152% higher for Motif B—highlighting its superior capacity to distinguish motif signals from noise. Since high SNR is critical for minimizing false discovery rates in regulatory motif identification, these findings underscore NucEL’s strength in fine-grained genomic modeling and its ability to capture long-range regulatory dependencies with greater precision and interpretability than existing models.

NucEL’s precise attention underscores its superiority in fine-grained genomic modeling and its ability to capture long-range regulatory interactions.

Limitations

While NucEL achieves state-of-the-art performance on human-genome benchmarks and demonstrates strong zero-shot generalization to non-human species (e.g., mouse, yeast, virus), its current pretraining remains confined to the human genome. Expanding pretraining to include diverse taxonomic groups (e.g., invertebrates, microbes) could enhance cross-species transferability, particularly for clade-specific regulatory elements or non-conserved motifs. This is a promising direction for future work.

Conclusion

This study introduces NucEL, the first ELECTRA-style pre-training framework for genomic sequences, shifting

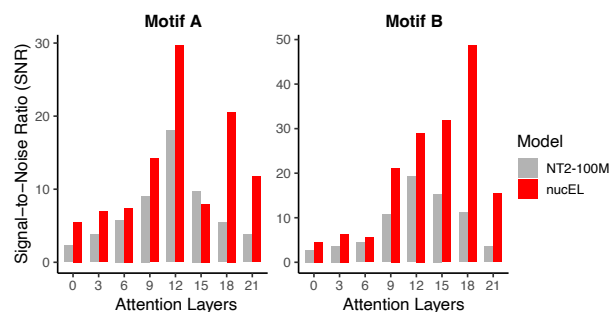


Figure 4: Signal-to-Noise Ratio (SNR) Comparison: Comparison of SNR for attention weights in NucEL and NT2-100M across global attention layers on the motif-order classification task, showing NucEL’s superior performance in detecting Motif A and Motif B with higher SNR values, indicating clearer motif identification and reduced background noise compared to NT2-100M.

from the conventional masked language modeling (MLM) paradigm. Extensive experiments demonstrate that NucEL’s replaced token detection (RTD) approach combined with single-nucleotide tokenization surpasses MLM, offering superior efficiency, accuracy, and interpretability in genomic representation learning. NucEL’s innovations include: (1) an ELECTRA-style generator-discriminator architecture for dense token-level supervision; (2) single-nucleotide tokenization, enabling precise modeling of fine-grained genomic features; and (3) ModernBERT’s hybrid attention mechanisms, capturing both local and global genomic dependencies. Combining single-nucleotide resolution with RTD enhances efficiency and precision, preserving base-level detail while dense supervision and hybrid attention mitigate computational costs. These advancements allow NucEL to achieve state-of-the-art performance across diverse genomic benchmarks with fewer parameters than competing models. Enhanced by fine-grained tokenization and attention mechanisms, NucEL provides clear insights into genomic features driving predictions, improving model interpretability. This work advances genomic language modeling by delivering a computationally efficient and interpretable tool, demonstrating the ability of new pre-training strategies to extract additional information in genomic analysis.

Acknowledgments

This work was supported by the Australian Research Council Centre of Excellence for the Mathematical Analysis of Cellular Systems (CE230100001), and the Talo Scholarship and the Talo Innovative Grant funded by Taiyang Zhang and Loong Wang. The authors acknowledge computational resources provided by the National Computational Infrastructure (NCI), Argonne National Laboratory, and Google Cloud, as well as the NCI HPC-AI Talent Program. We are especially grateful to Dr. Jingbo Wang and Dr. Arvind Ramanathan for their valuable discussions and constructive suggestions, which contributed substantially to this work.

References

- Abramson, J.; Adler, J.; Dunger, J.; Evans, R.; Green, T.; Pritzel, A.; Ronneberger, O.; Willmore, L.; Ballard, A. J.; Bambrick, J.; et al. 2024. Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature*, 630(8016): 493–500.
- Alipanahi, B.; Delong, A.; Weirauch, M. T.; and Frey, B. J. 2015. Predicting the sequence specificities of DNA-and RNA-binding proteins by deep learning. *Nature biotechnology*, 33(8): 831–838.
- Avsec, Ž.; Agarwal, V.; Visentin, D.; Ledsam, J. R.; Grabska-Barwinska, A.; Taylor, K. R.; Assael, Y.; Jumper, J.; Kohli, P.; and Kelley, D. R. 2021. Effective gene expression prediction from sequence by integrating long-range interactions. *Nature methods*, 18(10): 1196–1203.
- Beltagy, I.; Peters, M. E.; and Cohan, A. 2020. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.
- Clark, K.; Luong, M.-T.; Le, Q. V.; and Manning, C. D. 2020. Electra: Pre-training text encoders as discriminators rather than generators. *arXiv preprint arXiv:2003.10555*.
- Dalla-Torre, H.; Gonzalez, L.; Mendoza-Revilla, J.; Lopez Carranza, N.; Grzywaczewski, A. H.; Oteri, F.; Dallago, C.; Trop, E.; de Almeida, B. P.; and Sirelkhatim, H. 2025. Nucleotide Transformer: building and evaluating robust foundation models for human genomics. *Nature Methods*, 22(2): 287–297.
- Dao, T. 2023. Flashattention-2: Faster attention with better parallelism and work partitioning. *arXiv preprint arXiv:2307.08691*.
- Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, 4171–4186.
- Dotan, E.; Jaschek, G.; Pupko, T.; and Belinkov, Y. 2024. Effect of tokenization on transformers for biological sequences. *Bioinformatics*, 40(4): btac196.
- Grešová, K.; Martinek, V.; Čechák, D.; Šimeček, P.; and Alexiou, P. 2023. Genomic benchmarks: a collection of datasets for genomic sequence classification. *BMC Genomic Data*, 24(1): 25.
- He, P.; Gao, J.; and Chen, W. 2021. Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing. *arXiv preprint arXiv:2111.09543*.
- Hochreiter, S.; and Schmidhuber, J. 1997. Long short-term memory. *Neural computation*, 9(8): 1735–1780.
- Ji, Y.; Zhou, Z.; Liu, H.; and Davuluri, R. V. 2021. DNABERT: pre-trained Bidirectional Encoder Representations from Transformers model for DNA-language in genome. *Bioinformatics*, 37(15): 2112–2120.
- Jumper, J.; Evans, R.; Pritzel, A.; Green, T.; Figurnov, M.; Ronneberger, O.; Tunyasuvunakool, K.; Bates, R.; Židek, A.; Potapenko, A.; et al. 2021. Highly accurate protein structure prediction with AlphaFold. *nature*, 596(7873): 583–589.
- LeCun, Y.; Bengio, Y.; and Hinton, G. 2015. Deep learning. *nature*, 521(7553): 436–444.
- Nguyen, E.; Poli, M.; Faizi, M.; Thomas, A.; Wornow, M.; Birch-Sykes, C.; Massaroli, S.; Patel, A.; Rabideau, C.; and Bengio, Y. 2023. Hyenadna: Long-range genomic sequence modeling at single nucleotide resolution. *Advances in neural information processing systems*, 36: 43177–43201.
- Poli, M.; Massaroli, S.; Nguyen, E.; Fu, D. Y.; Dao, T.; Bacchus, S.; Bengio, Y.; Ermon, S.; and Ré, C. 2023. Hyena hierarchy: Towards larger convolutional language models. In *International Conference on Machine Learning*, 28043–28078. PMLR.
- Sanabria, M.; Hirsch, J.; Joubert, P. M.; and Poetsch, A. R. 2024. DNA language model GROVER learns sequence context in the human genome. *Nature Machine Intelligence*, 6(8): 911–923.
- Sasse, A.; Chikina, M.; and Mostafavi, S. 2024. Unlocking gene regulation with sequence-to-function models. *Nature methods*, 21(8): 1374–1377.
- Schiff, Y.; Kao, C.-H.; Gokaslan, A.; Dao, T.; Gu, A.; and Kuleshov, V. 2024. Caduceus: Bi-directional equivariant long-range dna sequence modeling. *arXiv preprint arXiv:2403.03234*.
- Warner, B.; Chaffin, A.; Clavié, B.; Weller, O.; Hallström, O.; Taghadouini, S.; Gallagher, A.; Biswas, R.; Ladhak, F.; and Aarsen, T. 2024. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. *arXiv preprint arXiv:2412.13663*.
- Zhou, Z.; Ji, Y.; Li, W.; Dutta, P.; Davuluri, R.; and Liu, H. 2023. Dnabert-2: Efficient foundation model and benchmark for multi-species genome. *arXiv preprint arXiv:2306.15006*.