BioToken and BioFM – Biologically-Informed Tokenization Enables Accurate and Efficient Genomic Foundation Models

Aleksandr Medvedev[1*†], Karthik Viswanathan[1†], Praveenkumar Kanithi[1*†], Kirill Vishniakov[1], Prateek Munjal[1], Clément Christophe[1], Marco AF Pimentel[1], Ronnie Rajan[1], Shadab Khan[#1,2*]

[1]M42, Abu Dhabi, United Arab Emirates.
[2]ADIA Lab, Abu Dhabi, United Arab Emirates.

*Corresponding author(s). E-mail(s): amedvedev@m42.ae; pkanithi@m42.ae; shadab.khan@adialab.ae;
†These authors contributed equally to this work.

Abstract

Genomic variation underlies human phenotypic diversity, disease susceptibility, and evolutionary adaptation. Although large-scale genomic sequencing has transformed our ability to map genetic variation, accurately modeling and interpreting this data remains challenging due to fundamental limitations in existing genomic foundation models (GFMs). Current genomic models typically treat DNA simplistically as linear textual sequences, overlooking critical biological context, such as genomic structural annotations, regulatory elements, and functional contexts central to genomic interpretation. As a result, these models are prone to positional memorization of common sequences, severely limiting their generalization to biologically meaningful tasks. Here, we introduce BioToken, a modular and extendable tokenization framework designed to encode genomic variants and biologically relevant structural annotations directly into genomic representations. By utilizing intrinsic inductive biases, BioToken facilitates meaningful representation learning and generalization across diverse molecular phenotypes, such as gene expression, alternative splicing, and variant pathogenicity prediction. Built on BioToken, our genomic foundation model, BioFM, achieves competitive or superior results relative to specialized models (e.g., Enformer, SpliceTransformer) across a comprehensive suite of genomic benchmarks, including noncoding pathogenicity, expression modulation, sQTL prediction, and long-range genomic interactions. Notably, BioFM achieves state-of-the-art performance with significantly fewer parameters (265M), substantially reducing training costs and computational requirements. Our findings highlight the substantial advantages of integrating biologically-informed inductive biases into genomic foundation modeling, providing a robust and accessible path forward in genomics. We provide our code and model checkpoints to support further research in this direction.

Keywords: GFM, Genomics, Variant, Foundation model, Splicing, Expression

Genomes encode life's biological complexity and underpin most aspects of organisms' function and phenotypic diversity through variations at the molecular level. With high-throughput sequencing, our ability to map and characterize genetic variants associated with diseases and phenotypic traits has improved. Despite these advances, accurately modeling and interpreting genomic data remains a challenge [1, 2]. This challenge largely arises from conceptual limitations inherent in existing modeling paradigms, specifically the widespread treatment of DNA simply as a linear textual sequence composed of nucleotides (A,C,T,G).

[#]S.K. was with with M42 (initially) and ADIA Lab (subsequently) during this research work.

Drawing inspiration from the success of Large Language Models (LLMs), recent research has adapted similar architectures and pretraining strategies to genomic data, producing Genome Foundation Models (GFMs) [3–9]. These models commonly employ tokenization and training objectives similar to those used for natural language processing, typically treating genomic sequences as strings of characters to be predicted sequentially. Although convenient, this approach overlooks critical differences between textual language and genomic information. DNA sequences, unlike human languages, do not encode semantics through syntactic rules alone. Rather, they encode functions through biochemical interactions, spatial chromatin configurations, and regulatory relationships. Hence, modeling genomes strictly as a linguistic text neglects vital inductive biases, the intrinsic biological structure, and the regularity inherent in genomic sequences.

This shortcoming is increasingly evident in the limitations observed in current GFMs. Human genomic variation is sparse [10, 11], and differences between individuals typically account for less than 1% of DNA. Thus, training models to predict the next nucleotide repeatedly, where deviations from a reference genome are quite infrequent, incentivizes models to memorize the constant background signal, the reference genome, rather than learning meaningful biological representation. As a result, while these models perform well in positional prediction tasks — where memorized sequence context directly informs the prediction — they struggle on tasks requiring deeper representation learning and generalization across biological contexts.

Recent works have highlighted these shortcomings. The recent studies [12, 13] demonstrate that pretrained genomic models frequently perform no better, and sometimes worse, than randomly initialized models on a range of downstream tasks. Such results question the utility of current genomic pretraining methods. These findings signal a clear need to revisit the assumptions of foundational modeling, developing architectures and data representations that explicitly leverage genomic structure.

To address these challenges, we introduce BioToken, a modular and extendable tokenization approach designed to encode genomic variants, including Single Nucleotide Variants (SNVs), insertions, deletions, and structural genomic features such as exons, introns, transcripts, and coding regions. By incorporating biologically relevant tokens directly into the genomic sequence representation, BioToken inherently utilizes genomic inductive biases, enabling models to learn biologically grounded representations.

Importantly, BioToken is highly modular and can be easily extended to incorporate additional biological modalities such as RNA transcript variants, epigenetic markers like methylation or histone modifications, and even chromatin accessibility signals. This makes BioToken a a robust approach for different modeling tasks in the rapidly evolving landscape of genomics research. Building upon BioToken, we developed BioFM, a genomic foundation model explicitly designed for computational and parameter efficiency, utilizing recent advances in transformer architectures optimized for biological sequences. BioFM is a decoder-only transformer architecture [14], with only 265 million parameters, significantly smaller than contemporary GFMs such as Nucleotide Transformer [3] with up to 2.5B parameters, Evo1 and Evo2 [4, 15], with up to 40B parameters.

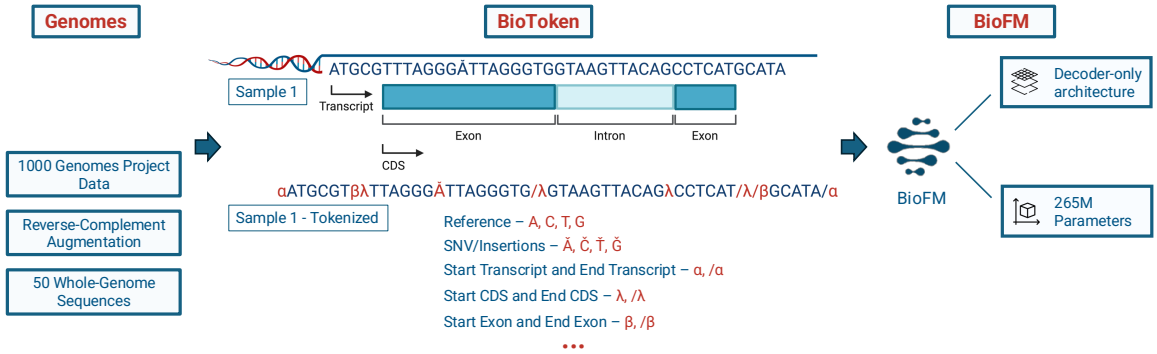
To rigorously validate BioFM and the effectiveness of BioToken, we conducted comprehensive benchmarking across critical genomic tasks, each representing distinct biological complexities. These evaluations included noncoding pathogenicity assessment (BEND benchmark [16]), gene expression modulation (DeepSea dataset [17]), sQTL prediction [18], coding variant pathogenicity (AlphaMissense [19]), and novel tasks such as local ancestry and rare variant classification. BioFM consistently achieved superior performance, matching or surpassing specialized models such as Enformer [20] in gene expression prediction and SpliceTransformer [21] in alternative splicing analysis, tasks previously considered challenging for generalized genomic models.

Further validating our approach, BioFM demonstrated exceptional performance in capturing long-range genomic dependencies, notably outperforming existing GFMs on the Genomic Long-Range Benchmark (GLRB) [22]. Moreover, BioFM exhibited promising generative capabilities, as demonstrated through generation of promoter sequences conditioned solely on gene inputs, a functionality previously uncharacterized among current genomic foundation models, and opening doors to synthetic generative biology with applications that can include discovery of novel regulatory structures given a gene as an input.

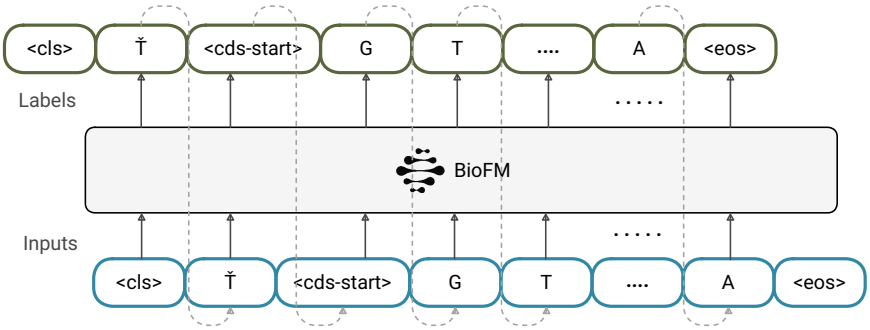
Our experimental results demonstrate that encoding genomic sequences with biologically meaningful structure and utilizing a highly efficient model architecture significantly enhances model performance and biological fidelity. Importantly, the computational efficiency afforded by BioFM

highlights BioToken's practical utility in offering significant performance gains while substantially reducing training cost and resource requirements.

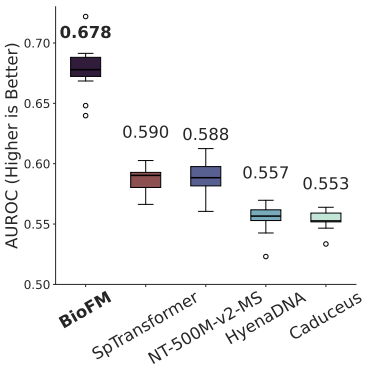
Ultimately, accurate genomic modeling profoundly impacts our understanding of molecular phenotypes, disease mechanisms, and personalized medicine [23]. By fundamentally revisiting DNA representation, BioToken and BioFM represent a meaningful advance, providing not only a biologically grounded modeling paradigm but also a computationally accessible framework for robust genomic research.



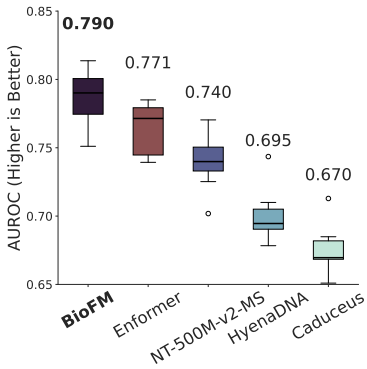
(a) Overview of our modeling approach: BioToken with variants and annotations, and a new foundation model BioFM



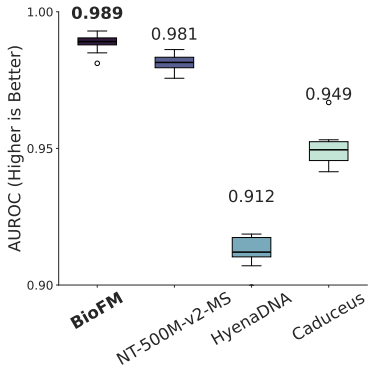
(b) DNA modelling with next token prediction



(c) sQTL Prediction



(d) Expression Prediction



(e) Pathogenicity Prediction

Fig. 1: **BioFM integrates biologically-informed tokenization, significantly outperforming specialized genomic models across a range of tasks. a)** Schematic representation of BioFM's architecture highlighting its key components and an illustrative example showing how DNA sequences are annotated. **b)** BioFM employs next token prediction to model standard nucleotides, variants and annotations. **c)** BioFM significantly outperforms Splice Transformer across 13 tissue-specific sQTL datasets, comprising 1M variants. Box plots show median and quartile distribution. **d)** BioFM significantly outperforms Enformer on expression prediction task. **e)** BioFM outperforms every other GFM on noncoding pathogenicity prediction.

Results

Biologically-Informed Tokenization of Genomes with BioToken

We developed BioToken, a modular tokenization framework designed that encodes variants and biological annotations directly within DNA sequences. Conventional genomic modeling simplifies DNA sequences as linear strings of nucleotides, ignoring biological context essential for accurate genomic interpretation. In contrast, to help the model focus on genetic variations, we developed a tokenization strategy that replaces standard nucleotide tokens (A, G, C, T, N) with special variant tokens (Ă, Ğ, Č, Ť) to capture single nucleotide variants (SNV) and insertions. Similarly, BioToken introduces special tokens that encode structural genomic annotations such as transcript boundaries, exon-intron architecture, and coding DNA sequences (CDS) (subsubsection A.2.1) (Figure 1a). Encoding these biologically meaningful annotations inherently leverages genomic inductive biases, shifting the learning objective from superficial positional memorization towards deeper functional representation learning. Notably, the modularity and extendibility of BioToken enables it to integrate additional genomic annotations not explored in this study, such as RNA secondary structures, epigenetic states (e.g., methylation, histone modifications), and chromatin accessibility patterns. BioToken therefore provides a foundation for genome-scale modeling across diverse downstream applications.

BioFM – Computationally Efficient Foundation Modeling

Using BioToken, we developed BioFM, a computationally efficient genome foundation model optimized for biologically-informed representation learning. BioFM employs a decoder-only transformer architecture comprising 265M parameters (Figure 1b), substantially smaller than contemporary GFMs, which significantly reduces computational resource demands. To systematically exploit DNA's inherent reverse-complement symmetry, we implemented reverse complement augmentation during training, randomly applying it to 50% of input sequences while retaining strand-specific annotations. This approach compels the model to learn symmetrical representations, which is essential given that many genomic regulatory and functional motifs exhibit strand symmetry. BioFM was pretrained on WGS data from 50 individuals randomly selected from the 1000 Genomes project [10], yielding approximately 160 billion tokens for pretraining. After evaluating multiple context lengths, we empirically determined 6K nucleotide as the ideal context window. Further, we adopted extended Rotary Position Embeddings (RoPE) but adjusted the positional embedding frequency parameter θ from 10^4 to 10^5, to mitigate positional biases and effectively model genomic dependencies over longer distances [24]. Our initial ablation analyses was conducted using a reduced-size 134M-parameter model, and all subsequent experiments were perfomed using a scaled-up 265M-parameter model.

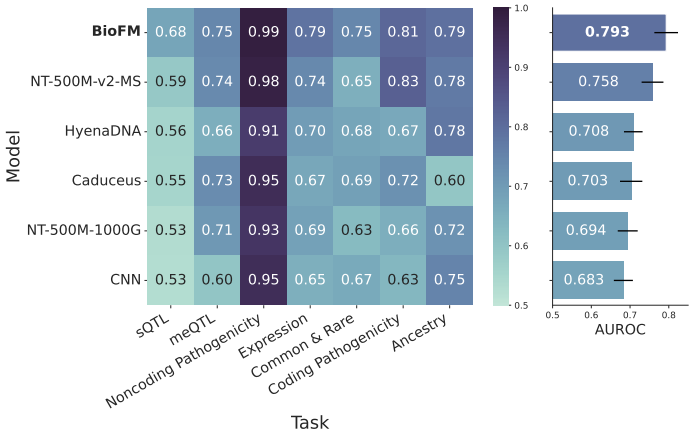
BioFM Accurately Predicts Variant Effects Across Diverse Molecular Phenotypes

We developed a new benchmark to evaluate BioFM's ability to effectively utilize variant information across multiple biological contexts, significantly extending beyond conventional genomic benchmarks [3, 16, 22, 25, 26], which primarily emphasize region classification tasks. Although existing assessments such as BEND [16] and the Genomic Long-Range Benchmark (GLRB) [22] measure critical functions like noncoding pathogenicity and tissue-specific expression, they do not represent the broader spectrum of variant-mediated molecular processes. In contrast, we organize our benchmark evaluations into two major categories to systematically assess BioFM's capabilities: **(1)** Functional Effects of Genetic Variants, focusing on the direct molecular and biological consequences of genetic variation, and **(2)** Population-Level Variation, assessing BioFM's sensitivity to broader genetic patterns across human populations.

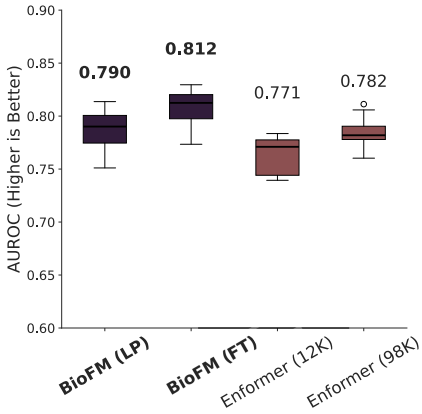
1. Functional Effects of Genetic Variants

We assessed BioFM's capability to model variant effects across five key molecular processes.

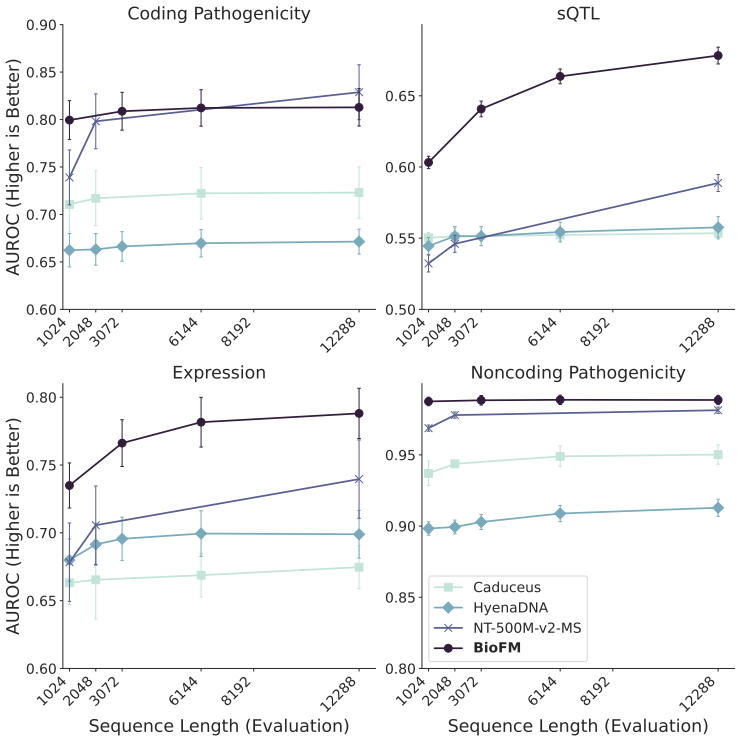
- **Coding pathogenicity assessment**: Accurate prediction of pathogenic coding variants is fundamental to precision medicine and clinical genomics. We assessed this using the AlphaMissense dataset [19], comprising a comprehensive catalog of coding variants annotated for pathogenicity.



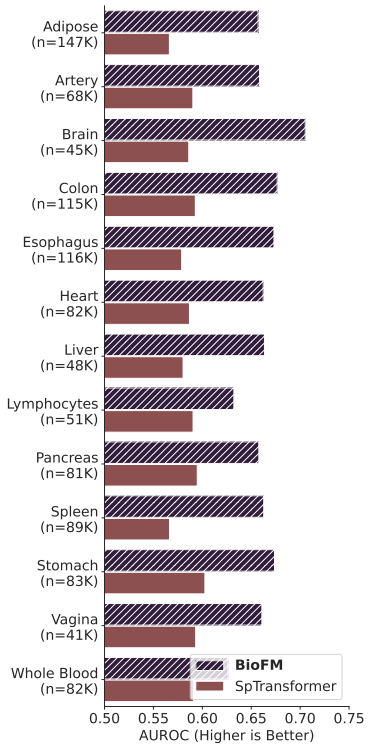
(a) AUROC on Variant Benchmark



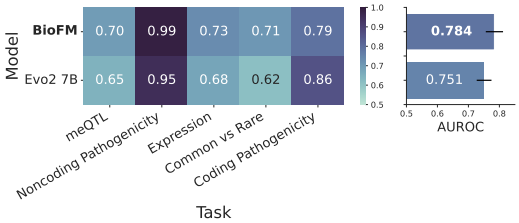
(b) BioFM linear probing (LP) and finetuned (FT) vs Enformer on Expression task



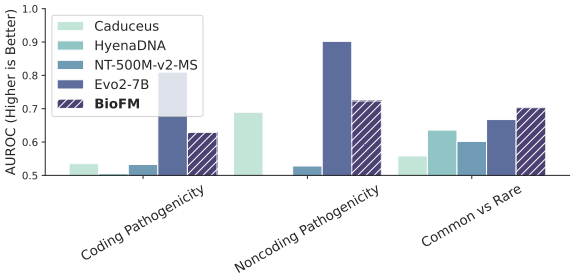
(c) Evaluation of context length efficiency



(d) sQTL per-tissue comparison with Splice Transformer



(e) Linear probing comparison with Evo2-7B model



(f) Zero-shot pathogenicity comparison with GFMs

Fig. 2: Extensive evaluation shows that BioFM does well in modeling variants, and efficiently using genomic context for superior accuracy and generalization. Comprehensive performance evaluation of BioFM across multiple genomic prediction tasks using 11-fold cross-validation. **a)** Across a broad spectrum of variant prediction tasks, BioFM outperforms other GFMs. **b)** BioFM with linear probing and finetuned versions both markedly improves expression predictions, exceeding Enformer performance even when Enformer utilizes significantly longer context windows. (*Contd.*)

Fig. 2: (*Continued from previous page*) **c)** BioFM demonstrates robust scalability and accuracy across increasing context lengths, consistently outperforming competing models, particularly in complex pathogenicity, splicing prediction, and expression prediction tasks. Error bands represent two standard deviations calculated across 11 cross-validation folds, demonstrating the robustness of predictions across different context sizes. **d)** BioFM achieves consistent, tissue-independent performance improvements over SpliceTransformer on tissue-specific splicing (sQTL) predictions. **e)** Despite substantially fewer parameters (25x smaller) , BioFM demonstrates competitive linear probing performance relative to the substantially larger Evo2-7B model, highlighting the computational efficiency and representation quality of BioFM's modeling approach. **f)** BioFM demonstrates competitive zero-shot variant pathogenicity prediction capabilities compared to other large-scale genomic foundation models, falling only behind Evo2 on 2 out of 3 tasks.

- **Noncoding pathogenicity assessment**: Pathogenic variants in noncoding regions significantly impact gene regulation, influencing many complex traits and diseases. To evaluate predictive accuracy, we employed the BEND dataset, containing 295,000 annotated SNVs in noncoding genomic regions [16].
- **Expression effect prediction**: Variant-driven changes in gene expression contribute to phenotypic diversity and disease processes. To test BioFM's representation of these regulatory effects, we utilized variant-associated gene expression data derived from the DeepSea dataset [17].
- **Alternative splicing**: Variant-induced alternative splicing contributes significantly to human proteomic diversity and affects biological processes and diseases. To evaluate modeling accuracy for splicing variations across multiple cellular contexts, we created an extensive splicing quantitative trait loci (sQTL) dataset based on sqtlSeeker2 data, which includes over one million variant-tissue pairs [18].
- **DNA Methylation**: Variant-driven changes in DNA methylation patterns modulate gene-environment interactions and influence diverse developmental and disease phenotypes. To test the model's ability to predict variant effects on DNA methylation, we utilized methylation quantitative trait loci (meQTL) data from the GRASP database [27].

2. Population-Level Variation

We next evaluated BioFM's ability to capture population-level insights. Specifically, we considered two tasks:

- **Ancestry classification**: Understanding how genetic variants collectively encode ancestral backgrounds informs studies of population structure, evolutionary biology, and disease susceptibility. To evaluate this capability, we used genomic segments labeled by five major superpopulations from the 1000 Genomes Project [10].
- **Common vs synthetic variants**: This task evaluates the model's ability to recognize biologically conserved genomic contexts characteristic of authentic common variants. We randomly sampled 100K common variants (MAF > 0.05) from GnomAD [11] and paired each with a synthetic control variant generated by randomly substituting a nucleotide within a ± 20-nucleotide local context window.

Benchmarking Protocol and Model Evaluation

We developed a standardized assessment protocol to fairly evaluate different model architectures across all tasks. For each variant, we use its DNA context up to the maximum length supported by the model, capped at 12K nucleotides. This limit was set because only HyenaDNA and Caduceus can process longer sequences, but their performance plateaus beyond this context length. Our evaluation approach differs slightly between encoder and decoder-only architectures to exploit their respective strengths. For encoder models, we center the variant within its context, using equal length of sequence information from both upstream and downstream regions. For decoder-only models, namely BioFM and HyenaDNA, we generate two complementary embeddings - one from the forward strand using upstream context, and another from the reverse complement of the downstream sequence. These complementary embeddings are then averaged to create the final variant representation. The final variant embeddings serve as inputs to a linear classifier. We avoid more complex nonlinear methods

or full fine-tuning, as these could obscure differences in embedding quality and require extensive hyperparameter tuning for each model.

We evaluated six GFMs—NT-500M-v2-MS, NT-500M-1000G, HyenaDNA, Caduceus, GENA-LM, and DNABERTv2—on the Variant Benchmark. As a baseline, we included a supervised convolutional neural network (CNN) with 120K parameters, trained separately for each task and fold. Performance results for NT-500M-v2-MS, NT-500M-1000G, HyenaDNA, and Caduceus are shown in Figure 2a. GENA-LM and DNABERTv2 are excluded from the figure, as their performance falls below that of the baseline CNN across all tasks, as shown in Figure A2.

BioFM consistently generated superior representations for variant effects prediction (overall $AUROC = 0.793$) relative to NT-500M-v2-MS ($AUROC = 0.758$), despite nearly half the parameter count. Notably, BioFM significantly outperformed NT-500M-1000G, highlighting limitations of fixed-length k-mer tokenization for variant-rich data. BioFM achieved the highest performance across six of seven tasks, with significant advantages particularly in distinguishing common versus rare variants ($AUROC = 0.746$ vs. 0.65), expression prediction ($AUROC = 0.786$ vs. 0.74), and sQTL prediction ($AUROC = 0.675$ vs. 0.590).

We found that BioFM produces the best features for variant effects prediction ($AUROC = 0.793$), compared to the second-best NT-500M-v2-MS, ($AUROC = 0.758$) which has almost twice as many parameters. Notably, BioFM significantly outperforms NT-500M-1000G which was trained on 1000 Genomes data. BioFM is the best model for six out of seven tasks except coding pathogenicity. It has especially large advantage for Common vs Rare variants differentiation ($AUROC = 0.746$ vs $AUROC = 0.65$), Expression ($AUROC = 0.786$ vs $AUROC = 0.74$) and sQTL prediction ($AUROC = 0.675$ vs $AUROC = 0.590$).

Context Length Utilization and Embedding Quality

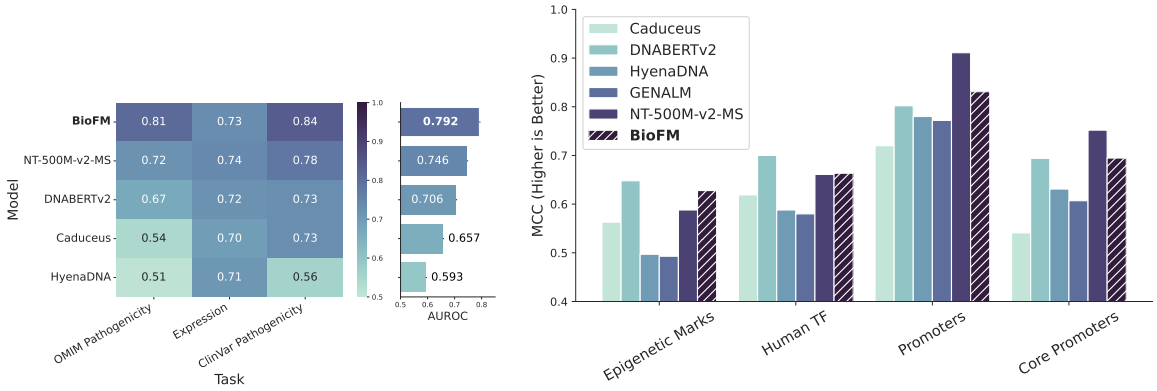
In general, transformer based GFMs are known to struggle to fully utilize their context windows [28]. To investigate how feature extraction is affected by context length, we evaluated model performance using variant contexts ranging from 1K to 12K nucleotides (Figure 2c). BioFM demonstrates strong performance even with a 1K-nucleotide context, significantly outperforming NT-500M-v2-MS across all tasks, including coding pathogenicity. However, NT-500M-v2-MS scales slightly better for pathogenicity tasks. For expression and sQTL prediction, BioFM's performance does not saturate, suggesting that longer-context models could further improve classification quality and that BioFM effectively utilizes its full 6K-nucleotide window. In contrast, the most technically capable long-context models, HyenaDNA and Caduceus, do not scale well. Their performance consistently plateaus at around 3K nucleotides, regardless of the task. Based on these findings, we limited evaluation contexts to a maximum of 12K nucleotides.

Direct Comparison with Specialized Genomic Models

Next, we compared BioFM to strong supervised baselines: Enformer for expression prediction and Splice Transformer for sQTL prediction. Since both models generate tissue- or cell type-specific predictions, and their prediction tracks do not always match with tracks in our data, we extracted their sequence embeddings for downstream linear probing to ensure a fair comparison. As shown in Figure 1d, BioFM matches Enformer's performance on expression prediction when both models use 12K context, making it the first ever GFM to do so. Next, we finetuned BioFM with parameter-efficient LoRA on expression prediction dataset and compared it with Enformer using 98K context (Figure 2b). BioFM significantly outperformed Enformer, median $AUROC = 0.812$ vs $AUROC = 0.782$ ($p = 0.002$, Wilcoxon signed-rank test). Moreover, BioFM is significantly better than Splice Transformer for sQTL prediction (Figure 1c), achieving median tissue $AUROC = 0.675$ compared to $AUROC = 0.590$ of Splice Transformer. Analyzing tissue-specific performance (Figure 2d) reveals that BioFM consistently performs better across all tissues, with no correlation between its performance and the number of tissue variants used for linear probing.

Computational Efficiency and Generalization Compared to Large Models

To assess how BioFM compares against a significantly larger model, we compared it with Evo2-7B model [15], which was trained on 3T nucleotide tokens sourced from all domains of life. Due to computational constraints, we excluded sQTL and ancestry tasks from this comparison, and limited context size to 1024 nucleotides. On Figure 2e, we show that BioFM is better on four out of five tasks,



(a) Genomics Long-Range Benchmark

(b) GUE MCC performance on unannotated data

Fig. 3: BioFM substantially improves predictive accuracy on genomic long-range benchmarks (GLRB). Without access to BioToken, in an evaluation of worst-case performance, it achieves competitive results across genomic functional annotations in the Genomic Understanding Evaluation (GUE). **a)** On GLRB, BioFM sets new standards, notably outperforming previous GFMs that required extensive fine-tuning [22] and longer genomic contexts, effectively demonstrating its ability to utilize long-range genomic dependencies. We used the same test set as [22]. **b)** Comprehensive performance evaluation on the GUE benchmark reveals BioFM consistently excels in functional genomic annotations even when specialized tokens are excluded, highlighting its generalization ability and understanding of of intrinsic genomic regulatory features. Detailed performance metrics are provided in Appendix, Table A3.

except coding pathogenicity. However, if we compare zero-shot performance of BioFM, Evo2-7B and other GFMs, Evo2-7B is significantly better on both pathogenicity prediction tasks. One possible takeaway is that incorporating annotations allows BioFM to be around two orders of magnitude compute-effective than Evo2-7B in producing variant representations, but given the large and diverse training dataset, and the large parameter-count, Evo2-7B is better in a zero-shot setting.

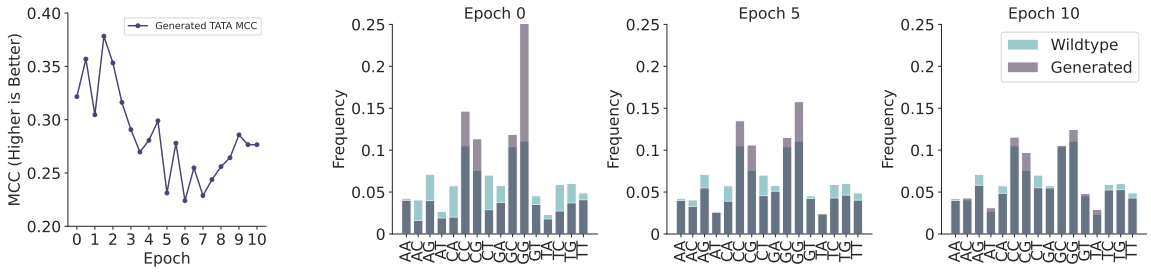
Further, recent studies [3, 15] have shown that optimal performance on downstream tasks typically requires selecting variant representations from intermediate layers, a process involving computationally expensive layer-wise evaluations. To determine whether BioFM similarly requires intermediate-layer selection, we compared the performance of variant representations from BioFM's intermediate layers to those from its last layer (Figure A1). Our analysis demonstrates that BioFM's last-layer embeddings are consistently superior or equivalent to intermediate-layer embeddings across a variety of downstream tasks. Thus, last-layer embeddings in BioFM serve as universally effective representations for downstream applications, avoiding the computational overhead of per-task layer selection.

In conclusion, these findings establish BioFM as a meaningful advance in genomic modeling - it is currently the only genomic foundation model that demonstrates practical utility for both variant expression and splicing effects, matching or exceeding the performance of specialized models designed specifically for these tasks.

BioFM Generalizes to Long-Range Genomic Contexts

Capturing long-range genomic interactions is essential for accurately modeling variant effects, as many regulatory mechanisms operate across extensive genomic distances. To rigorously assess BioFM's ability to utilize long-range context, we used the Genomics Long-Range Benchmark (GLRB)[22], selecting tasks directly relevant to predicting variant-mediated molecular consequences: ClinVar pathogenicity, expression effect prediction, and OMIM pathogenicity identification[29]. These tasks specifically evaluate the capacity to predict pathogenicity and regulatory consequences of genetic variants over extended sequence lengths.

BioFM achieved superior performance relative to all other evaluated genomic foundation models (GFMs) on these long-range variant effect tasks, achieving an overall AUROC of 0.792, notably surpassing NT-500M-v2-MS (AUROC=0.746), the second-best model, which was explicitly fine-tuned



(a) TATA MCC dynamics

(b) Change of 2-mer distribution during training

Fig. 4: BioFM effectively generates biologically plausible synthetic promoter sequences, progressively aligning with naturally occurring promoter features through fine-tuning. **a)** MCC tracking of TATA prediction accuracy across 10 training epochs. Performance is evaluated based on the model's ability to correctly generate TATA-box-containing promoters for genes naturally possessing wildtype TATA-box promoters (true positives) while avoiding incorrect TATA-box insertions in naturally non-TATA promoters (false positives). **b)** Comparative analysis of 2-mer frequency distributions between generated and wildtype promoter sequences, demonstrating progressive convergence of synthetic promoter sequence composition toward natural distributions during the finetuning process.

for ClinVar pathogenicity and expression tasks. Notably, BioFM utilized only linear probing without additional fine-tuning and employed a significantly shorter genomic context (12K nucleotides), whereas NT-500M-v2-MS required specialized context-extension training up to 66K nucleotides. These results highlight BioFM's capability to effectively utilize variant information in providing biologically meaningful variant representations even with long sequence length.

BioFM Captures Broad Functional Genomic Signals

To assess BioFM's general applicability as a foundation model, especially in scenarios where annotations are not available (and therefore annotation tokens are not included as part of the input), we evaluated its performance on the Genomic Understanding Evaluation (GUE) benchmark [25]. GUE benchmark encompasses human-centric genomic tasks, including prediction of epigenetic modifications, transcription factor binding sites, and promoter region characteristics, structured as binary region classification problems analogous to those in the NT benchmark [3]. Notably, these evaluations were conducted without BioFM's specialized features (BioToken's annotation tokens, variant tokens were not used, reverse complement augmentations was omitted), resulting in a lower-bound result for BioFM.

BioFM demonstrated consistently strong performance across all task categories within the GUE benchmark, ranking second overall. It was surpassed only by DNABERTv2 on epigenetic mark and transcription factor binding predictions, and by NT-500M-v2-MS on promoter-specific tasks. Nonetheless, BioFM consistently outperformed competing models such as HyenaDNA, Caduceus, and GENA-LM across all evaluated categories. Further, despite having nearly half the parameter count and lacking domain-specific adaptations, BioFM surpassed NT-500M-v2-MS on epigenetic mark and transcription factor binding tasks. These results demonstrate that BioFM remains a useful model even in applications where it doesn't fully utilize its strengths.

Instruction-Tuning Enables BioFM to Generate Regulatory Sequences

The generation of synthetic promoters represents a strong test for evaluating BioFM' generative capabilities. The ability to design synthetic promoters can enable precise control over transcriptional activity, with applications in optimizing gene expression for research and therapeutic interventions [30, 31]. As promoters are typically located upstream of the corresponding gene sequence, BioFM requires fine-tuning to generate them effectively. Applying the ideas of instruction tuning used in training language models for this task, we prompt model with a gene sequence and a special token $'P'$ after it, creating instruction set of ⟨`DNA sequence`⟩ ⟨`Task-identifier token`⟩ ⟨`Output`⟩. Training was done using instruction-answer pair of genes-promoters set and testing on a held-out set. Promoter sequences were drawn from the EPDNew dataset [32]

We systematically validated BioFM-generated promoters using three independent biological criteria: (1) compositional similarity to natural promoter sequences, and (2) accurate modeling of biologically significant promoter types (TATA versus non-TATA).

First, we assessed compositional similarity by analyzing k-mer distributions of generated promoter sequences. We found that after 10 epochs of fine-tuning, BioFM-generated promoters successfully captured the k-mer distributions of wildtype promoters, indicating convergence toward biologically realistic promoter compositions (Figure 4b).

Second, we measured BioFM's ability to model biologically significant promoter types accurately. Notably, the Matthews Correlation Coefficient (MCC) for predicting TATA-containing promoters was 0.28 after fine-tuning (Figure 4a), confirming the model's capability in discriminating genes more likely to possess TATA promoters. Interestingly, we observed an early training phenomenon where the TATA MCC peaked higher (approximately 0.38 within the first two epochs), coinciding with the generation of promoter sequences enriched in CC and GG dinucleotides compared to the wildtype distribution. This suggests an initial overfitting to easily identifiable motifs, which was subsequently refined toward wildtype distribution.

Collectively, these evaluations indicate that BioFM effectively internalizes the relationships between genes and promoters, applying this understanding to generate biologically plausible synthetic promoters for previously unseen genes.

Ablation, Scaling, Representation Analysis, and Attention Specialization

We assessed the contributions of BioToken and genomic variance data to BioFM's performance by performing systematic ablation experiments, embedding analyses, and evaluations of sample efficiency.

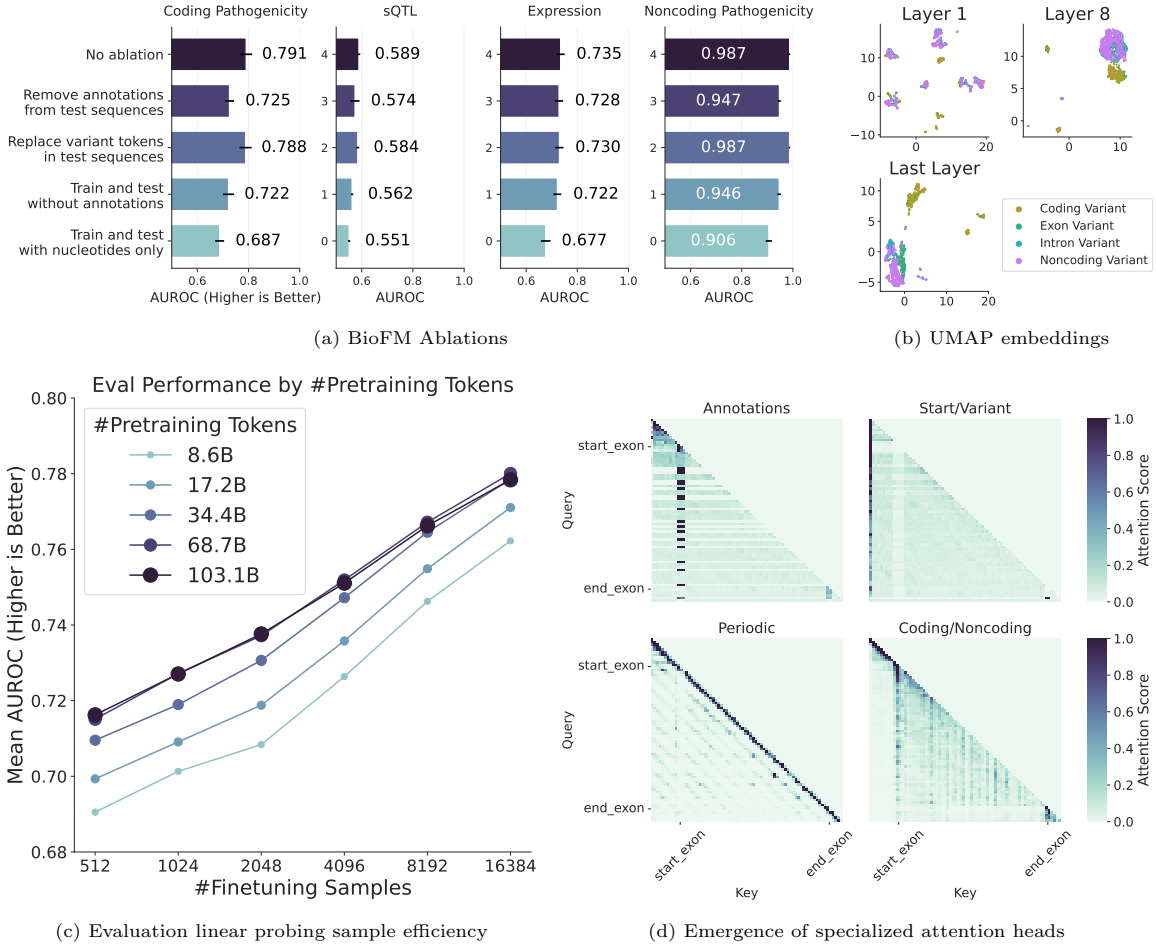
First, we quantified the impact of BioToken and genetic diversity through ablation experiments conducted with the 134M-parameter BioFM, where we ablated the variant and annotation tokens (Figure 5a). We established baseline comparisons by training models on the human reference genome alone and on 1000 Genomes data, each without specialized tokens. Further, we assessed the specific impact of annotation tokens by training an intermediate model that included variant tokens but excluded annotations. Additionally, during evaluation, we selectively removed annotation and/or variant tokens from test sequences to quantify their effect on model's performance.

Models trained without annotation tokens showed reduced performance across all evaluated tasks, with the coding pathogenicity task performance notably declining from 0.791 to 0.725 AUROC, highlighting the functional importance of the annotation tokens. Variant tokens specifically contributed significantly to sQTL prediction, where their removal resulted in a statistically meaningful decrease in AUROC (0.589 to 0.574). Collectively, these experiments demonstrate that the model's incorporation of genomic annotations and variants substantially improves performance, especially for variant-driven biological phenomena.

Next, we investigated the structure of variant representations learned by BioFM using UMAP embeddings derived from pathogenicity benchmark datasets (Figure 5b). This embedding analysis revealed progressive differentiation of variants in deeper model layers. While initial layers showed no biologically coherent separation, the final layer clearly distinguished coding from noncoding variants, with further sub-structuring between exonic and intronic variants. This aligns well with biological intuition, as coding variants primarily affect protein structure and function, whereas noncoding variants are predominantly involved in splicing and regulatory mechanisms.

We also evaluated BioFM's ability to utilize large-scale pretraining well by analyzing its sample efficiency with linear probing on variant benchmark tasks. We expected the downstream task performance differences to be larger when the number of finetuning samples is low. In agreement with this hypothesis, we observed that the difference between undertrained (8.6B, 17.2B) and trained models (68.7B, 103.1B) is particularly pronounced when the number of finetuning samples is low (Figure 5c). The difference reduces when we increase number of fine-tuning samples for linear probing. This observation illustrates that there is a clear benefit to scaling training data size, with the law of diminishing returns eventually becoming apparent.

Lastly, we probed BioFM's internal attention patterns to explore how biologically-informed annotations and variant tokens affect representation learning at a mechanistic level (Figure 5d). We hypothesized that specific attention heads may be specializing in processing annotation and variant tokens based on inherent sequence differences between coding and noncoding regions. This expectation aligns with the biological observation that the probability distribution of DNA k-mers differs



(a) BioFM Ablations

(b) UMAP embeddings

(c) Evaluation linear probing sample efficiency

(d) Emergence of specialized attention heads

Fig. 5: Detailed analysis shows that biologically-informed tokenization using BioToken in BioFM significantly enhances genomic modeling accuracy and computational efficiency. **a)** Ablation experiments clearly demonstrate that both annotation and variant-specific tokens in BioToken significantly contribute to improved genomic representation quality, with pronounced impacts on coding pathogenicity and splicing variant prediction. We removed annotations and variant tokens from test sequences and tested the original BioFM on them. Additionally, we trained a version of BioFM without annotation tokens and another version without annotation and variant tokens. **b)** UMAP embeddings illustrate progressive differentiation of variant types through model layers, indicating BioFM effectively learns biologically meaningful distinctions such as coding versus noncoding variants. **c)** Evaluation of linear probing efficiency highlights BioFM's sample efficiency and scalability, benefiting significantly from increased genomic training data while achieving superior downstream performance. **d)** Visualization of attention maps provides mechanistic insight into specialized attention heads, clearly demonstrating BioFM's ability to selectively attend to biologically significant genomic annotations and regions, further validating the functional interpretability and efficacy of the BioToken strategy.

between coding and noncoding regions–coding regions exhibit constrained sequence variation, while noncoding regions often contain repetitive elements. Attention mechanisms should thus help the model distinguish between these contexts by learning distinct probability distribution patterns.

To test this hypothesis, we selected a 64-base sequence that includes a short exon sequence in the middle, flanked by ⟨`start_exon`⟩ and ⟨`end_exon`⟩ tokens, with a variant token at the end. Visualization of attention maps for this sequence revealed distinct patterns among attention heads (Figure 5d). Some attention heads were dedicated to annotation tokens, while others focused on sequence start and variant tokens, potentially aiding in the recognition of contextual boundaries and mutation sites. Certain other attention heads exhibited periodic patterns, which may reflect the repetitive structure of noncoding regions, often associated with regulatory or structural genomic elements. Additionally, some attention heads segregated exonic and intronic regions, indicating a

learned distinction between protein-coding and noncoding sequences. These findings indicate that the model recognizes biologically meaningful sequence features, offering insights into how BioToken benefits BioFM in encoding genomic information.

Discussion and Conclusion

In this work, we introduced BioToken, a biologically-informed tokenization framework designed to improve the computational representation of genomic sequences and their variants. Using BioToken, we developed BioFM, a foundation model that achieves superior performance across a wide spectrum of downstream tasks. Unlike conventional genomic foundation models, BioFM explicitly encodes biological structures and variants into its token representations, enabling it to surpass existing GFMs of comparable size and even much larger models such as Evo2 7B across several tasks. Most significantly, BioFM matches or exceeds specialized models like Enformer and SpliceTransformer on their respective domain-specific benchmarks, bridging the longstanding gap between general-purpose genomic models and task-specific architectures. This makes BioFM the first GFM with a capability to model variant effects across the central dogma. We also showed that BioFM is competitive on region classification tasks and is able to generate promoters given a gene sequence.

The core innovation of BioFM stems from a fundamental shift in how genomic data is represented in machine learning models. Traditional approaches typically model DNA sequences as linear nucleotide strings devoid of biological context. BioToken corrects this limitation by encoding genomic variants (including SNVs, insertions, and deletions) and structural annotations (such as exons, introns, transcripts, and coding regions) directly into specialized tokens. Our comprehensive ablation studies confirm that integrating these biologically meaningful annotations is the primary factor driving BioFM's enhanced predictive performance.

A key strength of the BioToken framework is its modularity and extensibility. While our current implementation focuses on SNVs and basic genomic annotations, the framework can be readily extended to incorporate additional biological contexts. Future iterations could include tokens for chromatin 3D structure, methylation rates, histone modifications, or even cell- and tissue-type specific annotations. Another direction is to add information from single-cell foundational models [33–36] in special tokens similar to multimodal LLMs [37]. This extensibility ensures that BioToken can evolve alongside our growing understanding of genomic function and structure.

Another significant contribution of BioToken is its demonstration that encoding intrinsic biological biases directly into model inputs can dramatically improve computational efficiency without sacrificing performance. By providing the model with structured, biologically meaningful tokens, this method reduces the model's reliance on implicit learning of genomic patterns, significantly boosting computational efficiency. BioFM achieved its results using only 160 billion training tokens compared to trillions of tokens used to train other models [3, 15]. Another relevant finding was that BioFM does not require the expensive task-specific layer-wise search for best representation, as the last-layer embedding from BioFM leads to superior or equivalent performance on downstream tasks compared to the intermediate layers. Together, these results make genome foundation modeling more accessible for the community.

Moreover, the principles underlying BioToken are not limited to DNA sequences. The same approach can be adapted for protein language models by augmenting traditional amino acid tokenizers with domain annotations and variant tokens. This could significantly enhance our ability to model protein structure-function relationships and predict the functional consequences of amino acid substitutions. The potential applications extend to multi-modal models that integrate DNA, RNA, and protein data within a unified representation framework.

While our results are promising, several limitations merit consideration. First, the current implementation of BioToken and BioFM relies on high-quality annotations and variant data, which may not be equally available across all organisms. Application to non-model organisms or species with limited genomic resources would require adaptation and potentially alternative approaches to annotation. Second, the scaling properties of our model with increased variant data remains open for future works. While we trained on data from 50 individuals from the 1000 Genomes project, extending to the full scale of this dataset (nearly 3200 whole genomes) or biobank-scale data [38] could provide significant gains with a proportional increase in the model size.

In conclusion, BioToken and BioFM represent a meaningful advance in genomic modeling, providing a robust and extendable framework for modeling the complex relationships between genetic variation and molecular phenotypes. This work establishes a new paradigm for genomic foundation

models that prioritizes biological context and computational efficiency, offering a promising path forward for the field of computational genomics and its applications in precision medicine.

Code and Model Availability

- BioFM model is available on Hugging Face: https://huggingface.co/m42-health/BioFM-265M
- Python package for inference and embedding extraction is available on GitHub: https://github.com/m42-health/biofm-eval/
- Variant benchmark is available on Hugging Face: https://huggingface.co/datasets/m42-health/variant-benchmark
- The training code is adapted from Meta's llama-cookbook: https://github.com/meta-llama/llama-cookbook

Appendix A Methods

A.1 Data

A.1.1 Pretraining Dataset

We used WGS data from 1000 Genomes dataset. We selected 50 samples, 10 for each of 5 superpopulations presented in 3202 total samples. During training, we sampled sequences from regions of 6K nucleotides with overlap of 1K and random start shift from 0 to 1K. We filtered out any 6K regions which has more than 50% unknown nucleotides in human reference genome. For each sequence, we applied all SNVs and indels from randomly selected sample mapped to the corresponding sequence region. After that, mutated sequence was truncated to have exactly 6144 tokens.

We performed reverse complement augmentation with probability 50% for each regions. Transcript, Exon, and CDS boundaries were added only to the corresponding strand. We used chromosomes 1-22 and X for training.

A.1.2 Evaluation Datasets

Noncoding Pathogenicity: The noncoding pathogenicity prediction task focuses on determining whether a genetic variant from noncoding regions of the genome is pathogenic (disease-causing) or benign (non-disease-causing). We utilize the BEND benchmark dataset, derived from ClinVar, a widely used resource for clinically significant variants. The dataset contains 295,495 single-nucleotide polymorphisms (SNPs) exclusively from noncoding regions, of which approximately 5% are pathogenic, with the remaining 95% classified as benign. This makes the dataset highly imbalanced, presenting additional challenges for predictive modeling. The current state-of-the-art (SOTA) performance for noncoding pathogenicity prediction is achieved by CADD v1.7, with an AUROC of 0.992. However, it is essential to note that CADD is an aggregate classifier, combining multiple predictive features, some of which may have been derived from methods already trained on ClinVar data.
Coding Pathogenicity: The coding pathogenicity prediction task involves predicting whether genetic variants within coding regions of the genome are pathogenic or benign. Coding regions encode proteins and are critical to understanding disease mechanisms, making this task fundamental for medical genomics and precision medicine. The dataset used for this task is derived from the AlphaMissense, which filtered and curated variants from ClinVar. It contains 82,872 variants, with 37% classified as pathogenic. Compared to noncoding datasets, this dataset is significantly more balanced, providing a more robust foundation for model evaluation. Additionally, coding region data are often considered to be of higher quality due to better annotation and functional validation. Variants in coding regions often result in missense mutations that directly alter protein structure or function. While tools like CADD dominate the field, novel GFM approaches like GPN-MSA, are starting to outperform traditional methods.
Variant Expression: This task involves predicting whether a genetic variant influences gene expression. The ability to predict expression-altering variants has far-reaching implications for understanding the genetic basis of traits and diseases. Accurate predictions can guide experimental validation, prioritize variants for functional studies, and inform the development of gene therapies. The dataset for this task is sourced from the DeepSEA and has been preprocessed to ensure balance, containing a total of 156,162 variants, with exactly 50% having positive labels and other 50% randomly sampled from negative controls. This dataset does not include tissue-specific information, focusing on general expression-modulating effects.
sQTL Prediction: The splicing effect prediction task aims to determine whether a genetic variant influences alternative splicing patterns. Evaluation splicing effects is critical for understanding the mechanisms by which variants contribute to diseases, particularly for disorders caused by aberrant RNA processing. Existing methods primarily focus on predicting splicing sites, rather than the direct effects of specific variants on splicing, leaving a gap in accurate variant-level predictions. The dataset is derived from the sQTLSeeker2, which utilized splicing quantitative trait loci (sQTL) data from GTEx. We randomly selected 13 tissues from 50 tissues presented in GTEx to speed up the evaluation. Each selected tissue has less than 150,000 variants. We randomly subsampled negative controls to match the number of sQTL variants in the dataset.
Ancestry Prediction: The ancestry prediction task involves classifying individuals into one of five superpopulations based on genomic sequences. We took human reference genome (HG38) and applied mutations from every 1000 Genomes sample to get consensus sequences for eleven different

regions of the genome. We fixed the sequence length to 32K in each of the regions. Sequences differ by approximately 0.5% of positions and contain 33 variants on average, including SNPs, insertions, and deletions. These sequences, each of length 32,000, are sampled from 11 regions across different chromosomes in the 1000 Genomes dataset. Model embeddings are extracted, chunked, and linear classifier is used to classify ancestry using embeddings.

Common vs Rare Variants Classification: This task assesses whether model embeddings capture biologically conserved genomic contexts typical of authentic common variants. We randomly sampled 100,000 common variants (Minor Allele Frequency (MAF) ¿ 0.05) from GnomAD and, for each common variant, generated a synthetic negative control by randomly substituting a nucleotide within a ±20-nucleotide local sequence window. The resulting balanced dataset consists of exactly 200,000 single-nucleotide variants (SNVs), equally divided between genuine common variants and synthetic controls.

meQTL Prediction: This task involves predicting if a genetic variant influences nearby methylation rates. We downloaded meQTL dataset from GRASP [27] which contains 52418 datapoints. We filtered out duplicates by rsid, lifted variant coordinates to hg38 and obtained 25496 variants total. To create negative controls, we randomly sampled twice as much variants from variants with more than 5% MAF from Gnomad.

A.2 Modelling

ID	Token
0	D
1	N
2	A
3	T
4	G
5	C
6	[CLS]
7	[SEP]
8	[UNK]
9	[PAD]
10	[EOS]
11	[BOS]
12	[SPECIAL]
13	Ă
14	Ť
15	Ğ
16	Č
17	[START_CDS]
18	[END_CDS]
19	[START_TRANSCRIPT]
20	[END_TRANSCRIPT]
21	[START_EXON]
22	[END_EXON]

Table A1: BioToken Annotation Variant Vocabulary

A.2.1 BioToken

We experiment with three variants of BioToken to explore varying levels of genomic complexity. Each tokenizer incorporates unique aspects of genomic data, enhancing model understanding and performance. Figure 1 provides an overview of nucleotide-level points of interest and their integration into tokenization strategies. A brief description of the tokenizers is provided below:

BioToken-Base: This version represents each nucleotide (A, T, G, C) as a unique token, and is widely used in Genomic Foundation Models. By preserving precise positional and sequence-level details, it enables a granular representation of genetic information and simplifies preprocessing. However, the Character Tokenizer struggles treats nucleotides across the genome as same, and therefore does not factor in the inductive biases.

Parameter	Value
optimizer	AdamW
max_learning_rate	$7.2e-4$
min_learning_rate	$3.2e-5$
weight_decay	0.1
optimizer_momentum	$\beta_1, \beta_2 = 0.9, 0.999$
batch_size	64
context_size	6144
lr schedule	cosine with warmup
epochs	1
initializer_range	0.03
hidden_size	640
n_layers	23
n_heads	32
n_gqa	8

Table A2: Hyperparameters for BioFM 265M.

BioToken-Variant To address individual-specific variations, the this version of BioToken annotates variants explicitly relative to the reference genome. For each individual in the Genome1000 dataset, variants within the sampled location range are identified and integrated into the tokenization process. This approach enhances the model's ability to recognize regions of genomic variation. We assign special symbols to variant tokens: Ă, Ť, Ğ, Č. They account for insertions and SNVs in the genome.

BioToken–Annotation-Variant: Building upon the BioToken-Variant, this approach incorporates additional functional annotations crucial for understanding genomic patterns. Annotations include start and end of coding sequences, transcription regions, and exon regions providing context about regions critical to gene expression. These annotations enable the model to better capture biologically significant patterns. We use the GENCODE annotations version 38 [39]. We filter out every annotation which is not CDS, transcript or exon. We additionally remove duplicate annotations if both annotation type, start, and end are exactly the same. If two annotations of the same type start at the same position, we add both of them, i.e. annotated DNA sequence will have two subsequent start_exon tokens. It is important for the model to know the extent of alternative splicing of genes. The vocabulary for the annotation variant tokenizer is provided in Table A1.

A.2.2 BioFM

We utilize a decoder-only transformer architecture derived from Mistral [40]. It has 23 layers and 265M parameters in total. It uses grouped-query attention but we turned off sliding-window attention. We modified the Rotary Position Embeddings (RoPE) by adjusting the θ parameter from 10^4 to 10^5, following approaches for handling extended contexts [24]. All other hyperparameters are listed in Table A2.

A.3 Evaluation

Variant Benchmark

For each task in Variant Benchmark, except ancestry prediction, we divide variants into 11 folds. Each fold has 21 chromosomes as training set and 2 chromosomes as testing set, except fold 0, which has chromosomes 21, 22, and X as testing set. For encoder models, we put variant in the center of context window, construct embeddings of reference sequence and mutated sequence and concatenate them. For decoder-only models, we construct forward mutated sequence which ends with variant, and reverse complement mutated sequence which also ends with variant. Both forward and reverse sequences have length of exactly half of evaluation context size. After that, we average forward and reverse reference sequences, and forward and reverse mutated sequences, and concatenate the resulting two vectors. This approach ensures that each model has access to the exact same amount of context information and also works around the fact that decoder-only models are causal.

After obtaining embeddings for each variant, we concatenate them into a feature matrix for a linear regression. We call it linear probing. We decided against full or parameter-efficient finetuning to

Metric	Caduceus	HyenaDNA	NT-500m-v2-Multispecies	GenaLM	DNABERTv2	BioFM
EM Mean	0.563	0.497	0.588	0.493	0.648	0.628
H3	0.721	0.727	0.765	0.699	0.806	0.773
H3K14ac	0.513	0.406	0.565	0.455	0.653	0.604
H3K36me3	0.583	0.472	0.632	0.473	0.641	0.640
H3K4me1	0.459	0.412	0.552	0.380	0.537	0.537
H3K4me2	0.422	0.331	0.381	0.321	0.504	0.442
H3K4me3	0.445	0.341	0.444	0.343	0.545	0.572
H3K79me3	0.657	0.588	0.645	0.611	0.721	0.688
H3K9ac	0.513	0.515	0.601	0.517	0.655	0.607
H4	0.775	0.758	0.815	0.723	0.808	0.804
H4AC	0.544	0.424	0.480	0.410	0.606	0.612
PD Mean	0.720	0.780	0.911	0.772	0.802	0.831
all	0.844	0.841	0.931	0.828	0.778	0.869
notata	0.923	0.915	0.949	0.915	0.940	0.944
tata	0.394	0.584	0.854	0.573	0.687	0.680
TFH Mean	0.619	0.588	0.661	0.580	0.700	0.663
TF0	0.641	0.611	0.663	0.668	0.708	0.705
TF1	0.692	0.662	0.718	0.661	0.740	0.709
TF2	0.585	0.596	0.634	0.561	0.660	0.659
TF3	0.488	0.452	0.539	0.316	0.598	0.553
TF4	0.690	0.620	0.750	0.692	0.792	0.691
CPD Mean	0.541	0.631	0.752	0.607	0.694	0.695
all.1	0.621	0.633	0.719	0.599	0.663	0.652
tata.1	0.350	0.596	0.822	0.561	0.730	0.741
notata.1	0.653	0.664	0.715	0.660	0.688	0.691

Table A3: Full GUE Results.

mitigate the overfitting effects and effect of many additional degrees of freedom coming from having many hyperparameter combinations to select from. Linear probing of model embeddings also allows us to isolate the quality of obtained representations. We used default settings in a sklearn package, but with number of steps for lbgfs method set to 5000. We report AUROC for each test set in each fold, and then obtain average AUROC for Figure 2a or median AUROC for Figure 1c and Figure 1d. It allows us to get more robust performance estimates for each model.

For comparison with Enformer, we use an PyTorch Enformer implementation from EleutherAI github.com/lucidrains/enformer-pytorch. Enformer first compresses each sequence into blocks using convolutions. Following original evaluation approach in [20], we extract three hidden states for blocks centered about variant of interest, average them into one hidden state over sequence dimension and use it for the linear probing. Additionally, we ran an experiment where we used only block with the variant, but results did not change. We evaluated Enformer on expression prediction task from Variant Benchmark using context sizes of 12K, 24K, 48K, and 96K. Performance of Enformer saturated for 48K and 96K nucleotides, therefore we did not evaluate it on 192K context.

For comparison with Splice Transformer, we use an official implementation from authors at github.com/ShenLab-Genomics/SpliceTransformer. We always use the maximum available context size of 8K nucleotides, centered about variant of interest, extract embeddings for this variant only, and use linear probing on top.

GUE Evaluation We use the existing train-validation-test splits on the original GUE dataset for finetuning purposes. For each model, we ran a sweep through 4 values of learning rate and 2 values of maximum epochs. The validation dataset is used for model selection, and the test scores corresponding to model with the best validation score is reported. As stated in Vishniakov et al., we find that max-pooling performs better than cls/last token pooling during our initial experiments [13]. Our finetuning configuration is provided in Table A4.

Promoter Generation We use a high-quality experimental dataset EPDNew [32] for the Supervised Fine-Tuning (SFT). It contains 20,172 promoters for 9K genes. We split this dataset into 90% train genes and 10% test genes. After constructing gene sequences and accounting for reverse strands, we annotate each gene sequence with relevant metadata corresponding to the promoter dataset. Promoters are extracted using a -249,$+50$ bp window around the transcription start site, following the approach used in the GUE dataset [41]. Additionally, we collect gene-specific information, including gene symbol, start, and end positions, to provide better contextual representation of the generated promoters. We fine-tune our pre-trained model for promoter generation using the

config	value
optimizer	AdamW
learning_rate	3e-5, 5e-5, 7e-5, 1e-4
weight_decay	0.01
warmup_steps	40
optimizer_momentum	$\beta_1, \beta_2 = 0.9, 0.999$
batch_size	32
lr schedule	cosine
epochs	10 / 20

Table A4: Configuration for GUE fine-tuning.

SFTTrainer from the TRL library [42]. Our training data follows the structured response template: <gene_sequence> P <promoter_sequence>, which serves as inputs for the SFTTrainer.

Ablation, Scaling, Representation Analysis, and Attention Specialization

We conducted detailed analyses to better understand BioFM's internal representations and the influence of training scale on downstream task performance.

For representation analysis, we constructed two-component UMAP embeddings of variant representations from the smaller 134M-parameter BioFM model, which contains 15 layers. Sequences used for embedding were fixed at 1024 nucleotides in length for both forward and reverse strands, with each variant consistently positioned at the sequence end. We averaged embeddings from forward and reverse strands to obtain the final variant representation. UMAP embeddings were computed using the Python umap package, with parameters set to minimal distance 0.05 and number of neighbors 10. To assess biological distinctions captured by these embeddings, we subsampled 500 variants for each Ensembl VEP consequence type from pathogenicity datasets within the Variant Benchmark.

To evaluate how downstream linear probing performance scales with the number of training tokens and finetuning samples, we subsampled only the training sets from each fold across seven tasks in the Variant Benchmark. The test sets remained constant throughout these evaluations. We considered five model checkpoints trained on varying amounts of genomic data, ranging from $8.6B$ tokens (equivalent to approximately 2.5 genomes) to $103.1B$ tokens (approximately 35 genomes). For each subsampled training dataset, we computed the average AUROC across the full test sets of each task, thereby systematically quantifying the relationship between pretraining scale and downstream prediction accuracy.

Finally, to investigate attention specialization, we analyzed attention maps using a fixed 64-nucleotide sequence containing a small exon and coding sequence (CDS) at its center. We examined attention heads specifically selected for their hypothesized biological relevance: annotation and start/-variant token heads from the initial layer (layer 0), and periodic and coding/noncoding heads from the final layer. This analysis aimed to provide mechanistic insights into how BioFM learns biologically meaningful patterns at the nucleotide level.

config	value
train_batch_size	16
eval_batch_size	16
learning_rate	1e-3
weight_decay	0.01
warmup_steps	40
lr schedule	cosine
epochs	10
max_seq_length	6144

Table A5: Configuration for SFT step.

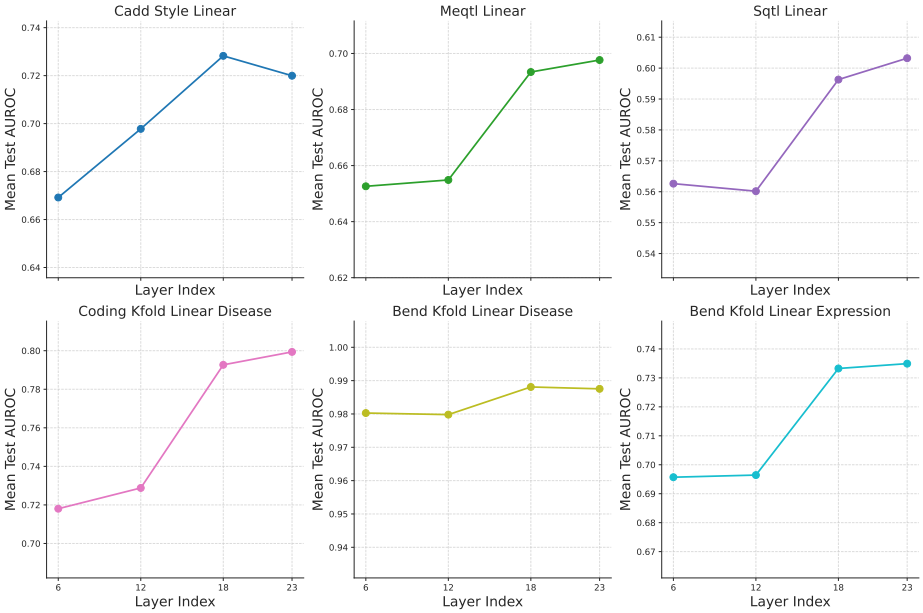


Fig. A1: Different layer embeddings performance of BioFM 265M.

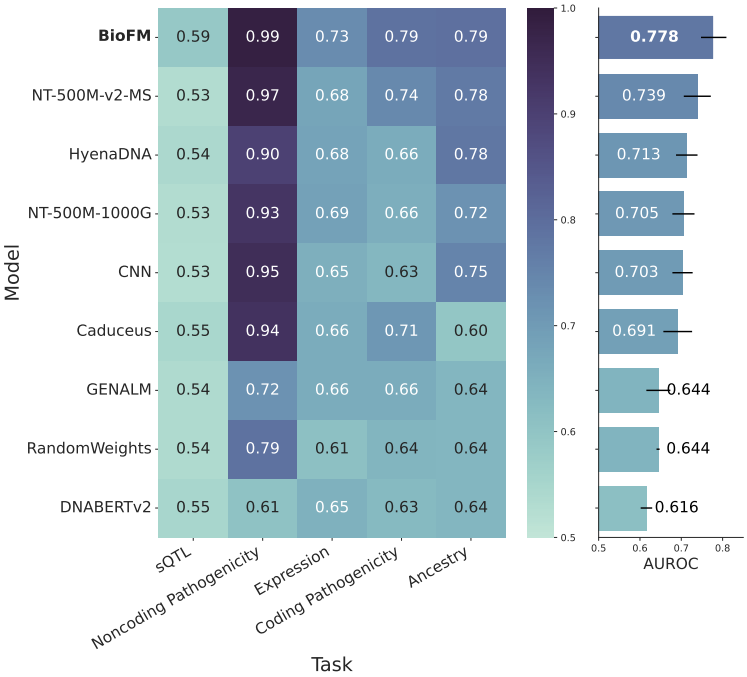


Fig. A2: Variant Benchmark linear probing GFM comparison using context length of 1K. RandomWeights model correspond to the untrained version of BioFM 134M, which essentially produces random embeddings.

References

[1] Marwaha, S., Knowles, J. W. & Ashley, E. A. A guide for the diagnosis of rare and undiagnosed disease: beyond the exome. *Genome medicine* **14**, 23 (2022).

[2] Cipriani, V. *et al.* Rare disease gene association discovery in the 100,000 genomes project. *Nature* 1–9 (2025).

[3] Dalla-Torre, H. *et al.* The nucleotide transformer: Building and evaluating robust foundation models for human genomics. *Nature Methods* (2024).

[4] Nguyen, E. *et al.* Sequence modeling and design from molecular to genome scale with evo. *Science* (2024).

[5] Benegas, G., Albors, C., Aw, A. J., Ye, C. & Song, Y. S. A dna language model based on multispecies alignment predicts the effects of genome-wide variants. *Nature Biotechnology* 1–6 (2025).

[6] Sanabria, M., Hirsch, J., Joubert, P. M. & Poetsch, A. R. Dna language model grover learns sequence context in the human genome. *Nature Machine Intelligence* **6**, 911–923 (2024).

[7] Nguyen, E. *et al.* *HyenaDNA: Long-range genomic sequence modeling at single nucleotide resolution* (2023).

[8] Fishman, V. *et al.* Gena-lm: A family of open-source foundational models for long dna sequences. *bioRxiv* (2023).

[9] Shao, B. & Yan, J. A long-context language model for deciphering and generating bacteriophage genomes. *Nature Communications* **15**, 9392 (2024).

[10] Consortium, . G. P. *et al.* A global reference for human genetic variation. *Nature* (2015).

[11] Chen, S. *et al.* A genomic mutational constraint map using variation in 76,156 human genomes. *Nature* **625**, 92–100 (2024).

[12] Xu, Z. *et al.* Specialized foundation models struggle to beat supervised baselines. *arXiv preprint arXiv:2411.02796* (2024).

[13] Vishniakov, K. *et al.* Genomic foundationless models: Pretraining does not promise performance. *bioRxiv* (2024). URL https://www.biorxiv.org/content/early/2024/12/20/2024.12.18.628606.

[14] Vaswani, A. *et al.* *Attention is all you need* (2017).

[15] Brixi, G. *et al.* Genome modeling and design across all domains of life with evo 2. *bioRxiv* 2025–02 (2025).

[16] Marin, F. I. *et al.* Bend: Benchmarking dna language models on biologically meaningful tasks. *ICLR* (2024).

[17] Zhou, J. & Troyanskaya, O. G. Predicting effects of noncoding variants with deep learning–based sequence model. *Nature methods* (2015).

[18] Garrido-Martín, D., Borsari, B., Calvo, M., Reverter, F. & Guigó, R. Identification and analysis of splicing quantitative trait loci across multiple tissues in the human genome. *Nature communications* (2021).

[19] Cheng, J. *et al.* Accurate proteome-wide missense variant effect prediction with alphamissense. *Science* **381**, eadg7492 (2023).

[20] Avsec, Ž. *et al.* Effective gene expression prediction from sequence by integrating long-range interactions. *Nature methods* (2021).

[21] You, N. *et al.* Splicetransformer predicts tissue-specific splicing linked to human diseases. *Nature Communications* **15**, 9129 (2024).

[22] Kao, C. H. *et al.* *Advancing dna language models: The genomics long-range benchmark* (ICLR, 2024).

[23] Reshef, Y. A. *et al.* Detecting genome-wide directional effects of transcription factor binding on polygenic disease risk. *Nature genetics* **50**, 1483–1493 (2018).

[24] Xiong, W. *et al.* Effective long-context scaling of foundation models. *arXiv preprint arXiv:2309.16039* (2023).

[25] Grešová, K., Martinek, V., Čechák, D., Šimeček, P. & Alexiou, P. Genomic benchmarks: a collection of datasets for genomic sequence classification. *BMC Genomic Data* (2023).

[26] Liu, Z. *et al.* Genbench: A benchmarking suite for systematic evaluation of genomic foundation models. *arXiv preprint arXiv:2406.01627* (2024).

[27] Leslie, R., O'Donnell, C. J. & Johnson, A. D. Grasp: analysis of genotype–phenotype results from 1390 genome-wide association studies and corresponding open access database. *Bioinformatics* **30**, i185–i194 (2014).

[28] Kuratov, Y. *et al.* Babilong: Testing the limits of llms with long context reasoning-in-a-haystack. *arXiv preprint arXiv:2406.10149* (2024).

[29] Amberger, J. S., Bocchini, C. A., Schiettecatte, F., Scott, A. F. & Hamosh, A. Omim. org: Online mendelian inheritance in man (omim®), an online catalog of human genes and genetic disorders. *Nucleic acids research* **43**, D789–D798 (2015).

[30] Mohamed, H., Chernajovsky, Y. & Gould, D. Assembly pcr synthesis of optimally designed, compact, multi-responsive promoters suited to gene therapy application. *Scientific reports* **6**, 29388 (2016).

[31] Tominaga, M. *et al.* Designing strong inducible synthetic promoters in yeasts. *Nature Communications* **15**, 10653 (2024).

[32] Meylan, P., Dreos, R., Ambrosini, G., Groux, R. & Bucher, P. Epd in 2020: enhanced data visualization and extension to ncrna promoters. *Nucleic acids research* **48**, D65–D69 (2020).

[33] Yang, F. *et al.* scbert as a large-scale pretrained deep language model for cell type annotation of single-cell rna-seq data. *Nature Machine Intelligence* **4**, 852–866 (2022).

[34] Cui, H. *et al.* scgpt: toward building a foundation model for single-cell multi-omics using generative ai. *Nature Methods* **21**, 1470–1480 (2024).

[35] Cui, Z., Xu, T., Wang, J., Liao, Y. & Wang, Y. *Geneformer: Learned gene compression using transformer-based context modeling*, 8035–8039 (IEEE, 2024).

[36] Fu, X. *et al.* A foundation model of transcription across human cell types. *Nature* 1–9 (2025).

[37] Team, C. Chameleon: Mixed-modal early-fusion foundation models. *arXiv preprint arXiv:2405.09818* (2024).

[38] Träuble, F. *et al.* Multi-megabase scale genome interpretation with genetic language models (2025). URL https://arxiv.org/abs/2501.07737. 2501.07737.

[39] Mudge, J. M. *et al.* Gencode 2025: reference gene annotation for human and mouse. *Nucleic acids research* (2025).

[40] Jiang, A. Q. *et al.* Mistral 7b. *arXiv preprint arXiv:2310.06825* (2023).

[41] Zhou, Z. *et al.* DNABERT-2: Efficient foundation model and benchmark for multi-species genomes. *ICLR* (2024).

[42] von Werra, L. *et al.* Trl: Transformer reinforcement learning. https://github.com/huggingface/trl (2020).