GENA-LM: A Family of Open-Source Foundational DNA Language Models for Long Sequences

Veniamin Fishman[1,2*†], Yuri Kuratov[1,3†], Maxim Petrov[1], Aleksei Shmelev[1,4], Denis Shepelin[1], Nikolay Chekanov[1], Olga Kardymon[1,4*], Mikhail Burtsev[5*]

[1]AIRI, Moscow, Russia.
[2]Institute of Cytology and Genetics, Novosibirsk, Russia.
[3]Moscow Institute of Physics and Technology, Dolgoprudny, Russia.
[4]HSE University, Moscow, Russia.
[5]London Institute for Mathematical Sciences, London, UK.

*Corresponding author(s). E-mail(s): minja-f@yandex.com; kardymon@airi.net; mb@lims.ac.uk;
†These authors contributed equally to this work.

Abstract

Recent advancements in genomics, propelled by artificial intelligence, have unlocked unprecedented capabilities in interpreting genomic sequences, mitigating the need for exhaustive experimental analysis of complex, intertwined molecular processes inherent in DNA function. A significant challenge, however, resides in accurately decoding genomic sequences, which inherently involves comprehending rich contextual information dispersed across thousands of nucleotides. To address this need, we introduce GENA-LM, a suite of transformer-based foundational DNA language models capable of handling input lengths up to 36,000 base pairs. Notably, integration of the newly-developed Recurrent Memory mechanism allows these models to process even larger DNA segments. We provide pre-trained versions of GENA-LM, demonstrating their capability for fine-tuning and addressing a spectrum of complex biological tasks with modest computational demands. While language models have already achieved significant breakthroughs in protein biology, GENA-LM showcases a similarly promising potential for reshaping the landscape of genomics and multi-omics data analysis. All models are publicly available on GitHub https://github.com/AIRI-Institute/GENA_LM and HuggingFace https://huggingface.co/AIRI-Institute.

1 Main

The encoding of genetic information by DNA is a principal subject in biology, involving both straightforward and complex systems of translation and epigenetic coding, respectively. While the translation of messenger RNA to amino acid sequences employs a widely-accepted genetic code, other forms of encoding, notably the epigenetic code, are more challenging [1]. DNA sequences dictate functional genome elements, including promoters, enhancers, and transcription factor binding sites, among others. However, the diversity and redundancy of their underlying motifs challenge their detection within vast eukaryotic genomes, complicating insights into non-coding genome evolution and interpretations of human genomic variants, given the yet-to-be-fully-unraveled complexity of the epigenetic code.

The advent of next-generation sequencing and additional high-throughput technologies has catalyzed the accumulation and public deposition of extensive databases, rich with functional genomic elements, enabling the broad application of computational methods to large-scale genomic data analysis [2]. We, along with others [3], have successfully employed machine-learning methods, including ensemble learning [4] and convolutional neural networks [5, 6], for this purpose. However, while potent, these approaches encounter constraints in identifying long-range dependencies within DNA sequences, a common phenomenon in human and other eukaryotic genomes [7]. Recent strategies employing transformer neural network-based approaches seek to surmount these constraints [8], with cutting-edge transformer architectures showcasing the capability to infer specific epigenetic properties and gene expression levels from DNA sequences with exceptional precision [8]. Nonetheless, a primary limitation exists in that training these specialized in domain models requires significant computational resources, and their inference capabilities are bounded by the features integrated into the training dataset.

Transfer learning, especially through pre-training, has been widely adopted in natural language processing for its capacity to enhance computational efficiency and performance in scenarios with limited target data [9–14]. Models pre-trained on substantial unlabeled datasets can be fine-tuned or utilized as feature extractors for new tasks, frequently outperforming models trained on task-specific datasets, particularly when those datasets are smaller. The application of this approach to bioinformatics is exemplified by the development of DNABERT [15], a BERT-like transformer neural network [14, 16] pre-trained on the human genome to predict subsequences from context, and subsequently fine-tuned for downstream tasks such as promoter activity prediction and transcription factor binding. While DNABERT signifies a promising advance, its applicability is hindered by an input size cap of 500 base pairs, which restricts its ability to capture the extended contexts vital for various genomic applications.

Enhancing input size for transformer models has recently been addressed through several developments, including sparse attention, effective attention, and recurrence. Sparse attention techniques, which utilize either predefined or learned attention patterns like sliding window or block-diagonal, linearize the quadratic dependency of full attention on input length [17–21]. Conversely, linear attention methods approximate full token-to-token interactions through softmax linearization [22, 23]. In the domain of recurrent models, inputs are segmented and sequentially processed, with inter-segment information relayed through prior hidden states [24, 25] or specialized memory [26–28]. Notably, the recently introduced Recurrent Memory Transformer architecture facilitates information aggregation from both long [28] and extremely long input sequences [29], spanning thousands to millions of elements respectively.

In this work, we showcase the successful application of advanced transformer-based neural networks for predictive analysis of various functional genomic elements within DNA sequences, encompassing promoter activity, splicing, polyadenylation sites, enhancer annotations, and chromatin profiles. We contribute to the research community by introducing GENA-LM, a family of open-source models available on GitHub [1] and pre-trained models (prefixed with `gena-lm-`) on https://huggingface.co/AIRI-Institute. We empirically demonstrate that, often, fine-tuning these models surpasses the results obtained from current state-of-the-art architectures. Moreover, we substantiate that augmenting GENA-LM with the Recurrent Memory Transformer (RMT) architecture facilitates increased input sequence length, thereby achieving enhanced performance on complex biological tasks.

2 Results

2.1 Family of pre-trained transformer based GENA-LM models

In this study, we introduce a new universal transformer model tailored for nucleic acid sequences, which offers several improvements over existing models such as DNABERT [15] and BigBird [20] (as depicted in Fig. 1, A). To ensure its versatility across various applications, we pre-trained our model using multiple datasets and diverse input sequence lengths.

In the data preprocessing phase, we have extended established pipelines by integrating Byte-Pair Encoding (BPE) for sequence tokenization (Fig.1, A, bottom). The essence of BPE is that it constructs a sequence dictionary to pinpoint the most frequently occurring subsequences within the genome. This results in tokens of diverse lengths, ranging from a single base pair up to 64 base pairs. In our tests, the median token length was determined to be 9 base pairs (Fig.1, C). Interestingly, our BPE vocabulary revealed tokens of significant biological relevance. For example, the longest tokens were often indicative of familiar repetitive elements, such as LINEs or simple repeats (Fig. 1, D). The tokenization approach we adopted is greedy, starting with the longest sequences in the dictionary and tokenizing them first. Employing BPE, as opposed to the overlapping k-mers used in earlier studies, allows for the analysis of more extended sequence fragments while maintaining the same model input size. To put it in perspective, 512 overlapping 6-mers represent 512 base pairs, but 512 BPE tokens can represent approximately 4.5 kb. This is a crucial factor when dealing with expansive and intricate genomes like that of humans. Nevertheless, it's worth noting that the model's granularity is confined to the resolution of these individual tokens, which might pose constraints for certain applications.

[1] https://github.com/AIRI-Institute/GENA_LM

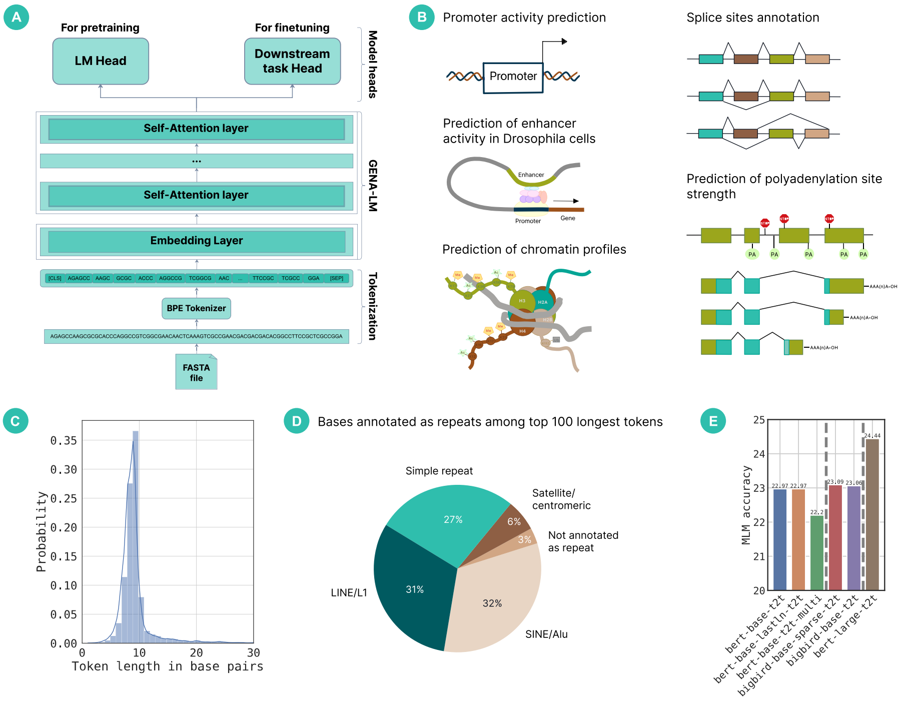

Fig. 1 The GENA-LM Family of Foundational DNA Language Models. **A**. The GENA-LM transformer-based architecture is pre-trained on DNA sequences using a masked language modeling (MLM) objective. DNA sequences are tokenized using a BPE algorithm before processing through the transformer layers. Post pre-training, this foundational DNA model incorporates a downstream task-specific head for fine-tuning. **B**. GENA's evaluation tasks include predictions related to promoter and enhancer activities, splicing sites, chromatin profiles, and polyadenylation site strength. **C**. Post-BPE tokenization, the median token length stands at nine base pairs, as reflected in the token length distribution. **D**. Illustration of repetitive element representation for the 100 longest tokens. **E**. Accuracies for pre-training on masked language modeling task demonstrate that models with a higher parameter count achieve superior performance.

Our second enhancement pertains to the diversification in the implementation of the attention mechanism. The foundational GENA models utilize a conventional attention mechanism, which empowers the model to discern relationships between every pair of tokens in the input sequences. Conversely, Sparse GENA models incorporate a sparse attention mechanism. This approach extends the permissible length of the input sequence by constraining the overall number of connections. Nevertheless, it retains the capability to understand relationships between distant sequence elements. In the case of recurrent GENA models, the transformer is supplemented with memory capabilities. This modification facilitates the processing of even longer inputs by segmenting them.

Through the integration of BPE tokenization and the sparse attention mechanism, we are able to train models that can handle input sequences of approximately 4.5 kb (512 tokens with full attention) and 36 kb (4096 tokens with sparse attention).

The incorporation of recurrent memory further expands this capacity, allowing for the processing of input sequences spanning hundreds of thousands of base pairs.

For model training, we utilized the masked language modeling task, a prevalent technique in natural language processing wherein the model predicts a masked token based on its surrounding sequence context. Unlike previous studies that used the hg38 genome assembly [15, 20], we trained all our models using the more recent human T2T genome assembly, setting our experiment apart. To mitigate the risk of overfitting to the reference genome, we incorporated common variants from the 1000-genome project database into some of our models. Additionally, we enriched our training dataset with genomes from diverse species, encompassing standard model organisms such as mice, fruit flies, nematode worms, and baker's yeast, as well as others covering the entire spectrum of eukaryotic taxa. For a detailed methodology see Section 4.

Throughout the manuscript, we collectively refer to our suite of developed models as GENA language models (GENA-LMs). Each specific model is designated by its label as shown in Table 2. While each model has its unique merits and constraints, we wish to highlight three:

1. The *gena-lm-bert-base-t2t* model: This model emulates the BERT transformer architecture, serving as a benchmark for subsequent models.
2. The *gena-lm-bert-large-t2t* model: With the most significant parameter count (336M) and an input capacity of 4.5 kb, it stands out in terms of complexity.
3. The *gena-lm-bigbird-base-sparse-t2t* models: These models, although having fewer parameters than the *gena-lm-bert-large-t2t*, boast a more extended input sequence length of 36 kb.

It's pertinent to note that the *gena-lm-bert-base* model was crafted during our early experimentation phase. It utilized a distinct tokenizer, train/test split, and other parameters that diverged from subsequent models. Nevertheless, we opted to retain this model in our comparative analysis. It bears the distinction of being the inaugural model our group made publicly accessible, and we aim to familiarize the research community with its potential.

Upon evaluating the performance of our models in the masked language modeling task (Fig. 1, E), we observed that models with sparse attention slightly outperformed their full-attention counterparts limited to 512 tokens. This underscores the role of contextual information in the training regimen. Nonetheless, it's imperative to note that while achieving commendable scores in the masked language modeling task is encouraging, it doesn't necessarily guarantee optimal translation of the learned DNA representations to downstream applications. Consequently, our study delves into the comprehensive assessment of GENA-LMs across a spectrum of biologically relevant tasks to explore their merits and constraints.

2.2 GENA-LM performance on different genomic tasks

To evaluate the foundational GENA-LM models, we selected a range of genomic challenges that have recently been addressed using artificial intelligence (Fig. 1, B). These challenges encompass: 1) prediction of polyadenylation site strength; 2) Estimation of DNA sequence promoter activity; 3) Identification of splicing sites; 4) Forecasting

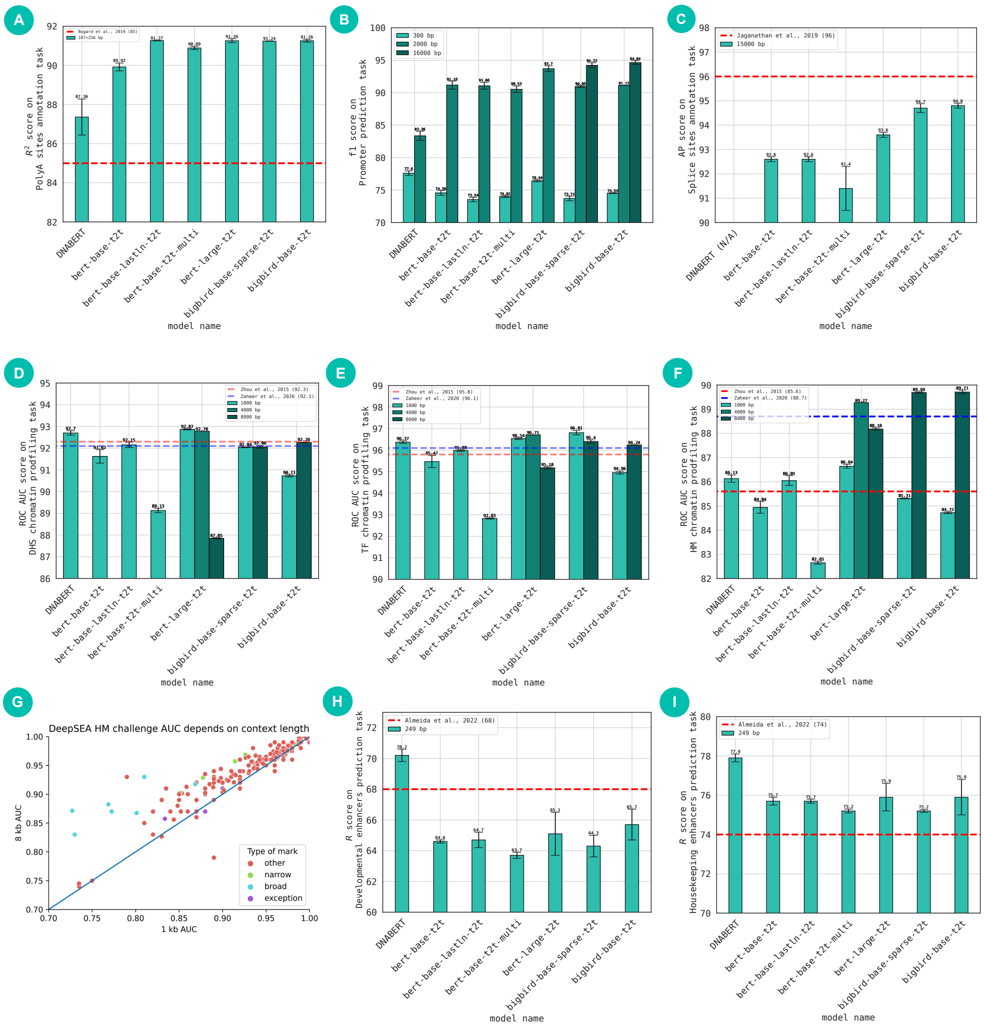

Fig. 2 GENA-LMs achieve state-of-the-art performance across various genomic tasks. **A**. Scores for polyadenylation site prediction. **B**. Promoter prediction. **C**. Annotations of splice sites. **D-F**. Annotations relating to epigenetic features. **G**. AUC score comparison for predicting chromatin occupancy, with a breakdown for each specific chromatin mark, contrasting GENA-LM trained on 1-kb versus 8-kb contexts. **H-I**. Drosophila enhancer classification: developmental (H) and housekeeping (I). For all panels except G, the Y-axis represents the evaluation metric for the models. A dashed vertical line indicates the performance of the previously reported state-of-the-art model, where available. Models were fine-tuned three times with distinct random seeds; standard deviations are indicated by error bars.

of chromatin profiles, which includes histone modifications, DNase I hypersensitivity sites, transcription factor binding sites, among others. The datasets for these tasks (1-4) are derived from human genomic data. To explore the versatility of models pretrained on human data when applied to non-vertebrate species, we introduced a fifth challenge: determining the activity of housekeeping and developmental enhancers in a STARR-seq assay using Drosophila sequences.

To benchmark our method against existing solutions, we established several comparative standards for each challenge. First, we fine-tuned the DNABERT transformer model, which is publicly available. Next, we referenced the performance metrics of state-of-the-art machine-learning models tailored for each specific challenge. Notably, we employed SpliceAI [30] for splicing prediction, APARENT [31] for polyadenylation, DeepSEA [32] for chromatin profiles and gene expression, and DeepSTARR [33] for predicting enhancer activity in Drosophila cells. We ensured that comparisons with the state-of-the-art methods were made only when the composition of these datasets could be replicated accurately. Furthermore, some evaluations incorporated extended versions of the original task. For instance, while the initial DeepSEA model was designed to predict chromatin occupancy of 200 bp loci using an 800 bp context, our study also introduced a model trained on 8 kb sequences. Such an approach facilitated our understanding of how the input DNA context length impacts model performance.

Prediction of polyadenylation site strength.

Polyadenylation in DNA plays a pivotal role, and variants that influence polyadenylation site selection have been linked to diseases. Recently, [31] introduced a reporter assay that quantifies the strength of the proximal polyadenylation signal across millions of short sequences. This study showcased specific nucleotide determinants of the polyadenylation signal strength and demonstrated that this strength could be predicted from the DNA sequence using the convolutional neural network, APARENT. In our work, we adapted GENA-LMs to predict polyadenylation signal strength. As illustrated in Fig. 2, A, all GENA-LM variants significantly surpassed APARENT, with the top-performing GENA-LM achieving a Pearson R^2 value of 0.91 ± 0.0002 compared to APARENT's $R^2 = 0.85$. The DNABERT model, when fine-tuned on the same dataset, also slightly outperformed APARENT ($R^2 = 0.87 \pm 0.01$ versus APARENT's $R^2 = 0.85$). However, its performance was notably lower than that of the GENA-LMs. These outcomes underscore the capability of the GENA-LM architecture to deliver state-of-the-art performance in specific genomic tasks. Nevertheless, it's worth noting that sequences profiled in the polyadenylation signal strength assay are relatively concise, comprising a 187 bp proximal section and a 256 bp distal part. Such short sequences may not leverage the full potential of long-input GENA-LMs. Consequently, we explored the performance of these models on challenges where the value of extended context could be better evaluated.

Promoter activity prediction.

We assessed our models using promoter sequences from the EPD dataset and juxtaposed them against non-promoter control samples. We observed that when the input sequence length was extended from 300 bp to 16 kb, there was a significant improvement in performance, as illustrated in Fig. 2, B. With 300 bp sequences, the DNABERT architecture emerged slightly superior, registering an f1 score of 78.5, compared to the top-performing GENA-LM's f1 score of 76.44 ± 0.16. However, when evaluating longer sequences, the GENA-LM clearly outperformed, recording an f1 score of 93.7 ± 0.44 for 2 kb inputs. This result was markedly higher than the DNABERT model's score, which, when fine-tuned for the same input sequence length, achieved an f1 score of 85.8.

In assessing the performance of GENA-LMs for predicting promoter activity, we observed the following: 1) Models with a greater number of parameters outperformed those with fewer. For instance, the *gena-lm-bert-large-t2t* surpassed the *gena-lm-bert-base-t2t*. 2) The ability to handle longer input sequences due to the sparse attention mechanism gave certain models an edge over traditional full-attention BERT models. As a result, the *gena-lm-bigbird-base-sparse* outperformed the *gena-lm-bert-base-t2t*. Interestingly, models with shorter inputs but more parameters, such as the *gena-lm-bert-large-t2t*, still had superior performance over the *gena-lm-bigbird-base-sparse*. 3) Incorporating multispecies training by using genomic sequences beyond just human data during pretraining did not result in improved performance, as seen when comparing the *gena-lm-bert-base-t2t-multi* with the *gena-lm-bert-base-t2t*. Significantly, all GENA-LM variants achieved better results than the DNABERT. This superiority is likely attributed to the BPE tokenization strategy adopted by GENA-LMs, allowing them to handle sequences around 4-5 kb in length. In contrast, DNABERT divides longer sequences into independent 512 bp segments for processing.

While models with more parameters outperformed those optimized for longer input sequences, this could stem from the 2 kb sequence being amenable to processing by both types of models without truncation. We further explored whether expanding the input length to offer more context, such as 16 kb, could enhance promoter classification accuracy (Fig. 2, B). As anticipated, the performance with these extended inputs outstripped any result observed on the 2 kb dataset: the *gena-lm-bigbird-base-t2t* model achieved an f1 score of 94.64 ± 0.3 on the 16 kb dataset, in comparison to 93.7 ± 0.44 for the *gena-lm-bert-large-t2t* on the 2 kb dataset. This highlights the critical role of context for this challenge. Importantly, the top score for the 16 kb dataset was achieved by a sparse-attention model with fewer parameters than its counterpart that excelled on the 2 kb dataset, suggesting that input context length can indeed outweigh the benefits of model complexity.

Splice site annotation.

We further optimized GENA-LMs to predict splice-donor and splice-acceptor sites within the human genome. We contrasted our results with the leading SpliceAI model, as detailed in Fig. 2, C. The task required analyzing large contexts: a 15 kb input comprised of a central 5 kb target flanked by 5 kb sequences on either end. Notably, the task-specific convolutional neural network, SpliceAI, marginally surpassed GENA-LMs, registering a mean PR AUC of 0.960 compared to 0.947 ± 0.002 for GENA-LMs.

For this task, models designed for longer sequence inputs, such as *gena-lm-bigbird-base-t2t*, outperformed those tailored for shorter inputs, even if the latter were equipped with more parameters, as in *gena-lm-bert-large-t2t*. This aligns with our earlier findings, suggesting that extending contextual information could be more beneficial than merely increasing the number of parameters. Consistent with our promoter analysis, multispecies models, like *gena-lm-bert-base-t2t-multi* (mean PR AUC of 0.914), did not enhance performance compared to their single-species counterparts, such as *gena-lm-bert-base-t2t* (mean PR AUC of 0.926).

Prediction of chromatin profiles.

Predicting a locus's epigenetic states based on its sequence remains a formidable challenge in genomics. To assess the capabilities of the GENA-LM transformers in addressing this, we used the renowned DeepSEA dataset (Fig. 2, D-F). This dataset encompasses over 900 cell-type-specific chromatin profiles, which are grouped into DNAse I hypersensitivity sites (DHS), histone marks (HM), and transcription factor binding sites (TF). In the foundational DeepSEA challenge, chromatin mark signals were predicted for each 200 bp genomic segment, informed by both its sequence and an additional 800 bp context derived from its flanking regions (±400 bp).

When deploying GENA-LMs for this challenge, we discovered that transformer models markedly surpassed the performance metrics previously achieved by the convolutional neural network, DeepSEA. Notably, for TF and DHS profiles, GENA-LMs delivered scores that eclipsed those reported for the BigBird architecture, even though BigBird utilized an expanded 8 kb context (leading GENA-LM average AUC on a 1 kb context for TF: 96.81 ± 0.1 vs. BigBird's 96.1; for DHS: 92.8 ± 0.03 vs. BigBird's 92.3). Furthermore, the performance metrics for GENA-LMs were either on par with or exceeded those recently cited by [34] for the Nucleotide Transformer. They also proved superior to the outcomes of the DNABERT architecture when trained on 1 kb input lengths.

To ensure a more equitable comparison between the BigBird and GENA-LM architectures, we adapted the DeepSEA dataset to incorporate expanded context sequences. This adaptation allowed us to match the 8 kb input length characteristic of the BigBird architecture. Intriguingly, the augmented context had differential effects on the prediction of various epigenetic profiles. For histone marks, a marked performance improvement was evident, with an AUC reaching 89.71 ± 0.08. This was notably superior to the shorter context's score of 86.64 ± 0.08, the original DeepSEA findings (85.6), and the BigBird's result (88.70). However, for TF and DHS predictions, the extension in input length yielded only marginal enhancements in performance.

We analyzed the AUC variations across individual histone marks to pinpoint which epigenetic profiles were responsible for the observed performance enhancement. Remarkably, there was a distinct divergence between narrow and broad histone marks. While the narrow marks demonstrated marginal AUC improvements, the broad marks exhibited a pronounced increase when the context length was extended (Fig. 2, G). These observations reinforce our prior observation [5] that broad histone marks necessitate an expansive context for precise prediction. This highlights the pivotal role of models capable of handling extended input lengths for such tasks.

The varying performance metrics of distinct GENA-LMs across diverse epigenetic profiles and context lengths underscore that no singular model universally excels across all challenges. For transcription factors (TFs), the *gena-lm-bigbird-base-sparse-t2t* stands out on 1 kb inputs, with performance diminishing marginally when the input size increases. In contrast, for DHS, the *gena-lm-bert-large-t2t* model, boasting the highest parameter count, emerges as the optimal choice. Surprisingly, extending the context for this model results in a notable performance dip. For histone marks (HM), the optimal approach hinges on processing extended contexts with the *gena-lm-bigbird-base-t2t* model. Collectively, GENA-LMs outstrips competing models like BigBird, DNABERT, and Nucleotide Transformer, marking a new performance state of the art for this task.

Prediction of enhancer activity in Drosophila cells.

Until now, our investigations focused solely on human data. Going beyond, we evaluated the adaptability of GENA-LMs to other species using the recently introduced DeepSTARR dataset [33]. This dataset presents enhancer activity for millions of short sequences gauged in *Drosophila* cells, classifying each sequence by its housekeeping (cell-type unspecific) and developmental (cell-type specific) enhancer strength. The original authors showed that the convolutional neural network, DeepSTARR, adeptly predicted these activities based on nucleotide sequences. When testing with GENA-LMs, our findings were nuanced (Fig. 2, H-I). In the realm of developmental enhancers, the specialized DeepSTARR model surpassed GENA-LMs (optimal GENA-LM Pearson R = 0.657 ± 0.01 vs. DeepSTARR's 0.68). Yet, for housekeeping enhancers, the tables turned with GENA-LM overshadowing DeepSTARR (optimal GENA-LM Pearson R = 0.768 ± 0.01 vs. DeepSTARR's 0.74). Crucially, against the benchmarks of DNABERT and the Nucleotide Transformer, GENA-LMs consistently showcased higher scores for both categories (Nucleotide Transformer reported R=0.64 for developmental and R=0.75 for housekeeping).

Despite the challenge relying on non-human sequencing data, the multispecies model didn't offer any enhanced performance. The standout performance came from the *gena-lm-bert-large-t2t* model, which boasts the highest parameter count. This suggests that GENA-LMs, even when trained on human data, yield sequence embeddings versatile enough for tasks involving non-human genomes.

Identifying functional elements inside long sequences with GENA-LMs.

Modern techniques for analyzing deep neural networks allow us to assess the contribution of each input element to a model's downstream task performance. Such analyses offer valuable insights into the underlying mechanisms of biological processes. Take the ChIP-seq technique, for instance, a prevalent method for chromatin profiling. Its resolution is approximately 100–200 bp. However, recognition motifs for the majority of DNA-binding proteins are significantly shorter, typically between 4–10 bp. Consequently, deducing precise binding locations from ChIP-seq data is challenging and often necessitates supplementary experiments [35].

To ascertain if GENA can enhance the resolution of experimental ChIP-seq data, we employed token importance scoring [36] on the *bigbird-base-sparse-t2t* model, which was fine-tuned using the DeepSEA dataset. This methodology assigns a significance value to each token within the input, based on its relevance to the prediction outcome. Here, we concentrate on the binding of three transcription factors: ATF1, CTCF, and GATA2 in human K562 cells. Each of these factors possesses well-established DNA recognition motifs (Fig.3, A). This allows for a comparison between tokens important

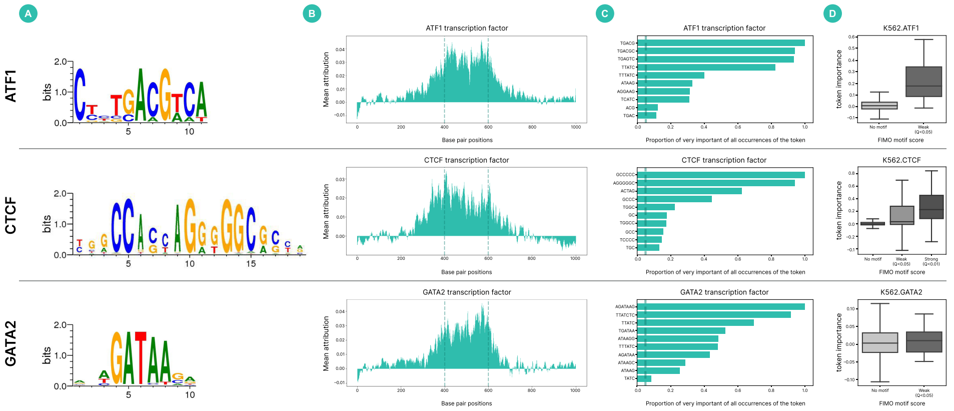

Fig. 3 Analysis with GENA-LM identifies DNA motifs essential for transcription factor binding. In panels A, B, C, and D, each row pertains to a distinct factor, labeled to the left. **A**. Logo representation of motifs for the three transcription factors considered in our analysis. **B**. Profile of average token importance scores over the sequence length. Vertical dashed lines demarcate the 200 bp prediction region. **C**. Bars represent the frequency of token occurrences in the "highly important" category (tokens with scores in the top 5th percentile). The X-axis shows the proportion of these occurrences relative to all occurrences for that token. A vertical reference line marks the 0.05 fraction threshold; only tokens exceeding this fraction are displayed. **D**. Boxplots detail the distribution of importance scores for tokens, categorized by different FIMO q-values. They display the median, interquartile range, as well as the 5th and 95th percentiles.

for GENA's predictions and the recognized sequence determinants associated with transcription factor binding.

First, we examined the distribution of importance scores across the sequence length, as depicted in Fig.3, B. It should be noted that during the fine-tuning process, the input consists of the DNA sequence from the 200 bp target region (where transcription factor binding is anticipated) accompanied by an 800 bp contextual sequence. Our analysis, presented in Fig.3, B, reveals a consistent pattern: the importance attributed to a token diminishes as its distance from the target region increases, a trend observed for all three transcription factors.

We subsequently sought to discern which sequences garnered high token importance scores. Tokens exceeding the 95th percentile of the importance score distribution were designated as "highly important." We then compiled tokens that consistently featured on this "highly important" list. Upon visual examination (Fig. 3, C), we observed that these tokens frequently encompassed full or fragmented motifs of the target transcription factors. For instance, ATF1, which has a core motif of TGACG, prominently displayed a token matching this exact sequence among its highly important tokens. In the case of the GATA2 factor (core motif: GATAA), the token AGATAAG, incorporating the GATA2 motif, was most prevalent among the highly important tokens. As for CTCF, which boasts a motif more intricate and extended than its counterparts, the most recurrent highly important tokens primarily featured GC-rich sub-sequences of the motif.

To more comprehensively assess the congruence between known transcription factor motifs and "highly important" tokens, we employed the FIMO tool to annotate all DNA samples. FIMO is a bioinformatics software designed to identify specific motifs by leveraging the motif's position weight matrix (PWM). As depicted in Supplementary Fig. 1, there's a discernible overlap between significant tokens and motifs detected by FIMO. Our statistical evaluation establishes a relationship between FIMO motif scores and token importance scores. Both robust motifs (FIMO q-value < 0.01) and more tenuous motifs ($0.01 <$ FIMO q-value < 0.05) manifest markedly elevated token importance scores compared to tokens absent of any discerned motif (FIMO q-value $>$ 0.01) (Fig. 3, D). It's worth noting, in the context of the GATA2 transcription factor which is characterized by a shorter motif length, sequences with high FIMO q-values are absent. Nonetheless, we observed that the majority of the "highly important" tokens encompass the core GATA2 motif, as delineated in Supplementary Fig. 2.

While the results affirm that token importance mirrors the presence of established motifs for DNA-binding transcription factors, the congruence between FIMO-detected motifs and tokens with high scores isn't absolute. This observation prompted us to delve into the nature of motifs encompassed by tokens vital for GENA model prediction, yet devoid of the target transcription factor's annotated motif as per FIMO. Utilizing the *de novo* motif discovery tool XSTREME, we analyzed a subset of important tokens lacking a canonical motif (with FIMO target factor motif q-value > 0.05) and examined the enriched motifs therein. Intriguingly, for both CTCF and GATA2, XSTREME predominantly identified their respective motifs. In the case of important ATF1 tokens sans ATF1 motif, the secondary most abundant motif discerned belonged to the ATF-family. This suggests that the rudimentary position weight matrix statistics employed by FIMO might overlook biologically pertinent motif variants that diverge notably from the consensus represented by the PWM. Conversely, GENA-LM exhibits a promising potential in recognizing these variant motifs. When consolidated during XSTREME analysis, these diverse motif representations converge to echo the canonical motif's PWM. Moreover, our analysis revealed a significant presence of GATA2 motifs within tokens deemed essential for the ATF1 factor, hinting at a possible functional synergy between these transcription factors in K562 cells — a nuance discerned by GENA-LM. Given that ATF1 is an integral part of the AP-1 complex, our findings resonate with, and potentially elucidate, prior experimental data evidencing cooperation between GATA2 and the AP-1 complex [37].

Having ascertained GENA's capacity at pinpointing DNA motifs crucial for transcription factor binding, we turned our attention to discerning sequence determinants associated with histone marks, which currently lack identifiable binding motifs. Our focus centered on H3K4me1, H3K9me3, and H3K27me3. These factors represent histone modifications with distinct and well-documented functional implications: H3K4me1 delineates active genomic regions, H3K9me3 signifies heterochromatin, and H3K27me3 demarcates the suppressed "facultative heterochromatin" including genes under developmental regulation, bound by polycomb group proteins. In our examination of these factors, we evaluated the distribution of importance scores relative to sequence length and accumulated tokens that frequently exhibited high importance values.

As depicted in Supplementary Fig. 3, the distribution of token importance scores across sequence lengths for these epigenetic markers mirrors that of the previously discussed factors. Notably, specific tokens consistently emerged as highly significant in predicting these histone marks, reminiscent of the motif-rich tokens previously identified as important for predicting associated transcription factors (see Supplementary Fig. 4).

Prompted by our observations, we sought to determine if motifs enriched among the tokens with high importance scores were shared across these three factors. Extending the sequences of our selected highly important tokens by 4 bp, we then undertook a rigorous motif analysis using XSTREME. Our examination revealed several motifs of significance (see Supplementary Table 4), each aligning with known transcription factors. For the active promoter mark, H3K4me1, the significant tokens were found to encompass motifs corresponding to the GATA, JUN, and FOSL transcription factors. These findings align with the documented roles of these factors in regulating transcription and influencing the oncogenic transformation of K562 cells [38, 39]. In the context of the H3K27me3 mark, which signifies facultative heterochromatin and designates functional elements repressed in specific cell lineages, our data from hematopoietic K562 cells indicated an enrichment of transcription factor motifs not typically associated with blood cells. Examples include SNAI2, pivotal in epidermal cell differentiation, and ASCLI, a critical regulator of neurogenesis. This suggests that within this setting, GENA discerned genomic motifs that might be activated in alternative cell types but are designated for repression in the K562 lineage. Lastly, for the H3K9me3 mark indicative of constitutive heterochromatin, our analysis highlighted an enrichment of the ZNF274 transcription factor motif. This is in agreement with its established role as a transcriptional repressor.

In summary, our findings indicate that beyond predicting the epigenetic profiles of a specific locus, GENA models can effectively identify the distinct subsequences that drive observed epigenetic signals. Such analysis holds potential to substantially augment the resolution of prevailing experimental approaches, like ChIP-seq, and to pinpoint transcription factors linked to particular histone modifications.

Species classification based on embeddings from GENA-LMs.

The universality of GENA-LMs in addressing various biological challenges suggests that the DNA embeddings of the pre-trained model encapsulate significant biological insights. Evolutionary distant species are known to exhibit divergence in their regulatory code and codon usage patterns. If GENA-LMs effectively capture these inherent biological characteristics during pre-training, one would expect that their embeddings could differentiate DNA sequences sourced from varying species without any additional fine-tuning. To evaluate this premise, we curated a set of 27 species spanning diverse taxonomic classifications, ranging from bacteria to humans (refer to Supplementary Table 3). These species also represent a broad spectrum of evolutionary divergence times, spanning from millions to over a billion years (as depicted in Fig. 4, A, B). We then examined the embeddings generated by inputting genomic DNA subsequences into pre-trained GENA-LMs. Our investigations encompassed a range of sequence lengths, beginning with the typical length of shotgun sequencing reads (100 bp) and

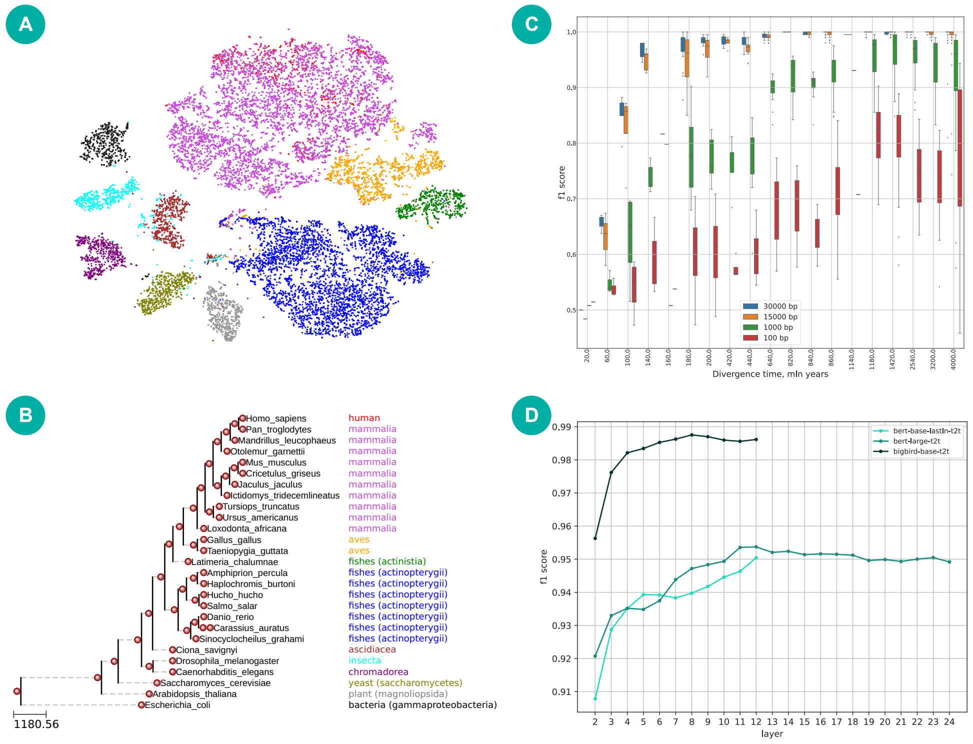

Fig. 4 Sequence embeddings from pre-trained GENA-LMs facilitate species classification. **A** and tSNE projections (A) of sequences sampled from 27 species (B), representing a spectrum across the tree of life. **C**. Classification performance for different sequence lengths plotted against divergence time. **D**. Classification performance of embeddings taken from different layers of three models. Data are presented for sequence lengths of 5 kbp (for *gena-lm-bert-base-lastln-t2t* and *gena-lm-bert-large-t2t*) and 30 kbp (for *gena-lm-bigbird-base-t2t*).

culminating at 30 kb, a size consistent with reads from third-generation sequencing platforms.

Initially, we employed tSNE to project sequence embeddings derived from all genomes into a 2D space. This visualization reveals discernible clusters that mirror the phylogenetic relations between the species. Notably, distinct groupings emerged for bacteria, plants, and yeasts, each isolated from the clusters representing animal genomes. Within the realm of animals, we could discriminate invertebrate species and various vertebrate classes. This evidence underscores that GENA-LM embeddings encapsulate nuances allowing for the differentiation of species based on their genomic sequences.

To delve deeper into these capabilities, we employed a Gradient Boosting algorithm for each of the 27 species pairs. Our aim was to achieve binary species classification leveraging the sequence embeddings.

The data in Fig. 4, C show the richness of information contained in GENA embeddings, enabling species differentiation based on their genomic DNA subsequences. The accuracy of classification is influenced by both the divergence time and the length of the input sequence, with the latter exerting a more pronounced effect. For species that are closely related (with divergence times $\leq$ 20 MYA), classification accuracy remains constrained (f1 score $\approx$ 0.7). However, for species diverging around 60–100 MYA-equivalent to the evolutionary separation among all mammalian species-employing the model that accepts longer sequence inputs boosts our classification capability, yielding an f1 score exceeding 0.8. Remarkably, for extensive divergence times ($\geq$ 200 MYA, reflecting the era of the last common ancestor of vertebrates), the classification's precision approaches perfection.

We next evaluated the classification efficacy of sequence embeddings derived from various layers and architectures of GENA-LMs. Across all models, embeddings sourced from the initial layers consistently delivered subpar performance. This performance incrementally improved, peaking around layers 9 to 12. Notably, for both *gena-lm-bert-large-t2t* (comprising 24 layers) and *gena-lm-bigbird-base-t2t* (with 12 layers), a minor performance dip was observed when utilizing embeddings from the final layers. This trend resonates with prior studies in Natural Language Processing (NLP)[40] and protein modeling[41]. Such studies have posited that in transformer-based language models, the terminal layer embeddings encapsulate information tailored to the specific model training task. In contrast, embeddings from intermediary layers are more versatile, proving advantageous in knowledge transfer for tasks not explicitly addressed during the pre-raining phase. The depth of the layer is also indicative of the abstraction degree of the representations. While preliminary layers prioritize local level representations, the advanced layers capture intricate global features, such as binding sites and contact maps [41].

Collectively, our findings demonstrate that the sequence embeddings from pre-trained GENA-LMs encapsulate abundant biological insights, enabling the resolution of genomic challenges without the necessity for fine-tuning.

2.3 Handling even longer sequences with recurrent memory

While the integration of sparse attention techniques and BPE tokenization in GENA-LMs has substantially expanded the permissible DNA input length, the current limit (about 36 kb) may not sufficiently capture certain biological dependencies. Notably, the prediction of chromatin interactions [7], enhancer-promoter associations [4], gene expression [8], and other genomic phenomena necessitate the processing of contexts that extend beyond 30 kb. Additionally, our empirical analyses show improvements in promoter and splice site predictions as the context size expands from 512 to 4096 tokens (Section 2.2). This indicates the potential benefits of further enhancing sequence length for these biological tasks.

To enhance the input capacity of GENA-LMs, we incorporated recurrent memory mechanisms. The Recurrent Memory Transformer (RMT) has been demonstrated as an efficient, plug-and-play method to handle extended input sequences using pre-trained Transformer models [29]. In this recurrent strategy, the input sequence is partitioned into segments which are processed one after the other (Fig. 5, A). Special

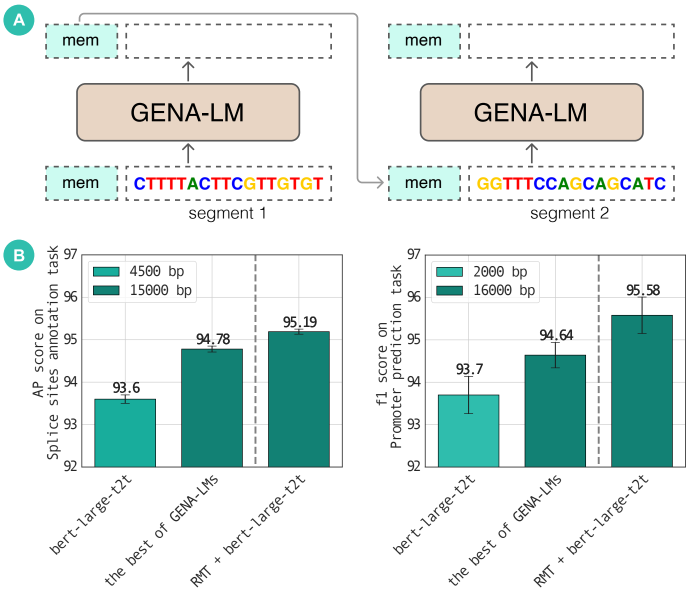

Fig. 5 Leveraging Recurrent Memory to enhance the input capacity of GENA-LM models yields improved performance in downstream tasks. **A**. The Recurrent Memory Transformer (RMT) architecture. A vocabulary of the model is augmented with a memory token denoted as *mem* on the diagram. Memory augmented model is finetuned to write relevant information in memory tokens and pass it to subsequent segments. **B**. The combination of RMT with *bert-large-t2t* with 3x–8x larger sequence lengths, outperforms the base *bert-large-t2t* model. Model with memory achieves superior results in splice site annotation and promoter prediction tasks when compared to all other GENA-LMs, including those utilizing sparse attention.

memory tokens are introduced to each segment to pass information between consecutive segments, allowing them to use information from all previous segments. Thus, the entire pre-trained Transformer effectively functions as a single recurrent unit.

For a comparative evaluation between RMT and other GENA models, we focused on tasks with inputs of moderate length (15–16 kb), which can be processed by sparse models. The *gena-lm-bert-large-t2t* model, when integrated with RMT, was fine-tuned on sequences of 15 kb for promoter prediction and 16 kb for splice site prediction. Inputs were divided into segments, with each segment comprising approximately 512 tokens or about 4.5 kb. These segments also included memory tokens as part of the input, with 10 memory tokens used for each task. When contrasted with the original *gena-lm-bert-large-t2t* model, the sequence length processed by the *gena-lm-bert-large-t2t* + RMT increased substantially: from 3 to 8 times (rising from 4.5 kb to 15 kb for the splice site prediction and from 2 kb to 16 kb for the promoter prediction).

The expansion in input length significantly enhanced the performance of the *gena-lm-bert-large-t2t* model, as depicted in Fig. 5. Notably, models employing the Recurrent Memory Transformer (RMT) outperformed all other GENA-LMs, including those sparse variants of GENA-LM. While these sparse models can accommodate the input lengths featured in the aforementioned tasks, they have fewer parameters compared to *gena-lm-bert-large-t2t*. Thus, RMT allows combining models with higher number of parameters and longer sequence inputs, achieving the best performance on the common biological tasks. Furthermore, RMT has no limit in a sequence length and could be used for even longer sequences. Sparse GENA-LMs, on the other hand, are limited to the lengths on which they were trained.

3 Discussion

Transformer architectures have garnered significant interest across diverse research domains, including genomics. They consistently achieve exemplary results in various biological tasks such as deciphering gene expression regulation in mammals [8] and *E. coli* [42], predicting phenotypes from gene expression [43, 44], deducing DNA methylation [45], and filling in missing genotypes [46], to name a few. Nevertheless, the challenge lies in training task-specific models for each distinct biological question. This process demands substantial time and resources. DNABERT and similar foundational DNA models such as BigBird and Nucleotide Transformer provide a solution by offering a platform for refining universally applicable models without starting from scratch. The Nucleotide Transformer v2 [34] has incorporated Rotary Embeddings and Gated Linear Units paired with swish activations, distinguishing it from its predecessor. This model has input size of 12 kb. DNABERT v2[47], while drawing upon the foundational DNABERT architecture, expands in terms of model parameters and employs BPE tokenization. However, its sequence input length remains below 1000 bp. Contrarily, HyenaDNA [48] introduces a novel architecture capable of handling vast DNA sequences, extending up to 1 million base pairs. Yet, benchmark results suggest an inverse relationship between HyenaDNA's performance and the input size used during its training, as noted by [34]. A unique feature of HyenaDNA is its decoder-only configuration. Unlike the encoder-centric GENA-LMs, HyenaDNA doesn't generate sequence embeddings directly. Instead, it produces DNA sequences, making the derivation of class labels (for classification purposes) or quantitative targets (for regression) from its outputs a complex task. To predict specific DNA states with the HyenaDNA model, the authors utilized a DNA-alphabet encoding, obliging the model to understand this biologically unrelated nucleotide sequence interpretation.

We introduce GENA-LMs, a collection of open-source models boasting the most extensive input capacity among all accessible DNA transformers[2]. The GENA-LM collection encompasses a spectrum of publicly accessible architectures, catering to researchers by offering tailored solutions for unique challenges. Our rigorous benchmarking affirms that GENA-LMs not only surpass earlier pre-trained models but occasionally even rival the precision of task-specific convolutional neural networks.

In our comparison of various GENA-LMs, we investigated the influence of context length and the total number of model parameters on predictive accuracy. We found that the optimal balance between these two factors varies depending on the specific task. For instance, an extended context is vital for predicting promoter activity or deciphering widespread histone mark distributions, as previously indicated by [5]. However, for certain tasks, a more concise context is adequate, making it more

[2] For models starting with the `gena-lm-` prefix, visit: https://huggingface.co/AIRI-Institute/

advantageous to augment the model's parameter count. The broad spectrum of GENA-LMs available offers researchers the flexibility to select a model best suited for their particular objective.

While GENA-LMs accommodate extensive input sizes, they occasionally fall short of the lengths required for peak accuracy in specific biological tasks. For example, research has shown that gene expression can be influenced by variants situated hundreds or even millions of base pairs distant from the promoter. This can be attributed to processes such as loop extrusion [49] and other 3D-genomic mechanisms [50]. There are several strategies to address this constraint in GENA.

First, the RMT technique facilitates the processing of extensive sequence inputs using powerful models with a large number of parameters. Our benchmarks reveal that this approach delivers superior results for tasks where the biological signal spans a lengthy context. Notably, unlike transformer layers that exhibit a quadratic memory dependence on the number of tokens, the computational resources needed for RMT training and inference scale linearly with sequence length. RMT can be integrated with GENA-LMs not only during the fine-tuning phase of downstream tasks but also throughout the MLM pre-training stage. This could enhance learning operations on extended sequences, particularly for downstream task datasets that are of limited size. Furthermore, RMT models pre-trained on multiple segments can be utilized for a greater number of segments during inference [29]. As such, RMT models are versatile enough to address a variety of downstream tasks, even for teams without access to cutting-edge computational infrastructure.

Second, the 3D proximity of chromatin can be determined using specialized models [7]. This information can then be directly incorporated into transformer models, enabling them to capture long-range associations between functional genomic elements.

One limitation of GENA-LMs arises from the granularity imposed by the use of BPE tokenization, which confines predictions to specific tokens. To overcome this, exploring alternative DNA tokenization methods and developing low-level nucleotide embeddings could offer solutions for certain applications.

Beyond merely predicting specific biological signals, we demonstrate that GENA-LMs can also be harnessed to decipher and understand the functions of sequences underpinning these signals. An analysis of token importance revealed that GENA-LM accurately detected motifs corresponding to known transcription factors. Furthermore, it pinpointed transcription factor binding sites crucial for specific histone modifications. There exists an array of factors that modify histones, termed histone "writers", many of which are cell-type specific. Determining these factors and their corresponding genomic binding sites is a complex endeavor. In this context, we illustrate how GENA-LMs can aid in this task by discerning motifs deemed "essential" for a specific histone mark within a particular cell type. However, it's pivotal to approach this method judiciously. The presence of enriched motifs may only indicate an association rather than a direct causal relationship. As an instance, while the enrichment of recognized activator factor motifs within H3K4me3-important tokens aligns with the understood biological roles of these factors, the presence of motifs specific to neural or dermal transcription factors within H3K27me3-important tokens in lymphoid K562 cells likely doesn't signify a direct causal role of these proteins in establishing the repressive H3K27me3 mark. We posit that these factors' targets were suppressed in blood lineage progenitors, implying that the enrichment of their motifs is a reflection of the developmental trajectory of these cells.

To sum up, our study provides compelling evidence that large language models trained on DNA sequences have the capability to generate useful biological insights. This not only presents an innovative method for solving an array of genomic challenges but also forges a pathway for a more nuanced understanding of genetic data. The transformative impact of language models has already been witnessed in protein biology, where they have brought about remarkable progress in predicting protein properties and engineering novel peptides with tailored functions ([51–53]). This is indicative of the potential these models hold, suggesting that their capabilities go beyond mere sequence analysis. With the exponential increase in multi-omics data—spanning genomics, transcriptomics, proteomics, and metabolomics—it's imperative to have advanced analytical tools that can seamlessly integrate and interpret these vast and complex datasets. Language models, as demonstrated by our findings, appear poised to fill this role. As the nexus between computational techniques and biology strengthens, it is foreseeable that language models will be pivotal in ushering in a new era of DNA-based technologies.

4 Methods

4.1 Datasets

4.1.1 Genomic datasets for language model pre-training

Human T2T v2 genome assembly was downloaded from NCBI (acc. GCF_009914755.1). Genomic datasets used to train multispecies models were downloaded from ENSEMBL release 106[3]. The list of species is provided in Supplementary Table 1. For the 1000-genome dataset, we used gnomAD v. 3.1.2 data.

4.1.2 Genomic datasets preprocessing

To prepare genomic datasets for our training corpus, we processed each record in the genomic fasta files. We excluded contigs with the substring "*mitochondrion*" in their identifiers and those shorter than 10 kb. From the remaining sequences, we divided them into 'sentences' spanning 500 to 1,000 base pairs (bp) — the sentence length being randomized — and compiled 'documents' with 50 to 100 consecutive sentences. This approach follows the data processing in BigBird [20]. Data augmentation incorporated reverse-complement sequences, and we applied a stochastic shift for some documents to include overlapping genomic sequences.

For the 1000-genome SNP augmentation, nucleotide substitution was executed, replacing reference alleles with alternative ones sourced from the gnomAD dataset. In order to maintain the haplotype structure, each gnomAD sample was processed individually. This meant that for every genomic region, multiple sequences were derived,

[3] https://ftp.ensembl.org/pub/release-106/

each resulting from swapping reference alleles with sample-specific alternative variants from gnomAD. We limited our focus to genomic regions where the proportion of positions with a noted variant for a given gnomAD sample exceeded 0.01.

4.1.3 Train and test split

For our initial models, *bert-base* and *bigbird-base-sparse*, we hold out human chromosomes 22 (CP068256.2) and Y (CP086569.2) as the test datasets for the masked language modeling task. In contrast, for subsequent models, identifiable by the "t2t" suffix in their names, we hold out human chromosomes 7 (CP068271.2) and 10 (CP068268.2) for testing. All remaining data was used for training. Models focusing exclusively on human data were trained using the pre-processed Human T2T v2 genome assembly combined with its 1000-genome SNP augmentations, totaling approximately $\approx 480\times10^9$ base pairs. On the other hand, multispecies models incorporated both the human-only and multispecies data, aggregating to roughly $\approx 1,072\times10^9$ base pairs. The data splitting strategy for downstream tasks was anchored to methodologies previously described in literature relevant for each particular downstream task. Comprehensive specifics for each task are provided in their respective dedicated sections.

4.1.4 Sequence tokenization

We employed Byte-Pair Encoding (BPE) tokenization [54] for our models, setting the dictionary size to 32,000 and initializing with a character-level vocabulary comprised of ['A', 'T', 'G', 'C', 'N']. Our study utilized two distinct tokenizers:

1. The first, trained exclusively on the human T2T v2 genome assembly, is denoted as 'T2T split v1' in Table 2.
2. The second tokenizer, trained on a mixture of human-only and multispecies data sampled equally, is labeled 'T2T+1000G SNPs+Multispecies'.

Both tokenizers incorporate special tokens: CLS, SEP, PAD, UNK, and MASK. Notably, the 'T2T+1000G SNPs+Multispecies' tokenizer integrates a preprocessing step to manage extensive gaps: sequences with over 10 consecutive 'N' characters are consolidated into a singular '–' token.

4.1.5 Downstream task datasets

A concise overview of the dataset parameters for downstream tasks is presented in Table 1. A comprehensive description follows.

Promoters prediction.

For the task of predicting promoters, we sourced human sequences located upstream of TSS (transcriptional start sites) from the EPDnew database[4]. Sequences of lengths 300 bp, 2,000 bp, and 16,000 bp were extracted, with each dataset being processed and assessed independently. We adopted the approach described in BigBird [20] to generate

[4] https://epd.epfl.ch/EPDnew_select.php

Table 1 Parameters of downstream tasks datasets.

Downstream task	Input length, bp	Number of targets	Task
Promoters prediction (300)	300	2	classification
Promoters prediction (2,000)	2,000	2	classification
Promoters prediction (16,000)	16,000	2	classification
Splice site prediction	15,000	3 per token / bp	multi-class classification
Drosophila enhancers prediction	249	2	regression
Chromatin profiling (1,000)	1,000	919	multi-label classification
Chromatin profiling (8,000)	8,000	919	multi-label classification
Polyadenylation sites prediction	443	1	regression

negative samples. Specifically, each promoter sequence was divided into 20 consecutive non-overlapping segments, and their order was then randomized. The entire dataset was segregated by sequence into training, validation, and testing sets. The objective of this task is a binary classification: determining the presence or absence of a promoter within a given region.

Splice site prediction.

To predict splice donor and acceptor sites, we replicated the dataset from [30], utilizing the original scripts provided by the authors. We adhered to the same training and testing splits as outlined in [30]. In this dataset, a central 5,000 bp target region is bracketed by 10,000 bp of context, with 5,000 bp on each side. Splice site annotations within the target region are aligned to token positions. Tokens overlapping with either splice-donor or splice-acceptor sites are designated as positive samples for their respective splicing annotation class. Subsequently, both the target and its context were tokenized independently. If the combined length diverged from the model's input size, adjustments were made through either padding or truncation. In the event of truncation, sequences furthest from the target region's midpoint were first removed. We demarcated the context and target sequences using SEP tokens. Through this procedure, the target's size matched the model's input token count. However, the computational loss did not account for tokens representing either context or padding. This challenge is a multi-class, token-level classification task encompassing three categories: splice donor, splice acceptor, and none.

Drosophila enhancers prediction.

Candidate sequences, along with their associated housekeeping and tissue-specific activity in *Drosophila* cells, were sourced from the Stark Lab repository[5]. These datasets are partitioned into training, validation, and testing sets, consistent with those used for training the DeepSTARR model [33]. The task at hand involves a two-class regression, wherein each 249-bp sequence is predicted to produce two continuous scores: one for housekeeping enhancer activity and another for developmental enhancer activity.

[5] https://data.starklab.org/almeida/DeepSTARR/Data/

Chromatin profiling.

We gained the DeepSEA dataset [32] from its original repository[6]. This dataset outlines the chromatin occupancy profiles of various genomic features, encompassing histone marks, transcription factors, and DNAse I hypersensitivity regions. The dataset comprises DNA sequences of 1,000 bp, with a central 200 bp target region flanked by 400 bp contexts on either side. Each feature's occupancy is quantified over this 200 bp target. Additionally, we trialed an expanded context of 7,800 base pairs (yielding a total input length of 8,000 bp). To elongate the DNA context, we aligned the input DNA segments to the hg19 genome using *bwa fastmap*. Surrounding sequences at mapped sites were then harvested. Sequences that either failed remapping or aligned too proximate to a chromosome's terminus to permit extension were omitted, though these comprised less than 1% of the dataset. Our partitioning for training, validation, and testing adhered to the divisions presented in the original DeepSEA dataset. The challenge is a multi-label classification, with class count reflecting the unique epigenetic profiles identified in DeepSEA (919 in total).

Polyadenylation sites prediction.

For predicting polyadenylation sites, we employed the APARENT dataset [31] (available at[7]). This dataset characterizes the frequency with which transcription machinery recognizes specific nucleotide sequences as polyadenylation signals. Utilizing the scripts published by the authors, we extracted the target values and delineated the training and testing datasets. Furthermore, we retrieved APARENT predictions (noted under the field *iso_pred*) to gauge the performance of the APARENT model. We tokenized the sequences from both upstream and downstream segments of the 5'-untranslated regions individually, and they were demarcated using a SEP token. This study focuses on regression analysis targeting 256 bp sequences.

4.2 Models architecture and training

4.2.1 DNA language models based on transformer architecture

We trained and expanded upon several transformer models, drawing inspiration from both BERT [14] and BigBird [20] architectures. These adapted models are consistently referred to as `GENA-LM` throughout this manuscript. Key distinctions between these architectures can be found in Table 2. A comprehensive breakdown of parameters and specific combinations for each model is available in Supplementary Table 2. Additionally, we enhanced BERT-based models with Pre-Layer normalization [55]. In instances where the layer normalization is applied even to the final layer output, it is distinctly mentioned as *lastln* in the model names. For precise parameter details, refer to Supplementary Table 2.

All models underwent training using the masked language modeling (MLM) approach. During this process, the sequence was tokenized and flanked by the special tokens, CLS and SEP. In alignment with the BERT pre-training methodology, 15% of the tokens were randomly selected for prediction. Among these, 80% were replaced

[6] http://deepsea.princeton.edu/media/code/deepsea_train_bundle.v0.9.tar.gz
[7] https://github.com/johli/aparent

with MASK tokens, 10% were swapped with random tokens, and the remaining 10% were retained unchanged. Training extended for 1-2 million steps, utilizing a batch size of 256 and operated on 8 or 16 NVIDIA A100 GPUs. We employed the FusedAdam implementation of the AdamW optimizer [56], made available through Nvidia Apex[8]. The initial learning rate was set at 1×10^{-4}, inclusive of a warm-up phase. For most models, we adopted a linear learning rate decay, but in instances where pre-training diverged, we manually adjusted the learning rate.

Table 2 Overview of the GENA-LM Foundational DNA Language Models. This table delineates the specifications of pre-trained GENA-LM models, highlighting variations in pre-training data, layer count, attention type, and sequence length. Models archived on the HuggingFace model hub adhere to a consistent naming convention, prefixed by `AIRI-Institute/gena-lm-`. Models based on the BERT architecture utilize Pre-Layer Normalization [55], with `lastln` indicating the application of layer normalization to the output of the terminal layer. 'T2T split v1' alludes to initial experiments using a non-augmented T2T human genome assembly split. The term '1KG' is shorthand for 1000G SNPs augmentations, while 'M' denotes the inclusion of Multispecies data. The designations 'DS Sparse' and 'HF Sparse' are references to the DeepSpeed sparse attention and HuggingFace BigBird implementations, respectively. The abbreviation 'RoPE' signifies the adoption of rotary position embeddings [57] as an alternative to BERT's absolute positional embeddings. The models were structured with either 12 (denoted BERT-12L) or 24 (denoted BERT-24L) layers, comprising 110M and 336M parameters, respectively.

Model	Architecture	Maximum seq len, tokens (≈ bp)	Tokenizer data	Training data
DNABERT	BERT-12L	512 (512)	3,4,5,6-mer	GRCh38.p13
GENA-LM models:				
bert-base	BERT-12L	512 (4,500)	T2T split v1	T2T split v1
bert-base-t2t	BERT-12L	512 (4,500)	T2T+1KG+M	T2T+1KG
bert-base-lastln-t2t	BERT-12L	512 (4,500)	T2T+1KG+M	T2T+1KG
bert-base-t2t-multi	BERT-12L	512 (4,500)	T2T+1KG+M	T2T+1KG+M
bert-large-t2t	BERT-24L	512 (4,500)	T2T+1KG+M	T2T+1KG
bigbird-base-sparse	BERT-12L, RoPE DS Sparse Att	4,096 (36,000)	T2T split v1	T2T split v1
bigbird-base-sparse-t2t	BERT-12L, RoPE DS Sparse Att	4,096 (36,000)	T2T+1KG+M	T2T+1KG
bigbird-base-t2t	BERT-12L HF Sparse Attention	4,096 (36,000)	T2T+1KG+M	T2T+1KG

4.2.2 GENA-LM fine-tuning

In our standard procedure, we tokenize input sequences and prepend and append them with the service tokens CLS and SEP, respectively. To ensure compatibility with the model's input requirements, sequences are either padded or truncated as needed. For datasets necessitating specialized tokenization, the specific preprocessing steps are detailed in the relevant dataset section.

[8] https://github.com/NVIDIA/apex

Tokenized sequences were provided as inputs to downstream models. These models utilized one of the pre-trained GENA-LM architectures, augmented with a single fully-connected output layer. The dimensions of this layer are denoted by `(hidden_size, target_size)`. Here, `hidden_size` refers to the hidden unit size specific to the GENA-LM model (refer to Supplementary Table 2), while `target_size` is specified in the description of each downstream task dataset discussed earlier. For single-label, multi-class classification tasks, we implemented a softmax activation function on the final layer, paired with cross-entropy loss. In contrast, multi-label, multi-class classification tasks employed a sigmoid activation function on the last layer, combined with a binary cross-entropy with logits loss. Regression tasks did not necessitate any activation function on the last layer and utilized mean squared error as the loss function. To address sequence classification and regression tasks, we used the hidden state of the CLS token from the final layer. Meanwhile, for token-level classification tasks, such as splice site prediction, all hidden states from the ultimate layer were employed. Both the weights of the final fully connected layer and the parameters of the entire GENA-LM were fine-tuned during this process. Learning rate warm-up [58] was consistently applied across all tasks. The optimal number of training and warm-up steps was determined empirically for each individual task.

4.2.3 GENA-LM fine-tuning with recurrent memory

The recently introduced Recurrent Memory Transformer (RMT) presents a novel approach to extend the context length of pre-trained models [28]. Unlike traditional transformers, which exhibit quadratic computational complexity in their attention layers, the RMT employs a recurrent mechanism to efficiently manage elongated sequences. This recurrent design ensures constant memory consumption and linear computational scaling with context length. To process input, the RMT divides the sequence into distinct segments, processing them in a sequential manner. Special memory tokens are integrated into the input of each segment. For a given segment, the outputs linked to its memory tokens are subsequently utilized as input vectors for memory tokens for the succeeding segment. By this method, a multi-layer transformer, such as the pre-trained GENA-LM, functions as a single recurrent cell, addressing one segment at a time.

For both promoter and splice site prediction tasks, we segmented the input sequence into units, with each containing 512 tokens (approximately 4.5 kb). The initial 10 tokens of every sequence were allocated for memory tokens. Segments were processed in a sequential manner, where outputs from the memory tokens of one segment are used as the input memory tokens of the subsequent segment. During the training phase, gradients were allowed to propagate from the final segment to the initial one through these memory tokens. We did not impose any restrictions on the number of unrolls in backpropagation through time (BPTT), allowing gradients to flow uninterrupted from the final to the initial segment. The initial states designated for memory tokens were randomly initialized, and further refined during the fine-tuning process. For the task of promoter prediction, we restricted loss computation to only the last segment. Conversely, for splice site prediction, the loss was determined for every individual segment. The training employed the AdamW optimizer and learning rates of {1e-04, 5e-05, 2e-05}. With a batch size set at 128, the training was terminated when there were no discernible improvements in validation scores. The results for the promoter prediction task are presented as averages over five folds. Meanwhile, the splice site prediction task results are averages across three runs, each employing a distinct random initialization. Training scripts are accessible within our provided codebase.

4.3 Token Attribution Analysis

We employed the Integrated Gradients algorithm [36] to conduct token attribution analysis. This study utilized the *bigbird-base-sparse-t2t* model, which was fine-tuned on the standard DeepSEA dataset with sequences of 1,000 bp. Despite the dataset comprising over 900 features, our analysis specifically targeted six key features: ATF1, CTCF, GATA2, H3K27me3, H3K9me3, and H3K4me1 ChIP-seq profiles from untreated K562 cells. For each genomic feature, we randomly chose 3,000 nucleotide sequences that encompassed ChIP-seq peaks. Subsequent tokenization of these sequences adhered to the same methodology as that applied in the chromatin profile fine-tuning task. With default parameters set, token attribution values were derived. For motif analysis, we leveraged the XSTREME tool [59]. Both FIMO and XSTREME assessments sourced motifs from the HOCOMOCO v11 database [60].

4.4 Phylogenetic analysis using GENA-LMs

For our phylogenetic analysis, we randomly sampled 500 subsequences from each genomic sequence, as detailed in Supplementary Table 3. To ensure representative sampling across entire genomes, the probability of selecting a sequence from a specific chromosome was proportionate to the chromosome's length. In instances not otherwise specified, we utilized the embedding of the CLS token from the final layer. Sequences shorter than 5 kb were processed using the *bert-large-t2t* model, whereas sequences exceeding this length were analyzed with the *big-bird-base-t2t* model to accommodate the extended context. For classifying species, we employed the HistGradientBoostingClassifier from the sklearn library, retaining its default parameters.

4.5 Code availability

The code to generate the findings of this manuscript is available in the *supplementary code* section and on our GitHub repository: https://github.com/AIRI-Institute/GENA_LM. Additionally, our trained models can be found on HuggingFace under the prefix "gena-lm": https://huggingface.co/AIRI-Institute/.

References

[1] Kim, S., Wysocka, J.: Deciphering the multi-scale, quantitative cis-regulatory code. Molecular Cell **83**(3), 373–392 (2023) https://doi.org/10.1016/j.molcel.2022.12.032 . Reimagining the Central Dogma

[2] Sean Whalen, W.S.N..K.S.P. Jacob Schreiber: Navigating the pitfalls of applying machine learning in genomics. Nature Reviews Genetics **23**, 169–181 (2022) https://doi.org/10.1038/s41576-021-00434-9

[3] Libbrecht, M.W., Noble, W.S.: Machine learning applications in genetics and genomics. Nature Reviews Genetics **16**(6), 321–332 (2015) https://doi.org/10.1038/nrg3920

[4] Belokopytova, P.S., Nuriddinov, M.A., Mozheiko, E.A., Fishman, D., Fishman, V.: Quantitative prediction of enhancer–promoter interactions. Genome research **30**(1), 72–84 (2020)

[5] Sindeeva, M., Chekanov, N., Avetisian, M., Shashkova, T.I., Baranov, N., Malkin, E., Lapin, A., Kardymon, O., Fishman, V.: Cell type–specific interpretation of noncoding variants using deep learning–based methods. GigaScience **12**, 015 (2023)

[6] Penzar, D., Nogina, D., Meshcheryakov, G., Lando, A., Rafi, A.M., Boer, C., Zinkevich, A., Kulakovskiy, I.V.: Legnet: resetting the bar in deep learning for accurate prediction of promoter activity and variant effects from massive parallel reporter assays. bioRxiv (2022) https://doi.org/10.1101/2022.12.22.521582 https://www.biorxiv.org/content/early/2022/12/23/2022.12.22.521582.full.pdf

[7] Belokopytova, P., Fishman, V.: Predicting Genome Architecture: Challenges and Solutions. Front. Genet. **11** (2021) https://doi.org/10.3389/fgene.2020.617202

[8] Avsec, Ž., Agarwal, V., Visentin, D., Ledsam, J.R., Grabska-Barwinska, A., Taylor, K.R., Assael, Y., Jumper, J., Kohli, P., Kelley, D.R.: Effective gene expression prediction from sequence by integrating long-range interactions. Nature methods **18**(10), 1196–1203 (2021)

[9] Pan, S.J., Yang, Q.: A survey on transfer learning. IEEE Transactions on Knowledge and Data Engineering **22**(10), 1345–1359 (2010) https://doi.org/10.1109/TKDE.2009.191

[10] Dai, A.M., Le, Q.V.: Semi-supervised sequence learning. In: Cortes, C., Lawrence, N., Lee, D., Sugiyama, M., Garnett, R. (eds.) Advances in Neural Information Processing Systems, vol. 28 (2015). https://proceedings.neurips.cc/paper_files/paper/2015/file/7137debd45ae4d0ab9aa953017286b20-Paper.pdf

[11] Peters, M.E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., Zettlemoyer, L.: Deep contextualized word representations. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), pp. 2227–2237. Association for Computational Linguistics, New Orleans, Louisiana (2018). https://doi.org/10.18653/v1/N18-1202 . https://aclanthology.org/N18-1202

[12] Howard, J., Ruder, S.: Universal language model fine-tuning for text classification. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 328–339. Association for Computational Linguistics, Melbourne, Australia (2018). https://doi.org/10.18653/v1/P18-1031 . https://aclanthology.org/P18-1031

[13] Radford, A., Narasimhan, K., Salimans, T., Sutskever, I.: Improving language understanding with unsupervised learning. Technical report (2018)

[14] Devlin, J., Chang, M.-W., Lee, K., Toutanova, K.: 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), pp. 4171–4186 (2019). https://aclweb.org/anthology/papers/N/N19/N19-1423/

[15] Ji, Y., Zhou, Z., Liu, H., Davuluri, R.V.: DNABERT: pre-trained Bidirectional Encoder Representations from Transformers model for DNA-language in genome. Bioinformatics **37**(15), 2112–2120 (2021) https://doi.org/10.1093/bioinformatics/btab083 https://academic.oup.com/bioinformatics/article-pdf/37/15/2112/39622303/btab083.pdf

[16] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is All you Need. In: Advances in Neural Information Processing Systems, pp. 5998–6008 (2017). http://papers.nips.cc/paper/7181-attention-is-all-you-need

[17] Guo, Q., Qiu, X., Liu, P., Shao, Y., Xue, X., Zhang, Z.: Star-Transformer (2019)

[18] Beltagy, I., Peters, M.E., Cohan, A.: Longformer: The long-document transformer. arXiv preprint arXiv:2004.05150 (2020)

[19] Ainslie, J., Ontanon, S., Alberti, C., Pham, P., Ravula, A., Sanghai, S.: ETC: Encoding Long and Structured Data in Transformers (2020)

[20] Zaheer, M., Guruganesh, G., Dubey, K.A., Ainslie, J., Alberti, C., Ontanon, S., Pham, P., Ravula, A., Wang, Q., Yang, L., Ahmed, A.: Big bird: Transformers for longer sequences. In: Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M.F., Lin, H. (eds.) Advances in Neural Information Processing Systems, vol. 33, pp. 17283–17297. Curran Associates, Inc., ??? (2020). https://proceedings.neurips.cc/paper_files/paper/2020/file/c8512d142a2d849725f31a9a7a361ab9-Paper.pdf

[21] Kitaev, N., Kaiser, L., Levskaya, A.: Reformer: The efficient transformer. In: International Conference on Learning Representations (2020). https://openreview.net/forum?id=rkgNKkHtvB

[22] Choromanski, K.M., Likhosherstov, V., Dohan, D., Song, X., Gane, A., Sarlos,

T., Hawkins, P., Davis, J.Q., Mohiuddin, A., Kaiser, L., Belanger, D.B., Colwell, L.J., Weller, A.: Rethinking attention with performers. In: International Conference on Learning Representations (2021). https://openreview.net/forum?id=Ua6zuk0WRH

[23] Katharopoulos, A., Vyas, A., Pappas, N., Fleuret, F.: Transformers are RNNs: Fast autoregressive transformers with linear attention. In: III, H.D., Singh, A. (eds.) Proceedings of the 37th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 119, pp. 5156–5165 (2020). https://proceedings.mlr.press/v119/katharopoulos20a.html

[24] Dai, Z., Yang, Z., Yang, Y., Carbonell, J., Le, Q., Salakhutdinov, R.: Transformer-XL: Attentive language models beyond a fixed-length context. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pp. 2978–2988. Association for Computational Linguistics, Florence, Italy (2019). https://doi.org/10.18653/v1/P19-1285 . https://aclanthology.org/P19-1285

[25] Rae, J.W., Potapenko, A., Jayakumar, S.M., Hillier, C., Lillicrap, T.P.: Compressive transformers for long-range sequence modelling. In: International Conference on Learning Representations (2020). https://openreview.net/forum?id=SylKikSYDH

[26] Wu, Q., Lan, Z., Qian, K., Gu, J., Geramifard, A., Yu, Z.: Memformer: A memory-augmented transformer for sequence modeling. In: Findings of the Association for Computational Linguistics: AACL-IJCNLP 2022, pp. 308–318. Association for Computational Linguistics, Online only (2022). https://aclanthology.org/2022.findings-aacl.29

[27] Hutchins, D., Schlag, I., Wu, Y., Dyer, E., Neyshabur, B.: Block-recurrent transformers. In: Oh, A.H., Agarwal, A., Belgrave, D., Cho, K. (eds.) Advances in Neural Information Processing Systems (2022). https://openreview.net/forum?id=uloenYmLCAo

[28] Bulatov, A., Kuratov, Y., Burtsev, M.: Recurrent Memory Transformer. In: Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., Oh, A. (eds.) Advances in Neural Information Processing Systems, vol. 35, pp. 11079–11091 (2022). https://proceedings.neurips.cc/paper_files/paper/2022/file/47e288629a6996a17ce50b90a056a0e1-Paper-Conference.pdf

[29] Bulatov, A., Kuratov, Y., Burtsev, M.S.: Scaling transformer to 1M tokens and beyond with RMT. arXiv preprint arXiv:2304.11062 (2023)

[30] Jaganathan, K., Kyriazopoulou Panagiotopoulou, S., McRae, J.F., Darbandi, S.F., Knowles, D., Li, Y.I., Kosmicki, J.A., Arbelaez, J., Cui, W., Schwartz, G.B., Chow, E.D., Kanterakis, E., Gao, H., Kia, A., Batzoglou, S., Sanders, S.J., Farh, K.K.-H.: Predicting splicing from primary sequence with deep learning. Cell **176**(3), 535–54824 (2019)

[31] Bogard, N., Linder, J., Rosenberg, A.B., Seelig, G.: A deep neural network for predicting and engineering alternative polyadenylation. Cell **178**(1), 91–10623 (2019)

[32] Zhou, J., Troyanskaya, O.G.: Predicting effects of noncoding variants with deep learning-based sequence model. Nat. Methods **12**(10), 931–934 (2015)

[33] Almeida, B.P., Reiter, F., Pagani, M., Stark, A.: DeepSTARR predicts enhancer activity from DNA sequence and enables the de novo design of synthetic enhancers. Nat. Genet. **54**(5), 613–624 (2022)

[34] Dalla-Torre, H., Gonzalez, L., Revilla, J.M., Carranza, N.L., Grzywaczewski, A.H., Oteri, F., Dallago, C., Trop, E., Sirelkhatim, H., Richard, G., Skwark, M., Beguir, K., Lopez, M., Pierrot, T.: The Nucleotide Transformer: Building and Evaluating Robust Foundation Models for Human Genomics. bioRxiv, 2023–01115236793 (2023) https://doi.org/10.1101/2023.01.11.523679 2023.01.11.523679v3

[35] Rhee, H.S., Pugh, B.F.: ChIP-exo: A Method to Identify Genomic Location of DNA-binding proteins at Near Single Nucleotide Accuracy. Current protocols in molecular biology / edited by Frederick M. Ausubel ... [et al.] **0 21** (2012) https://doi.org/10.1002/0471142727.mb2124s100

[36] Mukund Sundararajan, Q.Y. Ankur Taly: Axiomatic attribution for deep networks. arXiv preprint arXiv:1703.01365 (2017)

[37] Kawana, M., Lee, M.E., Quertermous, E.E., Quertermous, T.: Cooperative interaction of GATA-2 and AP1 regulates transcription of the endothelin-1 gene. Mol. Cell. Biol. **15**(8), 4225–4231 (1995) https://doi.org/10.1128/MCB.15.8.4225 7623817

[38] Menendez-Gonzalez, J.B., Sinnadurai, S., Gibbs, A., Thomas, L.-a., Konstantinou, M., Garcia-Valverde, A., Boyer, M., Wang, Z., Boyd, A.S., Blair, A., Morgan, R.G., Rodrigues, N.P.: Inhibition of GATA2 restrains cell proliferation and enhances apoptosis and chemotherapy mediated apoptosis in human GATA2 overexpressing AML cells. Sci. Rep. **9**(12212), 1–8 (2019) https://doi.org/10.1038/s41598-019-48589-0

[39] Eferl, R., Wagner, E.F.: AP-1: a double-edged sword in tumorigenesis. Nat. Rev. Cancer **3**, 859–868 (2003) https://doi.org/10.1038/nrc1209

[40] Rogers, A., Kovaleva, O., Rumshisky, A.: A Primer in BERTology: What We Know About How BERT Works. Transactions of the Association for Computational Linguistics **8**, 842–866 (2021) https://doi.org/10.1162/tacl_a_00349 https://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00349/1923281/tacl_a_00349.pdf

[41] Vig, J., Madani, A., Varshney, L.R., Xiong, C., socher, Rajani, N.: {BERT}ology meets biology: Interpreting attention in protein language models. In: International Conference on Learning Representations (2021). https://openreview.net/forum?id=YWtLZvLmud7

[42] Clauwaert, J., Menschaert, G., Waegeman, W.: Explainability in transformer models for functional genomics. Briefings Bioinf. **22**(5), 060 (2021) https://doi.org/10.1093/bib/bbab060

[43] Khan, A., Lee, B.: Gene Transformer: Transformers for the Gene Expression-based Classification of Lung Cancer Subtypes. arXiv (2021) https://doi.org/10.48550/arXiv.2108.11833 2108.11833

[44] Zhang, T.-H., Hasib, M.M., Chiu, Y.-C., Han, Z.-F., Jin, Y.-F., Flores, M., Chen, Y., Huang, Y.: Transformer for Gene Expression Modeling (T-GEM): An Interpretable Deep Learning Model for Gene Expression-Based Phenotype Predictions. Cancers **14**(19), 4763 (2022) https://doi.org/10.3390/cancers14194763

[45] Le, N.Q.K., Ho, Q.-T.: Deep transformers and convolutional neural network in identifying DNA N6-methyladenine sites in cross-species genomes. Methods **204**, 199–206 (2022) https://doi.org/10.1016/j.ymeth.2021.12.004

[46] Mowlaei, M.E., Li, C., Chen, J., Jamialahmadi, B., Kumar, S., Rebbeck, T.R., Shi, X.: Split-Transformer Impute (STI): Genotype Imputation Using a Transformer-Based Model. bioRxiv, 2023–0305531190 (2023) https://doi.org/10.1101/2023.03.05.531190 2023.03.05.531190

[47] Zhou, Z., Ji, Y., Li, W., Dutta, P., Davuluri, R., Liu, H.: DNABERT-2: Efficient Foundation Model and Benchmark For Multi-Species Genome. arXiv (2023) https://doi.org/10.48550/arXiv.2306.15006 2306.15006

[48] Nguyen, E., Poli, M., Faizi, M., Thomas, A., Birch-Sykes, C., Wornow, M., Patel, A., Rabideau, C., Massaroli, S., Bengio, Y., Ermon, S., Baccus, S.A., Ré, C.: HyenaDNA: Long-Range Genomic Sequence Modeling at Single Nucleotide Resolution. arXiv (2023) https://doi.org/10.48550/arXiv.2306.15794 2306.15794

[49] Kabirova, E., Nurislamov, A., Shadskiy, A., Smirnov, A., Popov, A., Salnikov, P., Battulin, N., Fishman, V.: Function and Evolution of the Loop Extrusion Machinery in Animals. Int. J. Mol. Sci. **24**(5), 5017 (2023) https://doi.org/10.3390/ijms24055017

[50] Fishman, V.S., Salnikov, P.A., Battulin, N.R.: Interpreting Chromosomal Rearrangements in the Context of 3-Dimentional Genome Organization: A Practical Guide for Medical Genetics. Biochemistry (Mosc.) **83**(4), 393–401 (2018) https://doi.org/10.1134/S0006297918040107

[51] Shashkova, T.I., Umerenkov, D., Salnikov, M., Strashnov, P.V., Konstantinova,

A.V., Lebed, I., Shcherbinin, D.N., Asatryan, M.N., Kardymon, O.L., Ivanisenko, N.V.: Sema: Antigen b-cell conformational epitope prediction using deep transfer learning. Frontiers in Immunology **13** (2022) https://doi.org/10.3389/fimmu.2022.960985

[52] Madani, A., Krause, B., Greene, E.R., Subramanian, S., Mohr, B.P., Holton, J.M., Olmos Jr, J.L., Xiong, C., Sun, Z.Z., Socher, R., et al.: Large language models generate functional protein sequences across diverse families. Nature Biotechnology, 1–8 (2023)

[53] Wang, Z., Combs, S.A., Brand, R., Calvo, M.R., Xu, P., Price, G., Golovach, N., Salawu, E.O., Wise, C.J., Ponnapalli, S.P., *et al.*: Lm-gvp: an extensible sequence and structure informed deep learning framework for protein property prediction. Scientific reports **12**(1), 6832 (2022)

[54] Sennrich, R., Haddow, B., Birch, A.: Neural machine translation of rare words with subword units. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 1715–1725. Association for Computational Linguistics, Berlin, Germany (2016). https://doi.org/10.18653/v1/P16-1162 . https://aclanthology.org/P16-1162

[55] Xiong, R., Yang, Y., He, D., Zheng, K., Zheng, S., Xing, C., Zhang, H., Lan, Y., Wang, L., Liu, T.: On layer normalization in the transformer architecture. In: III, H.D., Singh, A. (eds.) Proceedings of the 37th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 119, pp. 10524–10533 (2020). https://proceedings.mlr.press/v119/xiong20b.html

[56] Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: International Conference on Learning Representations (2019). https://openreview.net/forum?id=Bkg6RiCqY7

[57] Su, J., Lu, Y., Pan, S., Wen, B., Liu, Y.: Roformer: Enhanced transformer with rotary position embedding. ArXiv **abs/2104.09864** (2021)

[58] Goyal, P., Dollár, P., Girshick, R.B., Noordhuis, P., Wesolowski, L., Kyrola, A., Tulloch, A., Jia, Y., He, K.: Accurate, large minibatch SGD: training imagenet in 1 hour. CoRR **abs/1706.02677** (2017) 1706.02677

[59] Grant, C.E., Bailey, T.L.: Xstreme: comprehensive motif analysis of biological sequence datasets. BioRxiv, September 3 (2021)

[60] Kulakovskiy, I.V., Vorontsov, I.E., Yevshin, I.S., Sharipov, R.N., Fedorova, A.D., Rumynskiy, E.I., Medvedeva, Y.A., Magana-Mora, A., Bajic, V.B., Papatsenko, D.A., *et al.*: Hocomoco: towards a complete collection of transcription factor binding models for human and mouse via large-scale chip-seq analysis. Nucleic acids research **46**(D1), 252–259 (2018)

Declarations

- **Conflict of interest** The authors declare no competing interests
- **Code availability** Code is available as "Supplementary code" file and on github: GitHub: https://github.com/AIRI-Institute/GENA_LM. Pre-trained and fine-tuned models are available with `gena-lm-` prefix in https://huggingface.co/AIRI-Institute
- **Authors' contributions** M.B., O.K., D.S., V.F., and Y.K. conceived the study; Y.K. performed models pre-training and fine-tuning; V.F. performed bioinformatic analysis and prepared datasets; M.P. performed models fine-tuning and token importance analysis; A.S. performed models fine-tuning and species classification analysis; N.C. contributed to datasets preparation; M.B. and O.K. supervised the study; all authors contributed to manuscript preparation.

Supplementary information. Supplementary Figs. 1–4, Supplementary Tables 1–4, and Supplementary Code file.