Augmenting DNABERT Embeddings with Multimodal DNA Features for Improved Regulatory Sequence Interpretation

***Nimisha Papineni*[1], *Pratik Dutta*[1], *Max L. Chao*[1], *Orbin Acanto*[1], *Rekha Sathian*[1], *Pallavi Surana*[1], *Ramana V. Davuluri*[1, *]**

[1]*Department of Biomedical Informatics, Stony Brook University, New York, United States*

[*]*For correspondence:* ramana.davuluri@stonybrookmedicine.edu

Abstract:

While DNABERT leverages k-mer embeddings to model genomic sequences, its exclusive reliance on nucleotide k-mers can limit its effectiveness in capturing regulatory elements that lack distinct motif signals or display subtle, compositionally diffuse patterns. In this study, we explore a multimodal approach by augmenting DNABERT embeddings with DNA-intrinsic features—including nucleotide composition, purine–pyrimidine balance, CpG density, and structural properties such as minor groove width and electrostatic potential. These physicochemical and sequence-derived features offer complementary information about DNA shape and stability, often critical in regulatory regions such as non-TATA promoters and certain transcription factor binding sites (TFBS). By integrating these features with DNABERT representations, we show improved model performance in terms of overall prediction accuracy and ability to interpret pattern-depleted regulatory sequences. We applied this framework, DNABERT-CoreProm-MM (Core Promoter model with MultiModalities), to the task of promoter prediction, with a focus on both TATA and non-TATA Core promoter sequences. Our results demonstrate that the DNABERT-CoreProm-MM model improves prediction accuracy by 3.72% for TATA promoters and 22.56% for non-TATA promoters. These findings highlight the value of sequence-intrinsic and shape feature multimodalities in enhancing the interpretability and accuracy of transformer-based models, particularly for genomic sequences lacking strong motif structure. This approach offers a more comprehensive and biologically informed framework for modeling DNA regulatory elements.

Introduction:

Transformer-based models like DNABERT (1), Gena-LM (2), and Nucleotide Transformer (NT) (3) excel at decoding complex genomic patterns using bidirectional context, much like interpreting human language. Genomic language models have expanded the toolbox for understanding insights at the genomic, transcriptomic, and proteomic levels, which transformed the way we interpret biological sequences, utilizing concepts learned from natural language processing to decipher long-range interactions, make insights from sequences more accessible and applicable to broader contexts. Their ability to integrate DNA sequence information with epigenomic data, such as histone modifications and chromatin accessibility, and methylation patterns, makes them powerful tools for predicting gene regulation and function.

While examining the limitations of DNABERT in identifying sequences, it became evident that some non-TATA promoter models did not attain high accuracy. In identifying distinct patterns within DNA sequences, it encounters challenges when attempting to capture subtle and intricate properties that are not overtly structured. To delve deeper into this issue, we concentrated on the characteristics of these low-accuracy models and scrutinized the sequence properties they exhibited. Understanding physical properties, nucleotide fractions, and their associated sequence calculations is essential for identifying and analyzing disease and developmental stage-related genes. Some studies made use of DNA's nucleotide composition and its physical properties in classification tasks, identifying promoters by assessing genomic composition (4, 5) alongside mutational patterns like single nucleotide variants and TATA-Box motifs.

Nucleotide fractions offer a comprehensive overview of the genomic composition associated with cancer-related genes. Deviations from expected nucleotide ratios can indicate mutations or sequence changes associated with cancer, potentially marking sites of disease progression or mutational hotspots (6). The Purine-Pyrimidines Fraction assesses the balance between purines (Adenine (A), Guanine (G)) versus pyrimidines (Cytosine (C), Thymine (T)), reflecting structural DNA alterations, which are relevant for understanding genomic instability. Calculations related to CpG motifs hold particular significance in oncology due primarily to their influence over gene regulation mechanisms. CpG islands, rich in C and G, are commonly found near gene promoters and are critical for controlling gene expression. Changes within these areas have significant ramifications for tumorigenesis; notably, hypermethylation events affecting these islands often result in silencing tumor suppressor genes, promoting unregulated growth of malignant cells (7). Moreover, distinct mutational signatures manifest characteristic patterns indicative of specific nucleotide transformations tied closely with various types of cancers—for instance, C>T transitions prevalent among melanoma cases correlate strongly with ultraviolet radiation exposure, which aids investigations into environmental influences or inherent genetic vulnerabilities responsible for such mutations (8).

DNA shape features vary across the genome and are particularly enriched in regulatory regions like promoters, enhancers, and nucleosome-bound DNA. Derived from the intrinsic topology of DNA, these features offer deeper insights into how DNA-binding proteins (DBPs) recognize and interact with specific genomic sites (4). However, current high-throughput tools for discovering DNA shape motifs that incorporate multiple structural features remain limited. To address this, several models have been developed that integrate DNA sequence data with biophysical properties such as thermodynamic stability, shape, and flexibility to better understand their role in transcription factor (TF) binding (9). Among these features, a narrow Minor Groove Width (MGW) and strongly negative electrostatic potential (EP) are often associated with enhanced TF binding affinity. Roll and Helical Twist contribute to the structural flexibility of DNA, facilitating its wrapping around histones during nucleosome formation. Additionally, features such as Buckle, Shear, and Opening capacity reflect conformational instability and are frequently observed at mutation-prone regions. Together, these shape features highlight the structural complexity of DNA and underscore their importance in gene regulation and genome stability.

Our framework, DNABERT-CoreProm-MM, incorporates nucleotide fraction properties and some physical properties as an extension to DNABERT embeddings based on LLMs, alongside other multimodal genomic data, enhancing the models' ability to predict and interpret genomic regulation and variant impact. By providing insight into sequence composition and structural features, they help LLMs better understand the broader genomic context, uncover regulatory elements, refine variant predictions, and offer deeper insights into precision medicine applications. This integrated approach leads to a more holistic and nuanced understanding of how different genomic layers interact to regulate cellular function, providing potential solutions addressing existing challenges while boosting predictive accuracies relating especially to core promoter sequences.

Materials and Methods:

A. Data Retrieval:

I. Pattern-based (TATA box motif) dataset

The pattern-based dataset is curated with sequences that are identified with specific motifs (patterns) in a focused region. Here, promoter sequences are identified from transcription start sites (TSS) annotated in the GRCh38 genome by GENCODE (10). Promoter regions are defined as ±45 base pairs around each TSS, resulting in sequences of 90 base pairs. Focusing on the -35 to -25 bp region relative to the TSS, where the presence of the TATA-box motif exists (TATAWAW, where W = A or T) is analyzed. Promoters

containing this motif are labeled as "TATA" promoters, while those without are classified as "non-TATA" promoters.

II. Position Weight Matrix (PWM)- Based dataset

To classify core promoter regions as TATA or non-TATA, we employed a position weight matrix (11)(PWM) approach centered on the TATA box motif(12). A PWM is a scoring matrix representing the likelihood of each nucleotide (A, T, C, G) at every position within a motif. These matrices are typically derived from the observed frequencies of each nucleotide at each position, often obtained from resources like JASPAR [12] through multiple sequence alignments and can be expressed as motif logos [13] or probability matrices [14]. The frequencies are transformed into log-likelihood scores. To evaluate a sequence for a motif of length L, PWM scores for each nucleotide are summed across all positions, i.e.

$$PWM\ score = \sum_{j=1}^{L} S_{ij}\ ; \qquad where \quad S_{ij} = log_2\left(\frac{p_{ij}}{b_j}\right)$$

Where S_{ij} - Score for nucleotide *i* at position *j;* p_{ij} - The observed probability of nucleotide *i* at position *j*; b_j - Background frequency of nucleotide *j* in the genome. PWMs are used to detect motifs in sequences by sliding the matrix across the sequence and calculating cumulative scores; a score exceeding a defined threshold suggests the motif's presence.

We used a position weight matrix (PWM) built from the JASPAR TATA-box motif (POL012.1) to score each ±45 bp core-promoter region around annotated TSSs. A PWM encodes the log-likelihood of observing each nucleotide (A, T, C, G) at every motif position, allowing rapid motif detection. To evaluate all possible promoters, we extracted all annotated TSSs from Ensembl for the GRCh38 genome via BioMaRt(13) and extracted a ±45 base pair region of the template strand that is defined as our core promoter region. Next, we retrieved the POL012.1.pfm (TATA promoter position frequency matrix) from JASPAR (14). Using the position frequency matrix (pfm), we constructed a TATA position weight matrix (PWM) by assuming a uniform background frequency of 0.25 for each nucleotide (A, T, C, and G). We then applied the scoring method described above to evaluate each core promoter region, identifying the position within each sequence where the TATA PWM achieved the highest score.

To separate TATA versus non-TATA promoters, sequences with a TATA box PWM score above the threshold of 0.75, where the highest-scoring position fell between 25-35 bp upstream (-35 to –25 bp) relative to the TSS, were classified as TATA promoters. All other sequences within the TSS-centered regions were considered non-TATA promoters. The 0.75 threshold was validated against the Eukaryotic Promoter Database (EPD). This procedure yielded 10,080 TATA and 76,164 non-TATA promoters, which were used for model training and evaluation.

For the non-promoter set, we randomly sampled genomic regions (both positive and negative strands) that did not overlap any annotated TSS (±45 bp) as defined above. The same PWM scoring methodology was applied to categorize non-promoter sequences as either TATA or non-TATA. From genomic regions outside annotated TSSs, we selected 66,727 TATA non-promoters and 725,784 non-TATA non-promoters for further evaluation**.**

III. Eukaryotic Promoter Database (EPD)-new data set

The EPD (15, 16) employs a comprehensive curation strategy that combines experimental validation, extensive literature review, and advanced computational techniques to ensure the accuracy and integrity of promoter sequences. Data collection through experimental means utilizes high-throughput methodologies, including ChIP-seq (Chromatin Immunoprecipitation sequencing) and CAGE(17) (Cap Analysis of Gene Expression), for the confirmation of TSS. These methods facilitate the precise identification of promoter

sequences that consistently delineate TSS along with adjacent promoter regions, thereby establishing a solid foundation built on empirically derived data.

B. DNA Sequence Properties

We examined ten distinct features by analyzing the nucleotide composition exhibited by a DNA sequence (supplementary Fig.1). Among these, seven are derived from single-nucleotide composition, one is based on di-nucleotide composition, and the final two are rooted in tri-nucleotide composition.

A, C, T, G fractions (I, II, III, IV) represent the proportion of individual nucleotides in the sequence, indicating the relative abundance of each nucleotide, providing insight into the sequence composition.

I. $A\_Fraction = n_A/L$

II. $C\_Fraction = n_C/L$

III. $G\_Fraction = n_G/L$

IV. $T\_Fraction = n_T/L$

This *PurPyr* (V) fraction measures the difference between purine and pyrimidine content. This metric can indicate an imbalance in the nucleotide composition, which may have biological implications.

V. $PurPyr\_Fraction = (n_A + n_G - n_C - n_T)/L$

This *AmKe* (VI)fraction measures the balance of amino (A and C) versus keto (G and T) groups in the sequence. Differences in amino and keto group content can affect the chemical properties of the DNA.

VI. $AmKe\_Fraction = (n_A + n_C - n_G - n_T)/L$

This *WeSt* (VII)fraction measures the balance between weak (A and T) and strong (C and G) bonds. The proportion of weak versus strong bonds influences the melting temperature and stability of the DNA.

VII. $WeSt\_Fraction = (n_A + n_T - n_C - n_G)/L$

These CpG_1(VIII), CpG_2(IX), and CpG_3(X) calculations measure the frequency of CpG dinucleotides and trinucleotides, which are often involved in gene regulation and methylation processes.

VIII. $CpG_1 = (2n_{CG} + 2n_{GC})/(L-1)$

IX. $CpG_2 = (n_{ACG} + n_{AGC} + n_{CAG} + n_{CCG} + n_{CGA} + n_{CGC} + 2\,n_{CGG} + n_{CGT} + n_{CTG} + n_{GAC} + n_{GCA} + 2\,n_{GCC} + n_{GCG} + n_{GCT} + 2\,n_{GGC} + n_{GTC} + n_{TCG} + n_{TGC})/(L-2)$

X. $CpG_3 = (4\,n_{CAG} + n_{CCG} + n_{CGG} + 4\,n_{CTG} + 4\,n_{GAC} + n_{GCC} + n_{GGC} + 4\,n_{GTC})/(L-2)$

C. Shape Features:

Various physio-chemical properties (5) were analyzed to assess the impact on the model's performance. The following properties - A-Philicity, Base Stacking, Protein-DNA Twist, Protein-Induced Deformability, DNA Denaturation, and Z-DNA Stabilizing Energy were found to have no significant effect. Whereas Helical Roll, Helical Twist, Propeller Twist, Minor Groove Width (MGW), Electrostatic Potential (EP), Helical Rise, Helical Shift, Helical Slide, Helical Tilt, Buckle, Opening Capacity, Shear, Stagger, and Stretch contributed to performance improvements when extracted from the DNAshape prediction method (18); which computes shape features by sliding a window across the sequence, extracting each 5-mer, retrieving its corresponding shape value from all-atom Monte Carlo simulations in the R package DNAshapeR (19), and aggregating these values. The variation in shape values across the sequence provides a structural profile, capturing local conformational changes in the DNA. The package generates a set of quantitative descriptors for each sequence by summarizing these variations.

By augmenting these individual calculations of nucleotide sequence properties and shape features as additional feature embeddings to both the pretrained and fine-tuning tasks of DNABERT-CoreProm-MM, the model's performance can be enhanced, especially in cases where DNABERT struggles to accurately classify core promoters and non-promoters that lack specific patterns.

D. DNABERT- Driven embeddings for genomic analysis:

I. Task-Specific Fine-tuning DNABERT-CoreProm-MM with pre-trained DNABERT model

We finetuned DNABERT-CoreProm-MM for predicting core promoters using the pre-trained DNABERT model with the processed promoter and non-promoter data (Fig.1). Different fine-tuning promoter datasets were prepared based on the motif identification method. The data was then divided into a training dataset (80%) and an evaluation dataset (20%) for hyperparameter tuning.

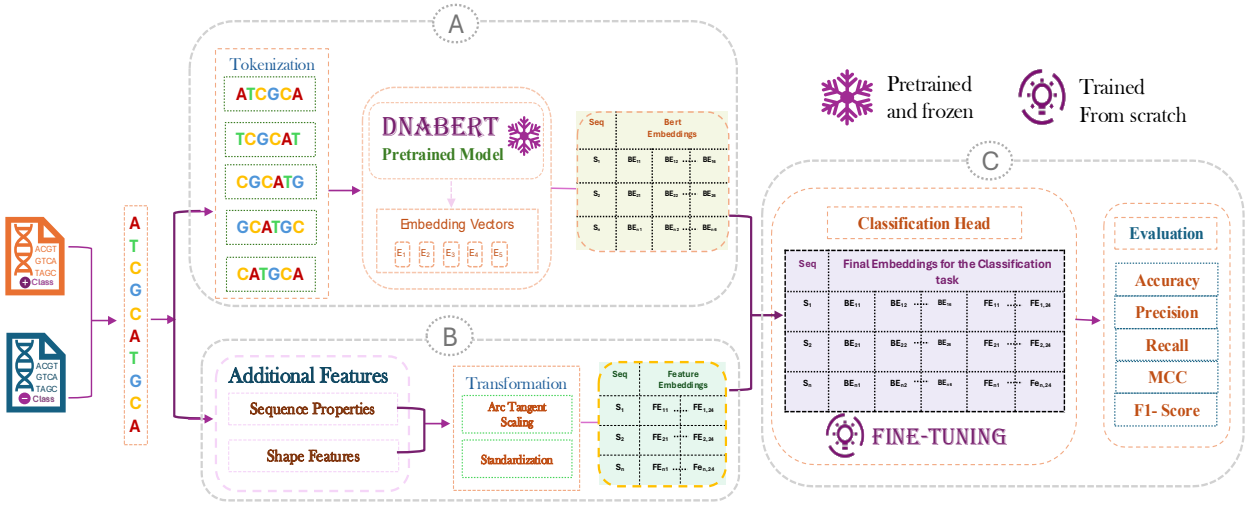

Fig.1: Overview of our multimodal framework DNABERT-CoreProm-MM, which integrates DNABERT embeddings with DNA-intrinsic features for promoter classification. **(A)** Input sequences are tokenized and passed through a pretrained (frozen) DNABERT model. **(B)** Additional sequence-derived features—including nucleotide composition and DNA shape properties—are transformed, standardized, and concatenated with DNABERT embeddings. **(C)** The combined representation is fine-tuned for classification. Evaluation metrics include accuracy, precision, recall, MCC, and F1-score.

DNABERT-CoreProm-MM uses a tokenized 6-mer sequence as an input to generate BERT embeddings with the pretrained DNABERT model, along with nucleotide sequence properties and shape features from the sequences in both the positive and negative classes. 6-mer tokens, with the DNABERT pretrained model, generate the embedding vectors with a self-attention mechanism, creating a Bert embeddings (BE) matrix (Fig.1A).

To integrate additional sequence-derived features into the model's input tensor, a transformation technique was applied to align their numerical range with that of the DNABERT embeddings (BE), enabling effective incorporation into the model's input space. Each feature value was first scaled using an arctangent-based nonlinear transformation, compressing the range and rescaling values to fall within the (0, 1) interval.

$$\text{Scaled value} = \frac{arctan(values)+\frac{\pi}{4}}{\frac{\pi}{2}}$$

This was followed by normalization, where the mean and standard deviation of the scaled values were calculated to standardize the features. This standardization centers the values around zero and adjusts for

their variation, producing a distribution more suitable for alignment with the DNABERT embedding range.

$$\text{Standardized features (SF)} = \frac{scaled\_value - mean}{std}$$

To further ensure consistency, the minimum and maximum values of the DNABERT embeddings were identified. The standardized features were then linearly mapped to this range (min (DE), max (DE)) to ensure scale compatibility. The processed features were then used to generate a feature embedding matrix (FE) (Fig. 1B).

$$rescaled_{features} = \left(\frac{SF - min(SF)}{max(SF) - min(SF)}\right) \times (max(BE) - \min(BE)) + min(BE)$$

For downstream analysis, the rescaled features were concatenated with the DNABERT embeddings (BE), preserving the model's original input structure. The combined BE and FE matrices (Fig. 1C) were used in a classification task, with the model trained across a grid of learning rates (from 1e-3 to 6e-6), warm-up steps (0, 1, 0.1, 0.01), and weight decay values (0.1 or 0.01), using a constant dropout rate to identify a highly accurate model with optimum parameters, resulting in evaluation metrics with accuracy, precision, recall, MCC, and F1 scores.

II. Additional Features incorporated in the Pre-training model

The model tokenizes input sequences into k-mers and computes sequence properties and shape features for each k-mer. These features are then merged with the numerical vectors derived from token embeddings. Each token is represented as a combined feature vector, resulting in a matrix **(**(I + Shape Feature + Sequence Property) $_{\text{k-mer}}$**)** that encodes the full sequence. The contextual information is captured by performing a multi-head self-attention mechanism. During pre-training, the model takes an input sequence with a maximum length of 512, which is tokenized as k-mers with a special token (CLS) at the beginning and (SEP) at the end, like the DNABERT pre-training model. The model, which learns DNA's basic syntax and semantics through self-supervision, randomly masks 15% of the sequence for the model to predict from the context of the unmasked regions with additional embeddings when the masked region's incorporated additional embeddings were hidden and tried to predict just the masked tokens. The architecture consists of 12 transformer layers with 768 hidden units and 12 attention heads in each layer (supplementary Fig. 2).

Results:

In this study, we conduct a comparative analysis between the task-specific fine-tuned model DNABERT-CoreProm-MM and the standard DNABERT on three distinct promoter datasets: those derived from a pattern-based approach (TATA box motif), a PWM-based approach, and the benchmarking EPD-new extracted datasets. This comparison focuses specifically on three classification tasks: (1) TATA promoters from TATA non-promoter sequences, (2) non-TATA promoters from non-TATA non-promoter sequences, and (3) All Promoters from all non-promoter sequences.

The benchmarking EPD-new dataset is compiled from computational analyses, experimental data, and literature reviews, providing limited promoter sequences. The pattern-based approach, grounded in canonical TATA box motif variations, identifies additional promoter sequences that exhibit the TATA pattern. To further enhance coverage, we employed a PWM-based approach in combination with GENCODE annotations. This integration enabled the identification and inclusion of a substantially larger set of TATA and non-TATA promoter sequences, along with their corresponding non-promoter counterparts. Notably, the non-TATA sequences—those lacking a TATA box around the transcription start site (TSS) region—demonstrated slightly reduced classification accuracy compared to their TATA-containing counterparts. Nevertheless, the dataset was significantly enriched, providing a broader and more

diverse representation of promoter types. All datasets were split into an 80:20 ratio for the train and evaluation split of the DNABERT model with features. When fine-tuned on this data, the DNABERT model with features exhibited improved performance, particularly for promoters with weak or non-canonical motif signatures.

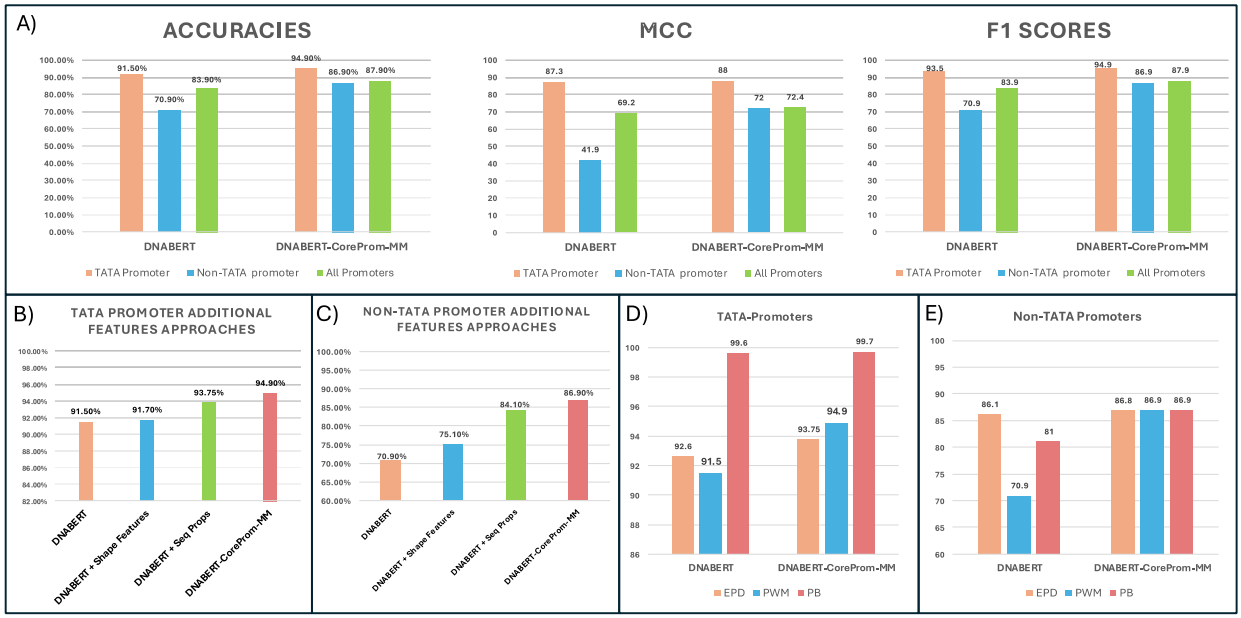

Fig.2: DNABERT-CoreProm-MM significantly outperforms DNABERT in identifying promoter regions that lack specific motifs. (A) (Left to Right) accuracy, MCC, and F1 scores of PWM-based dataset prediction in TATA promoter, non-TATA promoter, and all promoter datasets. **(B)**Accuracy of PWM-based TATA-promoters among DNABERT, DNABERT with 10 features (nucleotide sequence properties (seq props)), DNABERT with 14 shape features, and DNABERT-CoreProm-MM with 24 features (10 sequence properties and 14 shape features). **(C)**Accuracy of PWM-based non-TATA-promoters among DNABERT, DNABERT with 10 nucleotide sequence properties (seq props), DNABERT with 14 shape features, and DNABERT-CoreProm-MM with 24 features (10 sequence properties and 14 shape features). **(D)**Accuracy of DNABERT and DNABERT-CoreProm-MM within TATA-promoters among benchmarking EPD, Pattern-Based (PB), and PWM-based approaches. **(E)**Accuracy of DNABERT and DNABERT-CoreProm-MM within non-TATA-promoters among benchmarking EPD-new, Pattern-Based (PB), and PWM-based approaches.

The TATA-EPD-new dataset (Fig. 2D), comprising 4,862 training sequences and 1,216 testing sequences, achieves a notable accuracy of 92.6% using DNABERT. This dataset is derived from experimentally validated high-confidence promoter sequences. While its elevated accuracy of 93.75% with DNABERT-CoreProm-MM reflects data reliability, it is important to note that this stems from a relatively smaller sample size in comparison to other methods. The TATA-Pattern-Based Approach (Fig. 2D) identifies 13,580 sequences and yields 3,396 validated matches with a higher accuracy rate of 99.6% with DNABERT and 99.7% with DNABERT-CoreProm-MM. This enhanced precision can be linked to rigorous matching criteria applied to established promoter patterns, effectively reducing false positives but potentially overlooking non-canonical promoters. The TATA-PWM dataset (Fig.2A, Fig. 2B), which includes 16,128 training and 4,032 testing sequences, achieves an overall accuracy rate of 91.5% with DNABERT, 91.7% with DNABERT including shape features, 93.75% with DNABERT including Sequence properties, and 94.9% with DNABERT-CoreProm-MM. These results highlight the advantage of flexible sequence matching, accommodating variability within core-promoter regions and enhancing detection potential compared to more rigid approaches. When testing the TATA model with the Segment-NT and Gena-LM methods, the accuracies obtained were 68.2% and 69.6%**,** respectively. As our focus is on core-promoter

regions**,** these methods do not perform well for shorter sequence sizes. Although DeePromoter(20) is highly specific to the core promoter region and is designed for shorter sequences, its performance is outperformed by DNABERT in comparative studies such as *(21).* Since DNABERT demonstrated superior accuracy on promoter prediction tasks, we focused on extending DNABERT by incorporating additional sequence properties and DNA shape features within this narrow core promoter region.

The Non-TATA-EPD-new dataset (Fig. 2E) comprises a balanced training set of 27,537 sequences and a testing set of 6,885 sequences. This collection is sourced from experimentally validated promoters that exhibit high confidence levels, and its relatively smaller size compared to other datasets highlights its targeted and highly specific nature. Nevertheless, the models demonstrate impressive accuracy rates of 86.1%, reflecting both the quality and dependability of this data. This EPD-new non-TATA data contains either an Initiative motif (YYANWYY; (Y: C or T) ;(W- A or T)), GC box, or CCAAT pattern around the TSS, which can be the reason for not learning much from the sequence properties and just showing a little improvement of 86.8% with DNABERT-CoreProm-MM. The Non-TATA Pattern-Based dataset (Fig. 2E) is curated to include sequences lacking a TATA box around the transcription start site (TSS) region. It consists of a training set of 45,106 sequences and a testing set of 7,960 sequences. When evaluated using the standard DNABERT model, it achieves an accuracy of 81%, and 86.9% with DNABERT-CoreProm-MM. The non-TATA PWM-based dataset (Fig. 2A, 2C) comprises the largest training set (121,862 sequences) and testing set (30,466 sequences) and includes only sequences without a TATA motif around the TSS. This absence of a canonical motif leads to a lower accuracy of 70.9% when using the standard DNABERT model. Notably, performance improves with the inclusion of additional features: accuracy rises to 75.1% with shape features, 84.1% with sequence properties, and reaches 86.9% with DNABERT-CoreProm-MM when both sequence properties and shape features are incorporated during fine-tuning. When testing the non-TATA model with the Segment-NT and Gena-LM methods, the accuracies obtained were 70% and 70.1%, respectively.

TATA models leverage motif-based language logic and consistently achieve high accuracy even with the standard DNABERT architecture. Interestingly, the specialized fine-tuned model, DNABERT-CoreProm-MM, designed to enhance learning from sequences lacking well-defined motifs, does not offer substantial improvement over the standard model when applied to TATA sequences. While the incorporation of these features showed minimal effect during the pretraining stage—with fine-tuning on the feature-enhanced pretrained model resulting in a 2.8% drop in accuracy, their inclusion during fine-tuning led to substantial performance gains, particularly for non-TATA promoters, which are often compositionally diffuse and poorly captured by conventional motif-centric methods. In contrast, for non-TATA sequences where clear motif patterns are absent, the standard DNABERT model shows a modest drop in performance compared to its TATA-focused counterpart.

Conclusions:

We enhanced the DNABERT framework by integrating DNA sequence composition and shape-based features to improve the classification of regulatory sequences lacking distinct motif signals. While these features showed minimal effect during pretraining, their incorporation during fine-tuning led to substantial performance gains—particularly for non-TATA promoters, which are often compositionally diffuse and poorly captured by conventional motif-centric methods. Our approach improves both computational and sample efficiency by prioritizing biologically informative properties that extend beyond sequence motifs. This makes the model more robust in identifying regulatory signals embedded in noisy or low-pattern genomic regions. In future work, we aim to extend this framework to other challenging regulatory elements such as enhancers, boundary elements, transcription factor binding sites, and histone modification regions—many of which exhibit weak or variable sequence motifs, high cell-type specificity, or are defined by epigenetic states rather than sequence alone. By doing so, we hope to advance the application of transformer-based models for decoding gene regulation across a wider spectrum of non-coding genomic contexts.

Availability:

Source code and data are available at GitHub (https://github.com/NPAPINENI/Promoter-Model.git)

Acknowledgements:

We thank all members of the Davuluri lab (The State University of New York at Stony Brook)
This work was financially supported by grants from the National Library of Medicine/National Institutes of Health funding – [R01LM01372201 to R.D., R35GM128938 to F.A]

References:

1. Ji Y, Zhou Z, Liu H, Davuluri RV. DNABERT: pre-trained Bidirectional Encoder Representations from Transformers model for DNA-language in genome. Bioinformatics. 2021;37(15):2112-20.
2. Fishman V, Kuratov Y, Shmelev A, Petrov M, Penzar D, Shepelin D, et al. GENA-LM: a family of open-source foundational DNA language models for long sequences. Nucleic Acids Research. 2025;53(2).
3. Dalla-Torre H, Gonzalez L, Mendoza-Revilla J, Lopez Carranza N, Grzywaczewski AH, Oteri F, et al. Nucleotide Transformer: building and evaluating robust foundation models for human genomics. Nature Methods. 2025;22(2):287-97.
4. Chen N, Yu J, Liu Z, Meng L, Li X, Wong K-C. Discovering DNA shape motifs with multiple DNA shape features: generalization, methods, and validation. Nucleic Acids Research. 2024;52(8):4137-50.
5. Gupta R, Wikramasinghe P, Bhattacharyya A, Perez FA, Pal S, Davuluri RV. Annotation of gene promoters by integrative data-mining of ChIP-seq Pol-II enrichment data. BMC bioinformatics. 2010;11:1-9.
6. Bruhm DC, Mathios D, Foda ZH, Annapragada AV, Medina JE, Adleff V, et al. Single-molecule genome-wide mutation profiles of cell-free DNA for non-invasive detection of cancer. Nat Genet. 2023;55(8):1301-10.
7. Dietlein F, Weghorn D, Taylor-Weiner A, Richters A, Reardon B, Liu D, et al. Identification of cancer driver genes based on nucleotide context. Nat Genet. 2020;52(2):208-18.
8. Zhou X, Cheng Z, Dong M, Liu Q, Yang W, Liu M, et al. Publisher Correction: Tumor fractions deciphered from circulating cell-free DNA methylation for cancer early diagnosis. Nat Commun. 2023;14(1):328.
9. Abe N, Dror I, Yang L, Slattery M, Zhou T, Bussemaker HJ, et al. Deconvolving the recognition of DNA shape from sequence. Cell. 2015;161(2):307-18.
10. Mudge Jonathan M, Carbonell-Sala S, Diekhans M, Martinez Jose G, Hunt T, Jungreis I, et al. GENCODE 2025: reference gene annotation for human and mouse. Nucleic Acids Research. 2024;53(D1):D966-D75.
11. Sheu C-YL, Huang Y-C, Lin P-Y, Lin G-J, Chen P-Y. Bioinformatics of epigenetic data generated from next-generation sequencing. Epigenetics in human disease: Elsevier; 2024. p. 37-82.
12. Education SbN. TATA box. 2014.

13. Drost H-G, Paszkowski J. Biomartr: genomic data retrieval with R. Bioinformatics. 2017;33(8):1216-7.
14. Rauluseviciute I, Riudavets-Puig R, Blanc-Mathieu R, Castro-Mondragon Jaime A, Ferenc K, Kumar V, et al. JASPAR 2024: 20th anniversary of the open-access database of transcription factor binding profiles. Nucleic Acids Research. 2023;52(D1):D174-D82.
15. Meylan P, Dreos R, Ambrosini G, Groux R, Bucher P. EPD in 2020: enhanced data visualization and extension to ncRNA promoters. Nucleic Acids Res. 2020;48(D1):D65-D9.
16. Dreos R, Ambrosini G, Perier RC, Bucher P. The Eukaryotic Promoter Database: expansion of EPDnew and new promoter analysis tools. Nucleic Acids Res. 2015;43(Database issue):D92-6.
17. Morioka MS, Kawaji H, Nishiyori-Sueki H, Murata M, Kojima-Ishiyama M, Carninci P, et al. Cap Analysis of Gene Expression (CAGE): A Quantitative and Genome-Wide Assay of Transcription Start Sites. Methods Mol Biol. 2020;2120:277-301.
18. Zhou T, Shen N, Yang L, Abe N, Horton J, Mann RS, et al. Quantitative modeling of transcription factor binding specificities using DNA shape. Proceedings of the National Academy of Sciences. 2015;112(15):4654-9.
19. Chiu T-P, Comoglio F, Zhou T, Yang L, Paro R, Rohs R. DNAshapeR: an R/Bioconductor package for DNA shape prediction and feature encoding. Bioinformatics. 2015;32(8):1211-3.
20. Oubounyt M, Louadi Z, Tayara H, Chong KT. DeePromoter: robust promoter predictor using deep learning. Frontiers in genetics. 2019;10:286.
21. Zeng R, Li Z, Li J, Zhang Q. DNA promoter task-oriented dictionary mining and prediction model based on natural language technology. Scientific Reports. 2025;15(1):153.