TAFP-DNA: TEMPERATURE-ANNEALED FOCUSED POOLING FOR DNA FOUNDATION MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Fixed pooling strategies in current DNA foundation models poorly adapt to diverse Transcription Factor (TF) binding patterns and can make predictions vulnerable to sequence-level statistical biases such as Guanine-Cytosine (GC) content. To tackle these issues, we propose Temperature-Annealed Focused Pooling (TAFP), an attention pooling head with a temperature-scaled softmax. Inspired by simulated annealing, TAFP gradually lowers the temperature during fine-tuning to sharpen the attention distribution, moving from broad global aggregation to a focus on informative local regions; at inference, the temperature is kept fixed. When a known motif is available, an optional weak supervision loss aligns attention with motif sites. By exploiting the Reverse-Complement (RC) symmetry of DNA, RC training augmentation combined with averaging of forward- and RC-strand predictions yields RC-invariant predictions without an extra RC consistency loss. On ten TF binding prediction tasks from the Genome Understanding Evaluation (GUE) benchmark, the resulting TAFP-DNA achieves a state-of-the-art average Matthews Correlation Coefficient (MCC) of 0.7381 over seven baselines, and under the same backbone, focused pooling outperforms CLS and mean pooling by 0.044 and 0.034 MCC. With motif supervision, the motif recovered from attention weights matches the known MAF profile. TAFP is a lightweight module that can be added to pretrained DNA models without architectural changes.

1 INTRODUCTION

DNA foundation models (Ji et al., 2021; Zhou et al., 2024; Dalla-Torre et al., 2025) learn statistical patterns from large collections of genomic sequences and are now routinely fine-tuned for regulatory tasks such as Transcription Factor (TF) binding prediction. For these tasks, aggregate metrics alone are insufficient. Biologists also need to know which sequence segments drive a prediction, because binding motifs (Alipanahi et al., 2015), their combinations (Avsec et al., 2021b), and the positions of key regulatory segments (Avsec et al., 2021a) are what connect a model's output to gene regulation and to the interpretation of noncoding variants (Zhou & Troyanskaya, 2015).

Identifying these segments depends on how token-level representations are aggregated into a single sequence-level representation. Fine-tuned DNA foundation models commonly use a Classification (CLS) token, mean pooling, or max pooling, which apply the same aggregation rule regardless of which positions matter. Attention pooling instead learns a weight for each token and is a natural choice for TF binding, where a short motif within a longer sequence often determines the label. Using attention pooling for TF binding prediction, however, raises three challenges.

First, the concentration of attention must be controlled. Flat attention approaches mean pooling and dilutes the motif signal, whereas enforcing sharp attention from the start of fine-tuning forces the model to commit to a few tokens before its representations are informative, which can make optimization unstable. On mouse_3, for example, training at a fixed low temperature from the start drops the MCC below that of mean pooling (Section 4.4). Second, even well-concentrated attention is not guaranteed to fall on binding motifs. Sequence-level statistical signals such as Guanine-Cytosine (GC) content and repeats correlate with binding labels (Tourne et al., 2026), and attention can follow such shortcuts instead, so attention weights alone do not provide reliable motif-level interpretation. Third, a TF can bind either DNA strand, yet a standard fine-tuned model reads each

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

sequence in a single orientation and may give different predictions for a sequence and its Reverse Complement (RC) (Zhou et al., 2022).

We address these challenges with three components (Figure 1). For the first challenge, we propose temperature-annealed focused pooling: a linear layer scores each token, a temperature-scaled softmax converts the scores into pooling weights, and the temperature is annealed from 1.0 to 0.01 during fine-tuning, so that attention moves gradually from a broad aggregation to a focus on a few key tokens. For the second, when a known motif is available for the target TF, an optional weak supervision loss aligns the attention weights with motif sites in positive sequences. For the third, we train on both strands and average the forward- and RC-strand predictions at inference, which makes the predictions RC-invariant by construction. We refer to the resulting model as TAFP-DNA (Temperature-Annealed Focused Pooling for DNA).

We evaluate TAFP-DNA on the 10 TF binding prediction tasks of the Genome Understanding Evaluation (GUE) benchmark (Zhou et al., 2024) with a Nucleotide Transformer v2-500M backbone, where it reaches an average Matthews Correlation Coefficient (MCC) of 0.7381, the highest among the compared methods (Table 1). In ablations that replace only the pooling head under an identical training protocol, focused pooling improves the average MCC over CLS and mean pooling by 0.044 and 0.034, and annealing improves it by 0.018 over training at a fixed low temperature from the start (cold start). On mouse_3, annealed attention trained without motif supervision collapses onto a shortcut at the sequence end, whereas with motif supervision the motif recovered from the attention weights matches the Musculoaponeurotic Fibrosarcoma (MAF) profile in JASPAR. RC training augmentation and test-time averaging together improve the average MCC by 0.045. Our contributions are as follows:

1. **Temperature-annealed focused pooling.** To control attention concentration, we propose a lightweight pooling head that replaces CLS or mean pooling without architectural changes to the pretrained backbone. It anneals the attention temperature during fine-tuning, and its token-level weights are the starting point for motif analysis.
2. **Motif weakly supervised attention alignment.** To align attention with binding sites, we propose an optional weak supervision strategy based on known TF motif sites, and we use the resulting attention to recover sequence motifs with standard motif discovery tools.
3. **Reverse-complement dual-strand modeling.** To account for strand symmetry, we combine RC training augmentation with averaging of forward- and RC-strand predictions, which yields RC-invariant predictions without an additional RC consistency loss.

2 Related Work

DNA foundation models. DNABERT (Ji et al., 2021), DNABERT-2 (Zhou et al., 2024), and the Nucleotide Transformer (Dalla-Torre et al., 2025) are pretrained on large genomic corpora and fine-tuned for regulatory tasks, with GUE (Zhou et al., 2024) as a common benchmark. Later models vary the pretraining objective (Ding et al., 2026), representation geometry (Khan et al., 2025; Zeng et al., 2026; Shi et al., 2026), coding granularity (Mao et al., 2025), or architecture (Yang et al., 2025). While these methods strengthen pretraining and backbones, downstream fine-tuning still relies on static aggregation. Because such pooling is agnostic to TF-specific binding positions, a stronger backbone alone cannot ensure focus on relevant motifs.

Pooling and interpretability. Sequence-level predictions are usually obtained from a CLS token (Devlin et al., 2019), mean or max pooling, or attention pooling (Ilse et al., 2018). Interpretable TF binding models aim to identify motifs and their syntax (Alipanahi et al., 2015; Avsec et al., 2021b), while sequence statistics such as GC content can act as shortcuts (Tourne et al., 2026). Attention pooling can reweight positions, but it leaves the sharpness of its weights uncontrolled. Attention weights are also not guaranteed to align with binding motifs and may instead follow such shortcuts, which limits their use for motif-level interpretation.

Reverse-complement symmetry. RC data augmentation, post-hoc averaging over both strands, and RC-equivariant architectures have been compared by Zhou et al. (2022) and built into DNA models such as Caduceus (Schiff et al., 2024). RC augmentation alone does not guarantee strand-invariant predictions, post-hoc averaging acts only at inference, and RC-equivariant architectures

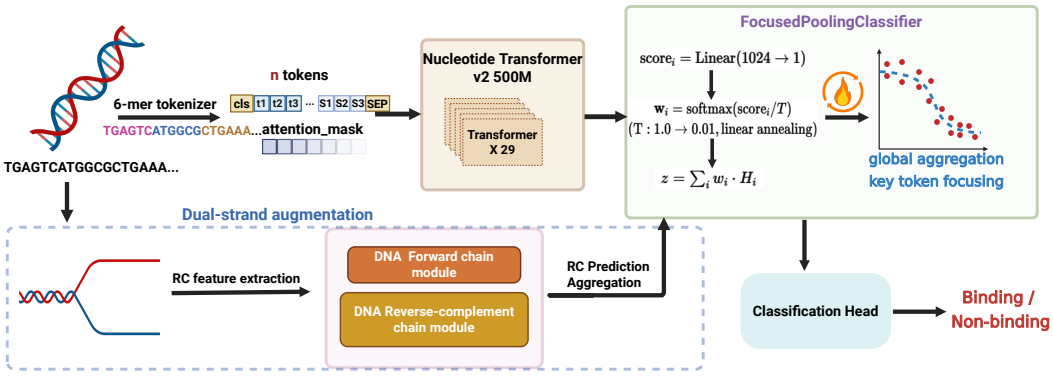

Figure 1: Overall architecture of the TAFP-DNA model.

require building the symmetry into the backbone, which does not directly apply to existing pretrained transformer models such as the Nucleotide Transformer.

3 METHOD

The overall framework is shown in Figure 1. The input DNA sequence is tokenized with a 6-mer tokenizer to obtain token embeddings, which are fed into the pretrained Nucleotide Transformer v2 (500M) backbone. The focused pooling head computes an importance score for each token and applies linear temperature annealing during training to generate a sequence representation through weighted aggregation. We further introduce dual-strand augmentation and RC prediction aggregation: during inference, predictions for the forward and reverse-complement strands are averaged to obtain a reverse-complement-invariant output.

3.1 BACKBONE: NUCLEOTIDE TRANSFORMER V2 500M

We use Nucleotide Transformer v2-500M (Dalla-Torre et al., 2025) as the DNA sequence encoding backbone. The model is based on the Evolutionary Scale Modeling (ESM) architecture (Rives et al., 2021) and uses 6-mer tokenization and Rotary Positional Embeddings (RoPE) (Su et al., 2024), with a hidden dimension of 1024. During training, the backbone, pooling head, and classification head are jointly optimized end to end. The learning rate is set to 3×10^{-5}, with linear decay and 50 learning-rate warmup steps.

3.2 TEMPERATURE-ANNEALED FOCUSED POOLING HEAD

Our design borrows the idea of annealing in metallurgy (Kirkpatrick et al., 1983). At high temperature, atoms move freely and can explore the whole energy landscape; if the temperature is lowered slowly, the system settles into the lowest-energy lattice. Rapid quenching, which corresponds to a fixed low temperature, instead freezes atoms in disordered, high-energy states. We map this analogy onto pooling as follows: the token scores play the role of atoms, T is the annealing temperature, and the classification loss defines the energy landscape. As T decreases, the pooling weights move gradually from a uniform distribution onto a few key tokens, which avoids the unstable optimization of a fixed low temperature. The resulting weights can be inspected at the token level without lowering binding prediction performance; with the motif supervision in Section 3.3, they can further be aligned with the motif segment that triggers TF binding.

The focused pooling head is lightweight and can replace CLS or mean pooling directly, with no change to the backbone architecture, tokenizer, or positional encoding. It maintains or improves predictive performance, and together with motif supervision, its attention weights can be used for motif localization.

3.2.1 TOKEN SCORING MAPPING

Let an input DNA sequence be encoded by the tokenizer into a token sequence of length L. The pretrained backbone produces a sequence hidden-state matrix:

$$\mathbf{H} = [\mathbf{h}_1, \mathbf{h}_2, \ldots, \mathbf{h}_L] \in \mathbb{R}^{L \times d}, \tag{1}$$

where d is the model's hidden dimension. Define $\mathbf{m} \in \{0,1\}^L$ as the attention mask, where $m_i = 1$ indicates a valid nucleotide token and $m_i = 0$ indicates an invalid padding position.

A learnable linear scoring layer predicts a scalar importance score for each token hidden state:

$$s_i = w_s^\top h_i + b_s, \qquad i = 1, \ldots, L, \tag{2}$$

where $w_s \in \mathbb{R}^d$ and $b_s \in \mathbb{R}$, corresponding to a $\mathrm{Linear}(d, 1)$ layer.

3.2.2 TEMPERATURE-SCALED ADAPTIVE ATTENTION WEIGHTS

We first mask the scores at padding positions, then scale the score distribution using the temperature coefficient T and apply softmax normalization to obtain token attention weights:

$$\widetilde{s}_i = \begin{cases} s_i, & m_i = 1, \\ -\infty, & m_i = 0, \end{cases} \qquad w_i = \frac{\exp(\widetilde{s}_i / T)}{\sum_{j=1}^{L} \exp(\widetilde{s}_j / T)}. \tag{3}$$

3.2.3 TEMPERATURE-ANNEALED FOCUSED POOLING

The normalized attention weights are used to aggregate all token hidden states into a global sequence-level representation:

$$\mathbf{z} = \sum_{i=1}^{L} w_i \mathbf{h}_i \in \mathbb{R}^d. \tag{4}$$

To enable dynamic focusing during training, we adopt a linear temperature annealing schedule that gradually concentrates the attention distribution:

$$T(t) = T_{\text{init}} - (T_{\text{init}} - T_{\min}) \cdot \frac{t}{t_{\text{total}}}, \tag{5}$$

where $T_{\text{init}} = 1.0$, $T_{\min} = 0.01$, t is the current training iteration, and t_{total} is the total number of training steps.

3.3 EXPLICIT WEAKLY SUPERVISED MOTIF ALIGNMENT

We add a cross-entropy weak supervision loss on motif positions. It constrains the attention weights with biological priors, improves motif localization, and supports motif recovery from attention weights. The supervision signal comes from experimentally validated, conserved TF motifs, which guide the focused attention toward known functional motif sites and strengthen biological interpretability. We enable motif supervision only in the interpretability analysis (Section 4.5); all other results, including Table 1, are obtained without it.

Specifically, for each positive DNA sequence, the International Union of Pure and Applied Chemistry (IUPAC) degenerate motif sequence of the corresponding TF (e.g., TGASTCA for Activator Protein 1 (AP-1)) is converted into a regular-expression matching rule. We scan both the forward and reverse-complement strands to locate all motif hits in the sequence. Using the tokenizer's offset mapping, nucleotide-level motif coordinates are converted into the corresponding set of valid tokens $\mathcal{M}_n$. Assigning equal weight to these tokens gives the target attention distribution $t_{n,i}$:

$$t_{n,i} = \begin{cases} \dfrac{1}{|\mathcal{M}_n|}, & i \in \mathcal{M}_n, \\ 0, & \text{otherwise.} \end{cases} \tag{6}$$

The loss is computed only on positive samples that contain at least one motif hit. Negative samples and positive samples without a hit are skipped, because they provide no motif target and would only

add noise. The loss is

$$\mathcal{L}_{\text{sup}} = \frac{\lambda_{\text{sup}}}{|\mathcal{S}_{\text{hit}}|} \sum_{n \in \mathcal{S}_{\text{hit}}} \left(-\sum_{i=1}^{L} t_{n,i} \log(w_{n,i} + \varepsilon) \right), \tag{7}$$

where $\mathcal{S}_{\text{hit}} = \{n : |\mathcal{M}_n| > 0\}$ is the set of samples containing valid motif sites, $\varepsilon = 10^{-9}$ is a numerical smoothing constant that prevents singularities in the logarithm, and λ_{sup} is the supervision loss weight.

The loss can also be read in information-theoretic terms. Let $H(t_n)$ denote the fixed entropy of the target distribution. The cross-entropy decomposes into an entropy term and a Kullback–Leibler (KL) divergence term (Kullback & Leibler, 1951):

$$-\sum_{i} t_{n,i} \log w_{n,i} = H(t_n) + D_{\text{KL}}(t_n \parallel w_n). \tag{8}$$

Because $H(t_n)$ is constant, minimizing $\mathcal{L}_{\text{sup}}$ amounts to minimizing (up to the smoothing constant ε) the KL divergence between the motif distribution t_n and the attention distribution w_n, which pulls the attention weights toward the motif positions.

3.4 Explicit Dual-Strand Modeling with Reverse Complements

DNA is double-stranded, and we use this strand symmetry during both training and inference to improve prediction robustness and classification accuracy (Zhou et al., 2022; Schiff et al., 2024). The RC module is independent of focused pooling and motif weak supervision and can be combined with either. The full system, which combines focused pooling, RC dual-strand modeling, inference-time Test-Time Augmentation (TTA), and multi-seed ensembling, improves overall performance and reaches the best average MCC of 0.7381.

A TF binding site can be read from either strand: a motif on the forward strand appears in reverse-complement form on the opposite strand, and both forms describe the same binding event. We therefore use both strands in training and at inference.

3.4.1 RC Data Augmentation during Training

For each input sequence, we additionally construct its reverse complement. Both strands carry the same label and are used for training, and the loss is the sum of their cross-entropy losses:

$$\mathcal{L}_{\text{train}}(\mathbf{x}, y) = \text{CE}(f(\mathbf{x}), y) + \text{CE}(f(\text{rc}(\mathbf{x})), y). \tag{9}$$

where CE denotes the Cross-Entropy (CE) loss and $\text{rc}(\cdot)$ the reverse-complement operation.

3.4.2 RC-TTA during Inference

At test time, we average the logits of the forward and reverse-complement inputs, so that the prediction does not depend on which strand is given:

$$\mathbf{z}_{\text{TTA}}(\mathbf{x}) = \frac{f(\mathbf{x}) + f(\text{rc}(\mathbf{x}))}{2}. \tag{10}$$

We use both RC augmentation during training and RC averaging at inference in the final model.

4 Experiments

Datasets

We evaluate our method on the TF binding site prediction tasks of the GUE benchmark (Zhou et al., 2024). These tasks comprise 10 TF binding datasets (five human and five mouse datasets). Each sample is a DNA sequence of 101 base pairs (bp) labeled according to whether it is a Chromatin ImmunoPrecipitation sequencing (ChIP-seq) binding peak of the corresponding TF, forming a binary classification task. Each dataset is divided into training, validation, and test sets, with training set sizes ranging from 1,904 (mouse_3) to 53,952 (mouse_1), as detailed in Appendix Table A1.

Table 1: MCC scores on the 10 TF binding classification tasks from the GUE benchmark.

Dataset	DNABERT-2	NucEL	HGE	HyCoSeq	ILNN	DNAZEN	SPACE	TAFP-DNA
human_tf_0	0.6673	0.6565	0.6233	0.6373	0.0380	0.6509	0.6357	**0.6967**
human_tf_1	0.7044	0.6908	0.6809	0.6754	0.0400	0.7314	0.7129	**0.7484**
human_tf_2	0.6161	0.6521	0.6738	0.5315	0.0762	0.6481	**0.7120**	0.6761
human_tf_3	0.5867	0.5039	0.4491	0.4501	-0.0381	0.5574	**0.6484**	0.6047
human_tf_4	0.7409	0.7452	0.7181	0.7690	0.0681	0.7849	**0.8139**	0.8088
mouse_0	0.5231	0.5015	0.1240	0.4047	0.5270	0.6242	0.6684	**0.7024**
mouse_1	0.8294	0.8424	0.7552	0.7383	0.8006	0.8525	0.8570	**0.8801**
mouse_2	0.8295	0.8002	0.7295	0.6177	0.7563	0.8373	**0.9099**	0.8477
mouse_3	0.7252	0.7656	0.7006	0.0721	0.7506	0.7332	**0.8873**	0.8666
mouse_4	0.3541	0.4847	0.3914	0.3871	0.4286	0.5085	0.5133	**0.5497**
Avg.	0.6577	0.6643	0.5846	0.5283	0.3447	0.6928	0.7359	**0.7381**
Std. ↓	0.1452	0.1301	0.2022	0.2090	0.3439	0.1149	0.1277	**0.1125**

We compare TAFP-DNA with seven existing state-of-the-art baseline methods: DNABERT-2 (Zhou et al., 2024), NucEL (Ding et al., 2026), HGE (Hyperbolic Genome Embeddings) (Khan et al., 2025), HyCoSeq (Zeng et al., 2026), ILNN (Intrinsic Lorentz Neural Network) (Shi et al., 2026), DNAZEN (Mao et al., 2025), and SPACE (Species-Profile Adaptive Collaborative Experts) (Yang et al., 2025). They cover alternative tokenization schemes and pretraining objectives, hyperbolic representation spaces, multigranularity encoders, and mixture-of-experts architectures. Detailed descriptions are provided in Appendix B.

Experimental Setup

We independently train models with three fixed random seeds (42/43/44). At test time, we ensemble the models by averaging their softmax output probabilities to reduce variation caused by training randomness. All ablation variants (CLS, mean, multi, focused pooling, etc.) reuse the Nucleotide Transformer v2-500M backbone and exactly the same training protocol: AdamW optimizer (Loshchilov & Hutter, 2019), a learning rate of 3×10^{-5}, three epochs, 16-bit floating-point (fp16) mixed precision, and a batch size of 64. Only the pooling head is replaced, so differences in performance can be attributed to the pooling module. Following the DNABERT-2 evaluation protocol, we use MCC (Chicco & Jurman, 2020) as the evaluation metric.

4.1 Comparison with Baseline Methods

As shown in Table 1, TAFP-DNA achieves an average MCC of 0.7381 across the 10 TF binding tasks in the GUE benchmark, the highest among the seven baselines (the runner-up, SPACE, reaches 0.7359), and it has the lowest cross-task standard deviation (0.1125). TAFP-DNA ranks in the top two on all 10 tasks and is the only method without a weak task. Compared with SPACE, its advantage comes mainly from tasks on which SPACE is relatively weak, namely mouse_0, mouse_1, and mouse_4 (an average improvement of +0.031 MCC), which we attribute to focused pooling and dual-strand augmentation.

Relative to the widely used DNABERT-2, TAFP-DNA improves the average MCC by 0.080, with the largest gains on mouse_0, mouse_3, and mouse_4 (+0.179, +0.141, and +0.196). These three tasks account for about two thirds of the total improvement.

4.2 Comparison of Pooling Methods

Figure 2 compares pooling methods. Over the 10 TF tasks, focused pooling reaches an average MCC of 0.7381, which is 0.044 higher than CLS pooling (0.6938) and 0.034 higher than mean pooling (0.7044). Multi pooling concatenates four signals (CLS, mean, softmax, and sigmoid) and has four times the feature dimension, yet it reaches 0.7361, still below the single focused head. Since the larger multi head does not do better, we attribute the gain to the focusing mechanism itself, which

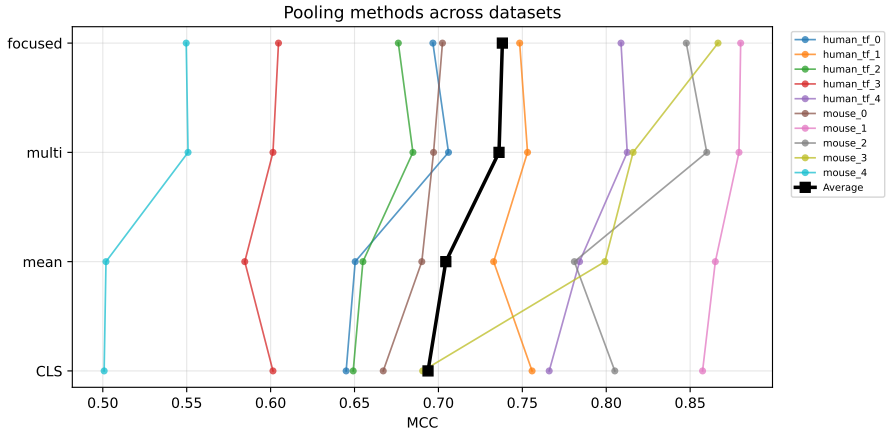

Figure 2: Comparison of MCC performance across pooling strategies in the GUE benchmark.

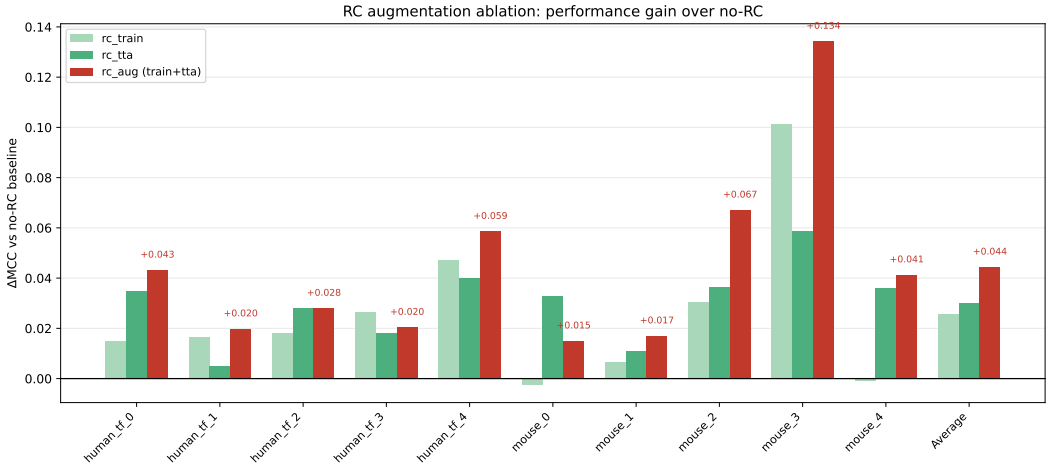

Figure 3: Ablation of the reverse-complement (RC) module: MCC gains relative to the no-RC.

concentrates attention on a few key tokens through the temperature-scaled softmax. Mean pooling beats CLS on average but is worse on human_tf_1, human_tf_3, and mouse_2.

4.3 RC Ablation Experiments

Dual-strand augmentation gives a large improvement. With RC augmentation at both training and test time (rc_aug), the average MCC over the 10 TF tasks is 0.7381, 0.045 higher than the no-RC baseline (0.6936). Training augmentation alone (rc_train) adds 0.026, and test-time augmentation alone (rc_tta) adds 0.030. Test-time averaging alone helps slightly more than training augmentation, at the cost of a second forward pass. The combined gain (0.045) is below the sum of the individual gains (0.056), so the two stages partly overlap, but their combination still outperforms either stage alone on average (though not on every task, e.g., human_tf_3 and mouse_0).

4.4 Temperature Annealing Experiments

Figure 4 compares mean pooling, cold-start training at a fixed temperature ($T = 0.01$), and temperature annealing ($1.0 \rightarrow 0.01$). The average MCC rises from 0.7044 to 0.7204 and then to 0.7381: curriculum-style (Bengio et al., 2009) annealing is 0.034 higher than mean pooling and 0.018 higher than cold start. Narrowing attention gradually is therefore more stable and more effective than fo-

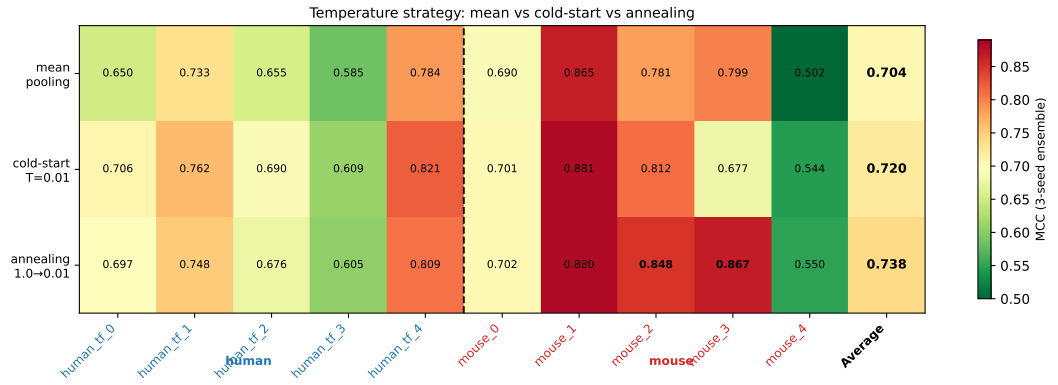

Figure 4: MCC heatmap for three training strategies: mean pooling, cold-start training at a fixed low temperature, and temperature annealing.

cusing at a fixed low temperature from the start, and these results show that annealing is effective in terms of average performance.

Annealing has a clear advantage on the mouse subset, and mouse_3 is the most representative case. Mean pooling reaches 0.7992 there, whereas cold-start training drops sharply to 0.6775 (-0.122). Annealing raises the MCC to 0.8666, 0.189 above cold-start training. We attribute the collapse to unstable optimization: a low-temperature cold start sharpens attention before the token representations are informative, which is especially harmful on this small dataset (1,904 training sequences), whereas annealing starts from a broad distribution and sharpens it gradually. Annealing alone does not guarantee that attention falls on the binding motif, however; without motif supervision, annealed attention on mouse_3 collapses onto a shortcut at the sequence end (Section 4.5). On the human subset, cold start is usually slightly better when the annealing endpoint is 0.01, because attention there tends to collapse onto a single token and noise hides the differences between strategies; with other endpoints, the annealing method achieves comparable performance (Table C1).

4.5 Case Study: Motif Localization and Weakly Supervised Motif Recovery

This analysis is the only experiment in which motif supervision is enabled. In the left panel of Figure 5, annealing with supervision reaches an MCC of 0.803 on mouse_3, 0.140 above cold start (0.663) but below the 0.8666 obtained without supervision in Table 1, so better motif alignment comes at some cost in classification performance. In the right panel, the localization curve of annealing follows the true motif peak more closely. Cold-start training also focuses on the motif, but its MCC is much lower, so its focusing is only weakly tied to its classification decisions. Annealing gets both the focusing and the classification right, and this combination is the main benefit of temperature annealing.

Figure 6 shows the motifs recovered by STREME (Bailey, 2021) from high-attention positions under each setting. Without motif supervision, annealing (second row, AGKGAMA) yields a degenerate motif because its attention collapses onto a shortcut at the sequence end, whereas cold-start training (third row, ACTCAGC) already recovers a motif consistent with the MAF-matching motifs obtained under supervision. With motif supervision, both annealing (fourth row, ASTCAGC) and cold-start training (fifth row, GCTGAST) recover motifs that TOMTOM (Gupta et al., 2007) matches to the MA1520.2 MAF profile of the basic Leucine Zipper (bZIP) family (Rauluseviciute et al., 2024) ($E = 0.0026$ and $E = 0.0059$, respectively). These two consensus sequences are reverse complements of each other and therefore describe the same binding site read from opposite strands. The two supervised settings thus recover essentially the same motif, and the difference between them lies in classification performance (Figure 5).

In summary, motif supervision is indispensable for annealed attention to align with binding sites on mouse_3. Annealing outperforms cold start (MCC +0.140) while both recover the MAF motif.

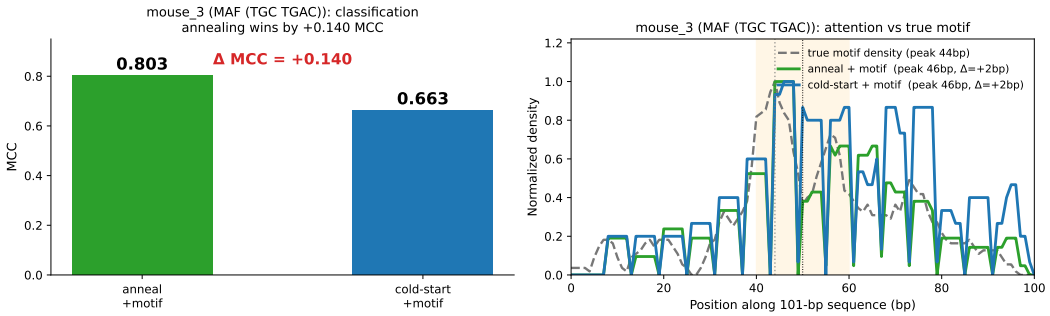

Figure 5: Annealing vs. cold start on mouse_3 (MAF) with motif supervision: annealing reaches an MCC of 0.803 (vs. 0.663 for cold start), while both lock onto the true motif. Left: classification performance. Right: attention distributions compared with the true motif density. Motif supervision is used only in this analysis and not for the results in Table 1.

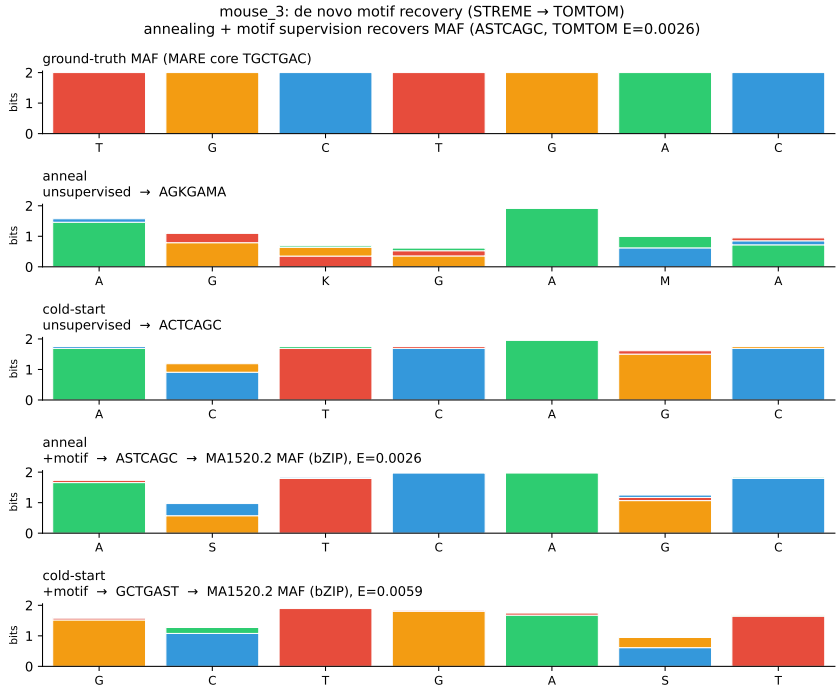

Figure 6: Motifs recovered from attention weights under different training settings.

5 Conclusion

TAFP-DNA addresses three challenges in aggregating DNA foundation model representations for TF binding prediction (attention concentration, motif alignment, and strand symmetry) with temperature-annealed focused pooling, optional motif supervision, and RC dual-strand modeling. On 10 TF binding tasks from the GUE benchmark, focused pooling outperforms CLS, mean, and multifeature concatenation pooling, and TAFP-DNA achieves the highest average MCC and the smallest cross-task variation among seven state-of-the-art baselines. Annealing avoids the performance collapse of low-temperature cold-start training, especially on the mouse datasets, and dual-strand modeling adds a consistent average gain. Aligning attention with binding motifs requires motif supervision: with it, the motif recovered from attention on mouse_3 matches the MAF profile, at some cost in classification performance. Motif supervision currently relies on known motif priors, and we plan to extend focused pooling to other genomic tasks and to longer regulatory sequences.

AI USE STATEMENT

We used generative AI tools only for language polishing, and grammatical correction in this work. All AI-assisted revisions were manually reviewed and verified to ensure technical accuracy and consistency with the original content.

ETHICS STATEMENT

This work uses only the publicly available GUE benchmark (Zhou et al., 2024), whose TF binding datasets consist of short human and mouse genomic sequences labeled from public ChIP-seq experiments. The data contain no individual-level genetic information, and the study involves no human subjects, personal data, or new data collection. The proposed method is intended for basic research in regulatory genomics. Attention weights and recovered motifs are model-derived hypotheses and should be validated experimentally before being used to support biological or clinical conclusions.

REPRODUCIBILITY STATEMENT

All datasets and motif analysis tools used in this work (GUE, STREME, TOMTOM, and JASPAR) are publicly available. Source code for training, evaluation, and motif analysis is available at the anonymous repository: `https://anonymous.4open.science/r/TAFP-DNA-141E`.

REFERENCES

Babak Alipanahi, Andrew Delong, Matthew T Weirauch, and Brendan J Frey. Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nature Biotechnology*, 33(8):831–838, 2015. doi: 10.1038/nbt.3300.

Žiga Avsec, Vikram Agarwal, Daniel Visentin, Joseph R. Ledsam, Agnieszka Grabska-Barwinska, Kyle R. Taylor, Yannis Assael, John Jumper, Pushmeet Kohli, and David R. Kelley. Effective gene expression prediction from sequence by integrating long-range interactions. *Nature Methods*, 18 (10):1196–1203, 2021a. doi: 10.1038/s41592-021-01252-x.

Žiga Avsec, Melanie Weilert, Avanti Shrikumar, Sabrina Krueger, Amr Alexandari, Khyati Dalal, Robin Fropf, Charles McAnany, Julien Gagneur, Anshul Kundaje, and Julia Zeitlinger. Base-resolution models of transcription-factor binding reveal soft motif syntax. *Nature Genetics*, 53 (3):354–366, 2021b. doi: 10.1038/s41588-021-00782-6.

Timothy L Bailey. STREME: accurate and versatile sequence motif discovery. *Bioinformatics*, 37 (18):2834–2840, 2021. doi: 10.1093/bioinformatics/btab203.

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pp. 41–48. ACM, 2009. doi: 10.1145/1553374.1553380.

Davide Chicco and Giuseppe Jurman. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics*, 21(1): 6, 2020. doi: 10.1186/s12864-019-6413-7.

Hugo Dalla-Torre, Liam Gonzalez, Javier Mendoza-Revilla, Nicolas Lopez Carranza, Adam Henryk Grzywaczewski, Francesco Oteri, Christian Dallago, Evan Trop, Bernardo P. de Almeida, Hassan Sirelkhatim, Guillaume Richard, Marcin Skwark, Karim Beguir, Marie Lopez, and Thomas Pierrot. Nucleotide Transformer: building and evaluating robust foundation models for human genomics. *Nature Methods*, 22(2):287–297, 2025. doi: 10.1038/s41592-024-02523-z.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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. Association for Computational Linguistics, 2019. doi: 10.18653/v1/n19-1423.

Ke Ding, Brian Parker, and Jiayu Wen. NucEL: Single-Nucleotide ELECTRA-Style Genomic Pretraining for Efficient and Interpretable Representations. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(1):220–227, 2026. doi: 10.1609/aaai.v40i1.36982.

Shobhit Gupta, John A Stamatoyannopoulos, Timothy L Bailey, and William Stafford Noble. Quantifying similarity between motifs. *Genome Biology*, 8(2):R24, 2007. doi: 10.1186/gb-2007-8-2-r24.

Maximilian Ilse, Jakub Tomczak, and Max Welling. Attention-based Deep Multiple Instance Learning. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 2127–2136. PMLR, 2018. URL https://proceedings.mlr.press/v80/ilse18a.html.

Yanrong Ji, Zhihan Zhou, Han Liu, and Ramana V Davuluri. DNABERT: pre-trained Bidirectional Encoder Representations from Transformers model for DNA-language in genome. *Bioinformatics*, 37(15):2112–2120, 2021. doi: 10.1093/bioinformatics/btab083.

Raiyan Khan, Philippe Chlenski, and Itsik Pe'er. Hyperbolic Genome Embeddings. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/b63ad8c24354b0e5bcb7aea16490beab-Abstract-Conference.html.

S. Kirkpatrick, C. D. Gelatt, Jr., and M. P. Vecchi. Optimization by Simulated Annealing. *Science*, 220(4598):671–680, 1983. doi: 10.1126/science.220.4598.671.

S. Kullback and R. A. Leibler. On Information and Sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86, 1951. doi: 10.1214/aoms/1177729694.

Ilya Loshchilov and Frank Hutter. Decoupled Weight Decay Regularization. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

Lei Mao, Yuanhe Tian, and Yan Song. DNAZEN: Enhanced Gene Sequence Representations via Mixed Granularities of Coding Units. *arXiv preprint arXiv:2505.02206*, 2025. doi: 10.48550/arXiv.2505.02206.

Ieva Rauluseviciute, Rafael Riudavets-Puig, Romain Blanc-Mathieu, Jaime A Castro-Mondragon, Katalin Ferenc, Vipin Kumar, Roza Berhanu Lemma, Jérémy Lucas, Jeanne Chèneby, Damir Baranasic, Aziz Khan, Oriol Fornes, Sveinung Gundersen, Morten Johansen, Eivind Hovig, Boris Lenhard, Albin Sandelin, Wyeth W Wasserman, François Parcy, and Anthony Mathelier. JASPAR 2024: 20th anniversary of the open-access database of transcription factor binding profiles. *Nucleic Acids Research*, 52(D1):D174–D182, 2024. doi: 10.1093/nar/gkad1059.

Alexander Rives, Joshua Meier, Tom Sercu, Siddharth Goyal, Zeming Lin, Jason Liu, Demi Guo, Myle Ott, C. Lawrence Zitnick, Jerry Ma, and Rob Fergus. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences*, 118(15):e2016239118, 2021. doi: 10.1073/pnas.2016239118.

Yair Schiff, Chia Hsiang Kao, Aaron Gokaslan, Tri Dao, Albert Gu, and Volodymyr Kuleshov. Caduceus: Bi-Directional Equivariant Long-Range DNA Sequence Modeling. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 43632–43648. PMLR, 2024. URL https://proceedings.mlr.press/v235/schiff24a.html.

Xianglong Shi, Ziheng Chen, Yunhan Jiang, and Nicu Sebe. Intrinsic Lorentz Neural Network. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/aefb164890f80762128efc135205a308-Abstract-Conference.html.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. RoFormer: Enhanced transformer with Rotary Position Embedding. *Neurocomputing*, 568:127063, 2024. doi: 10.1016/j.neucom.2023.127063.

Natan Tourne, Gaetan De Waele, Vanessa Vermeirssen, and Willem Waegeman. How negative sampling shapes the performance of transcription factor binding site prediction models. *Bioinformatics*, 42(2):btag048, 2026. doi: 10.1093/bioinformatics/btag048.

Zhao Yang, Jiwei Zhu, and Bing Su. SPACE: Your Genomic Profile Predictor is a Powerful DNA Foundation Model. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 71648–71667. PMLR, 2025. URL https://proceedings.mlr.press/v267/yang25ay.html.

Chenhao Zeng, Zhibin Pu, and Shufei Ge. HyCoSeq: Contextual Hyperbolic Representation Learning for Genomic Sequences. *arXiv preprint arXiv:2609.16925*, 2026. URL https://arxiv.org/abs/2609.16925.

Hannah Zhou, Avanti Shrikumar, and Anshul Kundaje. Towards a Better Understanding of Reverse-Complement Equivariance for Deep Learning Models in Genomics. In *Proceedings of the 16th Machine Learning in Computational Biology Meeting*, volume 165 of *Proceedings of Machine Learning Research*, pp. 1–33. PMLR, 2022. URL https://proceedings.mlr.press/v165/zhou22a.html.

Jian Zhou and Olga G Troyanskaya. Predicting effects of noncoding variants with deep learning–based sequence model. *Nature Methods*, 12(10):931–934, 2015. doi: 10.1038/nmeth.3547.

Zhihan Zhou, Yanrong Ji, Weijian Li, Pratik Dutta, Ramana Davuluri, and Han Liu. DNABERT-2: Efficient Foundation Model and Benchmark For Multi-Species Genomes. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/b633e7052970b8f5aa1a69164d99e9e8-Abstract-Conference.html.

A Dataset Details

Table A1 lists the number of training, validation, and test samples in each of the datasets.

Table A1: Sample counts for the training, validation, and test splits of the binding datasets.

Split	human_ tf_0	human_ tf_1	human_ tf_2	human_ tf_3	human_ tf_4	mouse_ 0	mouse_ 1	mouse_ 2	mouse_ 3	mouse_ 4
Train	32378	30672	19000	27294	19000	6478	53952	2620	1904	15064
Dev	1000	1000	1000	1000	1000	810	6745	328	239	1883
Test	1000	1000	1000	1000	1000	810	6745	328	239	1883

B Baseline Details

We summarize the seven baseline methods compared in Section 4.

1. **DNABERT-2** (Zhou et al., 2024) replaces fixed k-mer tokenization with Byte-Pair Encoding (BPE) and uses Attention with Linear Biases (ALiBi) positional encoding.
2. **NucEL** (Ding et al., 2026) adopts the ELECTRA (Efficiently Learning an Encoder that Classifies Token Replacements Accurately) paradigm, combining single-nucleotide tokenization, a ModernBERT backbone, and a Replaced Token Detection (RTD) objective that supervises every nucleotide.
3. **HGE** (Hyperbolic Genome Embeddings) (Khan et al., 2025) encodes genomic sequences with Lorentz Hyperbolic Neural Networks (HNNs), using hyperbolic Convolutional Neural Network (CNN) layers to capture hierarchical structure.
4. **HyCoSeq** (Zeng et al., 2026) constructs a context-aware network on hyperbolic manifolds and learns context-dependent sequence representations in hyperbolic space.
5. **ILNN** (Shi et al., 2026) designs specialized hyperbolic operators, including a native Point-to-hyperplane Lorentz Fully Connected (PLFC) layer and a Gyro Lorentz Batch Normalization (GyroLBN), to compute directly in hyperbolic space.
6. **DNAZEN** (Mao et al., 2025) builds a dual-branch encoder from multigranularity coding units to capture DNA features at different granularities.
7. **SPACE** (Yang et al., 2025) uses a Mixture-of-Experts (MoE) architecture to handle heterogeneity across species and epigenomic profiles.

C Additional Temperature Annealing Results

Table C1: MCC scores on the human TF datasets under cold-start training and different temperature annealing endpoints.

Dataset	**Cold start** $\boldsymbol{T = 0.01}$	Annealing 0.01	Annealing 0.1	Annealing 0.5	Annealing 0.7
human_tf_0	0.7063	0.6967	0.6958	0.7033	0.7004
human_tf_1	0.7617	0.7484	0.7550	0.7619	0.7670
human_tf_2	0.6900	0.6761	0.6703	0.6801	0.6801
human_tf_3	0.6092	0.6047	0.6064	0.6122	0.6082
human_tf_4	0.8205	0.8088	0.8024	0.8144	0.8323
Human average	0.7175	0.7069	0.7059	0.7144	0.7176