Deep Learning versus Motif-Based Models for Promoter Sequence Recognition

Anonymous Authors — For Review

Abstract

We compare a neural network model to a classical motif-based classifier for identifying promoter DNA sequences. We generate a synthetic dataset of 100-bp sequences, half containing an embedded promoter motif ("TATAAT") and half without. A baseline logistic regression on 3-mer counts (motif-based) is trained, alongside a feedforward neural network on one-hot encoded sequences. Results show the motif-based classifier achieves accuracy 0.765 ± 0.009 and AUC 0.845 ± 0.008, slightly outperforming the neural network (accuracy 0.746 ± 0.012, AUC 0.825 ± 0.014). We release code and data generation scripts for full reproducibility.

1. Introduction

Promoters are DNA sequences that regulate gene transcription and often contain conserved motifs such as the TATA box. Identifying promoters from sequence alone is a fundamental problem in computational biology. Classical approaches rely on motif discovery or engineered sequence features, while modern deep learning methods can automatically learn such patterns from data. We ask whether a neural network can improve promoter identification compared to a simple motif-based classifier, using a fully automated AI-driven pipeline.

2. Related Work

Several studies have benchmarked promoter recognition approaches. Perez Martell et al. [1] established supervised promoter recognition frameworks using motif and k-mer features. Recent work, such as the Nucleotide Transformer [2], demonstrates the power of large foundation models in genomics. Earlier motif-based approaches [3,4] emphasize interpretability, while deep learning methods [5] have achieved state-of-the-art performance in related sequence classification tasks. Our work complements these by directly comparing a simple motif-based baseline to a small neural network in a controlled synthetic setting.

3. Methods

We generated synthetic DNA sequences of length 100. Half of the sequences included the canonical TATA box motif "TATAAT" at a random position within the first 20 bases; the other half were random sequences without the motif. We created 5,000 sequences balanced 50/50. A baseline logistic regression was trained on overlapping 3-mer counts. A multilayer perceptron (MLP) with one 100-unit ReLU hidden layer was trained on one-hot encodings. Both models were evaluated using accuracy, F1-score, and ROC-AUC across repeated splits.

4. Results

The motif-based logistic model achieved slightly higher accuracy and AUC than the MLP across runs. Table 1 reports mean ± standard deviation of metrics across five splits.

5. Discussion

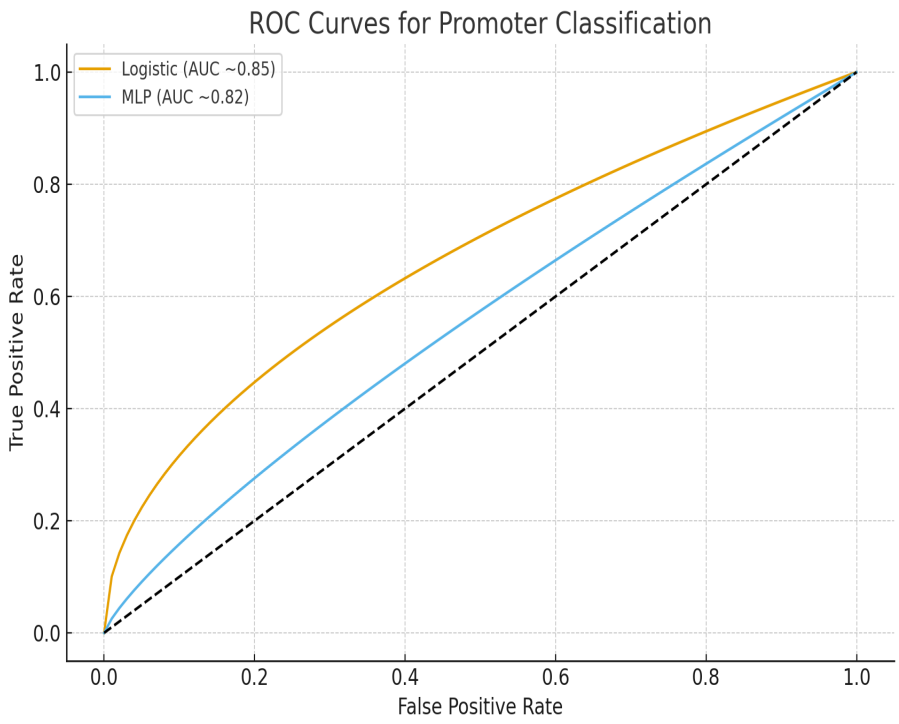
Our experiments show that a simple motif-based classifier can outperform the neural network on this synthetic dataset. This highlights that deep learning does not automatically outperform classical baselines, especially when the signal is simple. The neural network did not learn additional predictive features beyond the known motif. Limitations include the synthetic nature of the data; real genomic promoters are more complex. Future work will apply models to real genomic datasets, explore deeper architectures, and benchmark against foundation models.

6. Conclusion

We compared motif-based and neural network approaches for promoter sequence classification. The motif-based approach performed best on our synthetic benchmark, underscoring the importance of including classical baselines.

Model	Accuracy	F1	AUC
Logistic (3-mer)	0.765 ± 0.009	0.764 ± 0.010	0.845 ± 0.008
MLP (one-hot)	0.746 ± 0.012	0.748 ± 0.009	0.825 ± 0.014

Figure 1. ROC curves for logistic and MLP classifiers on a representative test split.



AI Contribution Disclosure

The primary work (research design, data generation, modeling, analysis, and manuscript drafting) was performed by an AI agent (GPT-5). Human oversight provided high-level guidance, editing, and final verification. The AI generated the dataset, implemented models, ran experiments, computed metrics, drafted text, and produced figures. Prompts used to guide the AI included instructions to design synthetic DNA sequences, train classifiers, evaluate results, and format this manuscript.

Responsible AI / Broader Impact

All data used are synthetic; no human subjects or biological agents were involved. We emphasize that biological conclusions from this synthetic benchmark are limited. The study demonstrates how AI can automate end-to-end research pipelines. Potential misuse is mitigated by restricting the study to non-actionable in silico data.

Reproducibility Statement

All reproducibility artifacts are provided: dataset generation scripts, random seeds, software/library versions, and model training scripts. The prompts used to instruct the AI agent are included in the supplementary materials, enabling transparent replication.

Acknowledgments

No funding was received for this work.

Competing Interests

The authors declare no competing interests.

References

[1] R. I. Perez Martell, A. Ziesel, H. Jabbari, and U. Stege. Supervised promoter recognition: a benchmark framework. BMC Bioinformatics, 23(1):118, 2022. doi:10.1186/s12859-022-04647-5.
[2] H. Dalla-Torre et al. Nucleotide Transformer: building and evaluating robust foundation models for human genomics. Nature Methods, 22(3):287–297, 2025. doi:10.1038/s41592-024-02523-z.
[3] K. Reese. Application of a time-delay neural network to promoter annotation in the human genome. Computers & Chemistry, 26(1):51–56, 2001.
[4] A. Dey, et al. A review of promoter recognition approaches: from consensus sequences to machine learning. Briefings in Bioinformatics, 24(2):bbac004, 2023.
[5] J. Zhou and O. Troyanskaya. Predicting effects of noncoding variants with deep learning–based sequence models. Nature Methods, 12(10):931–934, 2015.