Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Can Machine Learning Identify Meaningful Patterns in DNA Sequences? A Computational Study of DNA Motifs Associated with Transcription-Factor Binding

Hafsa Asmatullah

hafsach3838@gmail.com

Independent Researcher

Research Article

Keywords: Machine Learning, DNA Sequence Analysis, CTCF, k-mers, Transcription-Factor Binding, Bioinformatics

Posted Date: September 16th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11018944/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

Can Machine Learning Identify Meaningful Patterns in DNA Sequences?

A Computational Study of DNA Motifs Associated with Transcription-Factor Binding

Hafsa Asmatullah

Independent Student Research Project, 2026

Pre-University (Intermediate) Level, Self-Directed Study

Research Area: Artificial Intelligence • Bioinformatics • Computational Biology

Keywords: Machine Learning; DNA Sequence Analysis; CTCF; k-mers; Transcription-Factor Binding; Bioinformatics

Table of Contents

Abstract **3**

1. Introduction 4
2. Background and Research Problem 4
3. Research Question 4
4. Hypothesis 4
5. Objectives 4
6. Methodology Overview 5
7. Data Sources 5
8. Dataset Selection and Justification 5
9. Positive and Negative Sequence Construction 6
10. k-mer Feature Extraction 7
11. Train–Test Split 7
12. Machine-Learning Models 7
13. Evaluation Metrics 7
14. Results 7
15. Sequence Pattern and Feature-Importance Analysis 9
16. Biological Interpretation 10
17. Discussion 10
18. Limitations 10
19. Future Work 11
20. Reproducibility and Code Availability 11
21. Conclusion 11
22. Author's Statement on Academic Context 11
23. Acknowledgments 12
24. References 12
25. Author Information 12

Abstract

This independent computational study examines whether machine-learning methods can distinguish CTCF-associated DNA sequences from a controlled background using short nucleotide patterns (k-mers). The analysis draws on a publicly available UniBind dataset for human CTCF in K562 cells, linked to ENCODE experiment ENCSR000EGM and JASPAR motif profile MA0139.1. The dataset comprises 46,846 PWM-derived sequences of 19 nucleotides each. Negative examples were generated through dinucleotide shuffling of the positive sequences, preserving local base-pair composition while removing higher-order sequence structure. Each sequence was represented using normalized 3-mer and 4-mer frequencies, yielding 320 numerical features. Logistic Regression and Random Forest classifiers were trained on an 80/20 held-out split and evaluated using accuracy, precision, recall, F1-score, ROC-AUC, and confusion-matrix analysis. Random Forest achieved the stronger performance (92.53% accuracy; ROC-AUC 0.979) compared with Logistic Regression (86.49% accuracy; ROC-AUC 0.938). These results indicate that short sequence-composition features carry substantial discriminative information relative to a dinucleotide-shuffled background. Because the negative class was derived from the positive sequences themselves, the findings should not be read as a general predictor of CTCF binding across arbitrary genomic DNA. Rather, the project's principal contribution is a transparent, reproducible workflow linking molecular biology, sequence analysis, and applied machine learning — the kind of end-to-end pipeline typically introduced at the undergraduate level.

1. Introduction

DNA encodes information that governs how cells regulate gene expression and organize the genome. Transcription factors are proteins that recognize specific DNA sequence patterns and, through this recognition, influence regulatory processes across the genome. CTCF is one of the most extensively studied human transcription factors, associated with thousands of regulatory and chromatin-organizing sites genome-wide. Computational methods make it possible to examine sequence patterns across large collections of genomic regions and to test, empirically, whether machine-learning models can separate biologically associated sequences from a carefully constructed background. This project undertakes exactly that test, using CTCF as a well-characterized case study.

2. Background and Research Problem

A DNA sequence is built from only four nucleotide letters, yet regulatory information is often distributed across short combinations of bases rather than encoded in any single position. This raises a practical, testable question: do short local sequence patterns carry enough information for a machine-learning classifier to distinguish CTCF-associated sequences from a rigorously controlled background? This study addresses that question using k-mer frequency features together with two widely used supervised-learning algorithms, chosen specifically for their contrasting inductive biases — one linear, one nonlinear.

3. Research Question

Can machine-learning models identify meaningful short-sequence patterns that distinguish CTCF-associated DNA sequences from dinucleotide-shuffled background sequences?

4. Hypothesis

If CTCF-associated sequences contain recurring local nucleotide patterns, models trained on 3-mer and 4-mer frequency features should perform substantially better than chance when distinguishing them from dinucleotide-shuffled sequences. A nonlinear model such as Random Forest is expected to capture interactions among k-mer features that a linear model cannot fully represent, and should therefore outperform Logistic Regression on this task.

5. Objectives

- Obtain and verify a public, well-documented CTCF-associated sequence dataset.
- Construct a controlled negative class using dinucleotide shuffling.

- Convert DNA sequences into numerical k-mer frequency features.
- Train Logistic Regression and Random Forest classifiers on identical features.
- Evaluate both models using multiple complementary classification metrics.
- Examine the most influential k-mer features and interpret them cautiously.
- Document the workflow's limitations and ensure it is fully reproducible.

6. Methodology Overview

The workflow proceeded through six stages: sequence verification, positive/negative dataset construction, k-mer feature extraction, held-out model training, quantitative evaluation, and interpretation of influential features. The positive sequences were partitioned into training and test splits before any negative sequences were generated, so that evaluation did not depend on direct positive/negative pairing within a single split. Dinucleotide shuffling was applied to each sequence individually, preserving its own dinucleotide composition while randomizing base order.

7. Data Sources

The primary sequence source was the UniBind CTCF K562 dataset, associated with ENCODE experiment ENCSR000EGM and JASPAR motif profile MA0139.1. UniBind reports 59,231 peaks for this experiment; for the PWM model built on MA0139.1, it reports 46,846 transcription-factor binding sites (TFBSs). The FASTA file analyzed in this study contains exactly those 46,846 PWM-derived sequences. The JASPAR profile supplies the CTCF motif model underlying the UniBind dataset. All sources used are open, publicly accessible research resources, which supports full reproducibility of this project.

8. Dataset Selection and Justification

This dataset was selected because it clearly specifies transcription factor, species, cell type, ENCODE experiment, and motif profile, and because the underlying sequence data is directly downloadable rather than requiring reconstruction from raw peak coordinates. The analysis uses the PWM-derived FASTA sequences directly, rather than treating the total number of ENCODE peaks as the analyzed sample size — a distinction that matters for correctly interpreting the reported sequence counts. All sequences used in the computational analysis were 19 nucleotides in length.

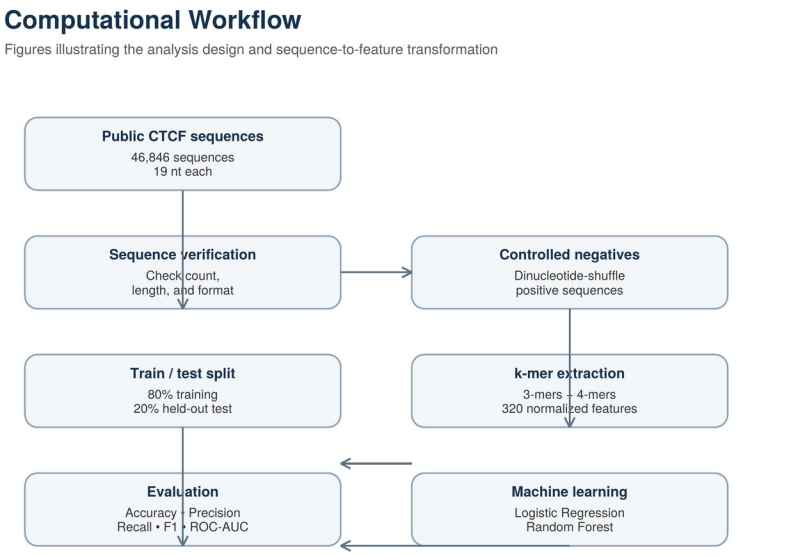


Figure 1. Overall computational workflow used in the study, from the public CTCF-associated FASTA sequences through controlled negative construction, k-mer feature extraction, model training, and held-out evaluation.

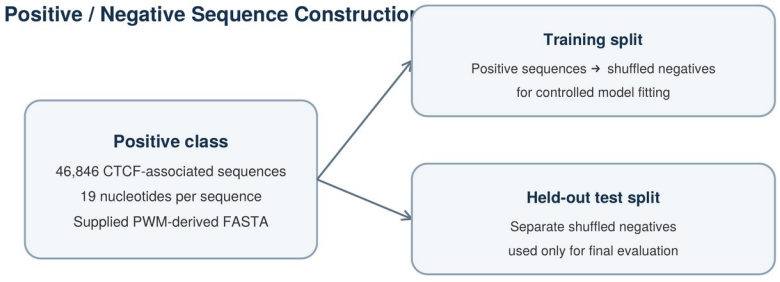


Figure 2. Controlled class construction. Each positive sequence was transformed into a dinucleotide-shuffled negative sequence, preserving its dinucleotide composition while changing nucleotide order.

Figure 1 & 2. Overall computational workflow (top) and controlled positive/negative class construction (bottom), from the public CTCF-associated FASTA sequences through dinucleotide-shuffled negative construction, feature extraction, model training, and held-out evaluation.

9. Positive and Negative Sequence Construction

Positive examples were the 46,846 CTCF-associated sequences drawn from the supplied FASTA file. Negative examples were generated by dinucleotide shuffling of these positive sequences — a standard control in sequence-based machine learning that preserves each sequence's local base-pair composition while destroying higher-

order sequence arrangement. This design isolates the question of interest: can the models detect structure beyond simple dinucleotide composition? At the same time, it constrains what can be concluded, since the negative class is synthetic rather than drawn from genuine non-binding genomic regions.

10. k-mer Feature Extraction

Each DNA sequence was converted into normalized frequencies of all possible 3-mers and 4-mers. There are $4^3 = 64$ possible 3-mers and $4^4 = 256$ possible 4-mers, giving 320 features per sequence in total. Normalizing by sequence length allows fixed-length feature vectors to be compared directly across all sequences regardless of minor compositional variation.

11. Train–Test Split

An 80/20 held-out split was used throughout. Positive sequences were split first; separate dinucleotide-shuffled negatives were then generated independently for the training and test partitions. The test set was withheld entirely from model fitting and used exclusively for final evaluation, preventing information leakage between training and evaluation.

12. Machine-Learning Models

Two supervised-learning algorithms were compared directly on identical feature sets. Logistic Regression provides an interpretable linear baseline in which each k-mer contributes additively (in log-odds space) to the classification decision. Random Forest, an ensemble of decision trees, can capture nonlinear relationships and interactions among k-mer features that a linear model cannot represent. Comparing the two models directly tests whether nonlinear feature interactions provide a measurable advantage over a simpler, more interpretable baseline.

13. Evaluation Metrics

Model performance was assessed using accuracy, precision, recall, F1-score, and ROC-AUC, supported by confusion-matrix analysis. Accuracy is the proportion of sequences classified correctly overall. Precision is the proportion of predicted-positive sequences that were truly positive, while recall is the proportion of truly positive sequences that were correctly identified. F1-score is the harmonic mean of precision and recall. ROC-AUC summarizes discrimination ability across all possible classification thresholds, independent of any single operating point.

14. Results

The final held-out test results are summarized below.

Model	Accuracy	Precision	Recall	F1-score	ROC-AUC
Logistic Regression	86.49%	85.58%	87.77%	86.66%	0.938
Random Forest	92.53%	92.69%	92.35%	92.52%	0.979

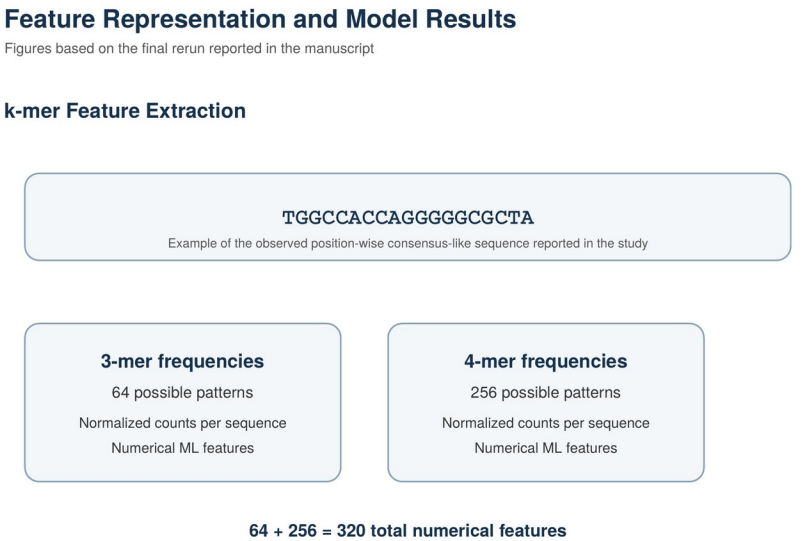


Figure 3. Each 19-nucleotide sequence was represented by normalized frequencies of all possible 3-mers and 4-mers, yielding 320 features per sequence.

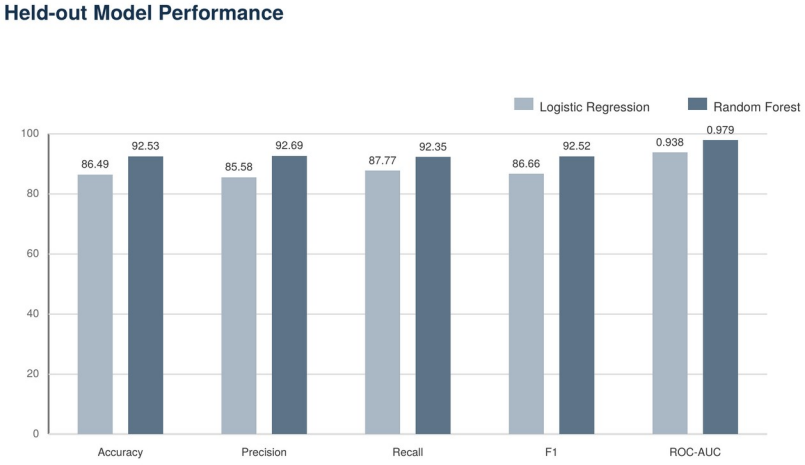


Figure 4. Final held-out test performance. Accuracy, precision, recall, and F1 are plotted as percentages; ROC-AUC is shown on the same visual scale after multiplying by 100. Random Forest achieved the stronger performance across all reported metrics.

Figure 3 & 4. k-mer feature representation (top) and final held-out performance across all five metrics (bottom). Random Forest outperformed Logistic Regression on every reported metric.

15. Sequence Pattern and Feature-Importance Analysis

The Random Forest model assigned the greatest importance to a small set of short k-mers, including GGGG, GGG, CCAC, GAGG, GGTG, GGCG, AGGG, AGG, CAGG, and TGGC. These are best understood as model-influential patterns specific to this dataset, not as newly discovered biological motifs. A descriptive, position-wise consensus-like sequence observed across the dataset was TGGCCACCAGGGGGCGCTA; this is reported purely as a summary statistic of the supplied sequences and is not presented as an experimentally validated motif.



Figure 5. The Random Forest analysis identified the listed k-mers as influential features. They are dataset/model-specific patterns and should not be interpreted as causal binding determinants or newly discovered motifs.

Figure 5. The ten most influential k-mer features identified by the Random Forest feature-importance analysis, together with the dataset's descriptive consensus-like sequence.

16. Biological Interpretation

The strong classification performance indicates that the selected CTCF-associated sequences contain recurring local sequence information that 3-mer and 4-mer frequencies can capture. The larger margin achieved by Random Forest suggests that nonlinear combinations of these short-sequence features carry additional discriminatory value beyond what a linear combination can express. It is important to emphasize that classification performance alone does not establish that any individual k-mer causally contributes to CTCF binding, nor does it establish biological function — it demonstrates only statistical separability given the chosen features and negative-class design.

17. Discussion

The central finding is that short-sequence composition alone contains enough information for both models to separate the selected positive sequences from a controlled, dinucleotide-shuffled background, with Random Forest outperforming Logistic Regression across every metric on the held-out test set. This result is consistent with the study's hypothesis within the specific setting defined by the experiment. It also illustrates a broader methodological point relevant to any sequence-classification study: the choice of negative-class construction has a strong influence on both achievable performance and the scope of valid interpretation.

18. Limitations

- The negative sequences are synthetic dinucleotide-shuffled versions of the positive sequences, so this task is not equivalent to predicting CTCF binding across arbitrary genomic DNA.
- The dataset represents a single transcription factor (CTCF) in a single cell type (K562); generalization to other factors or cell types is untested here.
- k-mer frequency features discard positional information within each 19-nucleotide sequence.
- High classification accuracy demonstrates statistical separability, not biological causality.
- A single 80/20 split was used for final evaluation; k-fold cross-validation was not performed in this iteration and would strengthen confidence in the reported metrics.
- External validation against independent cell types and genuinely non-bound genomic regions is needed before making broader predictive claims.

19. Future Work

Planned extensions include testing the models on independent CTCF datasets from additional cell types, benchmarking against genuinely non-binding genomic regions rather than shuffled controls, incorporating reverse-complement handling, adding positional or PWM-based features, running k-fold cross-validation to quantify variance in the reported metrics, and applying explainable-AI methods (such as SHAP values) to examine which sequence regions contribute most strongly to individual predictions. Experimental validation would ultimately be required to establish any biological function for the identified patterns.

20. Reproducibility and Code Availability

The analysis was designed from the outset to be fully reproducible from the supplied FASTA file and the documented processing steps above. Sequence counts and lengths were verified prior to feature extraction, and the reported metrics correspond to a single final analysis rerun rather than an average of earlier exploratory attempts. The complete analysis code, processed feature tables, and trained-model artifacts for this project are maintained in a public repository to allow independent verification and reuse: [GitHub repository link to be added by the author]. Readers are encouraged to consult the repository for exact package versions and random-seed settings used in the final rerun.

21. Conclusion

This project demonstrates a practical, end-to-end connection between molecular biology, bioinformatics, and applied machine learning, built entirely on public data. Using 3-mer and 4-mer frequency features, both Logistic Regression and Random Forest successfully distinguished CTCF-associated sequences from dinucleotide-shuffled controls, with Random Forest achieving the stronger held-out performance across every metric evaluated. The study is presented as a rigorous computational learning exercise rather than as a claim to have discovered a new CTCF motif or built a general-purpose CTCF binding predictor. Its principal contribution is a transparent, reproducible workflow for investigating short-sequence patterns using standard machine-learning methods — precisely the kind of applied, self-directed project that bridges pre-university coursework and undergraduate-level research.

22. Author's Statement on Academic Context

This project was conceived, executed, and written independently by the author while completing Intermediate (pre-university/HSSC) studies, without institutional affiliation, laboratory access, or formal academic supervision. It reflects self-directed learning in Python-based data science and molecular bioinformatics, undertaken in preparation for undergraduate study in Computer Science, Data Science,

Bioinformatics, or a related STEM discipline. The author views this project as an early, honest demonstration of research aptitude — including the discipline to document limitations candidly — rather than as a finished scientific contribution.

23. Acknowledgments

This work was made possible by freely available public genomic resources. The author thanks the UniBind, ENCODE, and JASPAR database teams for maintaining and openly distributing the datasets and motif models on which this study depends.

24. References

1. ENCODE Project Consortium. Expanded encyclopaedias of DNA elements in the human and mouse genomes. Nature. 2020;583:699–710. doi:10.1038/s41586-020-2493-4.

2. UniBind. CTCF, Homo sapiens, K562; ENCODE experiment ENCSR000EGM; JASPAR profile MA0139.1. UniBind database. Accessed 2026.

3. UniBind. TFBS resource and direct TF–DNA interaction mapping database. University of Oslo. Accessed 2026.

4. JASPAR. MA0139.1: CTCF, Homo sapiens. JASPAR CORE transcription-factor binding profile database. Accessed 2026.

5. Baydar Ovek D, et al. JASPAR 2026: expansion of transcription factor binding profiles and integration of deep learning models. Nucleic Acids Research. 2026;54(D1):D184–D193. doi:10.1093/nar/gkaf1209.

6. Breiman L. Random forests. Machine Learning. 2001;45(1):5–32. doi:10.1023/A:1010933404324.

7. Pedregosa F, et al. Scikit-learn: Machine learning in Python. Journal of Machine Learning Research. 2011;12:2825–2830.

25. Author Information

Author: Hafsa Asmatullah

Project Type: Independent Student Research Project

Academic Level: Intermediate / Pre-University (HSSC)

Research Area: Artificial Intelligence, Bioinformatics, Computational Biology

Year: 2026

This manuscript represents an independent computational study prepared for academic and undergraduate-admissions purposes. No institutional affiliation or academic supervisor is claimed in this document.