

# Explainable AI for Genomic Tasks with Genome Language Model

**CAP 5610 - Machine Learning**

Spring 2025

Instructor - **Mengxin Zheng**

**Presented By**

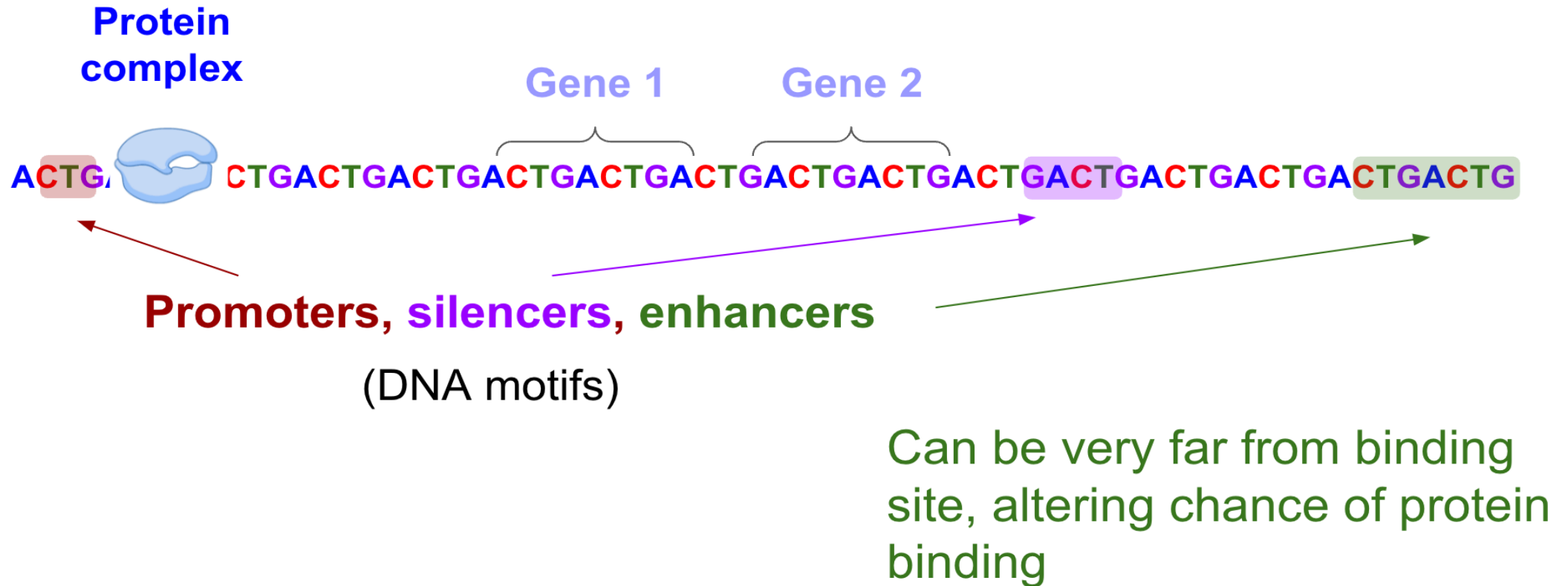
Sourav Saha

Abdur Rouf

Nowfel Mashnoor

Sagor Biswas

# Lets Talk DNA



# Motivation



- **Advanced Genome Language Models:** Transformer-based models such as DNABERT, DNABERT2, Nucleotide Transformers, and HyenaDNA have demonstrated strong performance on various genome-specific classification tasks.
- **Knowledge Gap:** Despite their success, the depth of these models' understanding of genomic concepts remains unclear.
- **Explainability Focus:** This project aims to investigate how well these models can explain their predictions by applying them to a downstream interpretability task.

# Task Description



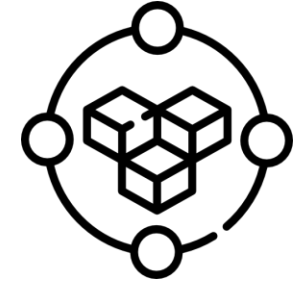
- **Objective:** Use advanced genomic language models (e.g., DNABERT, HyenaDNA) for promoter detection.
- **Focus:** Investigate whether these models truly leverage the presence of TATA boxes for classification.
- **Hypothesis:** TATA boxes are key signals that significantly influence the model's decision-making in identifying promoters.
- **Explainability Goal:** Demonstrate how attention or feature-attribution methods can reveal the role of TATA box motifs in the models' predictions.

# Dataset



Category	Details
Dataset Used	Genome Understanding Evaluation (GUE) for human promoter detection
Sequence Length	300 Nucleotides
Dataset Split	<b>Training Set:</b> 47.4k samples <b>Validation Set:</b> 5.2k samples <b>Test Set:</b> 5.2k samples
Class Distribution	Balanced (Positive:Negative = 50:50)

# Models for Promoter Detection



Model	Purpose
<b>DNABERT</b>	Transformer-based model for DNA sequence analysis
<b>DNABERT2</b>	Improved version of DNABERT with enhanced performance
<b>Nucleotide Transformers (NT)</b>	General-purpose nucleotide sequence transformer
<b>HyenaDNA</b>	A novel architecture designed for genomic data

# Results

Model	Parameters	Architecture	Tokenizer	Accuracy	MCC	F1 Score	FT Time
DNABERT	1.6 M	Transformers	K-mer	0.969	0.909	0.959	230 m
DNABERT2	117 M	Transformers	BPE	0.932	0.863	0.931	150 m
NMT	2.5 B	Transformers	K-mer	0.930	0.853	0.927	120 m
HyenaDNA	1.6 M	Implicit Conv.	Single Nucleotide	0.940	0.862	0.931	35 m

# Model Analysis

- **Human-specific pretraining:** Models trained on the human genome excel in species-specific tasks.
- **Multispecies challenge:** Multispecies-pretrained models underperform due to insufficient parameter scaling.
- **Scaling advantage:** Larger models yield better results.
- **Runtime efficiency:** Alternatives to transformer architectures offer faster runtimes.





# Tools for Explainability in ML

- **LIME (Local Interpretable Model-Agnostic Explanations)**

Works with any black-box model and provides human-interpretable explanations. Used on text classification(NLP), image classification, and Tabular Data (fraud detection, medical diagnosis)

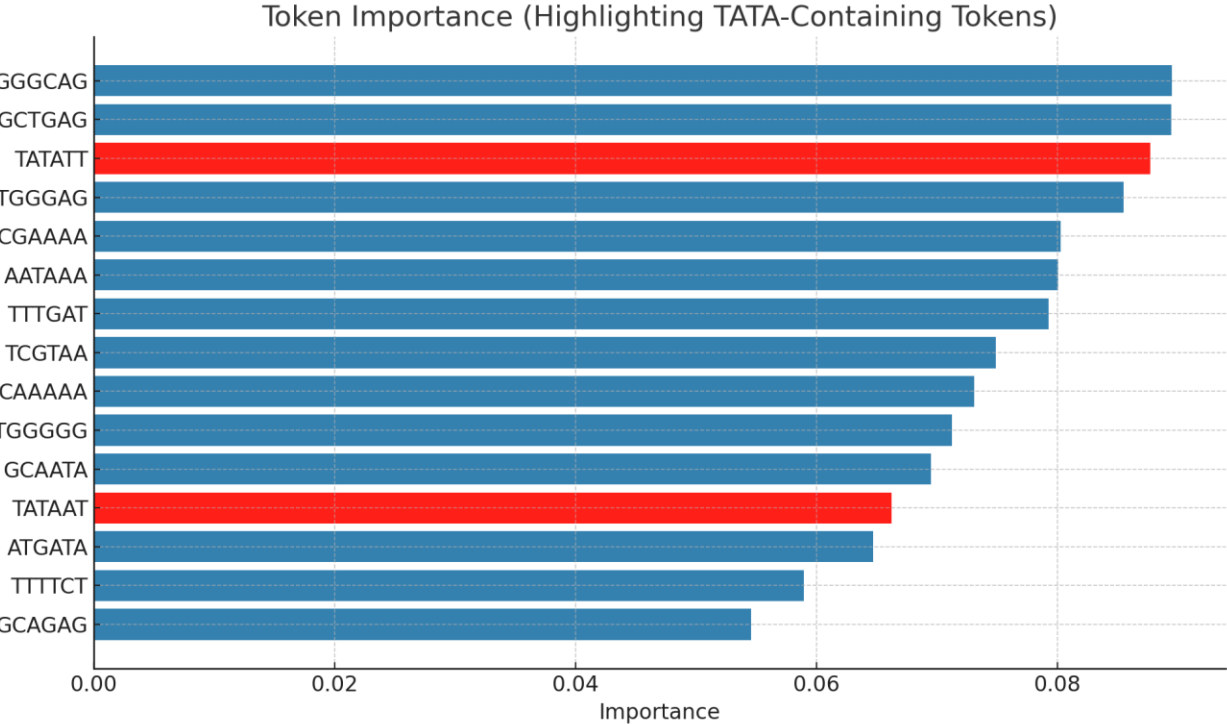
- **SHAP (SHapley Additive Explanations)**

Works with various model types (tree-based, deep learning, linear models) and provides both local and global explanations)

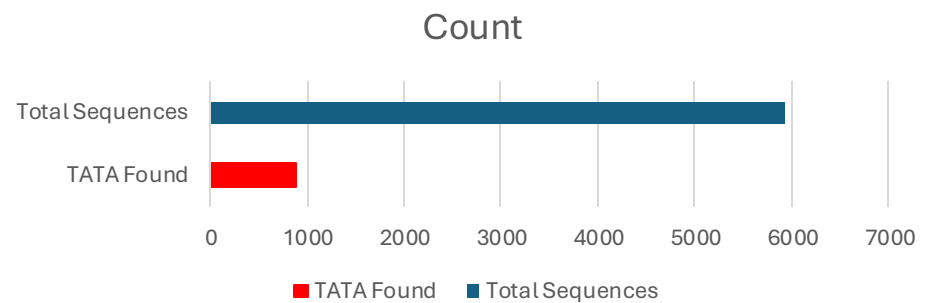
- **ANCHOR (High-Precision Model-Agnostic Explanations)**

More stable model-agnostic tool that provides high-confidence explanations. Used for Text classification, Image classification, Fraud detection, and healthcare )

# LIME - Results

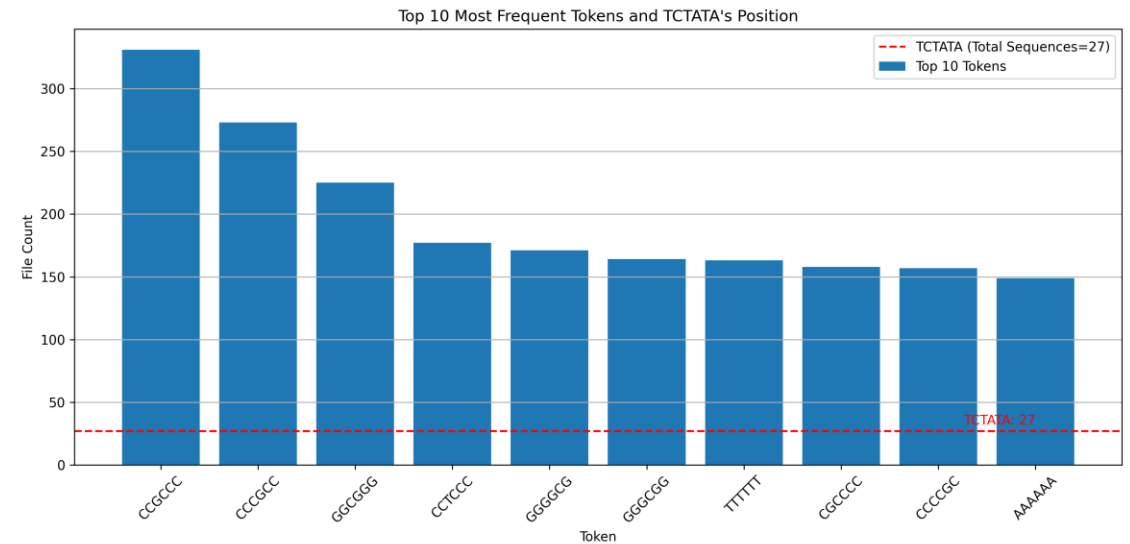
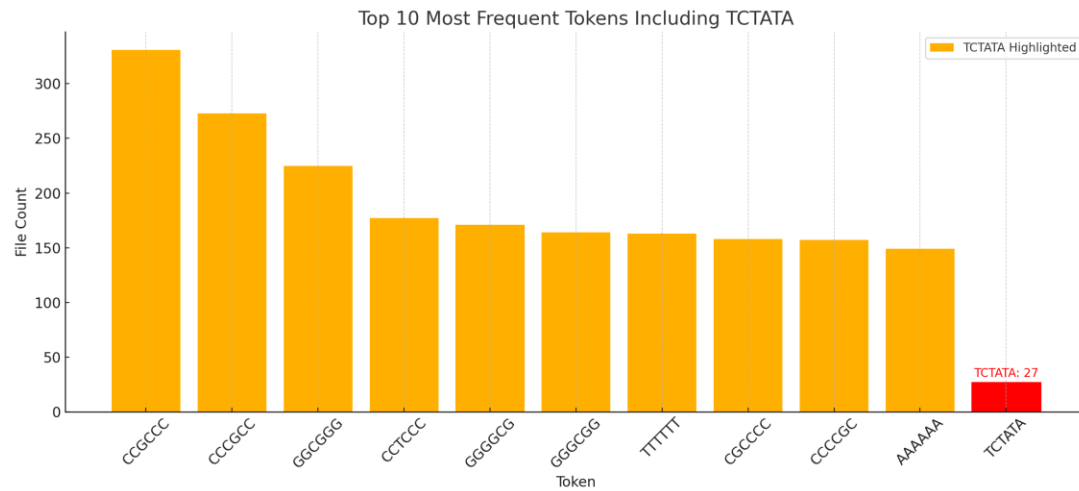


Importance/Label	0	1
Importance	54	99
Negative	639	84



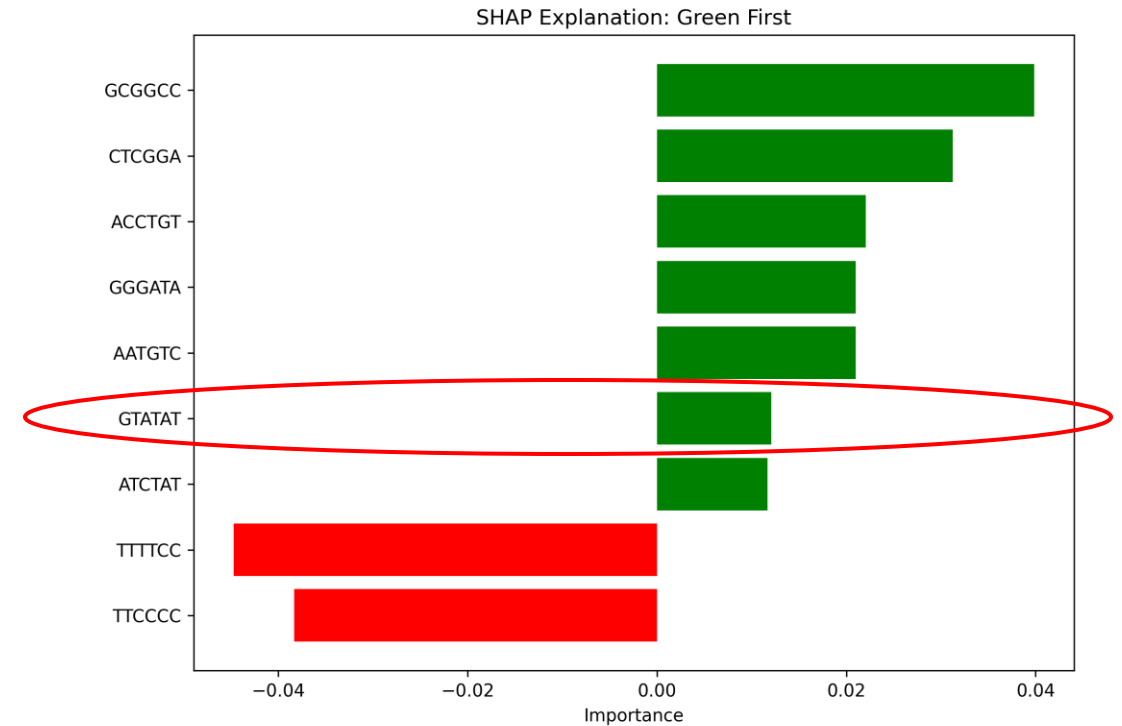
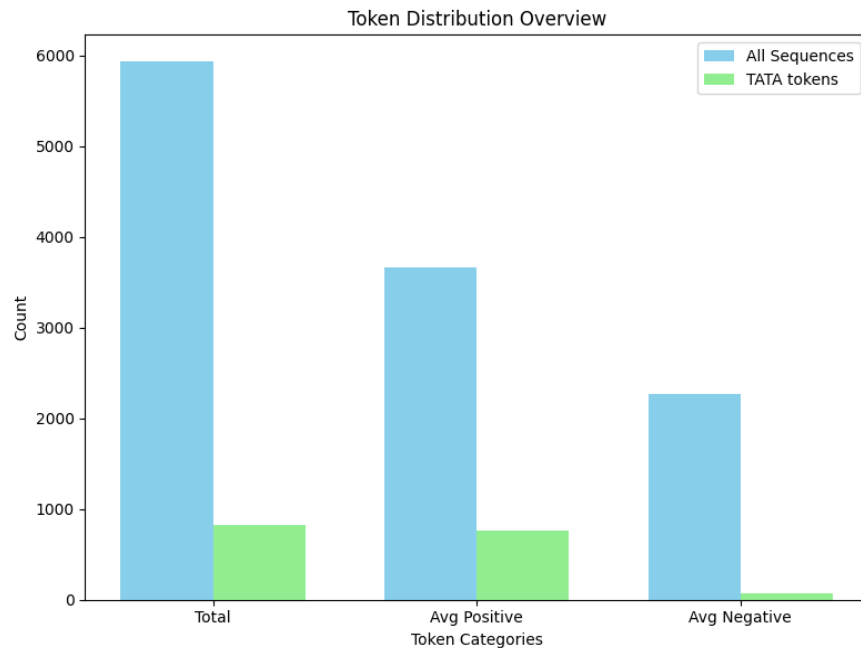
# LIME – Results – cont.

**TCTATA**, the most common **TATA-containing token**, appears in **27** sequences (**20%**), whereas the top tokens average around **150** appearances—indicating TCTATA's moderate presence.



# SHAP (SHapley Additive exPlanations)

- Model-agnostic : Can explain any machine learning model
- Unlike LIME utilizes Global understanding
- Creates set of features to find importance of each



- ☐ Total sequence = 5930
- ☐ Tokens analyzed per sequence = 10
- ☐ Total sequences = 5930
- ☐ Total TATA tokens = 828
- ☐ Avg Positive tokens = 3662
- ☐ Avg Negative tokens = 2268
- ☐ Positive TATA = 763
- ☐ Negative TATA = 65

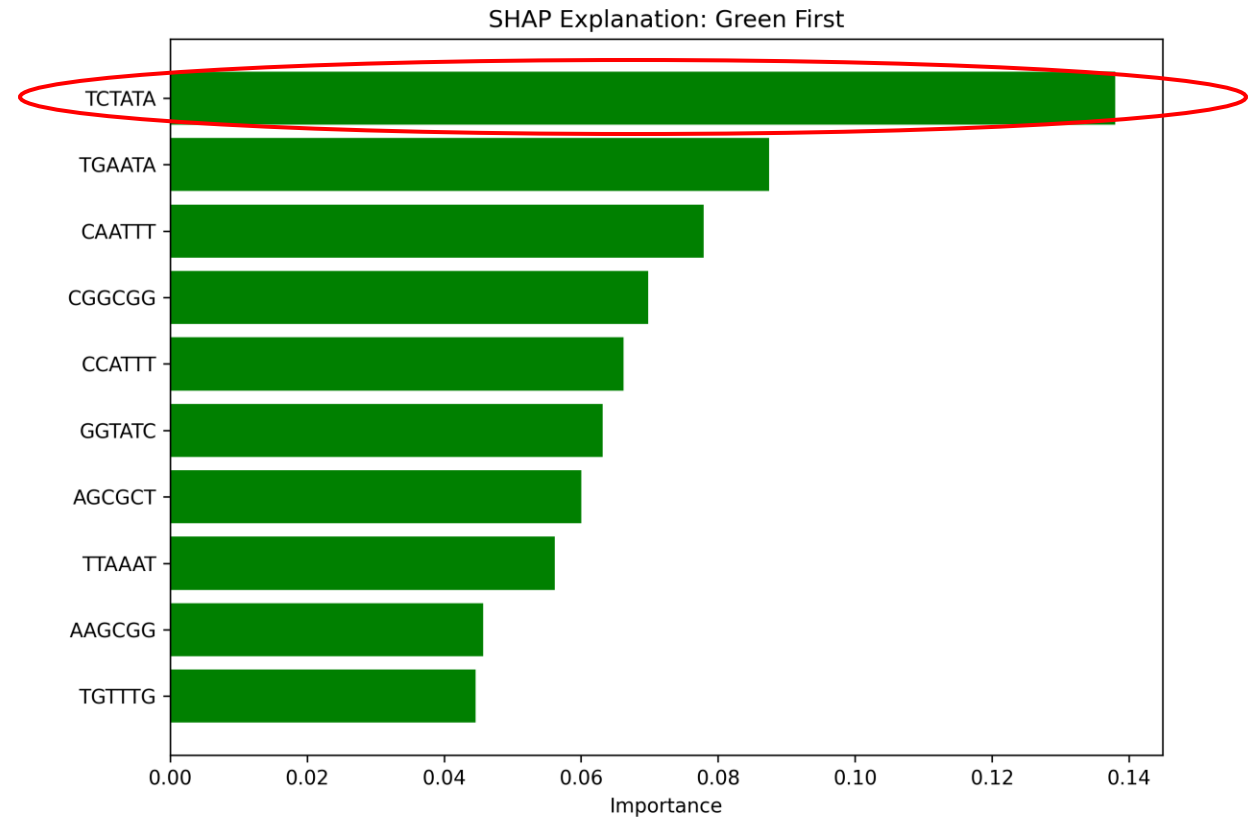
# SHAP

➤ Top 10 tokens with positive importance

<b>TCTATA</b>	<b>118</b>
<b>CTCCCC</b>	<b>78</b>
<b>GGCGGG</b>	<b>78</b>
<b>GGTGGA</b>	<b>65</b>
<b>GCGGCG</b>	<b>65</b>
<b>GCCTGG</b>	<b>65</b>
<b>TGAATA</b>	<b>65</b>
<b>CCAGAG</b>	<b>65</b>
<b>CCTCCC</b>	<b>65</b>
<b>TATTTT</b>	<b>52</b>

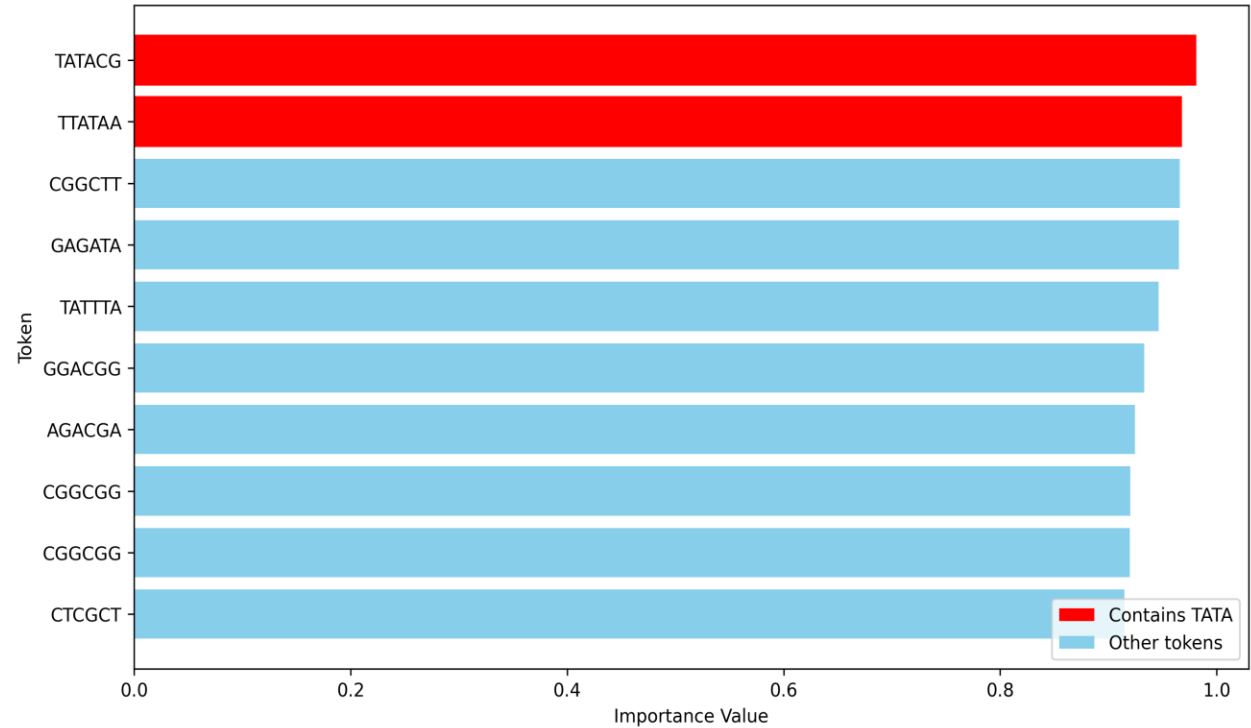
➤ Model Used : Nucleotide Transformers

➤ Token size = 6

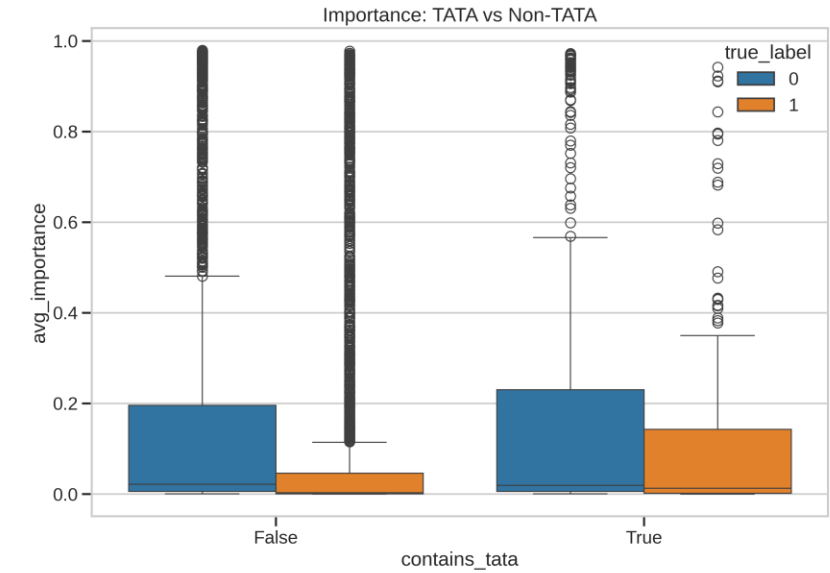


# ANCHOR

Row 459 — True: 0, Pred: 0



Category	Count
Contains TATA	554 (9.34%)
Lable 0 with TATA	407
Label 1 with TATA	147
TATA in Label 0, Negative	0
TATA in Label 0, Positive	407
TATA in Label 1, Negative	0
TATA in Label 1, Positive	147



# Limitations

- Resource
- Time
- Others

# Conclusion

- Not only "TATA" sequence is enough for detecting promoter



**What??**  
**You Want**  
**More??**



**THANK YOU**

**END OF PRESENTATION**