Predicting Transcription Levels of Bacterial Promoters Using Deep Learning

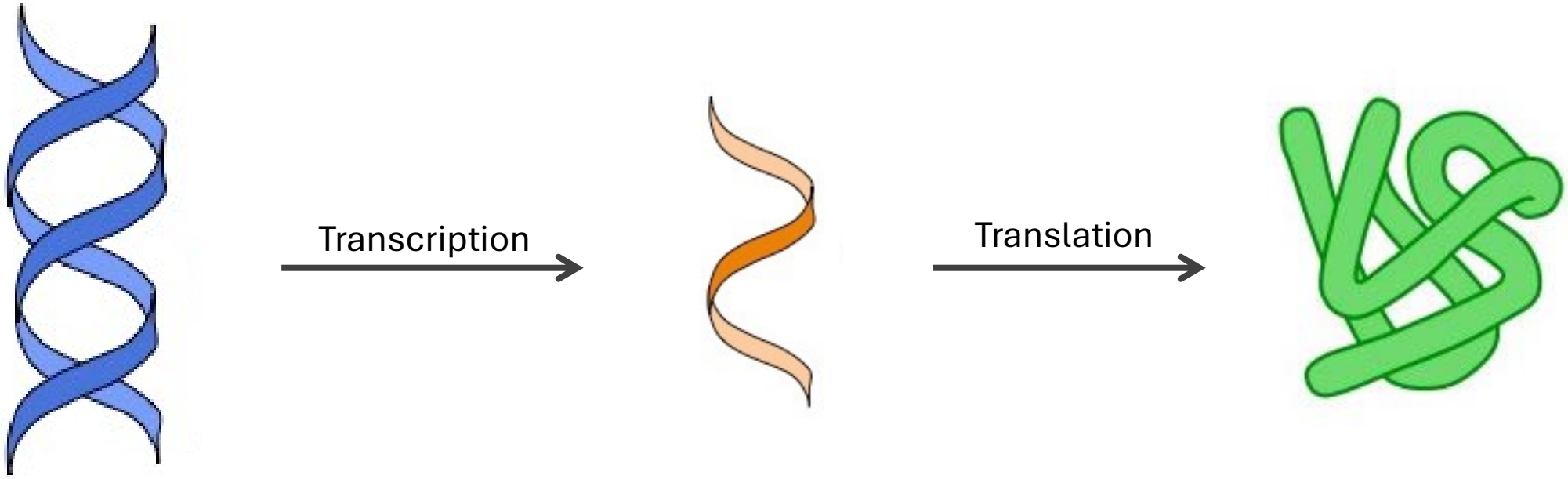
Sergio Salgado Briegas

Data Science and Machine Learning Bootcamp

31 July 2024

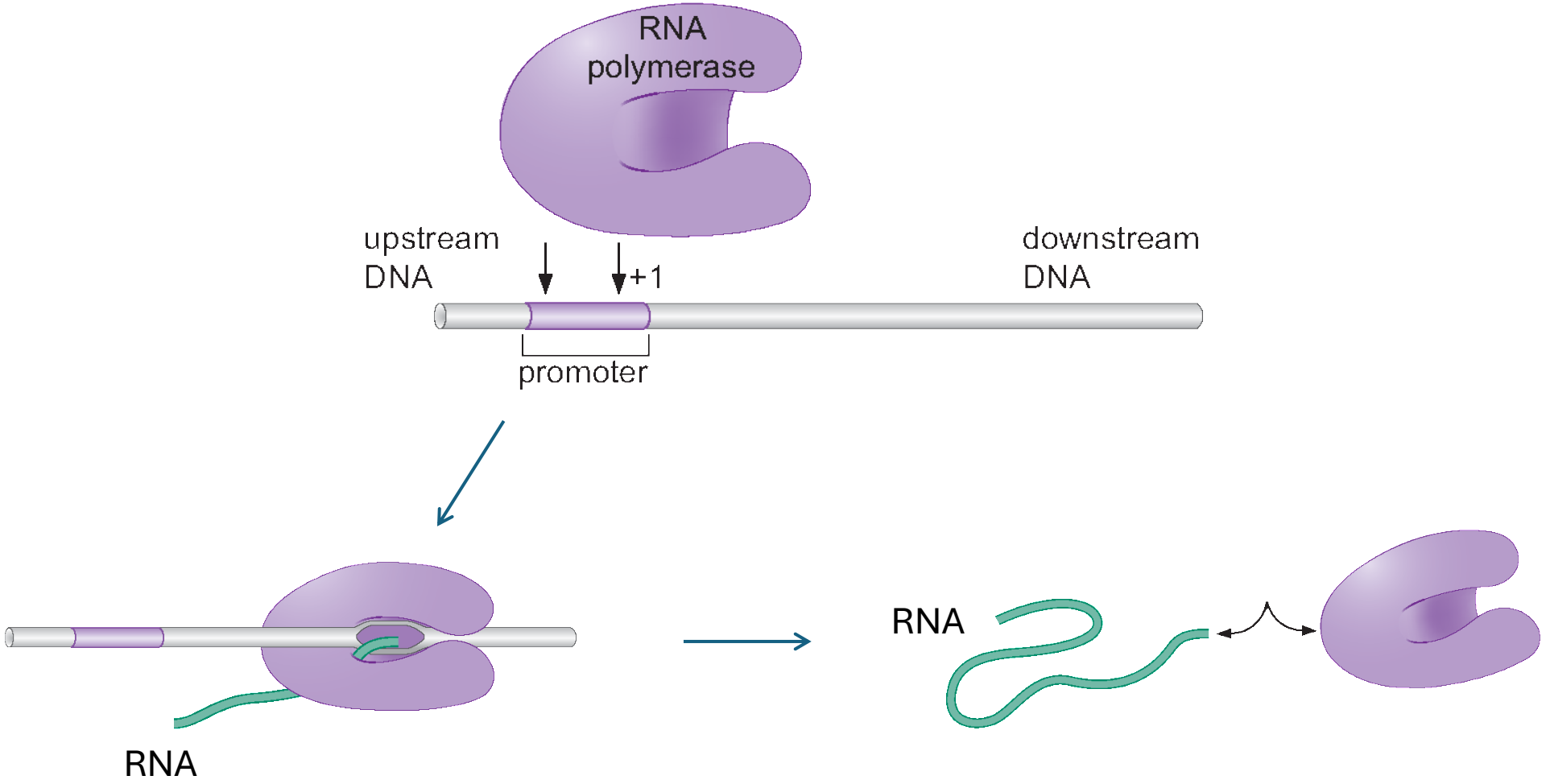
Central Biology Dogma

DNA **RNA** **Protein**

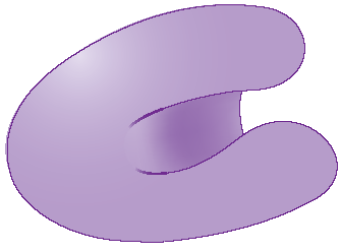


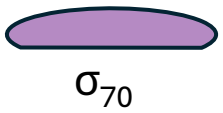
GENE EXPRESSION

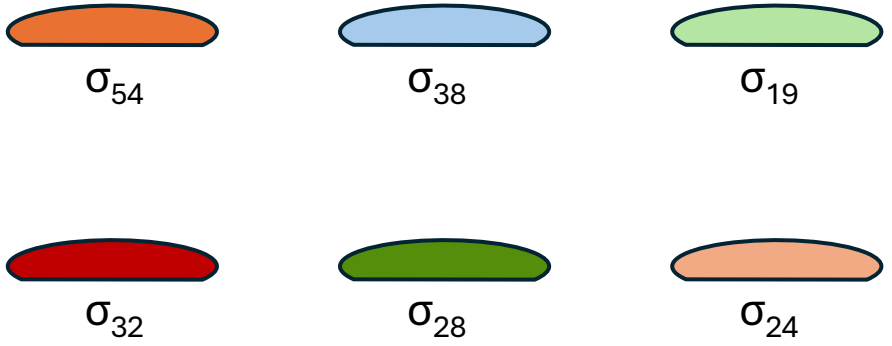
Transcription

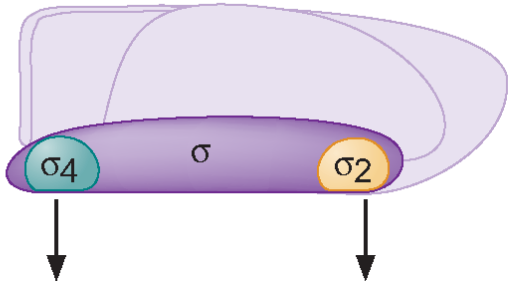


RNA polymerase. Sigma Factors (σ)

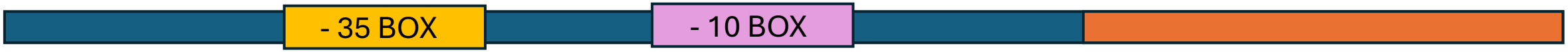




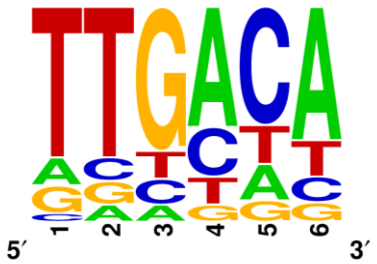




Alternate between different growth phases or stress response

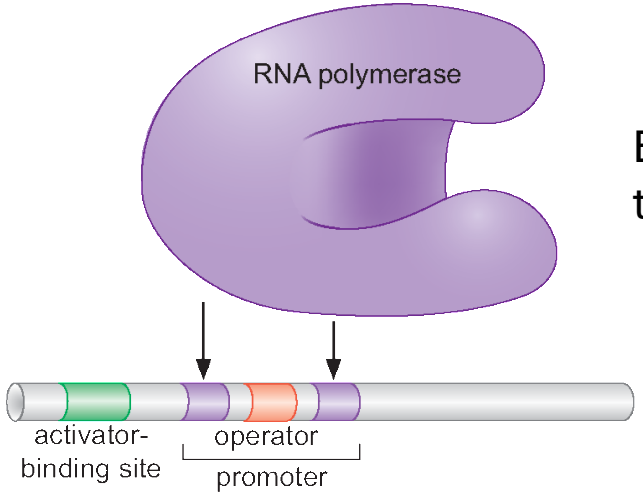


Promoter **Coding Sequence**



Sequence conservation is related to promoter strength

RNA polymerase. Regulators



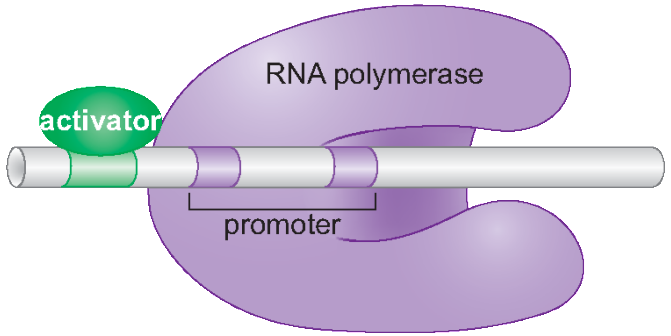
Basal levels of transcription

Repressors

15-25 nucleotides

Non-specific location

Decreased transcription



Activators

15-25 nucleotides

Before the promoter

Increased transcription

For what purpose??

Bacteria can be modified to produce relevant compounds...

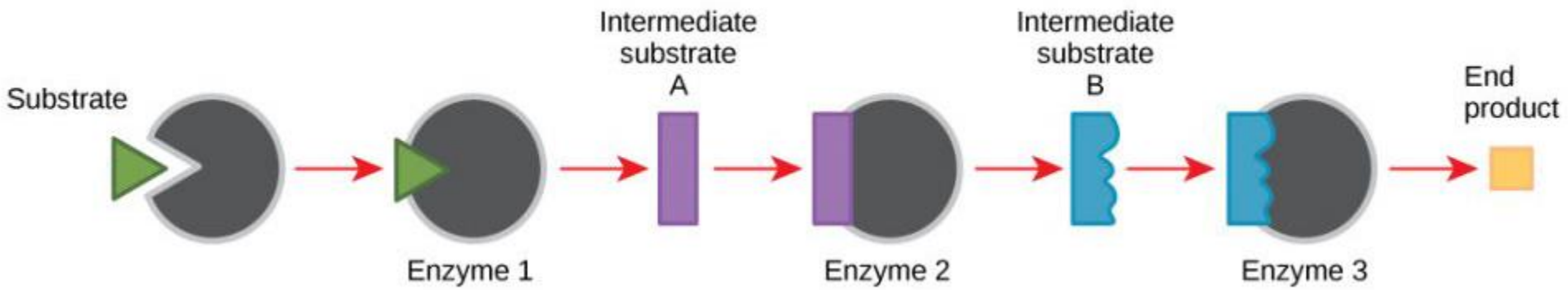
Agroindustry

Food Industry

Pharmaceutical Industry

Biotechnological Industry

...through metabolic pathways (sometimes quite complex)



Regulate the concentration of each enzyme by controlling gene expression combining promoters

Objectives

Predict *in vivo* transcription levels using bacterial promoter sequences

Correlate -35 and -10 boxes sequences with RNA concentration

Identify activator and repressor sequences

Dataset construction

RegulonDB

Comprehensive DB of *Escherichia coli*

Sigma factor (σ)	Promoter's sequences
σ_{70}	1075
σ_{38}	216
σ_{32}	96
σ_{24}	78
σ_{54}	52
σ_{28}	33
σ_{19}	1
Unknown	696
Total	**2247**

iModulonDB

Transcription Datasets for *E. coli* genes

Transcription levels of
4257 genes
166 conditions

Preprocessing

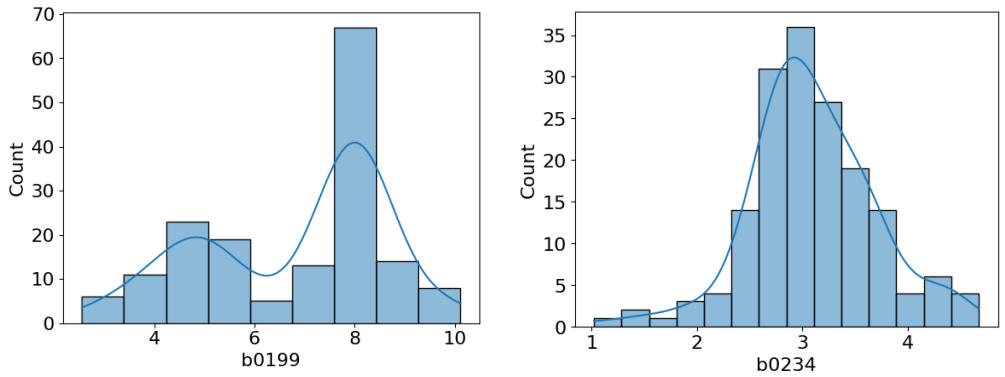
Promoter's sequences

σ_{70} promoters

Genes controlled by single promoters

Transcription levels

Differential transcription levels of genes



Mode of each gene as value

Categorize as 'Low', 'Medium' and 'Strong'

Dataset of 660 σ_{70} promoters

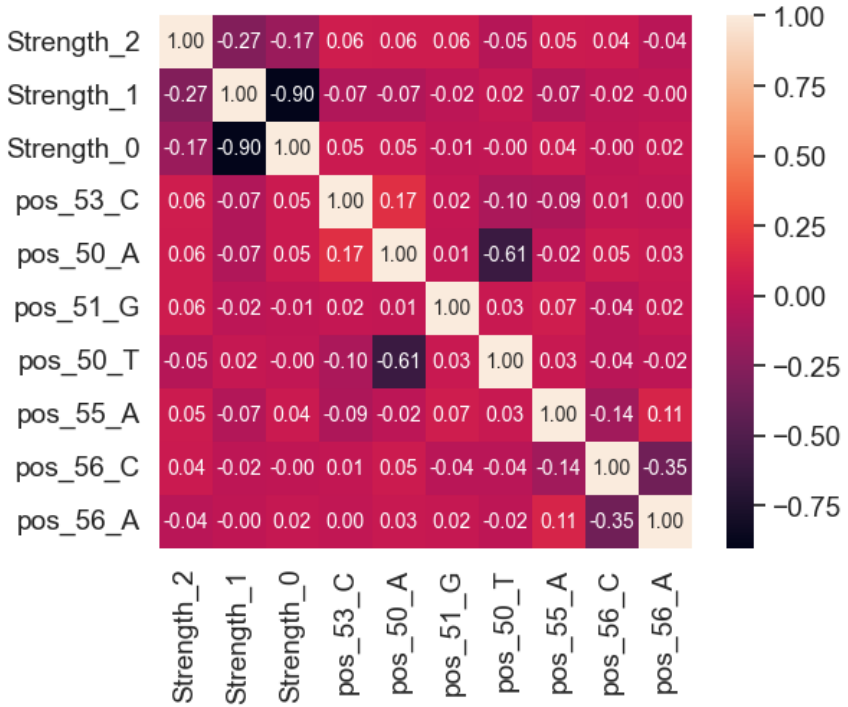
Exploratory Data Analysis

ATACATATAATAATTTAATCTTAAATGAAATTTATTAAAATTTGCAAAC**TATAAT**TTTGTGTATAAAAATATAAATGCACA appYp3

50

TTATTCACCTTTTGGCTACTTATTGTTTGAAATCACGGGGGCGCACCG**TATAAT**TTGACCGCTTTTTGATGCTTGACTCTA atplp

49

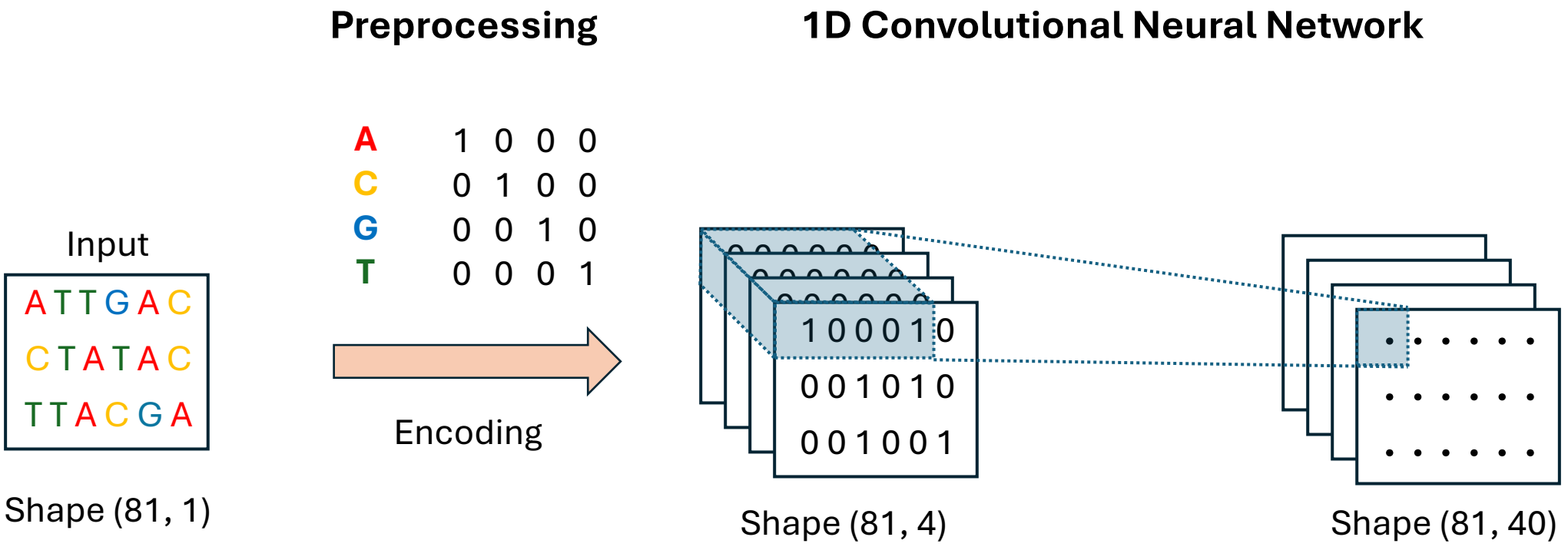


Although motifs can be conserved in sequence, they can be displaced

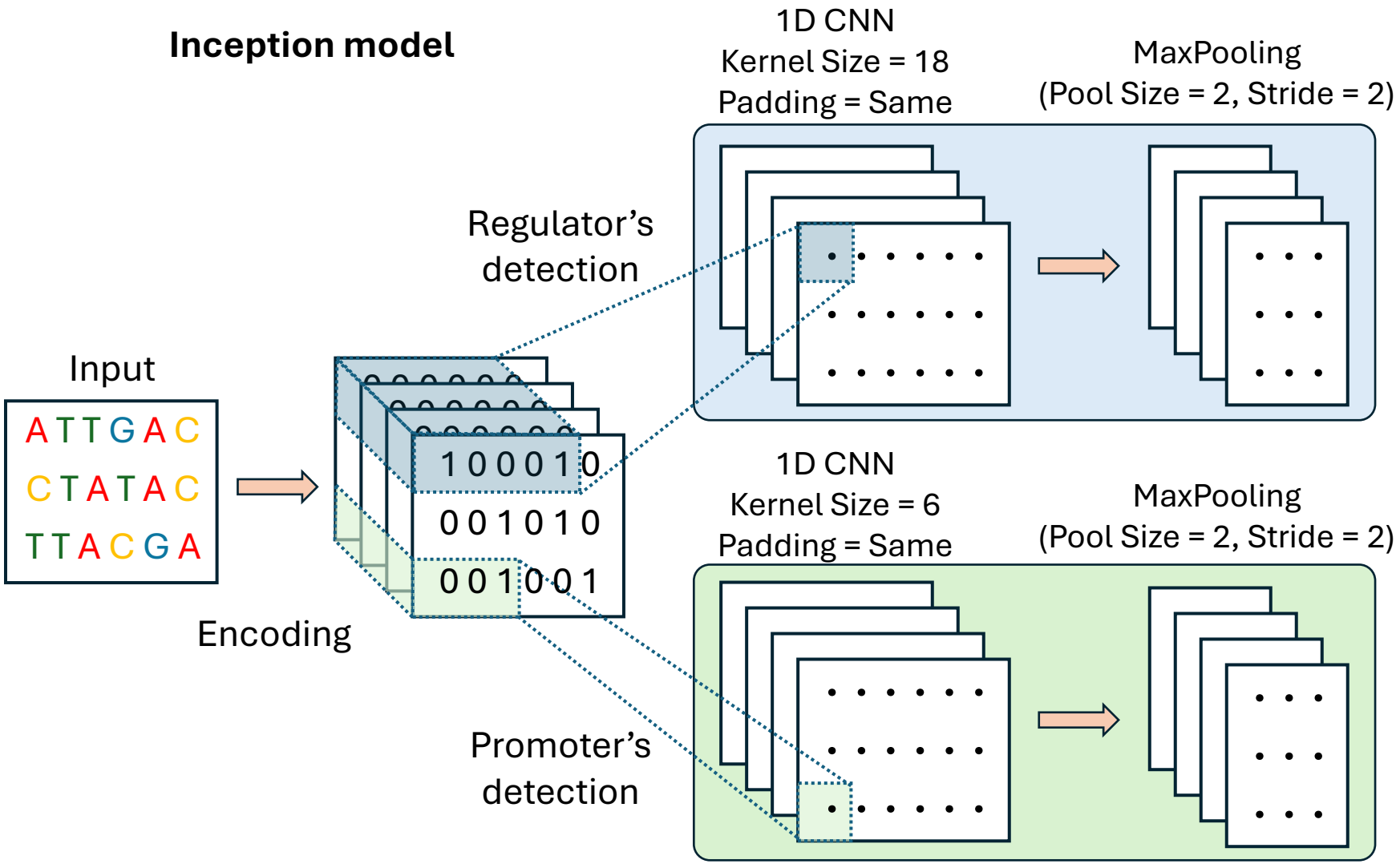
Correlations are difficult to establish

Needs for a model for pattern recognition

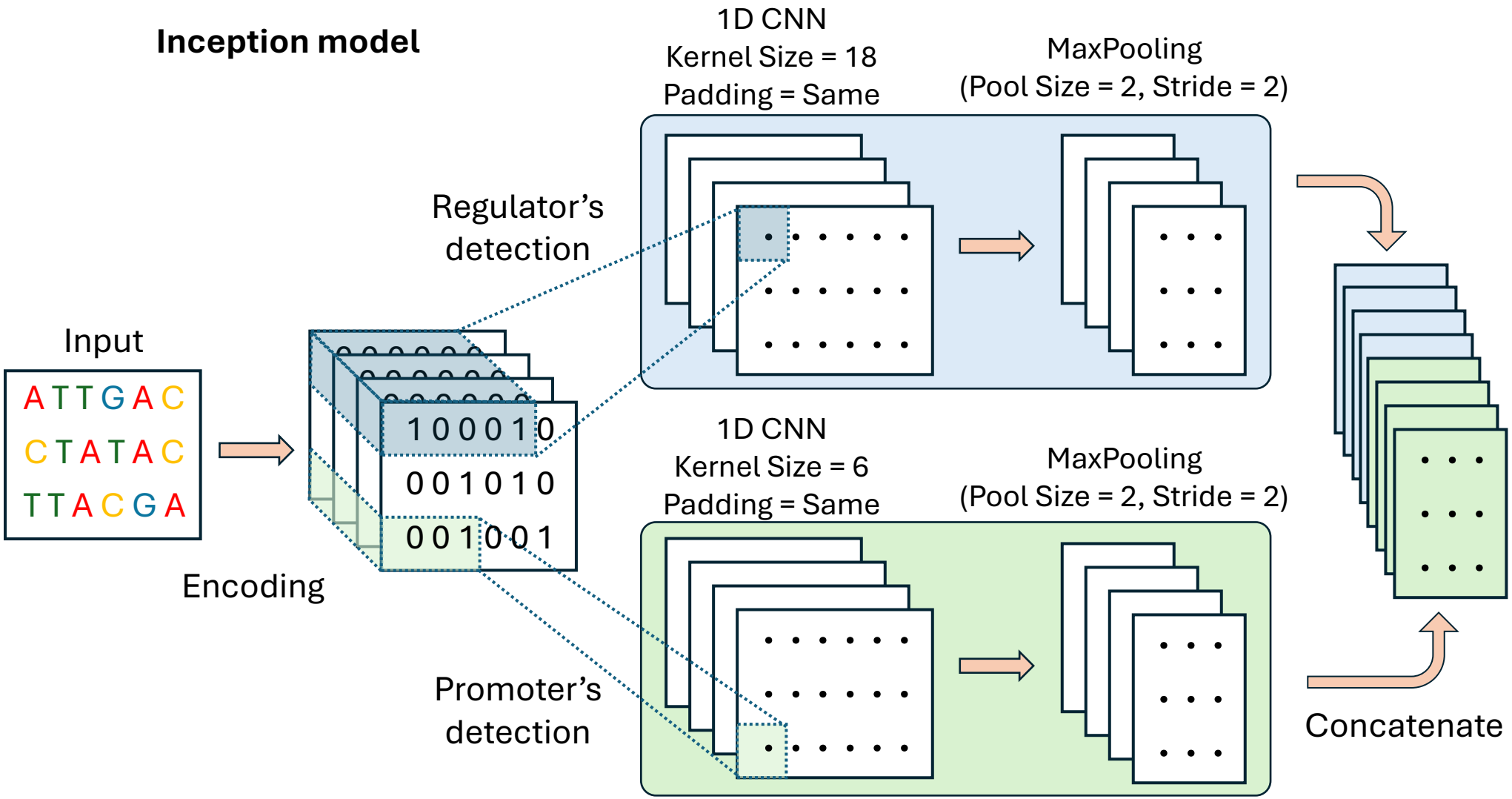
Convolutional Neural Network



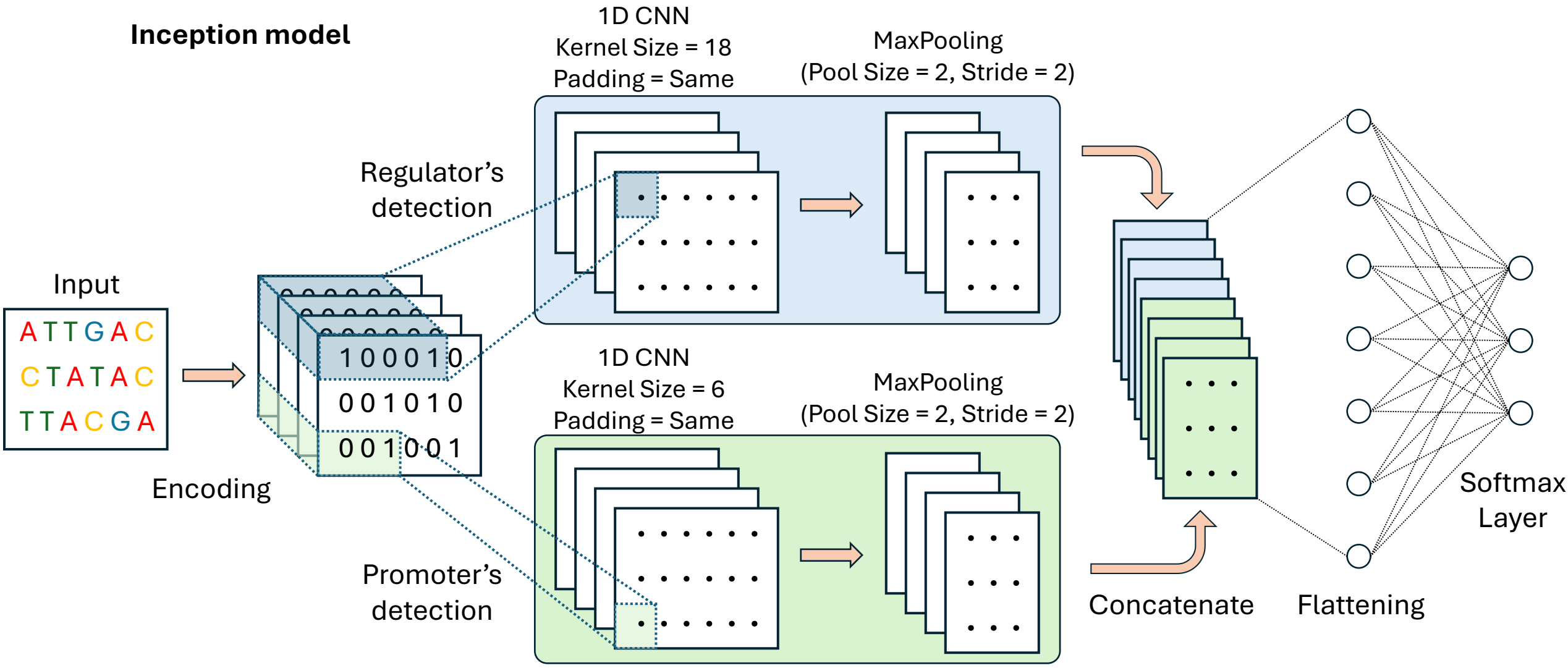
Convolutional Neural Network



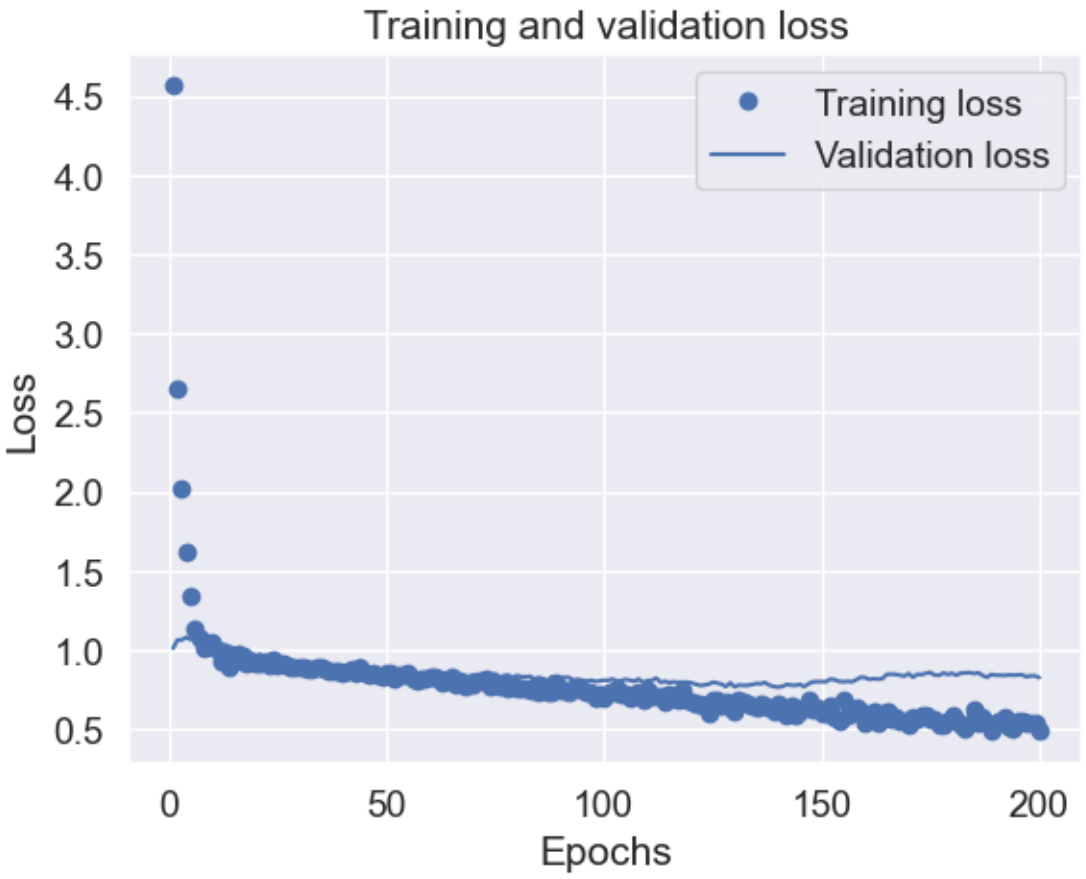
Convolutional Neural Network

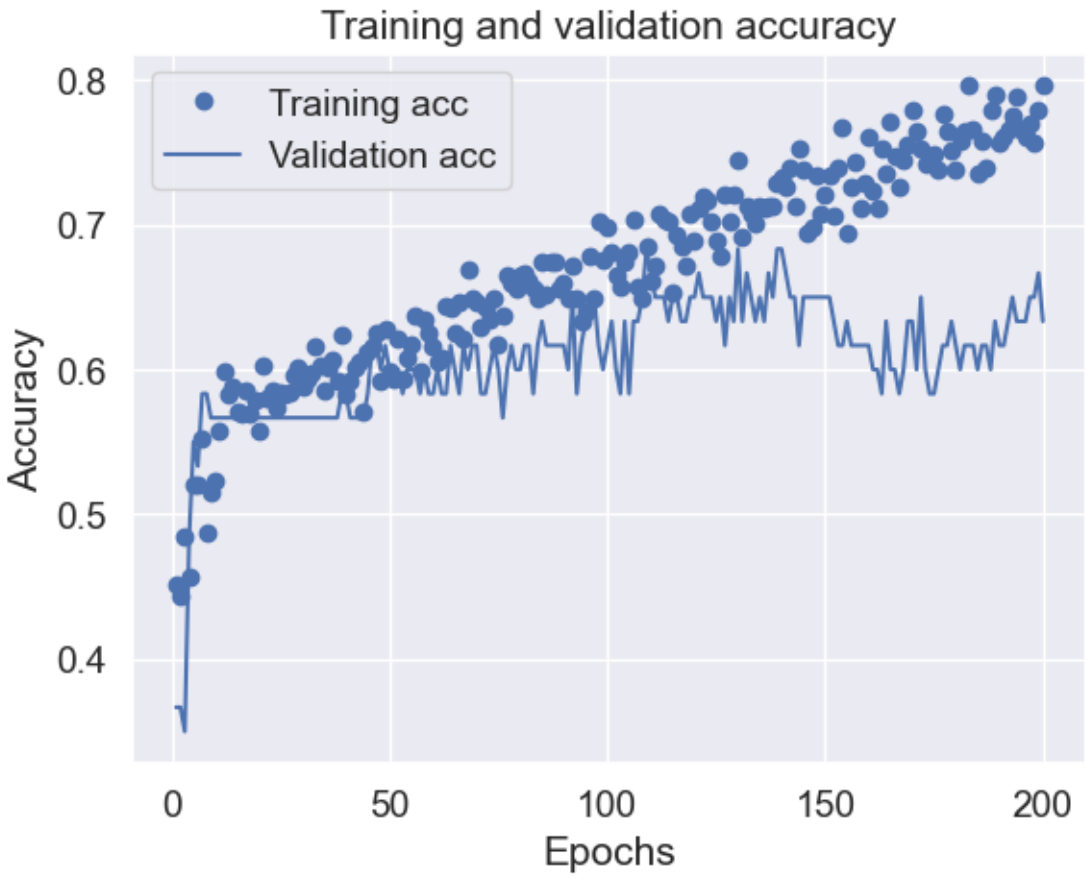


Convolutional Neural Network

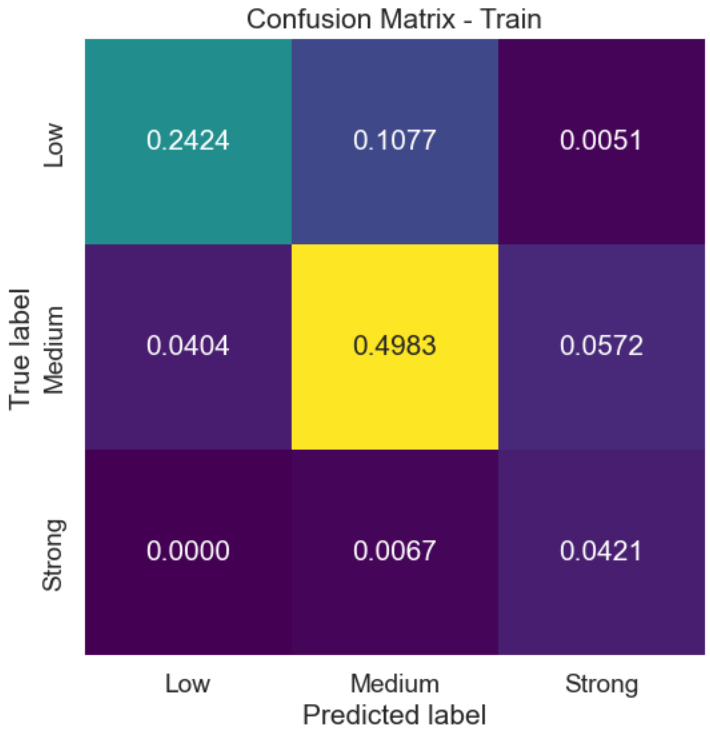


Convolutional Neural Network. Results

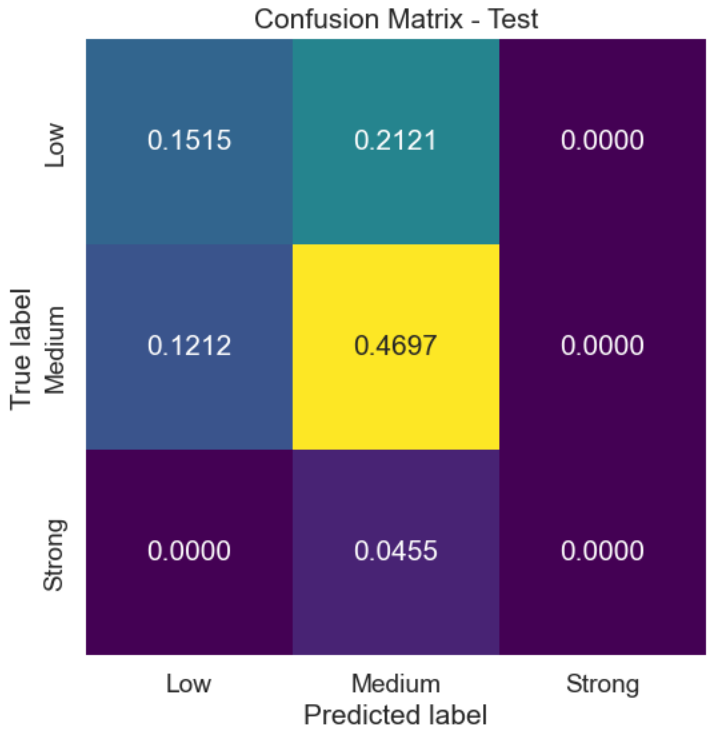




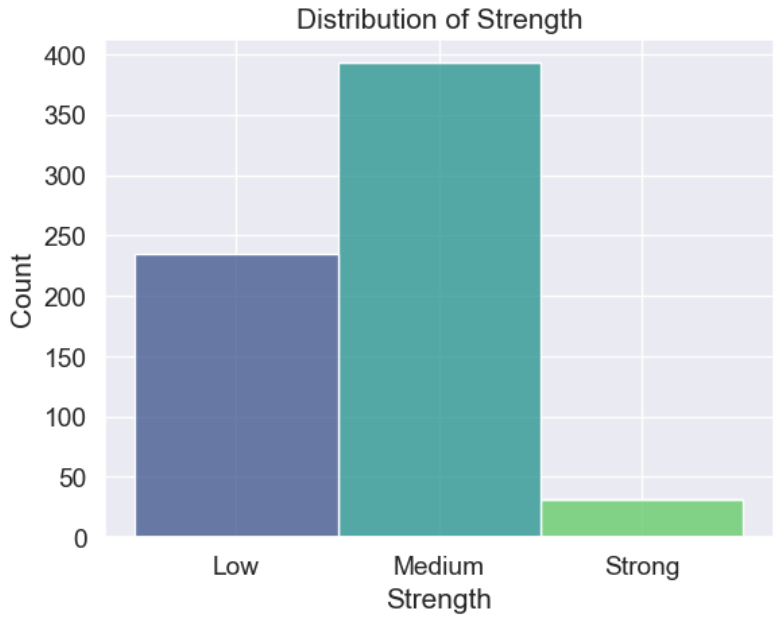
Convolutional Neural Network. Results



Loss: 0.31
Accuracy: 0.95



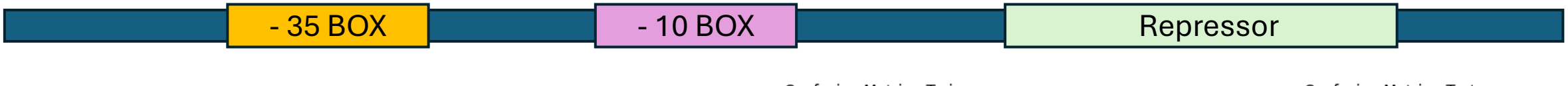
Loss: 1.05
Accuracy: 0.59

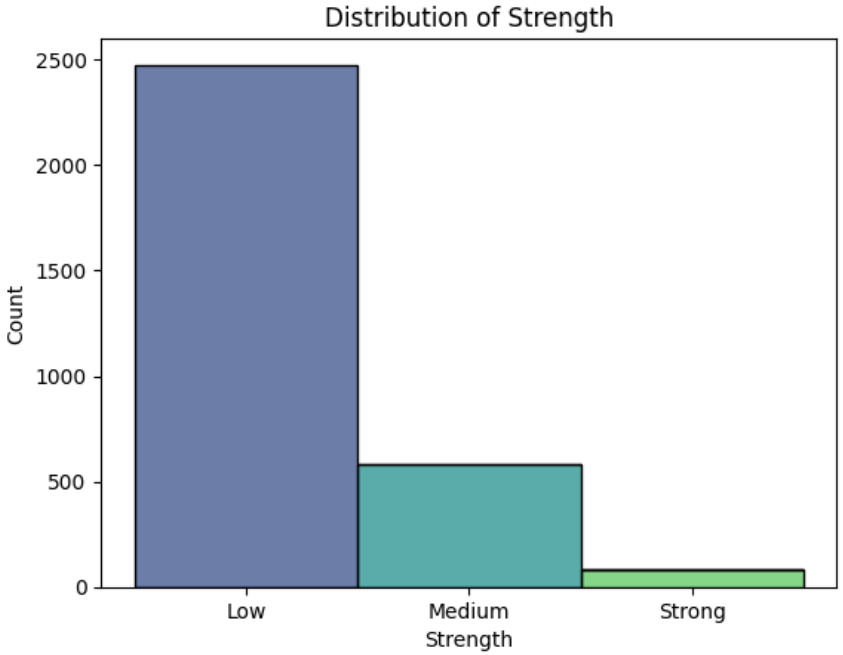


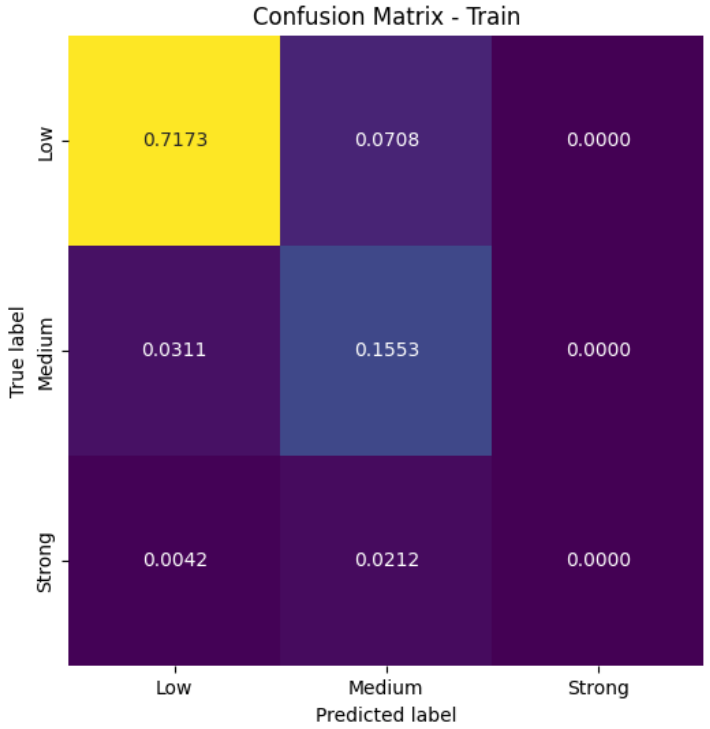
Very little data on
'Strong' promoters

Reducing complexity

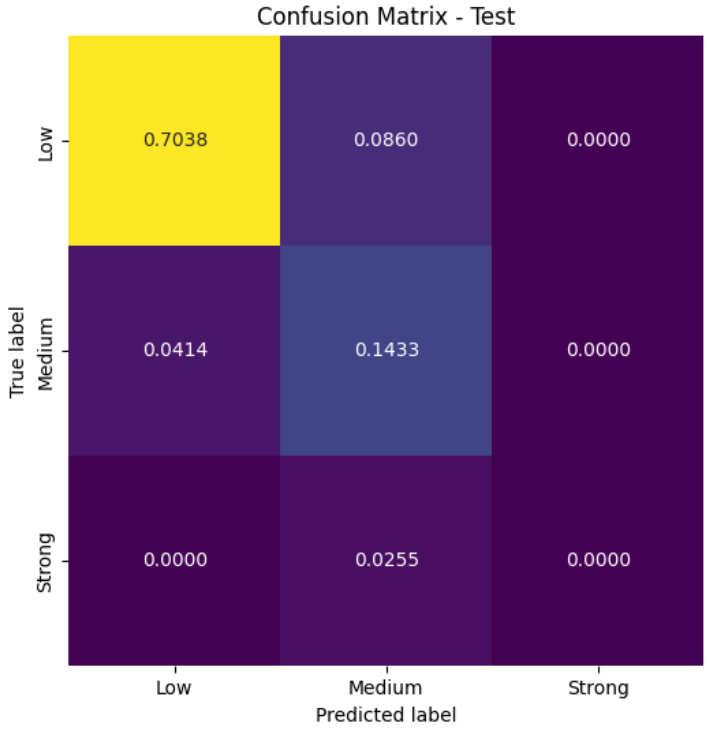
Random mutations in pTrc promoter – 3140 sequences







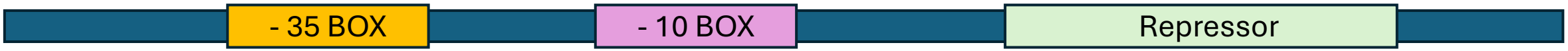
Loss: 0.30
Accuracy: 0.87



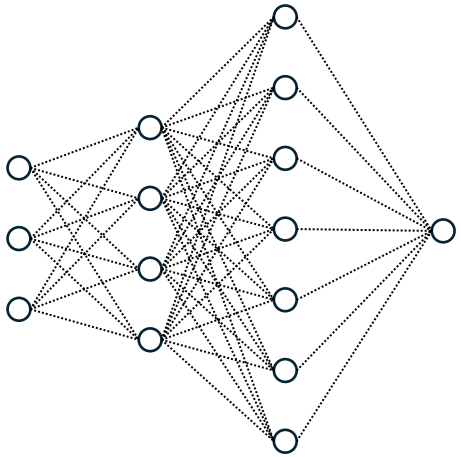
Loss: 0.40
Accuracy: 0.85

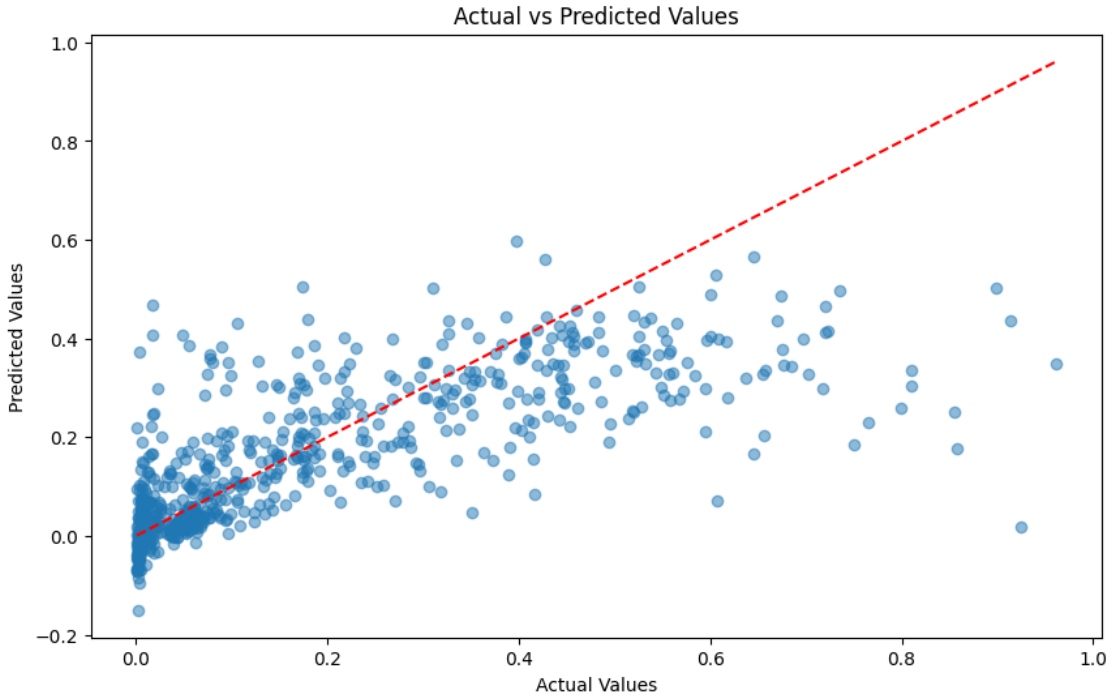
Reducing complexity

Random mutations in pTrc promoter – 3140 sequences



Regression with a Sequential Neural Network



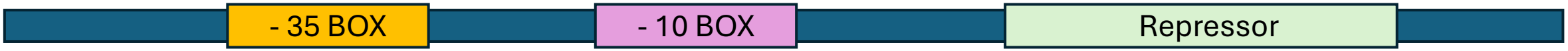


Train
MSE: 0.014
R^2 score: 0.645

Test
MSE: 0.021
R^2 score: 0.525

Reducing complexity. Machine Learning Model

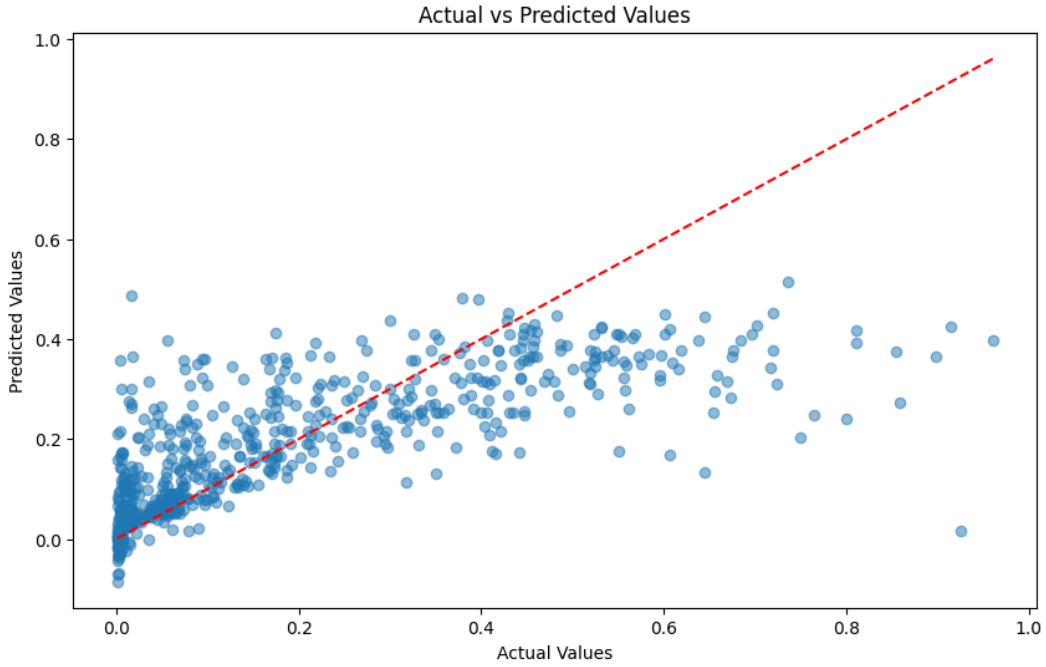
Random mutations in pTrc promoter – 3140 sequences



Regressions with a Machine Learning Algorithms

Model	R^2 score Train	R^2 score Test
Linear	0.53	0.47
KNN	0.62	0.40
DecisionTree	1.0	0.30
RandomForest	0.92	0.52
GBoost	0.54	0.48
XGBoost	0.83	0.52

XGBoost hyperparameter tuning using GridSearch CV

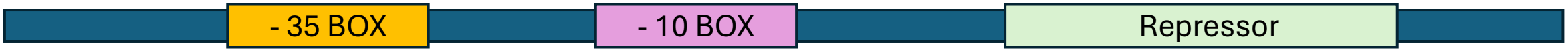


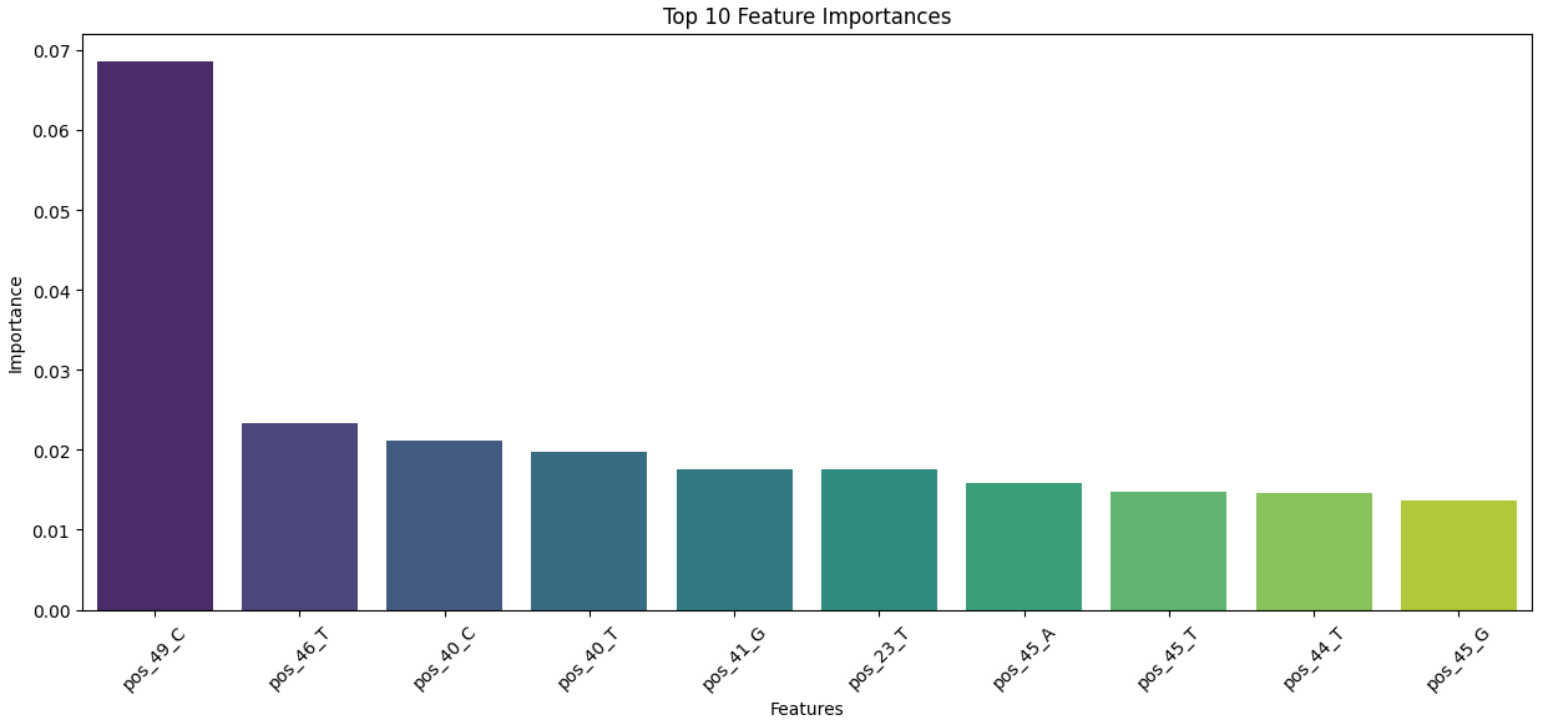
Train
MSE: 0.009
R^2 score: 0.741

Test
MSE: 0.020
R^2 score: 0.548

Reducing complexity. Machine Learning Model

Random mutations in pTrc promoter – 3140 sequences





The more important mutation (pos_49_C) is in the repressor

Conclusions

- Convolutional Neural Networks allow for the identification of patterns that facilitate the classification of promoters using their sequence and transcription data.
- Unbalanced classes lead to the incorrect identification of the underrepresented classes
- Localizing the problem in a single gene and increasing the data amount would allow for the identification of contributing nucleotides in promoter strength.

Future ideas

- Use Convolutional Neural Networks for the identification of regulators.
- Quantify the effect that regulators have on transcription.
- Incorporate regulator's transcription data into the X.

Predicting Transcription Levels of Bacterial Promoters Using Deep Learning

Sergio Salgado Briegas

Data Science and Machine Learning Bootcamp

31 July 2024

linkedin.com/in/salgado-sergio

@ salgado.sergio@protonmail.com