# Predicting mRNA Abundance Directly from Genomic Sequence Using Deep Convolutional Neural Networks

•••

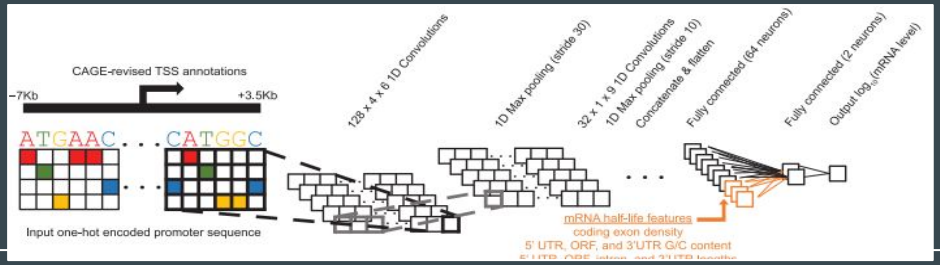
Deep Learning
IIT Jodhpur

Princy Gautam
B20BB051

# Introduction

Harnessing the power of deep neural networks to unravel the intricate relationship between genetic information encoded in promoter sequences and the resulting mRNA expression levels. This groundbreaking study not only showcases the potential of deep learning but also sheds light on the underlying mechanisms that govern gene expression.

# Implementation



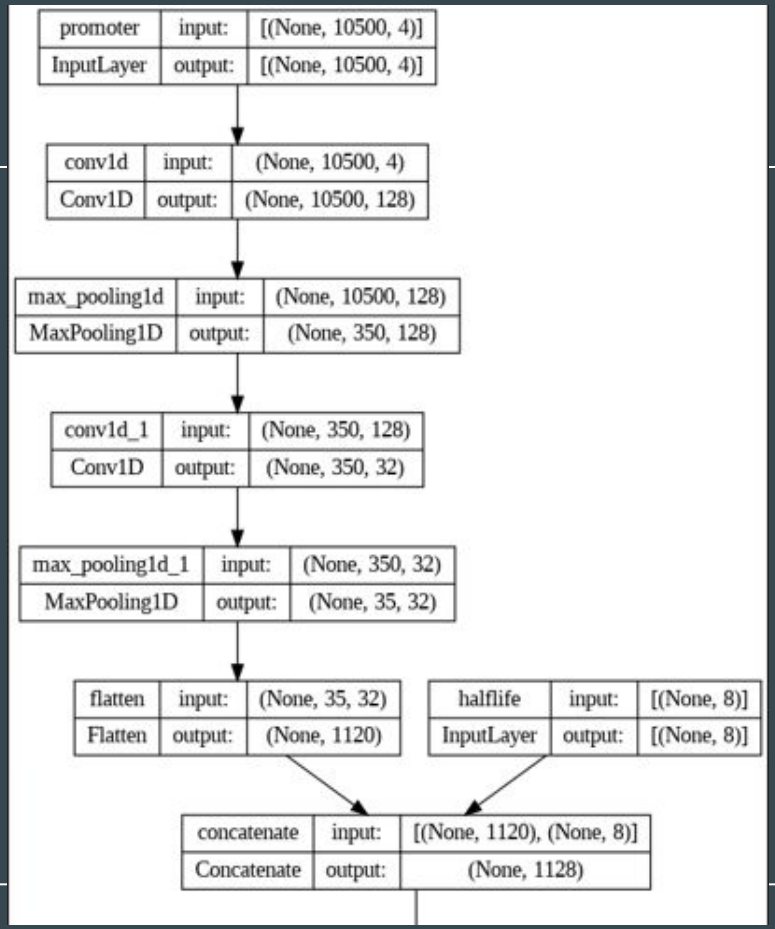

A nested dictionary structure to define different choices and ranges for hyperparameters. The values of hyperparameters are determined by the hyperopt optimization algorithm during the search process.

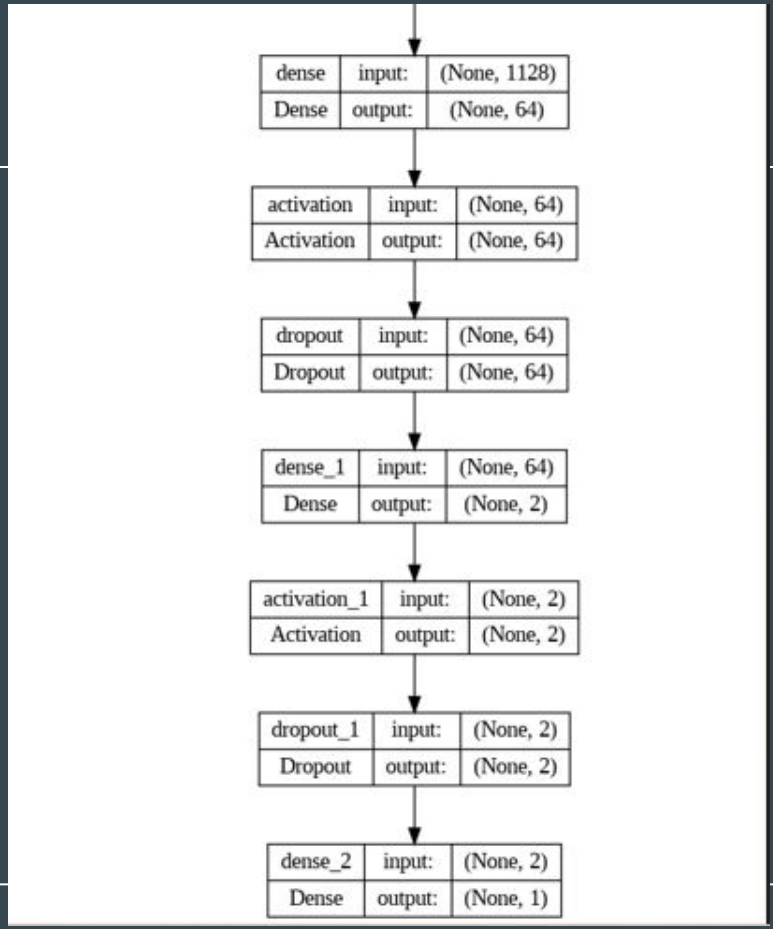
The objective function takes dictionary of parameters as input, defines model architecture, performs training and evaluation and aim to minimize the MSE loss on the validation set.

DNA sequence is taken as input and converts it into a one-hot encoded representation using a NumPy array.

Predictions are generated for input sequences and saves the predictions to an output file.

# Model


promoter | input: | [(None, 10500, 4)]
InputLayer | output: | [(None, 10500, 4)]
conv1d | input: | (None, 10500, 4)
Conv1D | output: | (None, 10500, 128)
max_pooling1d | input: | (None, 10500, 128)
MaxPooling1D | output: | (None, 350, 128)
conv1d_1 | input: | (None, 350, 128)
Conv1D | output: | (None, 350, 32)
max_pooling1d_1 | input: | (None, 350, 32)
MaxPooling1D | output: | (None, 35, 32)
flatten | input: | (None, 35, 32)
Flatten | output: | (None, 1120)
halflife | input: | [(None, 8)]
InputLayer | output: | [(None, 8)]
concatenate | input: | [(None, 1120), (None, 8)]
Concatenate | output: | (None, 1128)


dense | input: | (None, 1128)
Dense | output: | (None, 64)
activation | input: | (None, 64)
Activation | output: | (None, 64)
dropout | input: | (None, 64)
Dropout | output: | (None, 64)
dense_1 | input: | (None, 64)
Dense | output: | (None, 2)
activation_1 | input: | (None, 2)
Activation | output: | (None, 2)
dropout_1 | input: | (None, 2)
Dropout | output: | (None, 2)
dense_2 | input: | (None, 2)
Dense | output: | (None, 1)

# Results

Observed R-squared value for the predictions made by the model on the test set = 0.008 and Mean squared error = 1.019

```
Test R^2 = 0.008
Rows & Cols: (1000, 3)
Best Validation MSE = 1.019
```

Predictions generated on a small dataset:

predictions.txt

| | Gene | Pred | Actual |
|---|---|---|---|
| 2 | b'ENSG00000102547' | b'0.005055092' | b'0.18310068655135697' |
| 3 | b'ENSG00000254901' | b'0.005055092' | b'0.476182929606854 43' |
| 4 | b'ENSG00000138380' | b'0.005055092' | b'-0.1387473122816179' |
| 5 | b'ENSG00000103489' | b'0.005055092' | b'0.4328246299405266' |

# Future prospects

The features most strongly associated with increased steady-state mRNA abundance in both the human and mouse corresponded to ORF exon density and 50 UTR GC content, followed by weaker associations to 50 UTR length and ORF length. Thus mRNA abundance with ORF and GC content features might produce benchmarking results.

The model can be used as a hypothesis generation engine to uncover additional gene regulatory mechanisms that further explain outliers.

# Applications

In synthetic biology, the model could be used to design synthetic promoters with tunable levels of transcriptional activity.

For the medical genetics purposes, methods such as Xpresso could be used to interpret the functional consequences of genetic mutations or indels within promoters on gene expression levels,

# References

https://www.sciencedirect.com/science/article/pii/S2211124720306161

https://egg2.wustl.edu/roadmap/data/byDataType/rna/expression/57epigenomes.RPKM.pc.gz

https://www.semanticscholar.org/paper/A-deep-auto-encoder-model-for-gene-expression-Xie-Wen/7b0832a59aa2dd6059665cd1062928444ab29f07

https://github.com/vagarwal87/Xpresso

https://www.semanticscholar.org/paper/A-new-LSTM-based-gene-expression-prediction-model%3A-Wang-Li/97dcdc4186c3eab15042ed942376bde2cf2a1847

# Thank You