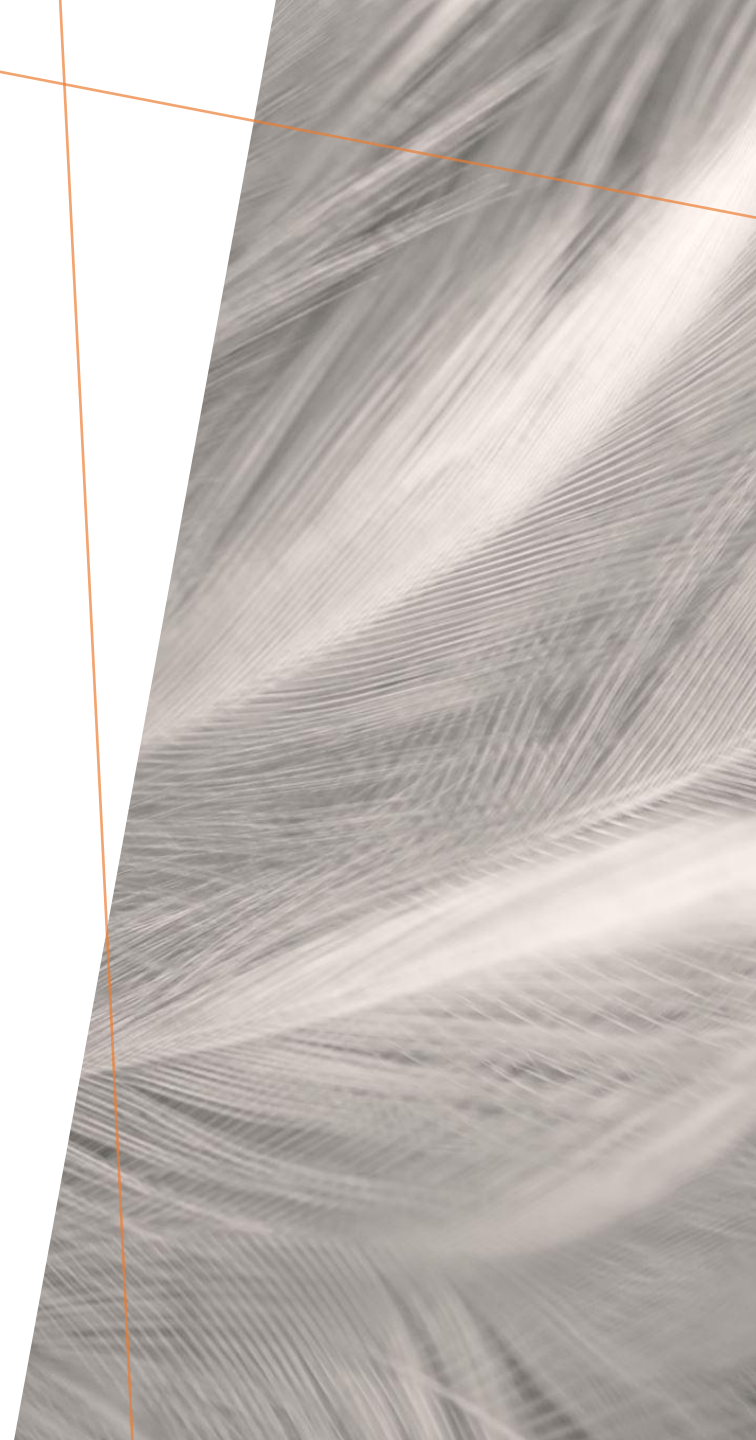
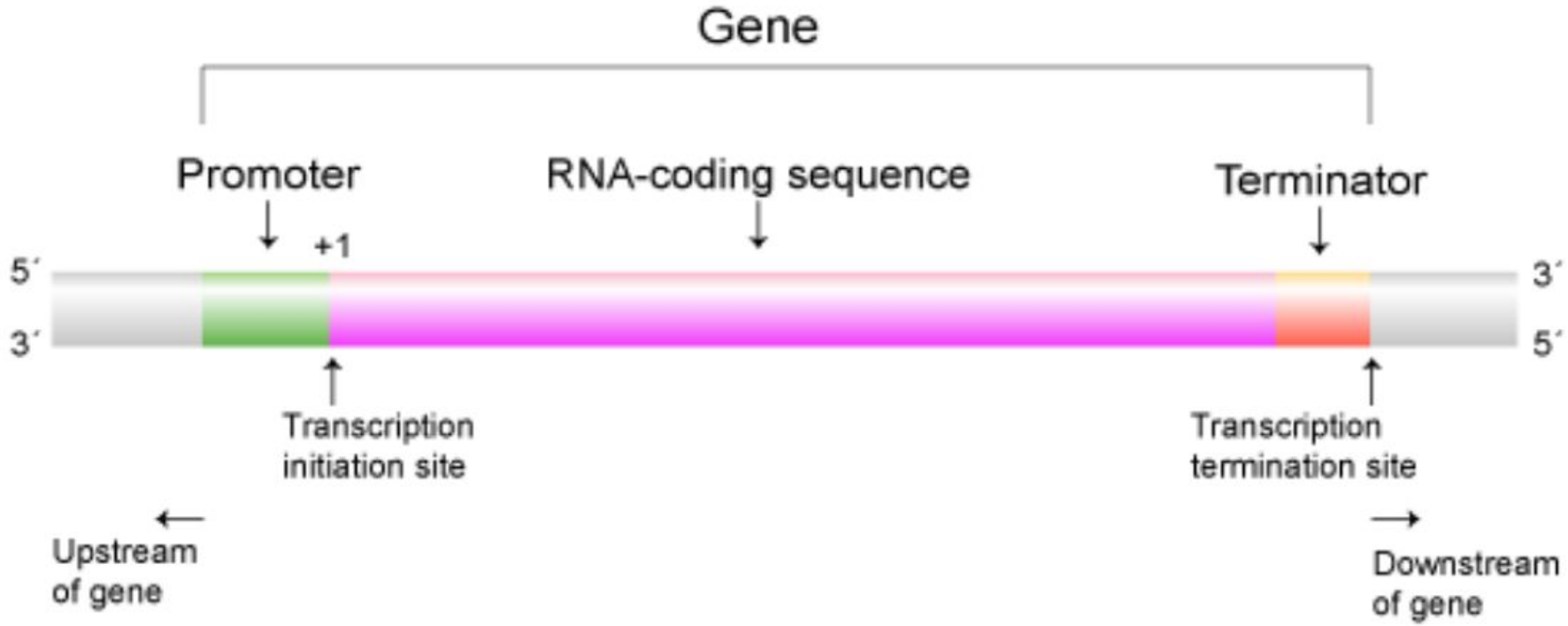
CS6024 Project:

ResNet Models for Promoter Classification

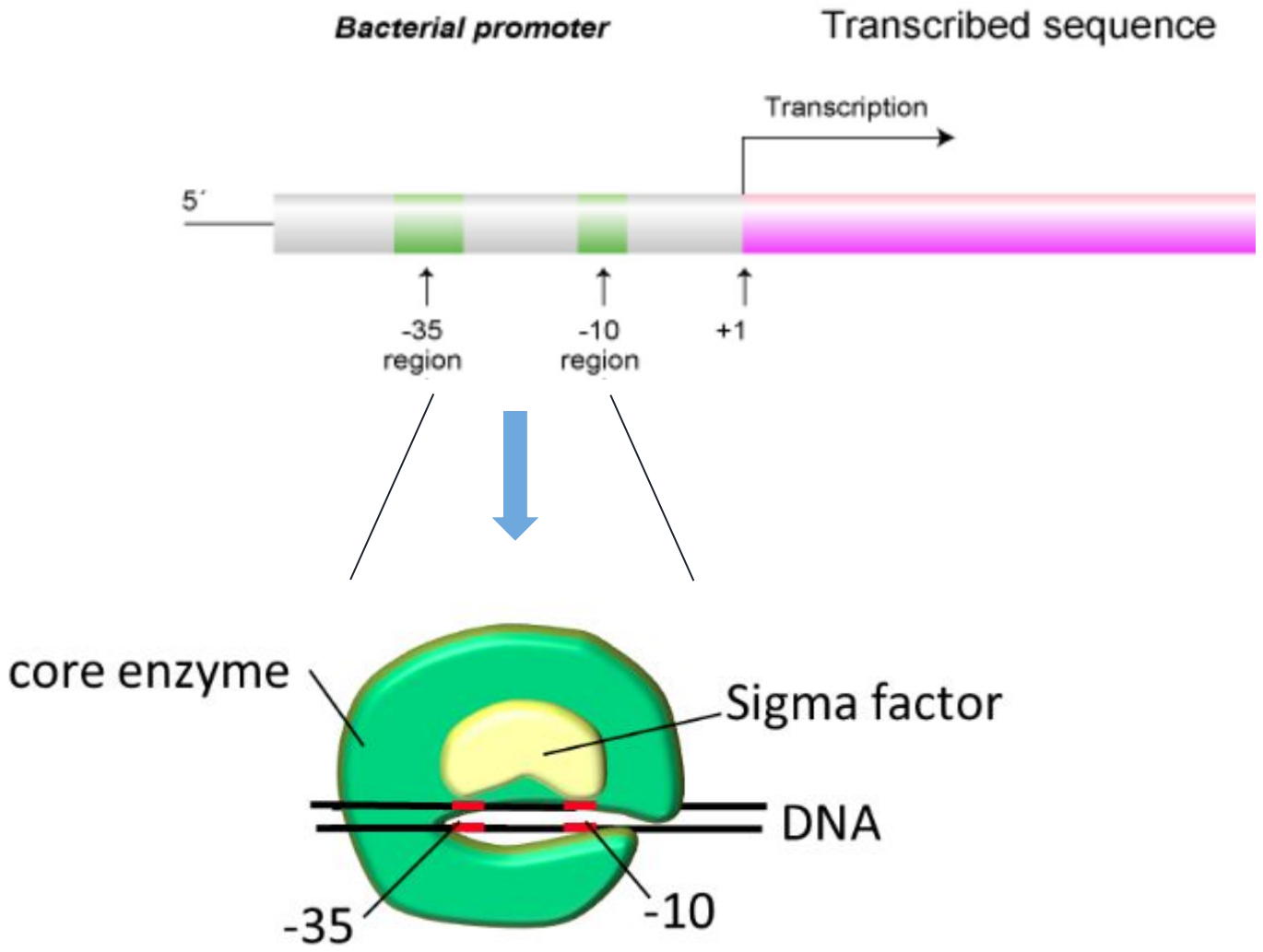


What are Promoters?



Promoters are short sequences in DNA which are responsible for initiating RNA transcriptions of a downstream gene. They are usually ~81 nucleotides long.

What are Sigma Promoters?



Promoters that σ-factors bind to are called σ-promoters.

Based on the σ-factor that binds to the promoter, it performs different functions such as

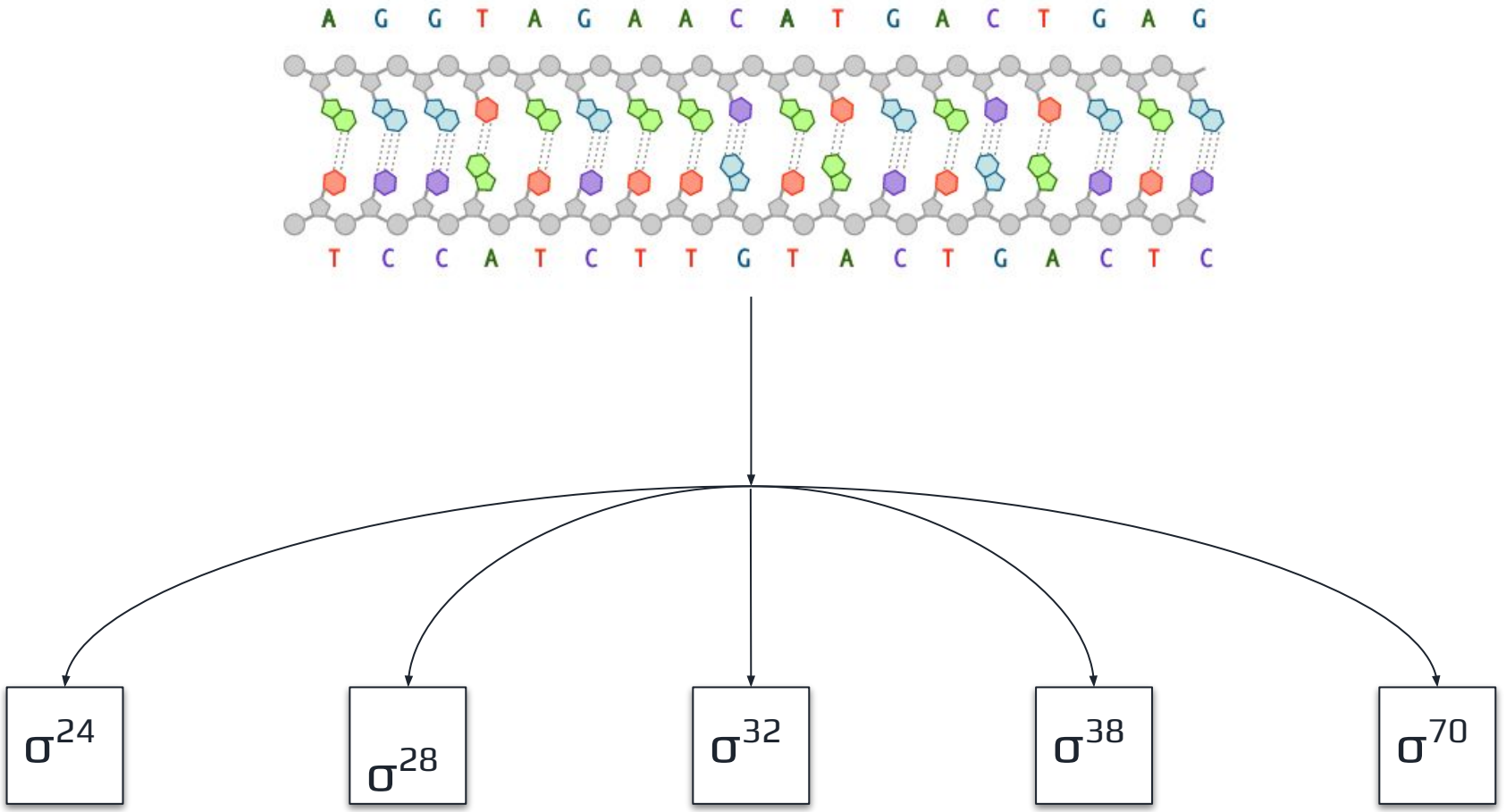
- heat stress response
- flagellar synthesis
- nitrogen limitation

In this project we attempt to classify the promoters based on the sigma factor that bound to them.

Arnhav Datar(CS18B003)

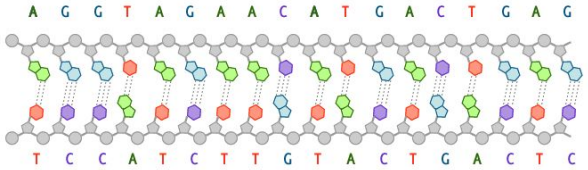
Anirudh Ajith(CS18B070)

Is this a computational Problem?



Our Method

Formatting the input



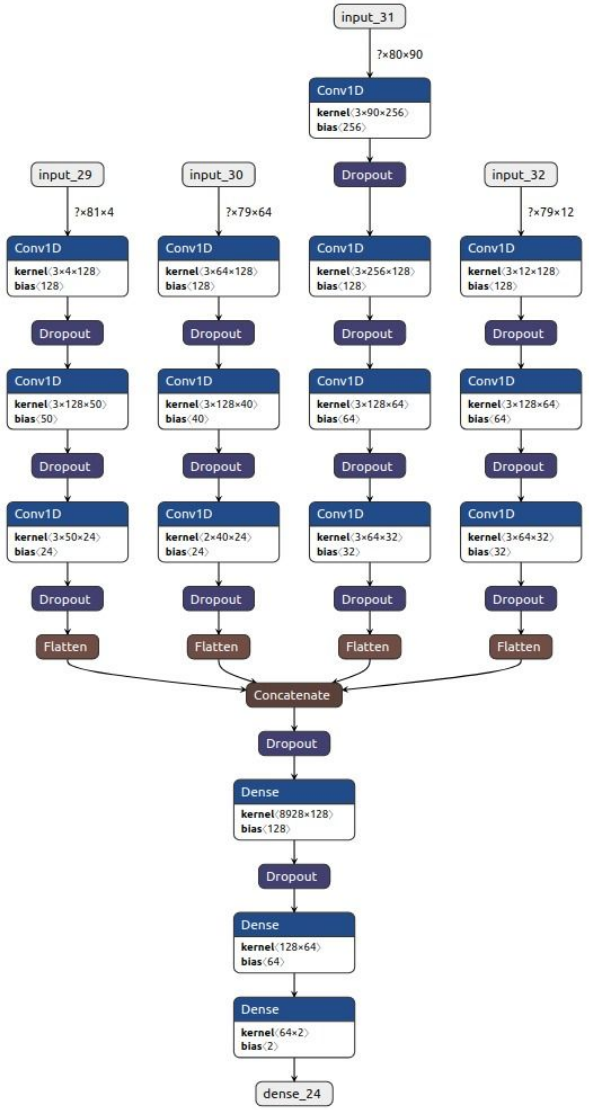
- Shape (81,4); One hot matrix for each one-mer
- Shape (79,64); One hot matrix for each trimer

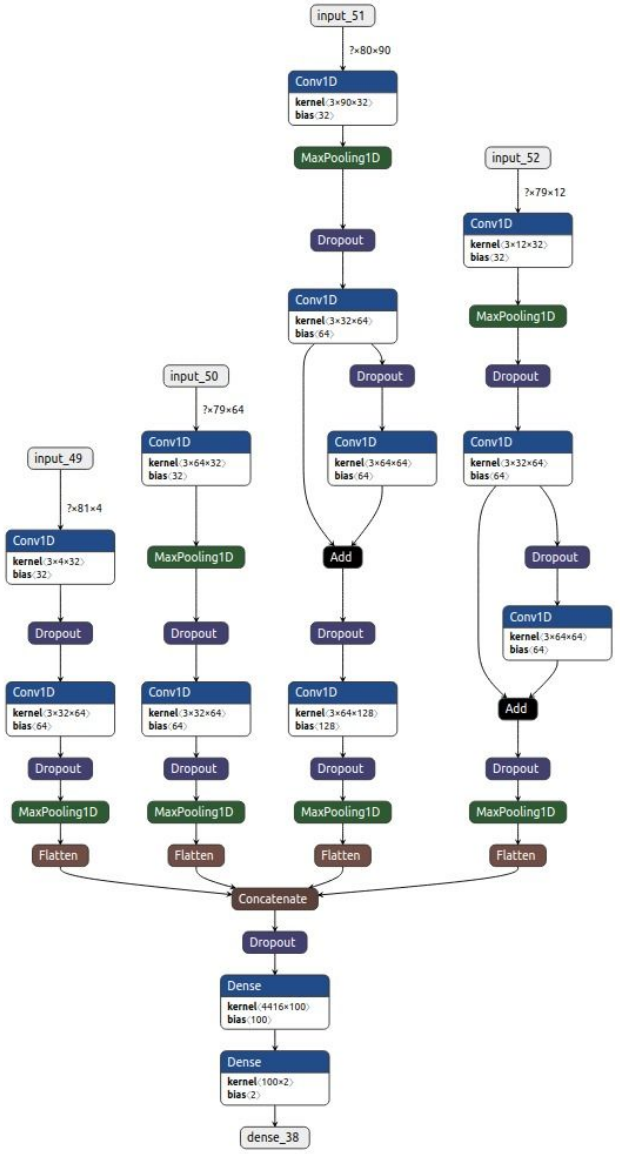
One Hot encoding Matrices

- Shape (80,90); 90 structural properties for each of the 80 dimers.
- Shape (79,12); 12 structural properties for each of the 79 trimers.

Structural properties Matrices

Our Method





Original Paper

Proposed Network

Arnhav Datar(CS18B003)

Anirudh Ajith(CS18B070)

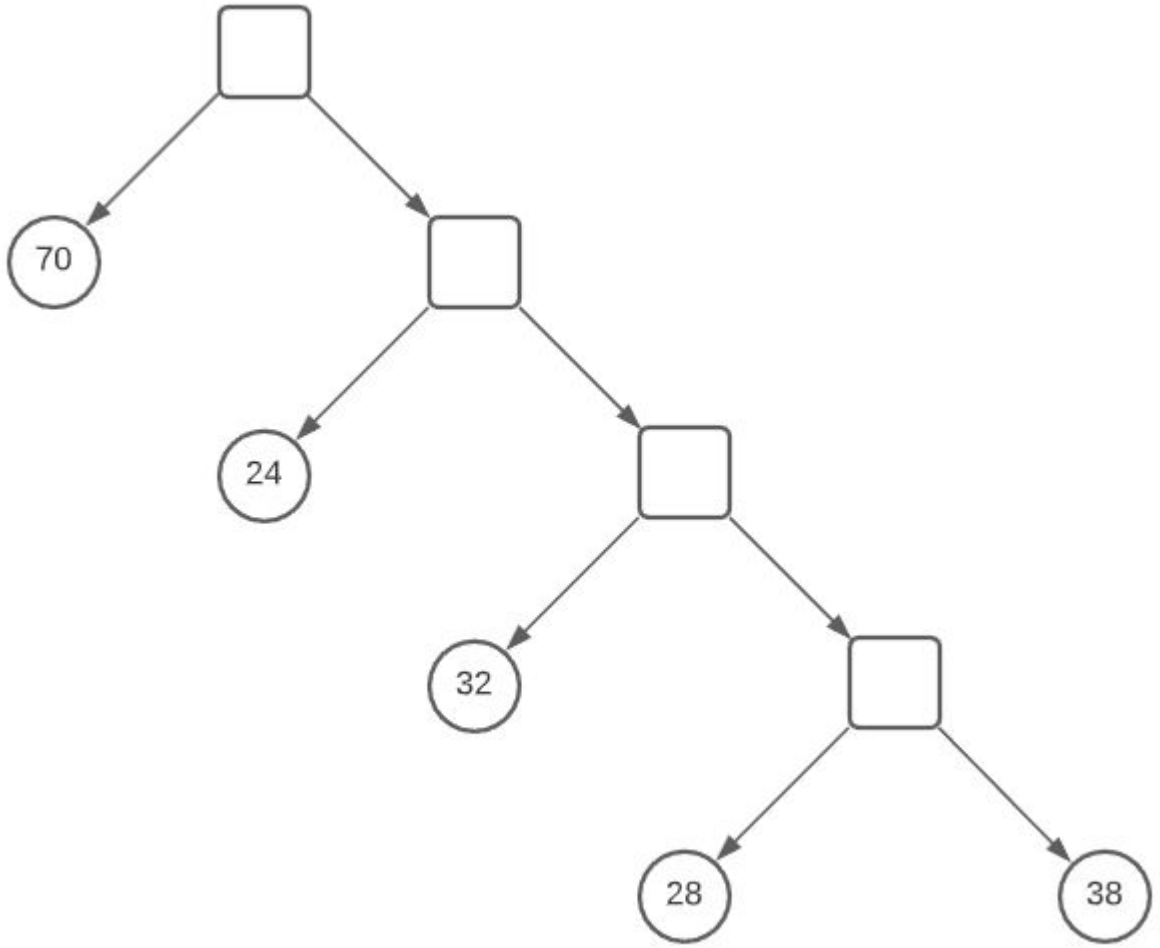
Our Method

Because our dataset was heavily imbalanced, we could not do a 5-way classification, as it leads to lower accuracy.

We therefore do the 5-way classification by doing 4 binary classifications.

The order (70,24,32...) was chosen as the descending order of the size of the promoter datasets.

This helped ensure that each binary classifier had positive and negative classes of roughly equal size.

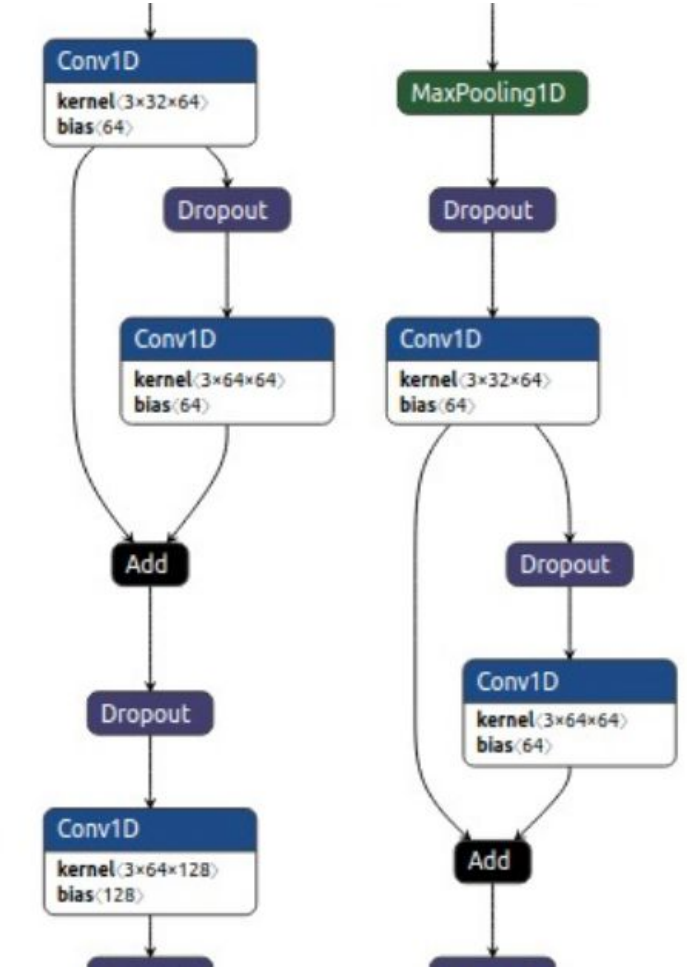


Salient Features

Addition of Skip Connections

As proposed, we add skip connections to the Convnet. However it was observed that the performance improves significantly only when we add skip-connections to the dimer and the trimer structural properties branches.

We did not see any improvement in performance when we added skip connections the to the one-hot-encoding branches.

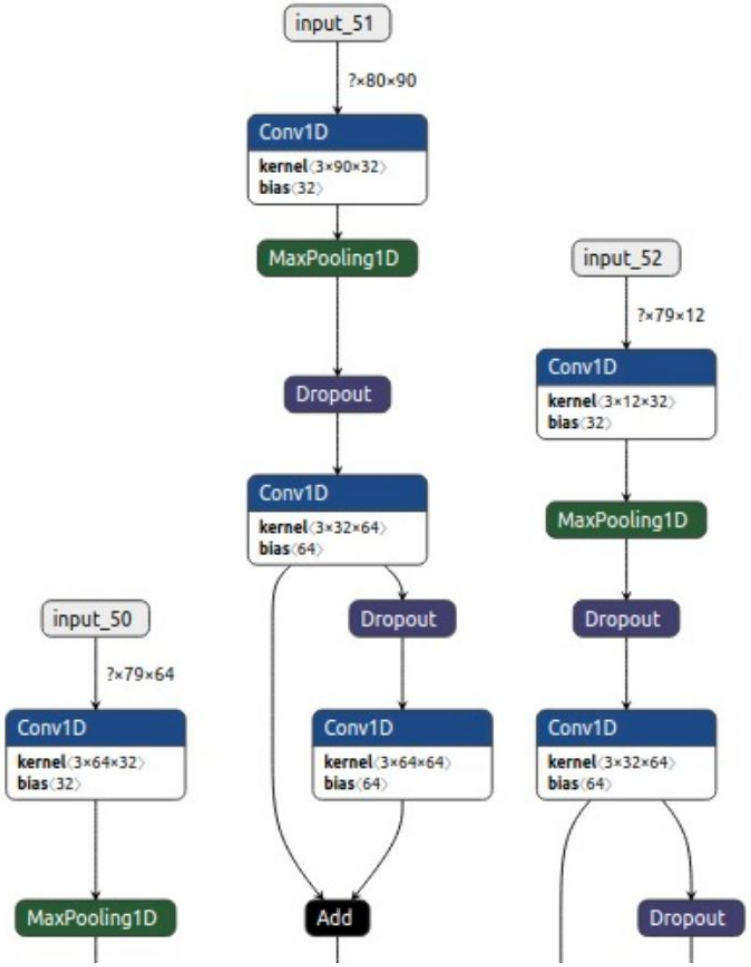


Salient Features

Max-Pooling Layers at the start based on structure

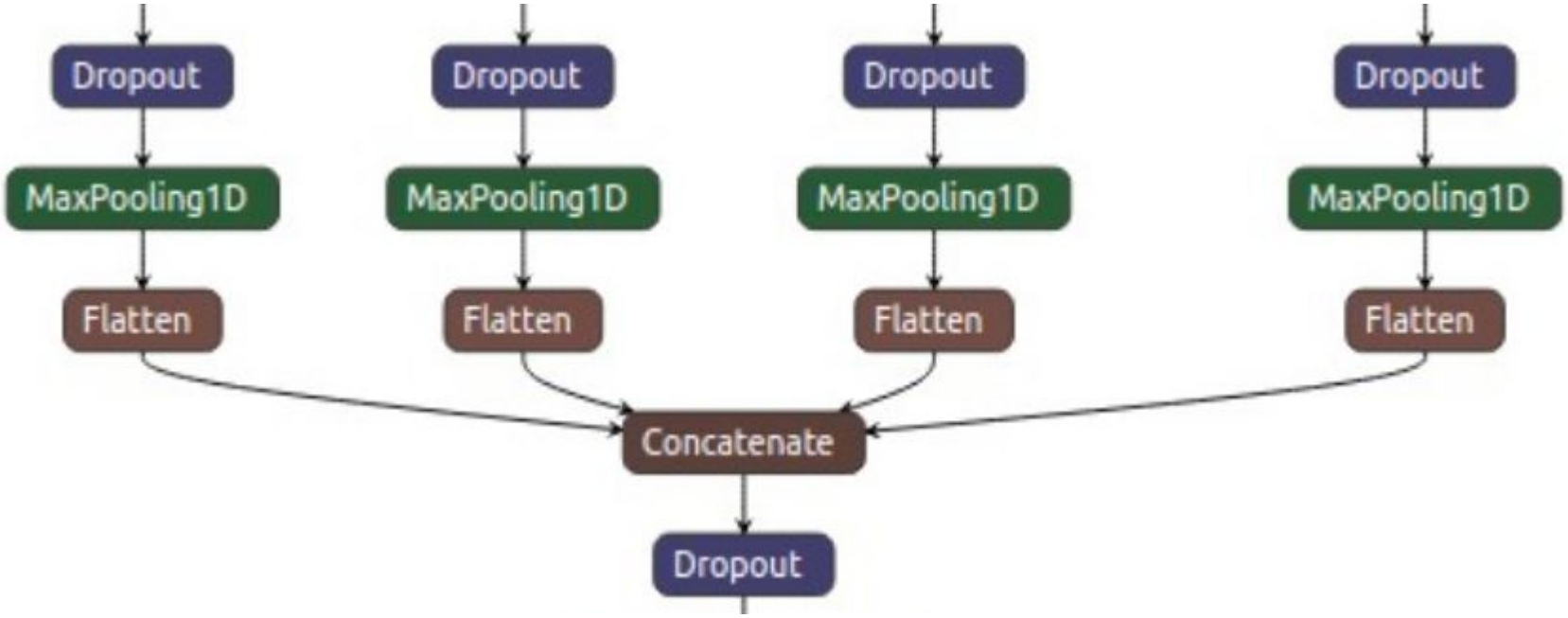
Another observation was addition of max-pooling layers to the dimer properties, trimer properties and one-hot trimer matrices improved performance.

We used pool size and stride size of 3 for the tri-mer properties and one-hot tri-mer matrices, and pool size and stride size of 2 for the dimer properties matrix.



Salient Features

Max-Pooling Layers at the end



We also add max-pooling layers at the end of each convolutional path before concatenation to reduce the number of parameters.

Results

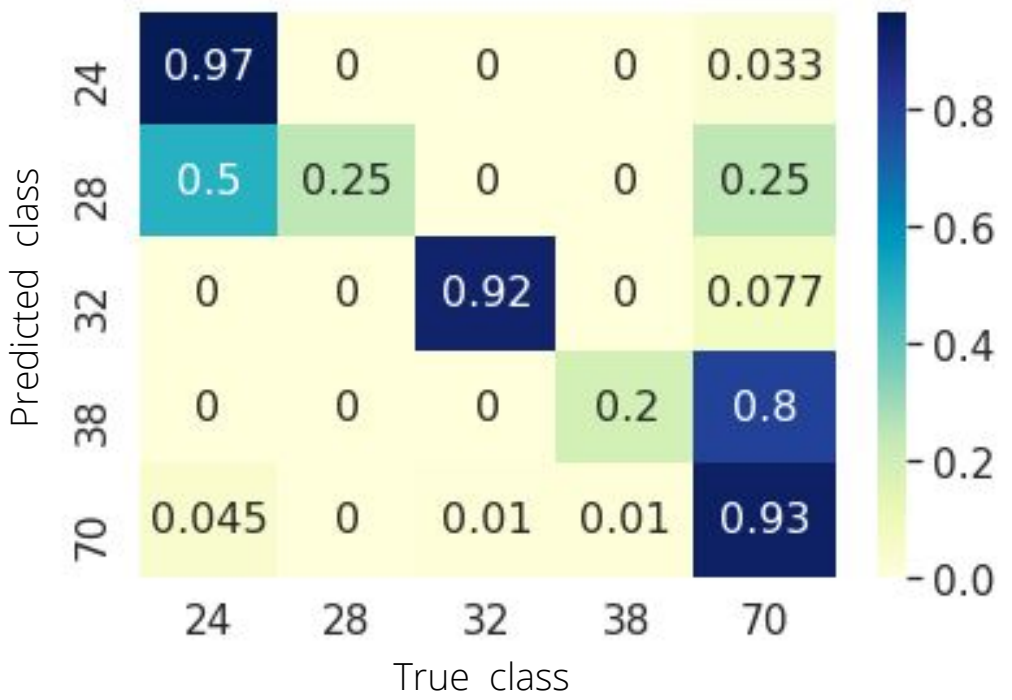
Improved Accuracy on the test-set

Why have σ^{28} and σ^{38} not performed well?

1. Lack of training and test data
   a. Test size of 4 and 10.
   b. Training size of 134 and 163.

2. By comparison the Sigma 70 promoter has train and test sizes of 1694 and 199 respectively.

3. This lead to a massive class imbalance.



Confusion Matrix with 90% Accuracy

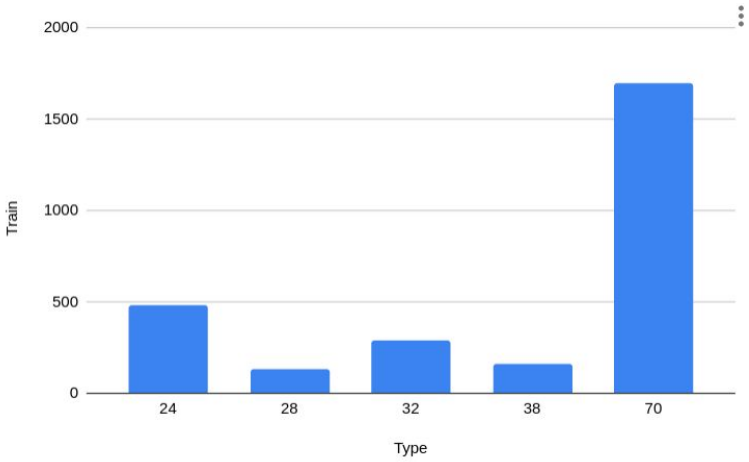
Further Work

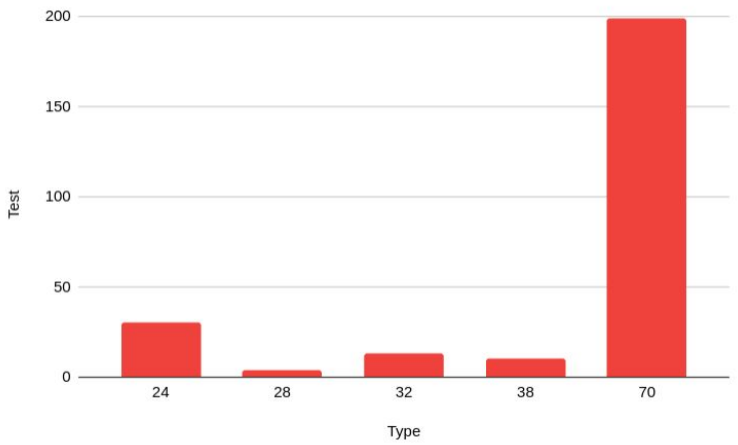
1. We are yet to do a k-fold cross validation to get the best hyper-parameters. This may further improve the accuracy.

2. The dataset which was used as a test dataset was a rather small dataset. To further solidify the results, a larger dataset is required

3. There is a significant imbalance in the dataset right now. A more balanced dataset is required.

4. Attention models which have shown promise in the field of sequence based classification can be used.





Arnhav Datar(CS18B003)

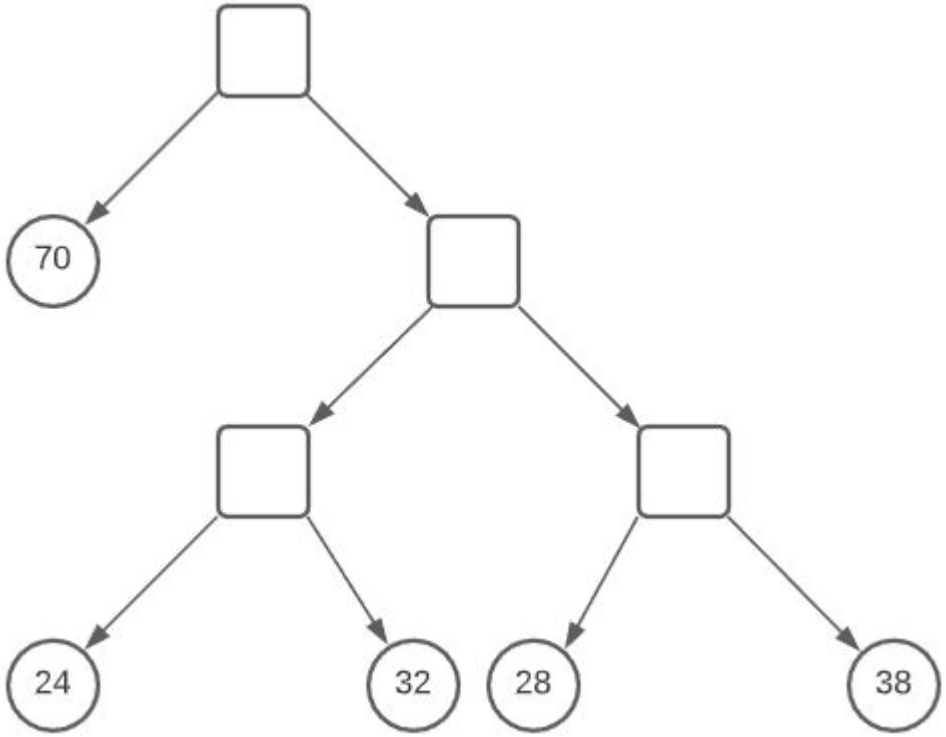
Anirudh Ajith(CS18B070)

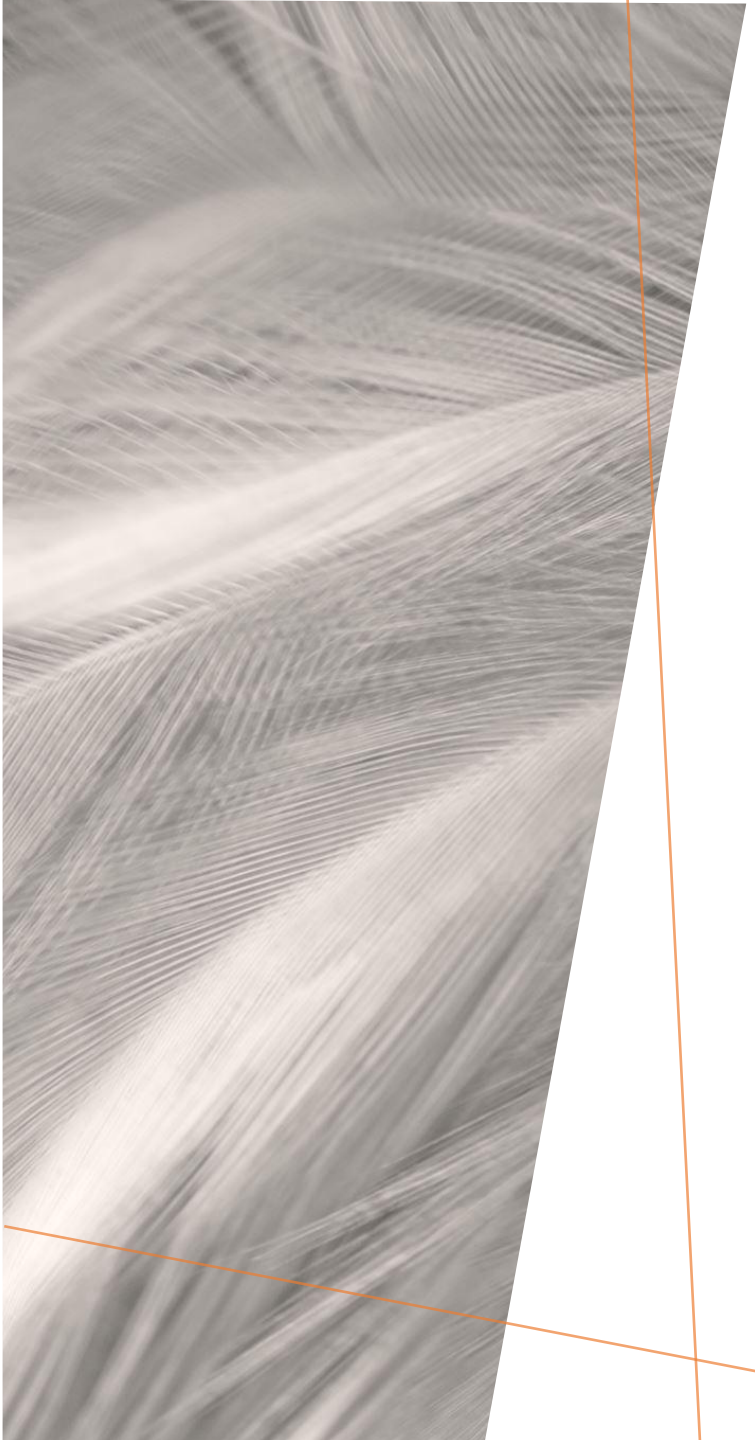
Other Attempts

Since the σ^{24} and σ^{32} are responsible for heat stress response and heat shock respectively, we thought there may have been a chance, that they were structurally similar.

However the accuracy of the model fell upon using this kind of a classification tree.
Possible reasons:-

1. They are structurally different and the property matrices were unable to take them into account.
2. The test-set was too small to conclude.





Thank You