KU LEUVEN

Integrated Bioinformatics Project

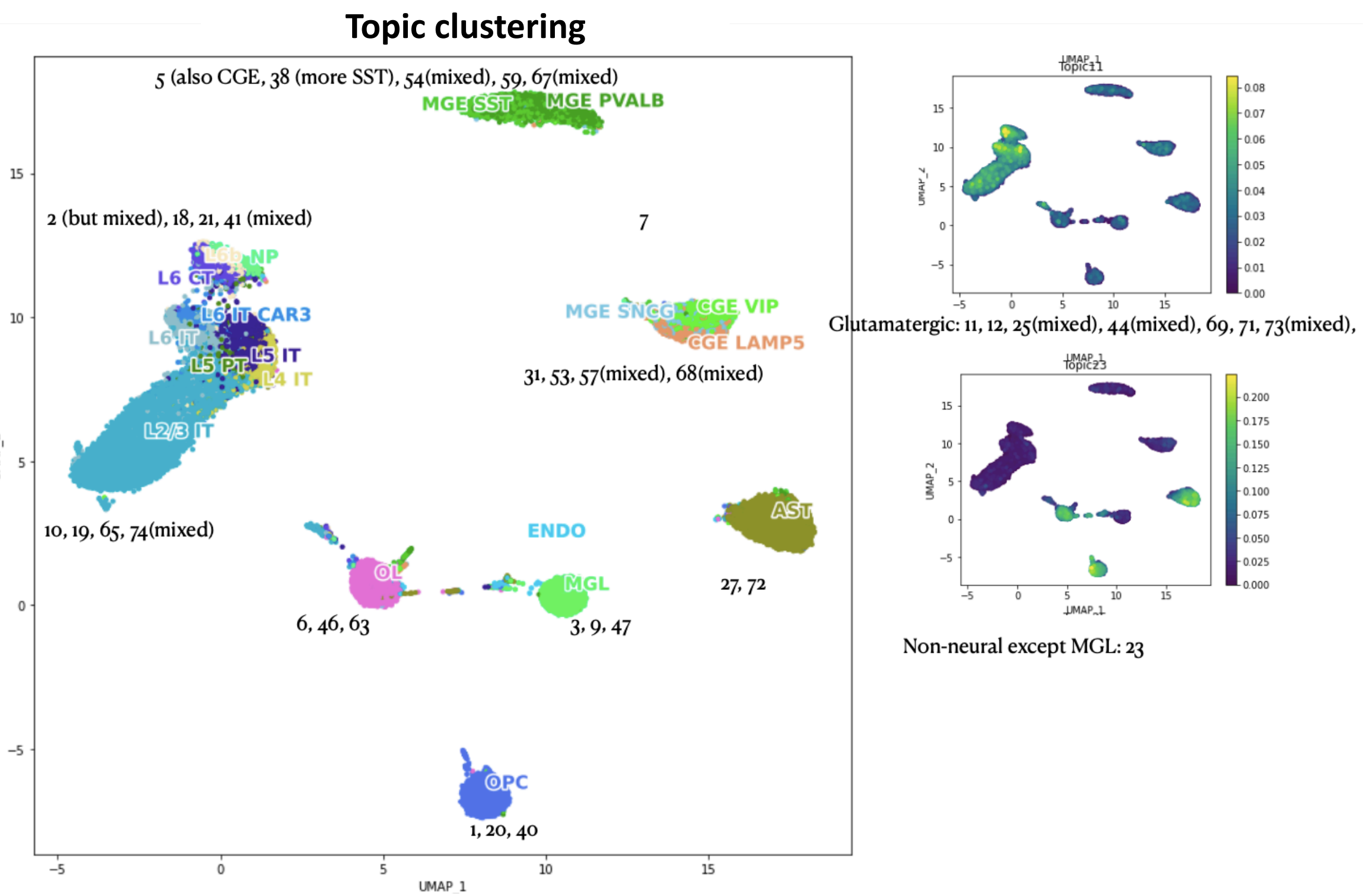
Brain regulatory sequence interpretation using deep learning

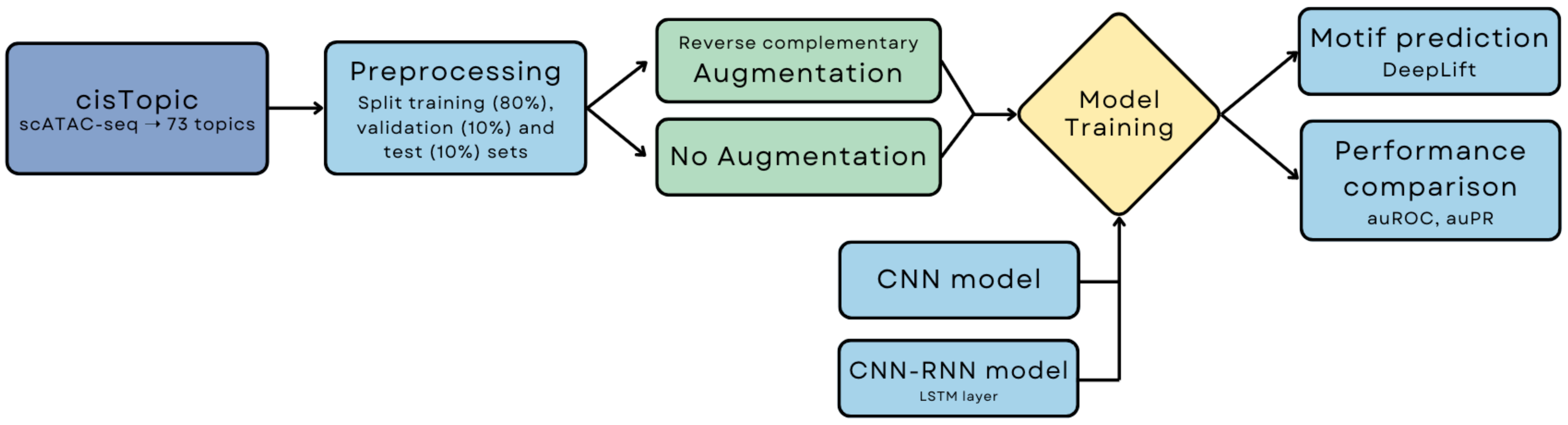
Ilias Theodoros Koulalis   Adrián García Herrero   Guillem Campo Fernández   Laura Lanzas Olsina

Introduction

Functional cell diversity stems from fine-tuned regulatory programs controlled by transcription factor expression and genomic architecture. Assays like scATAC-seq allow us to identify chromatin locations accessible to transcription factors at the single cell level. This data is used by cisTopic, which categorizes accessible regions into topics to which different cell types can be assigned. In order to decipher the regulatory systems that cause differential gene expression and analyze these topics, deep learning-based methods such as DeepMEL have been developed. Here, we develop 2 different deep learning architectures with convolutional and recurrent elements in order to perform a multi-label multi-class classification task. These models are expected to extract as learning features the transcription factor binding regions of the sequences. We analyze the inner workings of the models using DeepLift to try discovering sequences of novel transcription factor binding sites involved in the differential expression pattern of the cells. The dataset consists of 209.000 sequences extracted from scATACseq peaks, with a uniform length of 500bp, all of which come from human motor cells. Through pre-processing using cisTopic 73 topics were obtained.

Topic clustering





Convolutional Neural Network

DeepCLIC

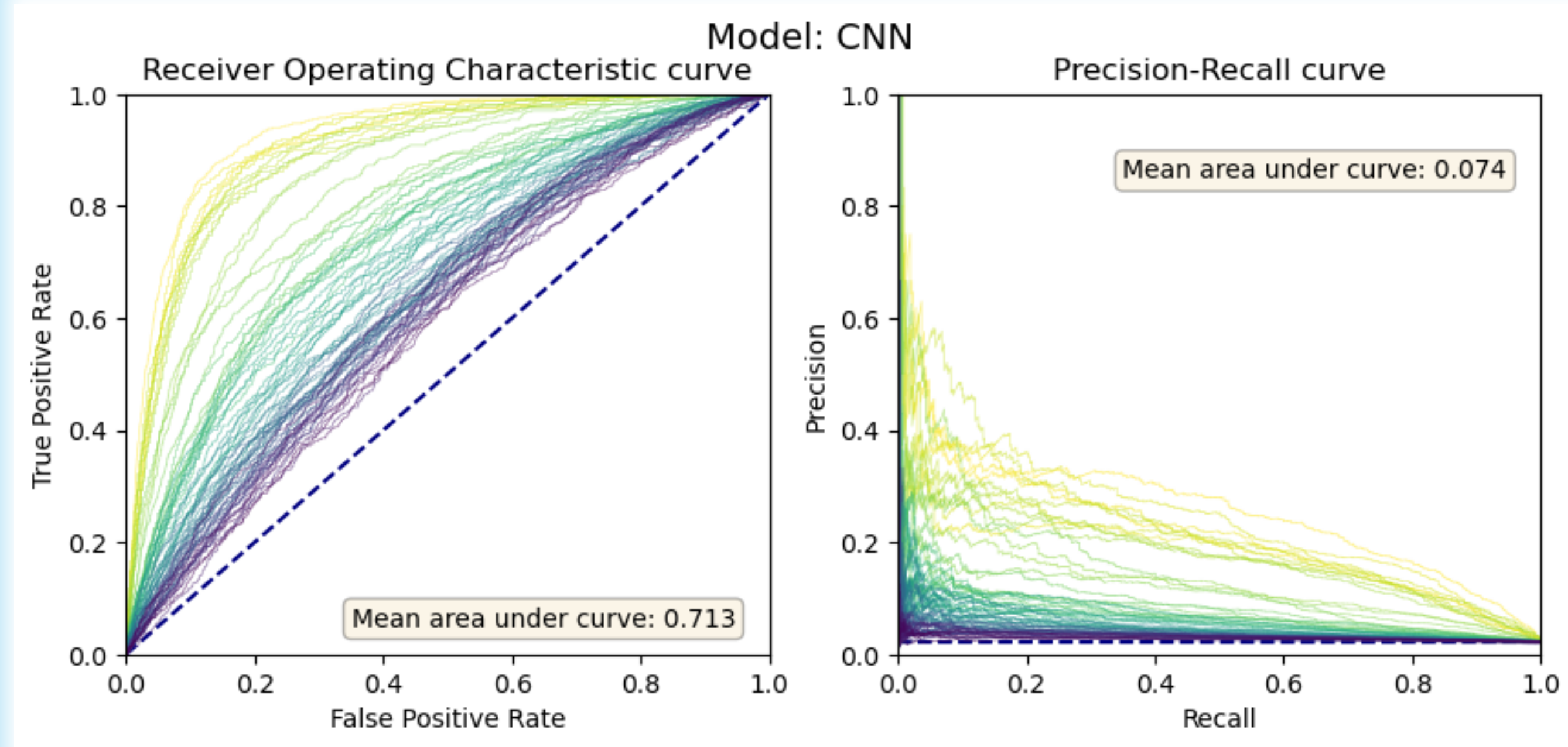
Dataset: augmented

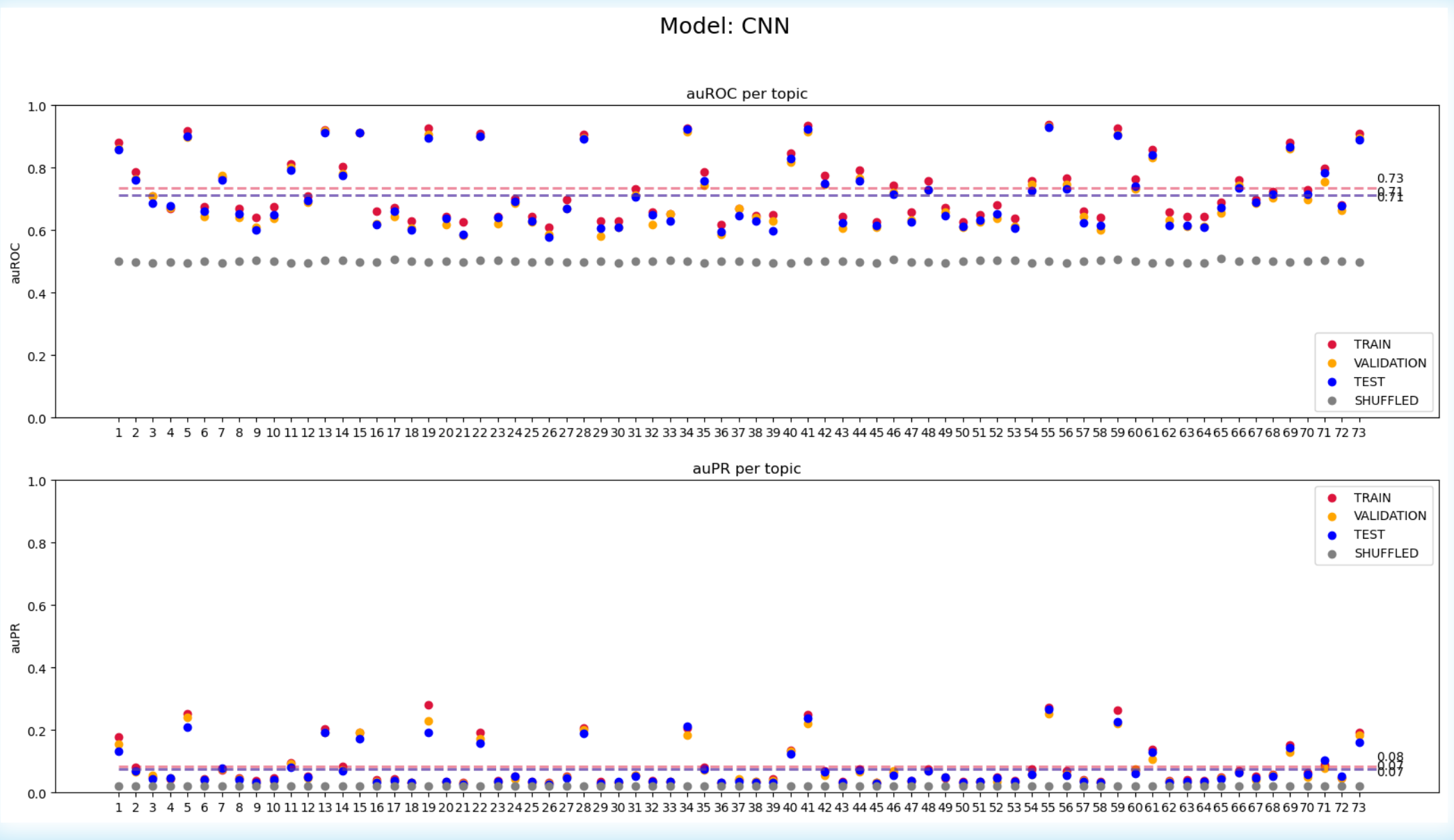
Average auROC

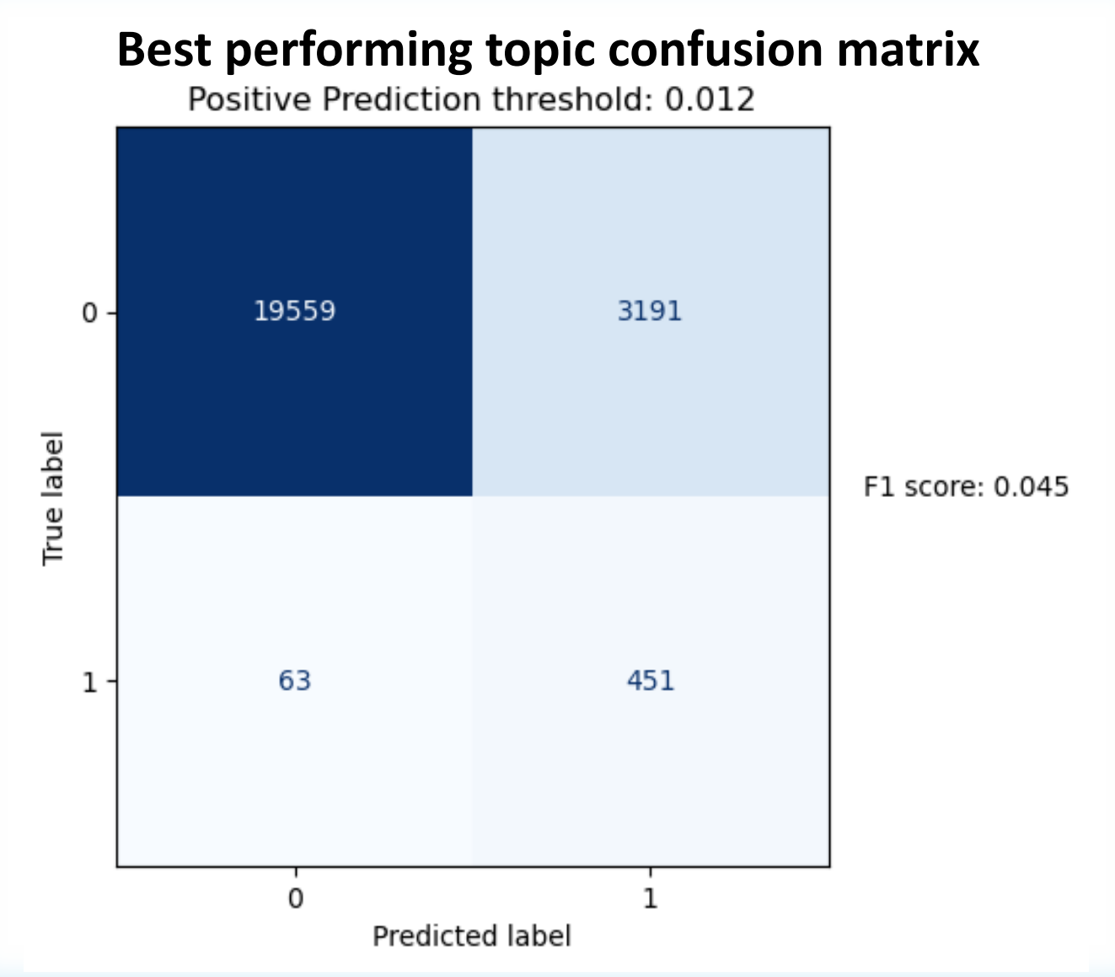
- Training: 0.734
- Validation: 0.713
- Test: 0.713

Average auPR

- Training: 0.083
- Validation: 0.074
- Test: 0.074







Convolutional - Recurrent Neural Network

DeepGLIA

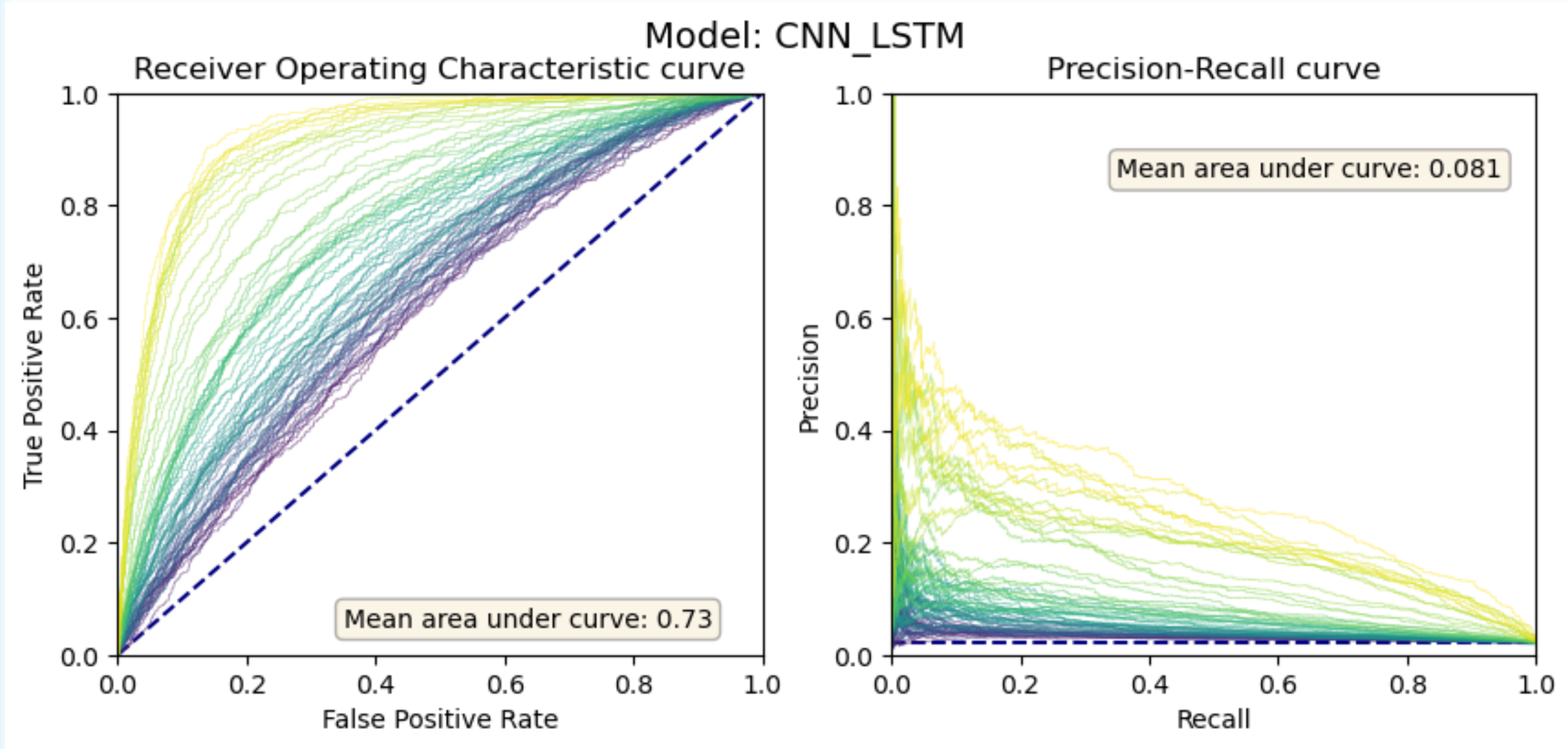
Dataset: augmented

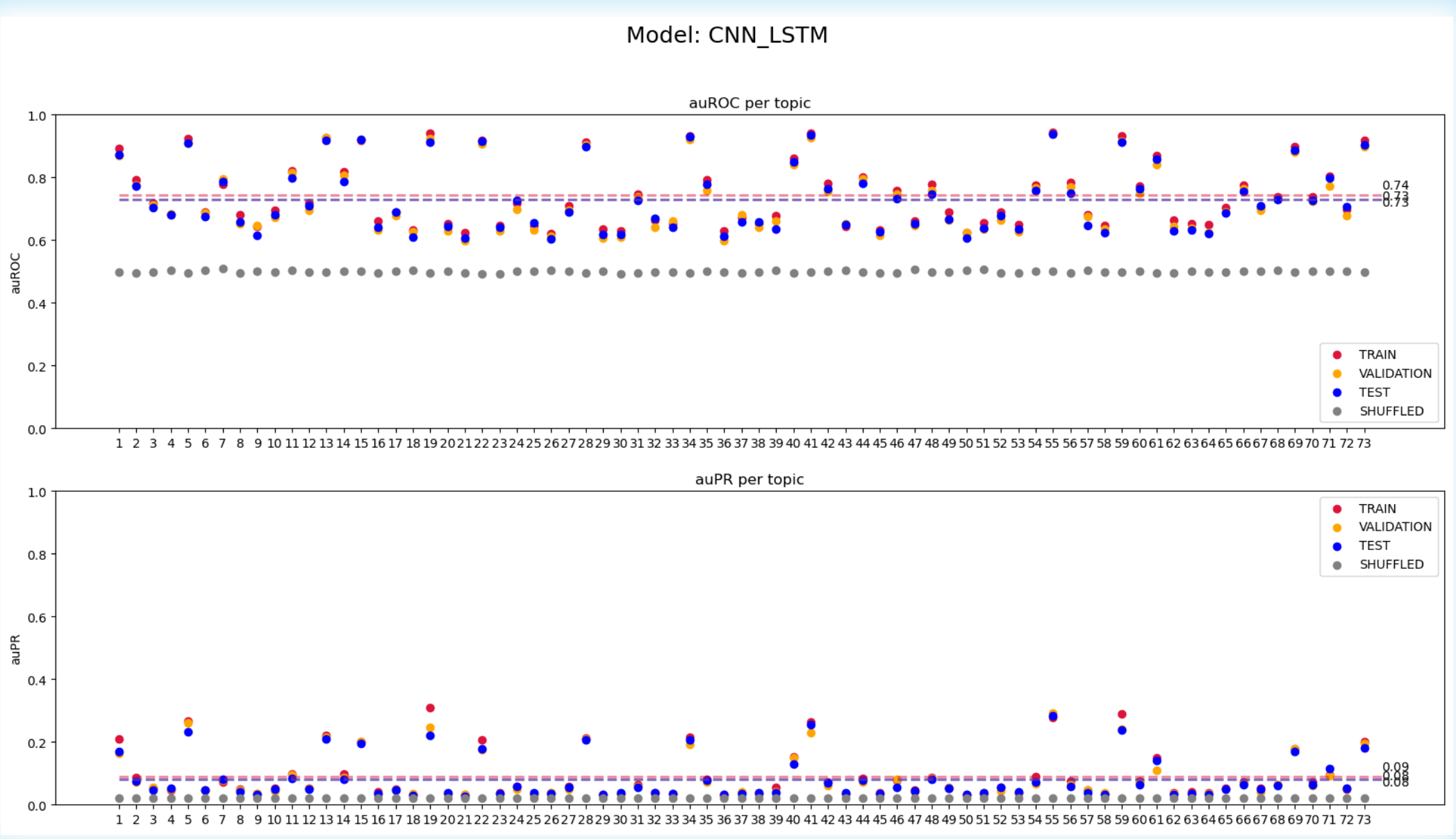
Average auROC

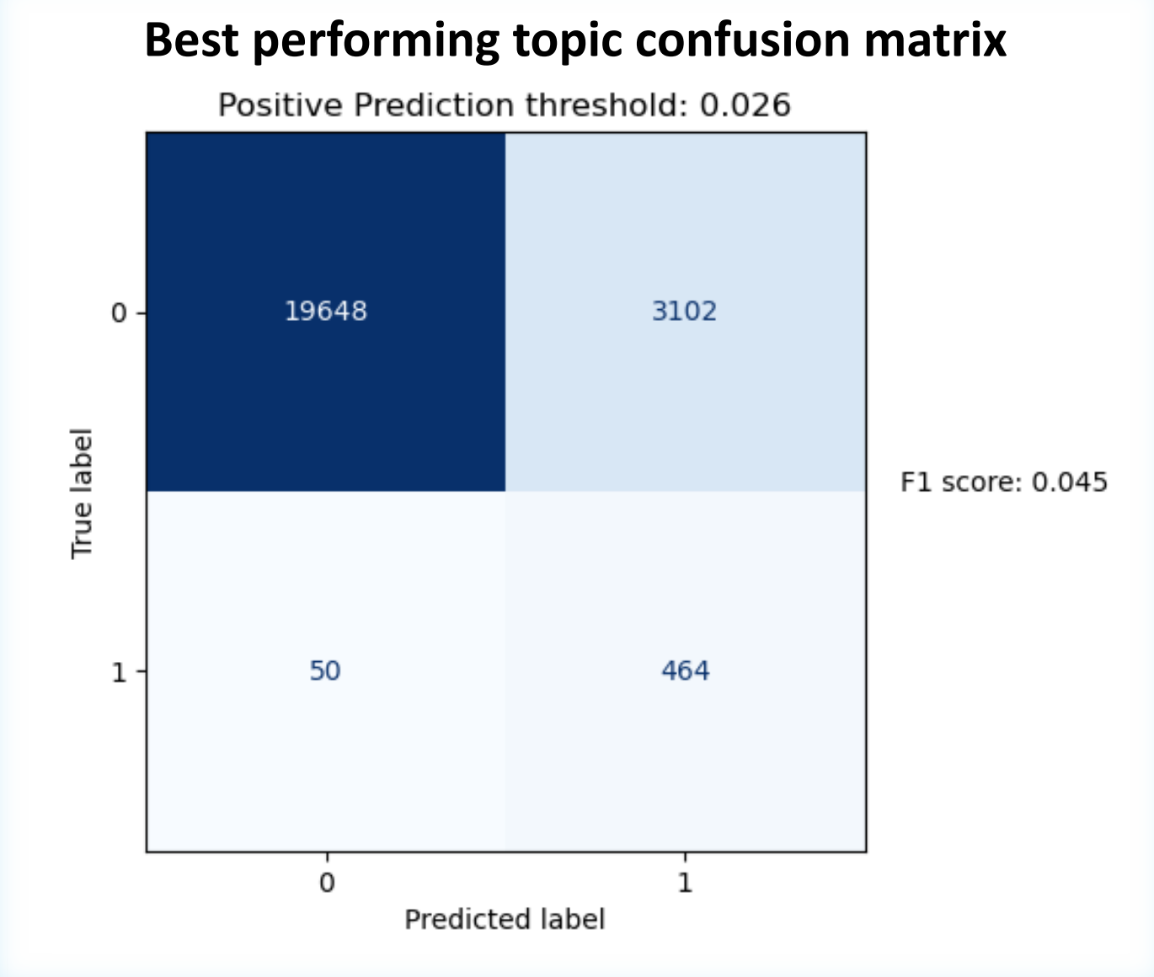
- Training: 0.743
- Validation: 0.729
- Test: 0.729

Average auPR

- Training: 0.088
- Validation: 0.081
- Test: 0.081



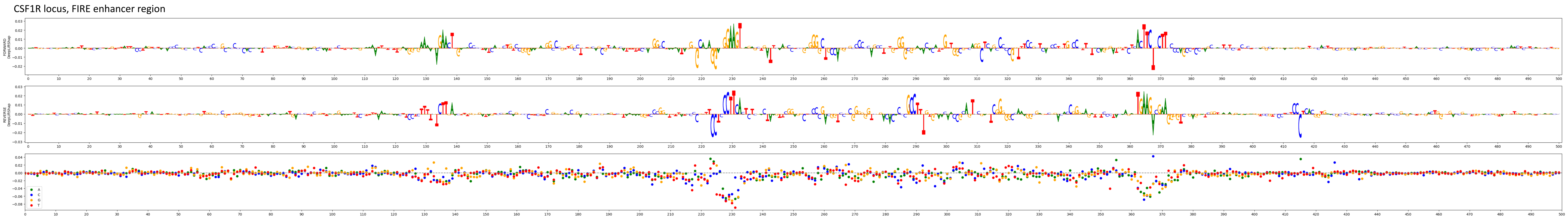




Extraction of learning featrues (DeepLift)

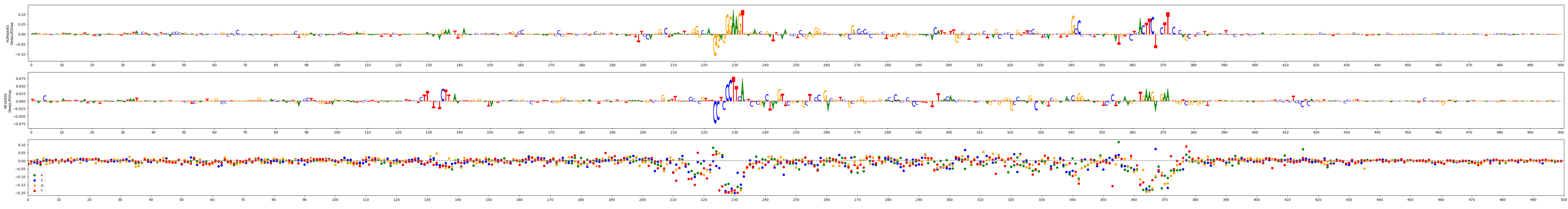
CNN-RNN

CSF1R locus, FIRE enhancer region



CNN

CSF1R locus, FIRE enhancer region



Conclusions

By using the augmented (reverse complementary) data we obtained some remarkable improvement in both models' performance. For DeepGLIA (CNN-LSTM) we obtained a 1.25% increase in the AuROC values and 2.7% increase in the AuPR. For DeepCLIC the improvement was even better with a 2.8% increase in the AuROC and a 10.1% increase in the AuPR.

Nevertheless, we have very low precision in our models. Therefore, to improve that precision in incoming studies we suggest using a bigger input dataset, additional augmentation methods, and integrating additional omic data further than just ATAC-seq.

In order to make their predictions, both models used common features shared by the sequences extracted from ATACseq peaks belonging to co-accessible regions. These common motifs learned by the model and visualized using DeepLift seem to have biological significance, and correspond to transcription factor binding sites.

References

- C. Bravo González-Blas et al., "cisTopic: cis-regulatory topic modeling on single-cell ATAC-seq data," Nature Methods 2019 16:5, vol. 16, no. 5, pp. 397–400, Apr. 2019, doi: 10.1038/s41592-019-0367-1.
- D. Quang and X. Xie, "DanQ: a hybrid convolutional and recurrent deep neural network for quantifying the function of DNA sequences," Nucleic Acids Res, vol. 44, no. 11, Jun. 2016, doi: 10.1093/NAR/GKW226.
- L. Minnoye et al., "Cross-species analysis of enhancer logic using deep learning," Genome *Res*, vol. 31, no. 12, pp. 1815–1834, Jul. 2020, doi: 10.1101/GR.260844.120/-/DC1.
- J. Zhou and O. G. Troyanskaya, "Predicting effects of noncoding variants with deep learning–based sequence model," *Nature Methods 2015 12:10*, vol. 12, no. 10, pp. 931–934, Aug. 2015, doi: 10.1038/nmeth.3547.