CB-DNN

YEHU SAPIR AND ODED YECHIEL

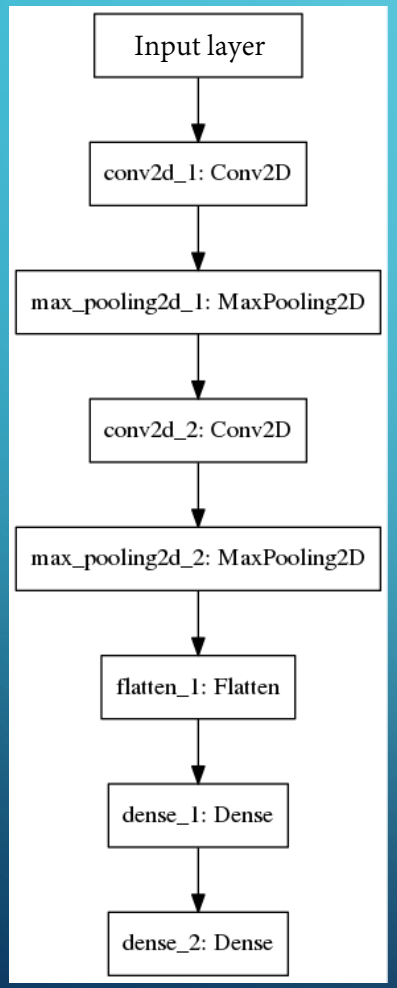
INTRODUCTION

- Goal: rank the binding strength of PBM sequences based on SELEX trial.
- Two network architectures of Deep Neural Network (DNN) were tested:
  - PBM25
  - inception
  - CpGenie [1]
  - DeepBind [2]
- Training computer systems: Colfax and PC with NVidia Titan GPU.
- Python, Keras, Tensorflow.

DATA PRE-PROCESSING

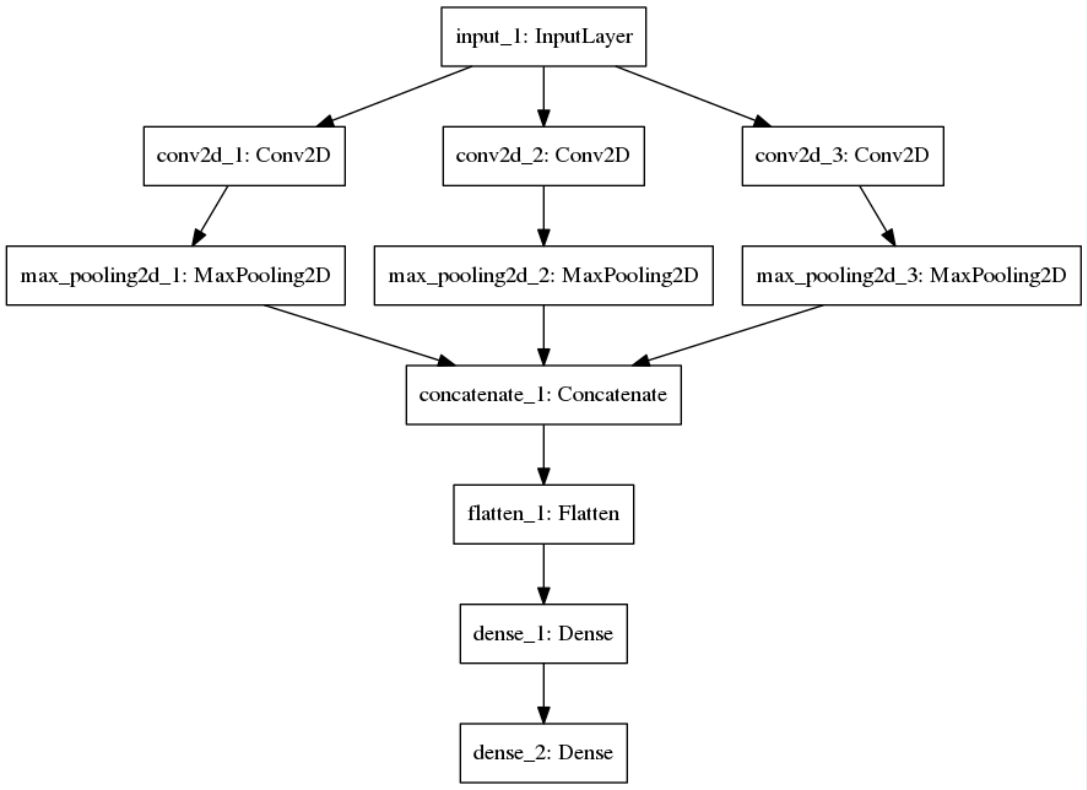
- Randomly choose T training samples from each Selex file
- Randomly choose S testing samples from each Selex file
- One-Hot to Lx4
- Uniform symmetric padding to 36x4x1 matrix
- Create the reverse complement of the vector
- Concatenate all data and shuffle

DEEP-BIND LIKE ARCHITECTURE – PBM25



Layer	Size
Input layer	36x4x1
conv2D 1	32x6x4
MaxPool	5x1
conv2D 2	16x8x4
MaxPool	5x1
Dense 1	128
Dense 2	5 Or 2

INCEPTION



Layer	Size
Input layer	36x4x1
conv2D 1	16x3x4
conv2D 2	16x5x4
conv2D 3	8x10x4
Dense 1	128
Dense 2	2 Or 5

HYPER-PARAMETERS

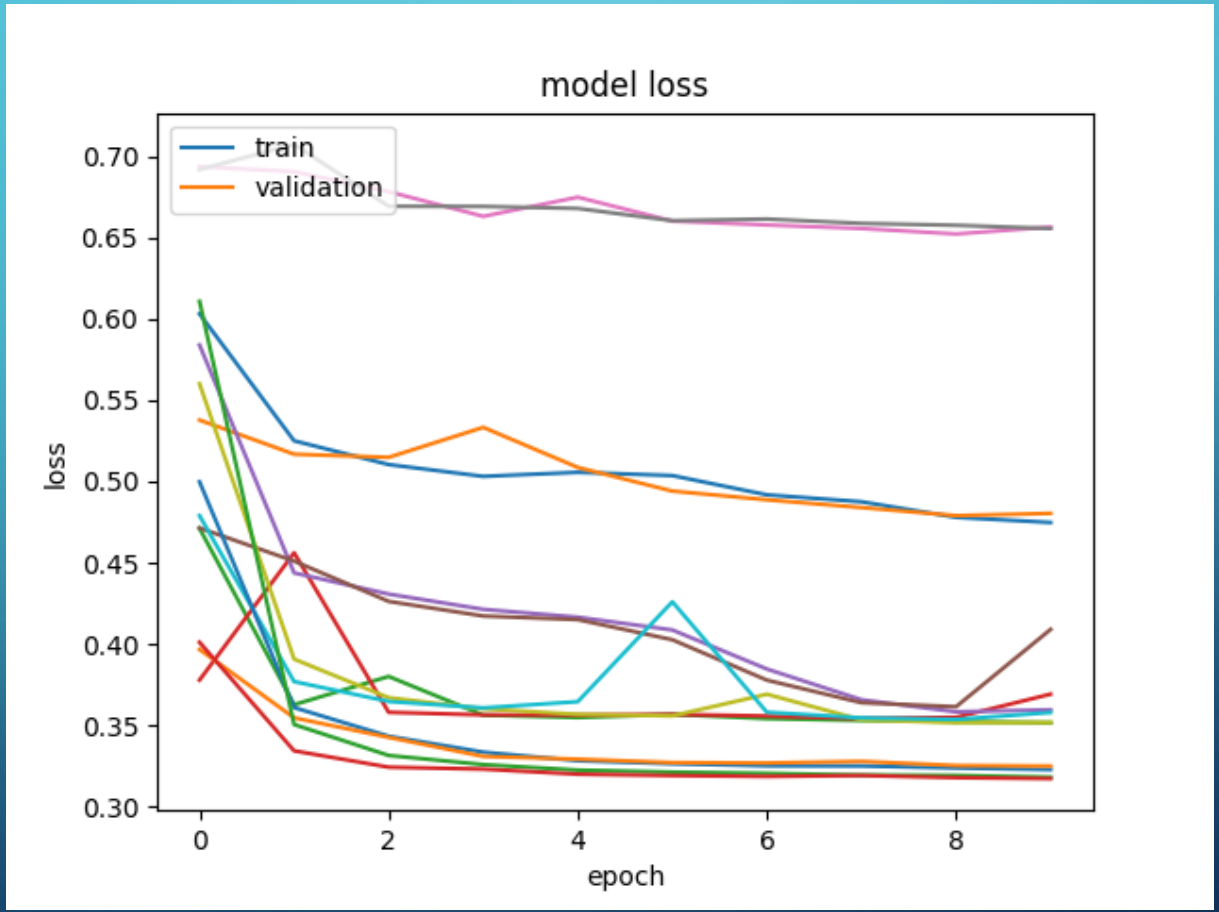
- Optimizer – adam
  - Learning rate – 0.001
- Loss function – categorical cross-entropy
- Metrics – accuracy
- Validation split – 30%
- Epochs - 12
- Batch size - 512

PBM - DECISION RULE

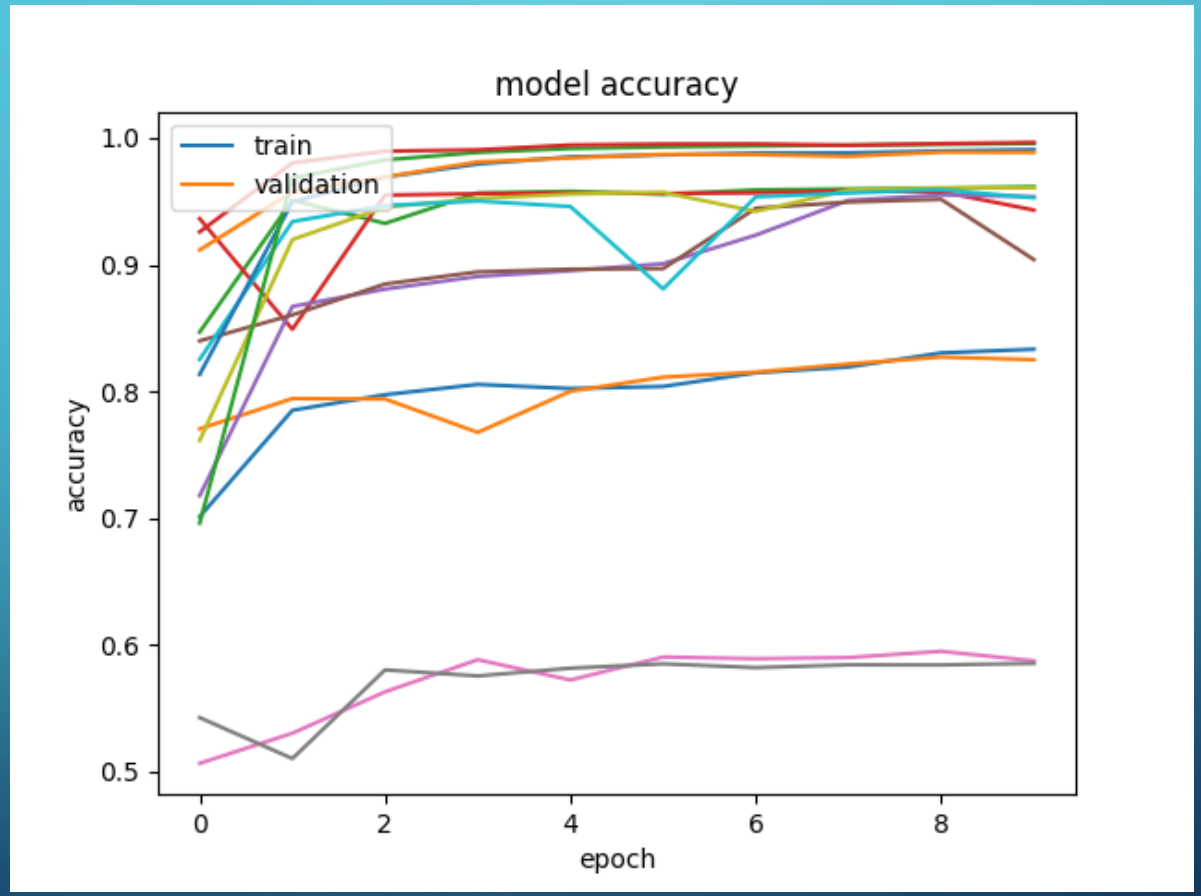
- Network input – 36x4x1 PBM strings
- Network output – 2/5 class Selex group relevance probability

- Rank rule – sort the PBM strings by selex_0 probability
  - Smallest value is ranked first
  - Largest value is ranked last

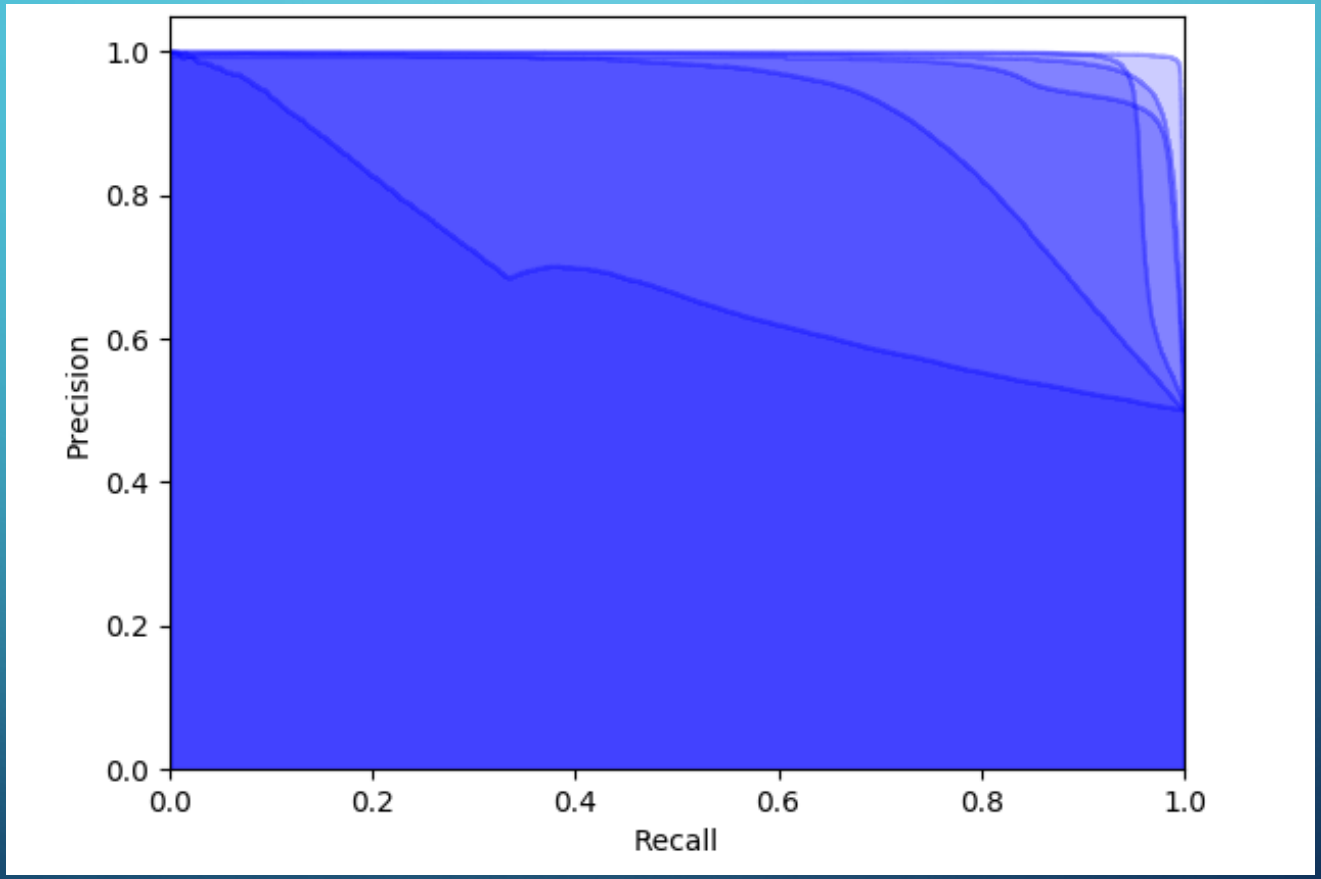
RESULTS – SELEX CLASSIFICATION - PBM25-5



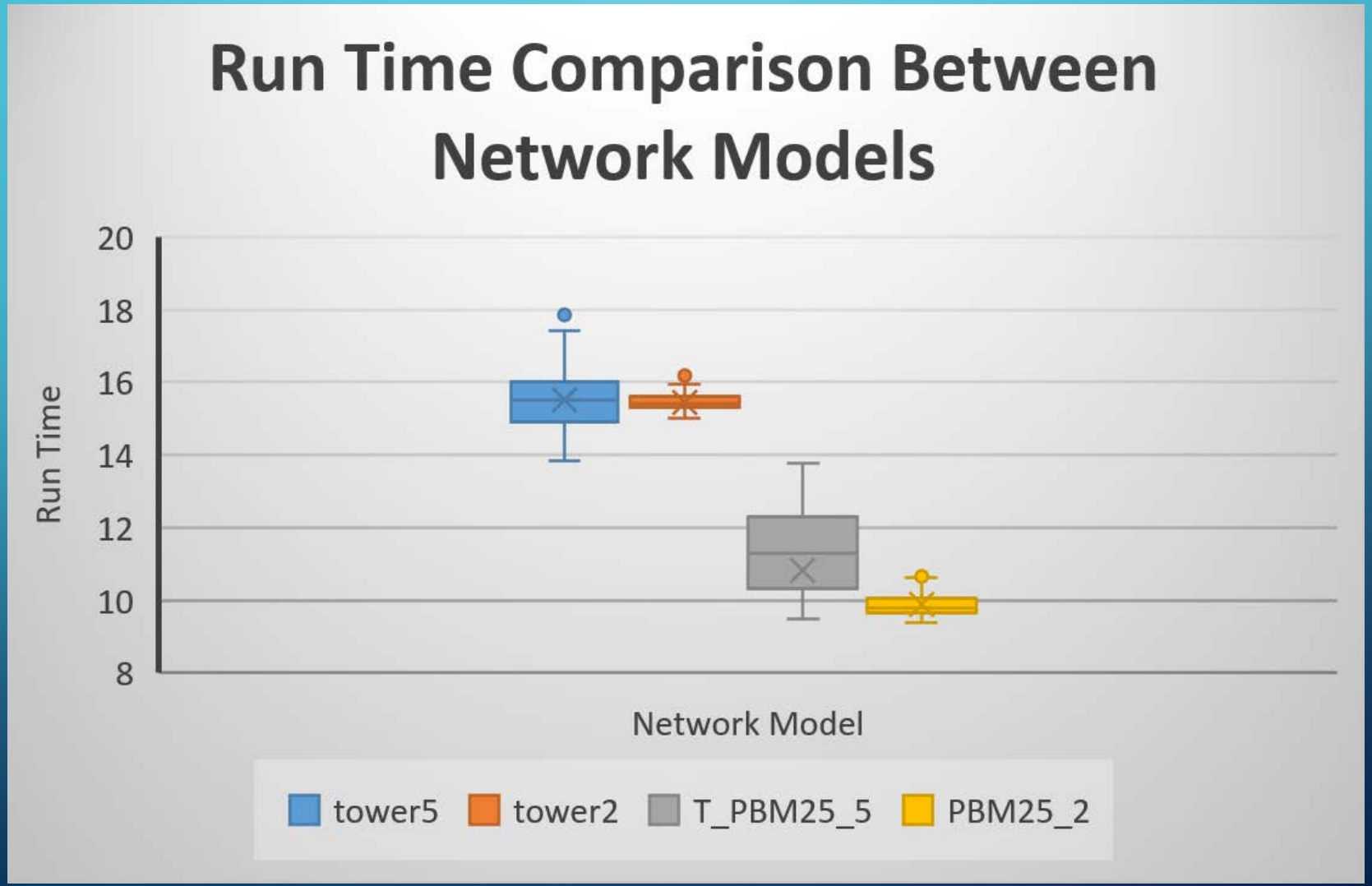
RESULTS – SELEX CLASSIFICATION – PBM25-5



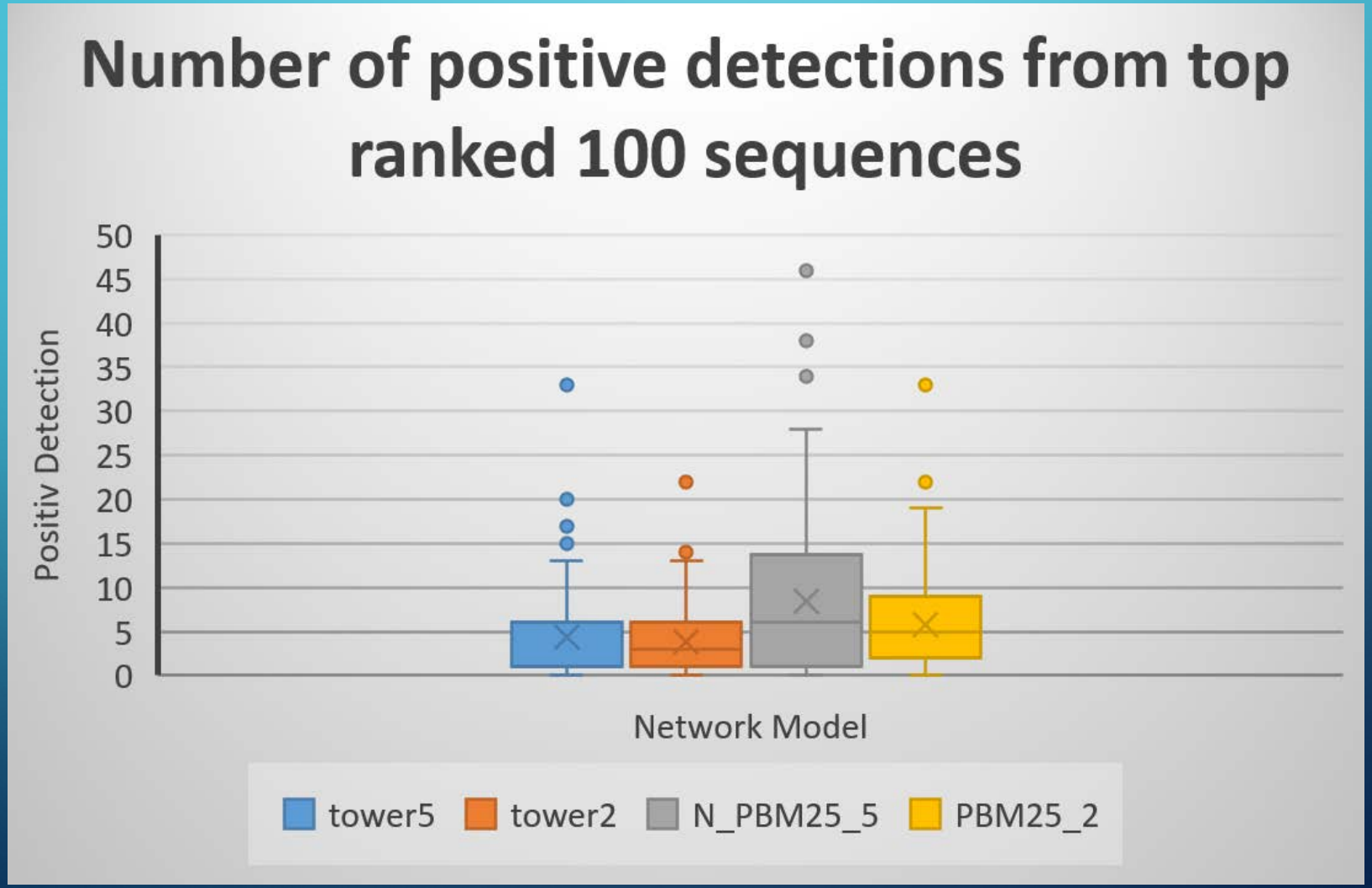
RESULTS – SELEX CLASSIFICATION - PBM25-5



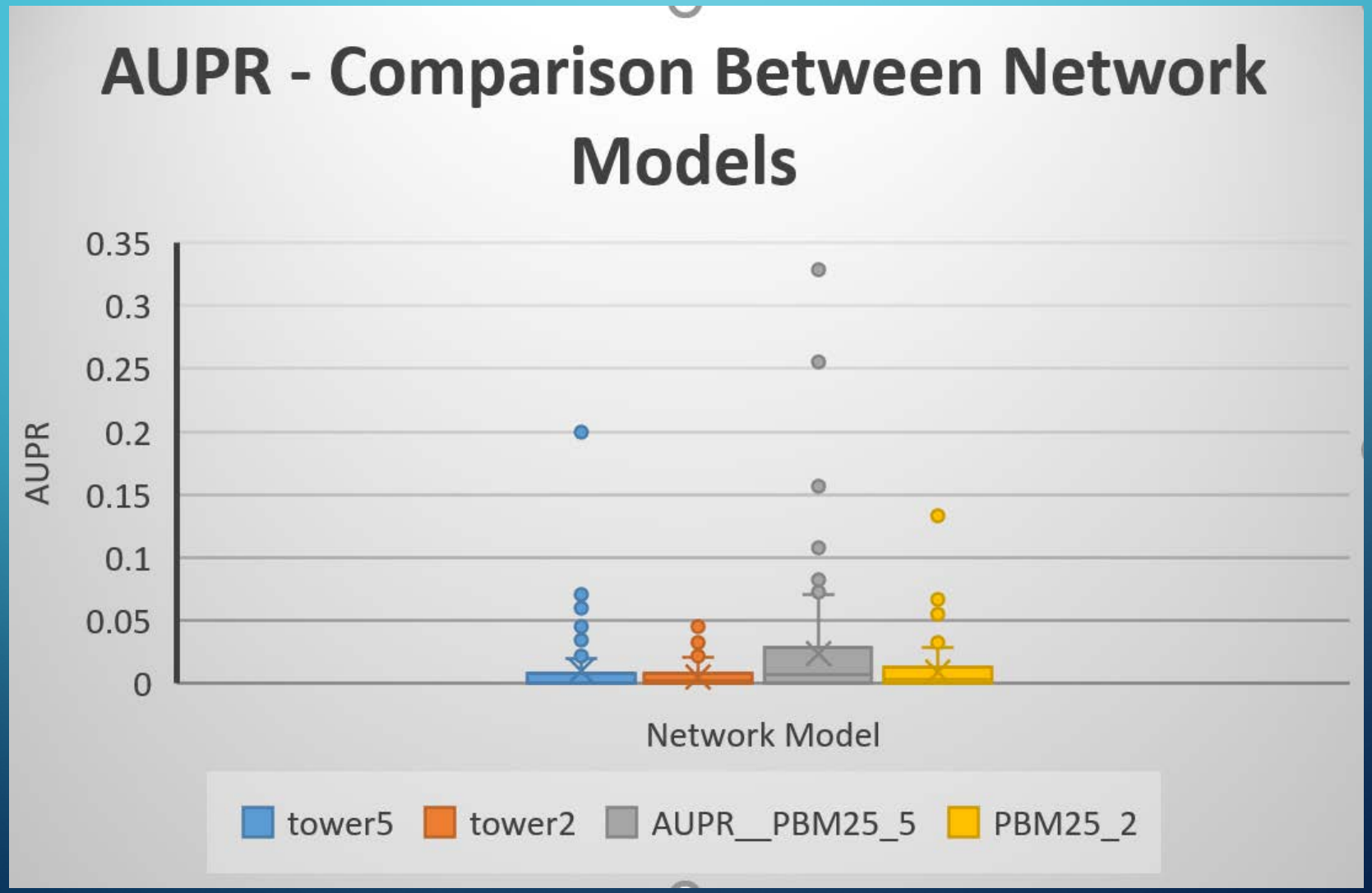
RESULTS – PBM RANKING – RUN ON PC



RESULTS – PBM RANKING



RESULTS – PBM RANKING



CONCLUSIONS

- The input data plays a crucial role. Better understanding is need on how it was generated, and expected features.
- Reducing number of sequences in training boosts the overall time dramatically
- AUPR of PBM is extremely low opposed to classical techniques
- Although the AUPR is low, #seq/100 is much higher than random choice

REFERENCE:

- [1] Zeng H, Gifford DK., "Predicting the impact of non-coding variants on DNA methylation", Nucleic Acids Res, 45 (11): e99.
- [2] Alipanahi B, Delong A, Weirauch MT, Frey BJ., "Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning", Nat. Biotechnol. [Internet]. Nature Publishing Group, a division of Macmillan Publishers Limited. All Rights Reserved.; 2015 [cited 2015 Jul 27];33:831–8. Available from: https://doi.org/10.1038/nbt.3300
- [3] Orenstein, Yaron, and Ron Shamir. "HTS-IBIS: fast and accurate inference of binding site motifs from HT-SELEX data." bioRxiv (2015): 022277.
- [4] Szegedy, Christian, et al. "Going deeper with convolutions." Proceedings of the IEEE conference on computer vision and pattern recognition. 2015.
- [5] Sergey I, Christian S, "Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift", arXiv:1502.03167v3 [cs.LG] 2 Mar 2015