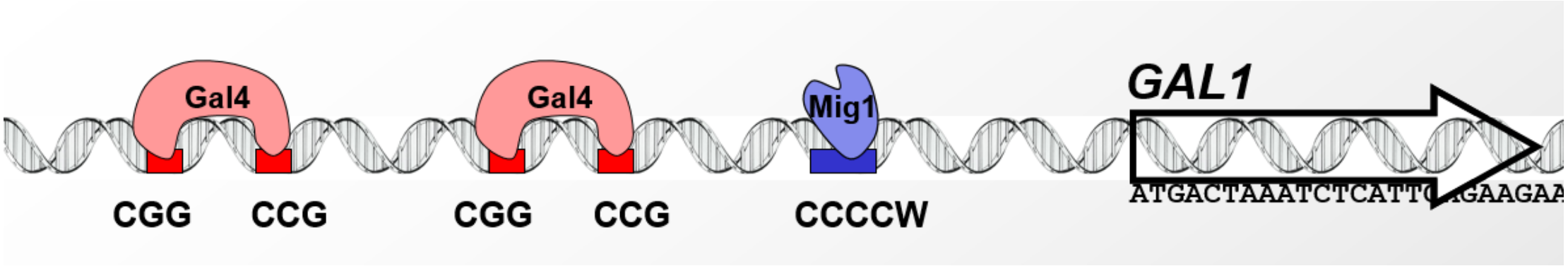
Transcription Factor Bound Regions Prediction: WordToVector Technique with Convolutional Neural Network

Mingye Wang (Benjamin)
Ruoxi Dai (Daniel)
Rixin Chen (Chris)

1. Introduction: Motifs & Bound Regions



- **Motifs:** A pattern, with important biological functions e.g: Promoters

- **Transcription Factor Bound Regions (TFBRs)**: Not only related with motifs

1. Introduction: Relate Bound Regions with Natural Language

- Motif ⟶ 'Word' ,
- Combination of Various Motifs ⟶ 'Sentence'
- Bound Region ⟶ 'Meaning of the Sentence'

Meaning

Sentence

SUZ12_N	chr21	9437245	9437298	AAAGGCTCT	PPARA_3;IKZF2_2;RARG_4
SUZ12_N	chr21	9437434	9437516	GTGTTTTTTG	ZNF384_1;MYC_disc10;YY1_d
MAFK_He	chr21	9437491	9437516	TCCAGATGG	EGR1_known4;TAL1_known3;
MAFF_He	chr21	9438106	9438108	AGC	

Bound Regions

Motifs in this Bound Region

2. Methods: Dataset OVERFLOW

- Comes from 21st chromosome of human

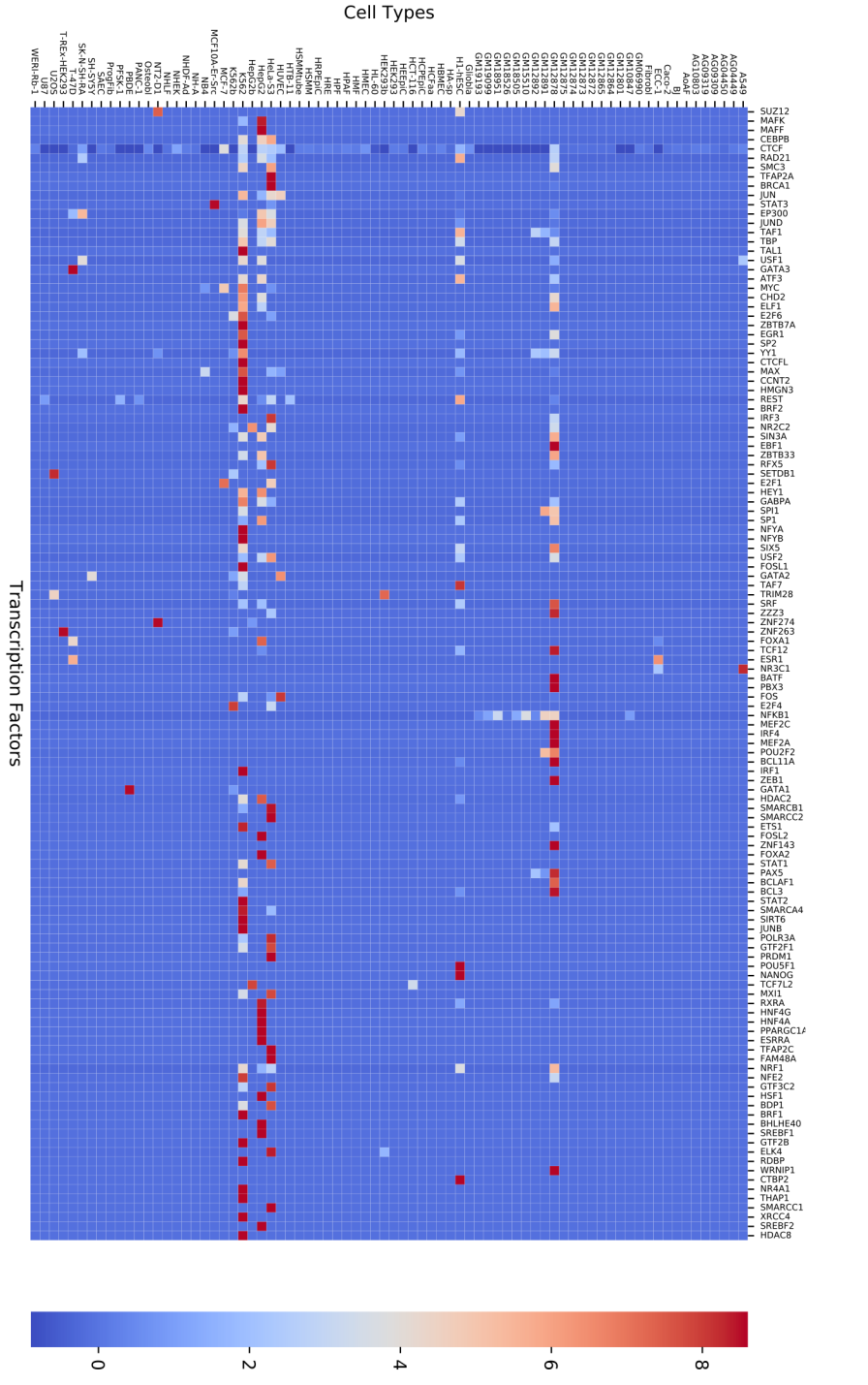
To Process :

- Group motif instances into one motif pattern:

CTCF_known1;COMP1_1;CTCF_known2

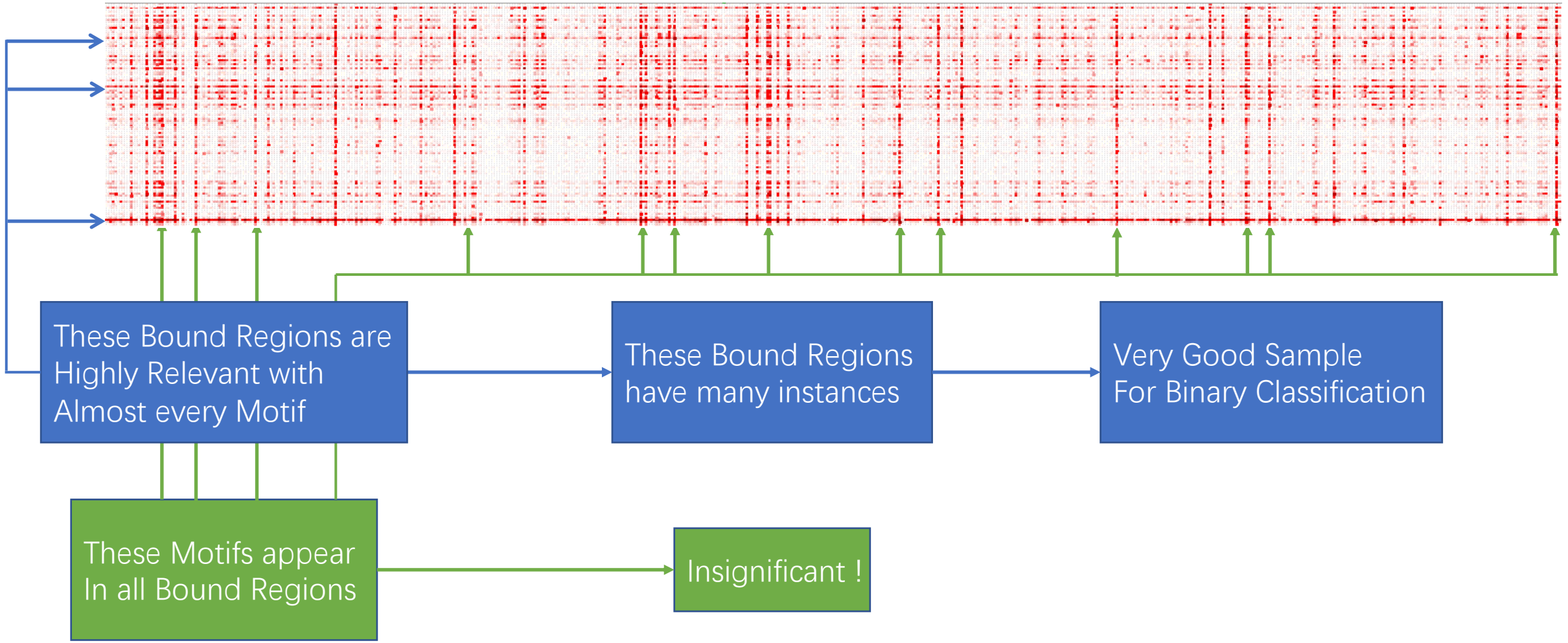
- Group TF bound regions with cell types:

CTCF_GM12878_encode-Bernstein_seq_hsa
CTCF_HBMEC_encode-Stam_seq_hsa
CTCF_HEEpiC_encode-Stam_seq_hsa
CTCF_HPAF_encode-Stam_seq_hsa



2. Methods: Motifs Distribution

Row		Bound Regions	122
Column		Motifs	602



2. Methods: Word2Vec Algorithm

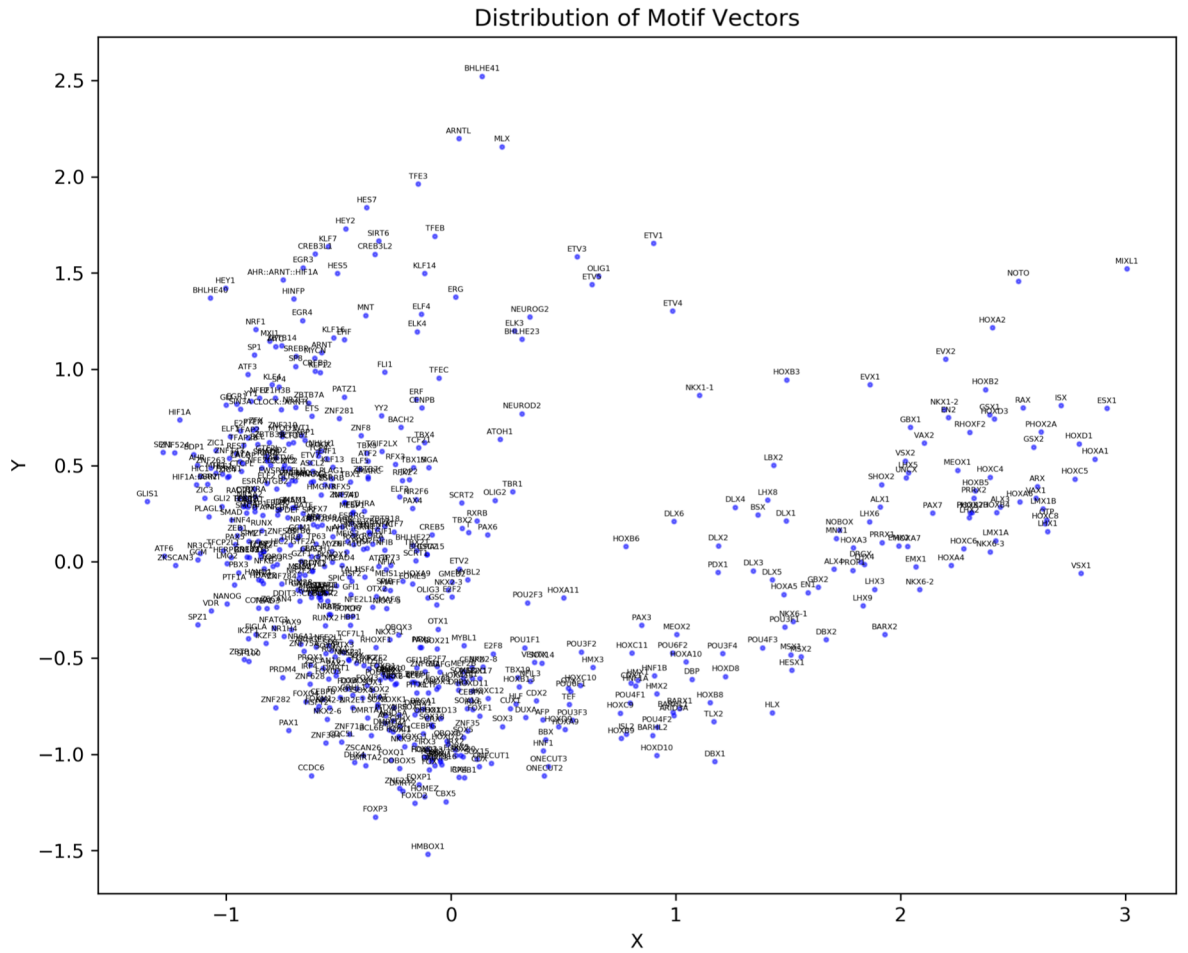
- Each motif ➔ A 200-dimensional vector
- Motifs constantly bound together ➔ Closer in Distance in the Scatter plot

e.g: 'Word' : CTCF

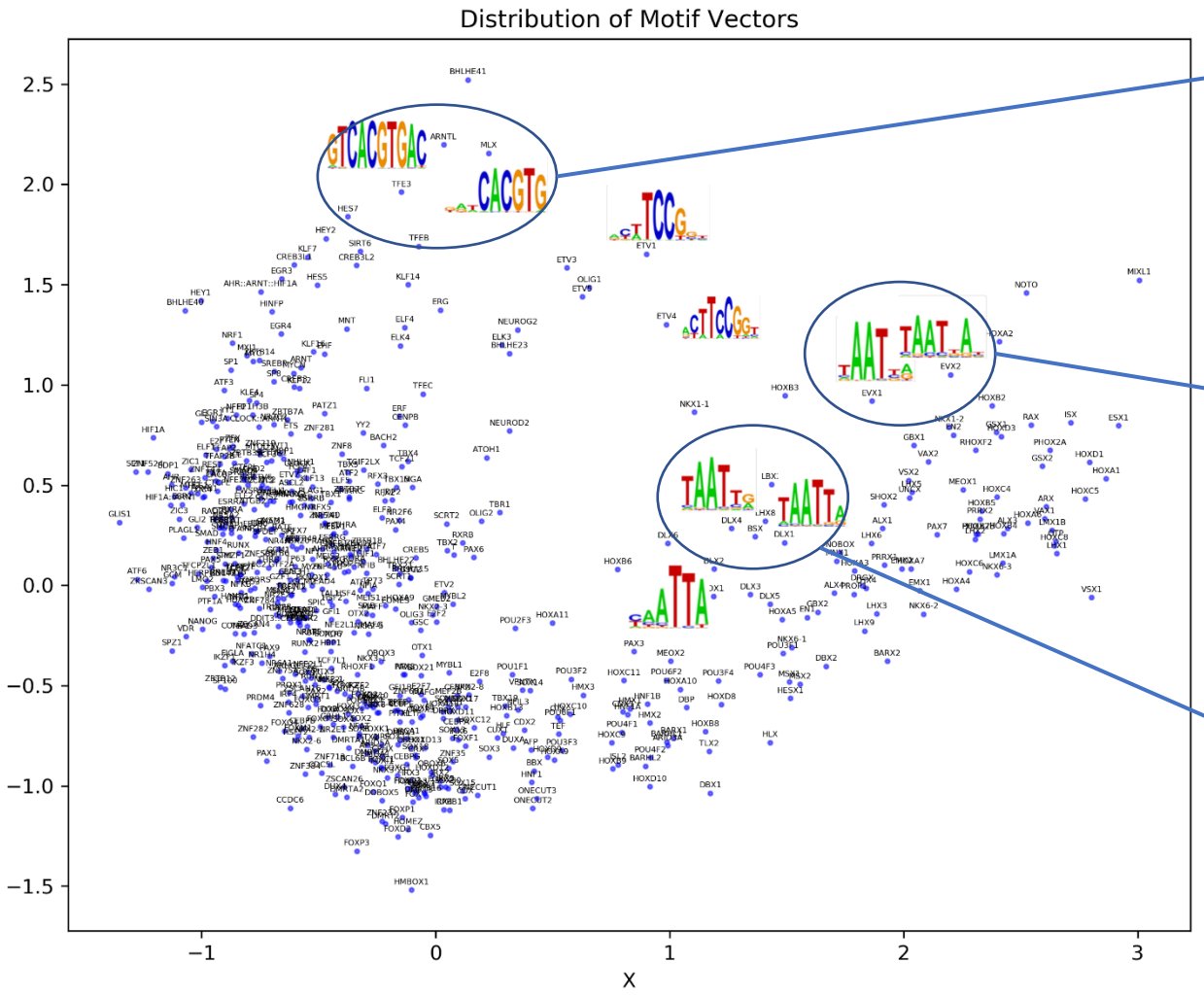
200-dimensional vector

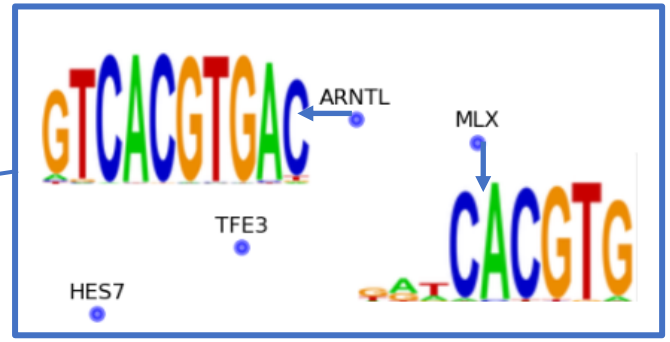
PCA

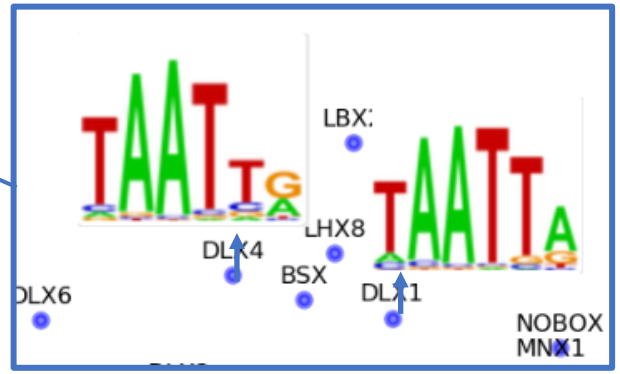
2-dimensional vector



2. Methods: Distribution of Motifs







2. Methods: Label of the Dataset

- Binary Classification of Bound Regions:

CTCF_GM12878_encode-Bernstein_seq_hsa
CTCF_HBMEC_encode-Stam_seq_hsa
CTCF_HEEpiC_encode-Stam_seq_hsa
CTCF_HPAF_encode-Stam_seq_hsa
CTCF_HSMM_encode-Bernstein_seq_hsa
CTCF_NH-A_encode-Bernstein_seq_hsa
CTCF_NHEK_encode-Bernstein_seq_hsa
CTCF_NHEK_encode-Stam_seq_hsa
CTCF_Osteobl_encode-Bernstein_seq_hsa
CTCF_AG09309_encode-Stam_seq_hsa

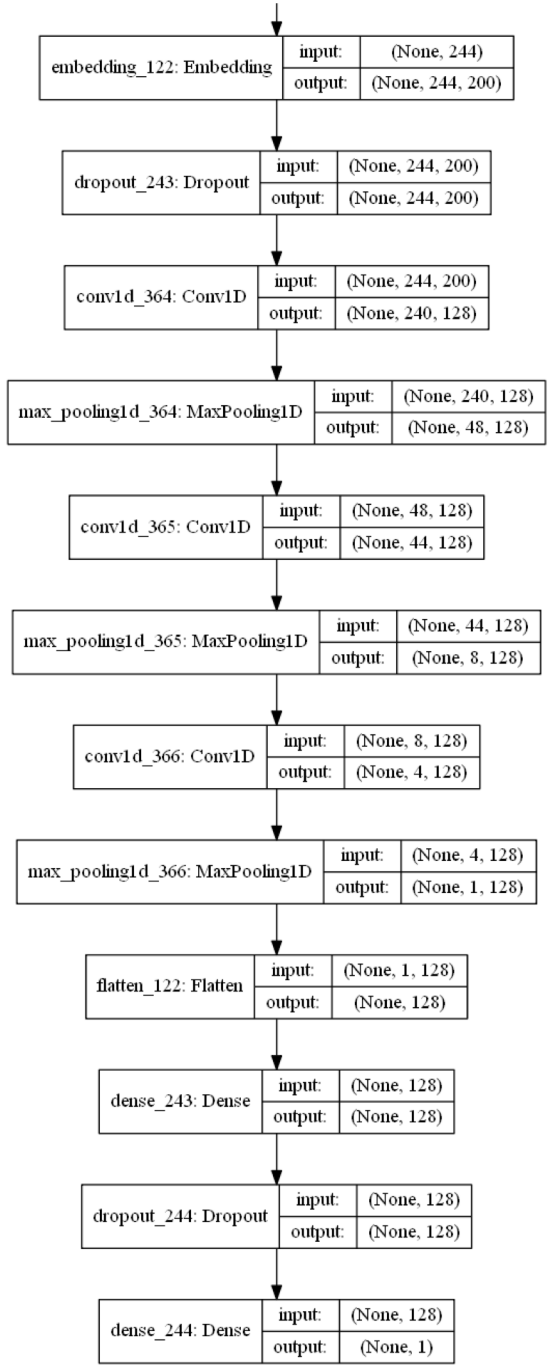
RAD21_GM12878_encode-Myers_seq_hsa_v041610.
RAD21_HeLa-S3_encode-Snyder_seq_hsa_IgG-rab
RAD21_K562_encode-Snyder_seq_hsa
EGR1_K562_encode-Myers_seq_hsa_v041610.1
SP2_K562_encode-Myers_seq_hsa_v041610.2-SC-64

(+): CTCF, 26211, label: 1

(-): Non-CTCF, 38983 ,label: 0

2. Methods: Model Structure

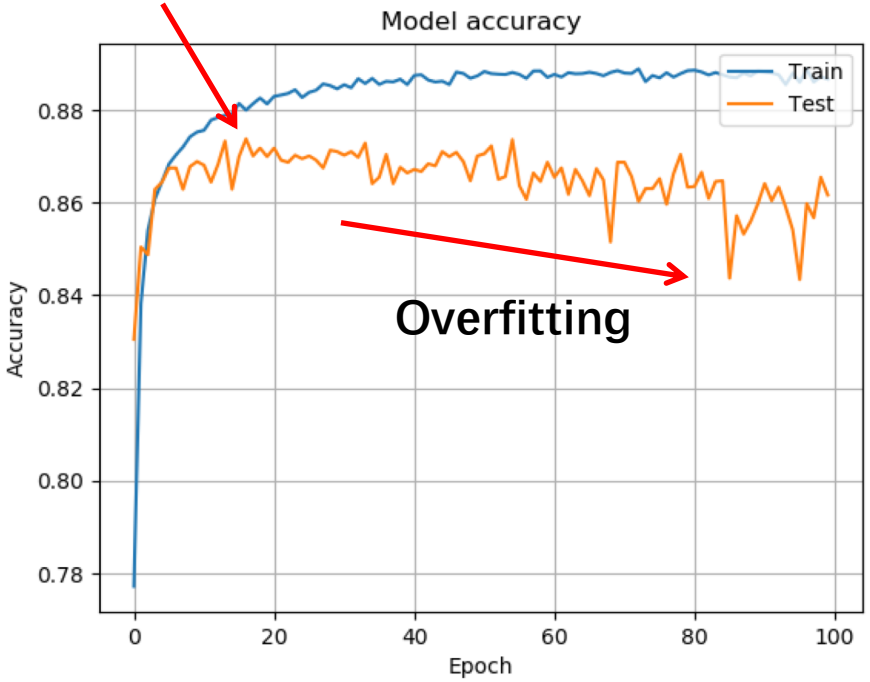
- Sequential
- Embedding: Transfer word index
- CNN: Feature Extraction
- Dropout: Overfitting Prevention
- Sigmoid: Binary Classification

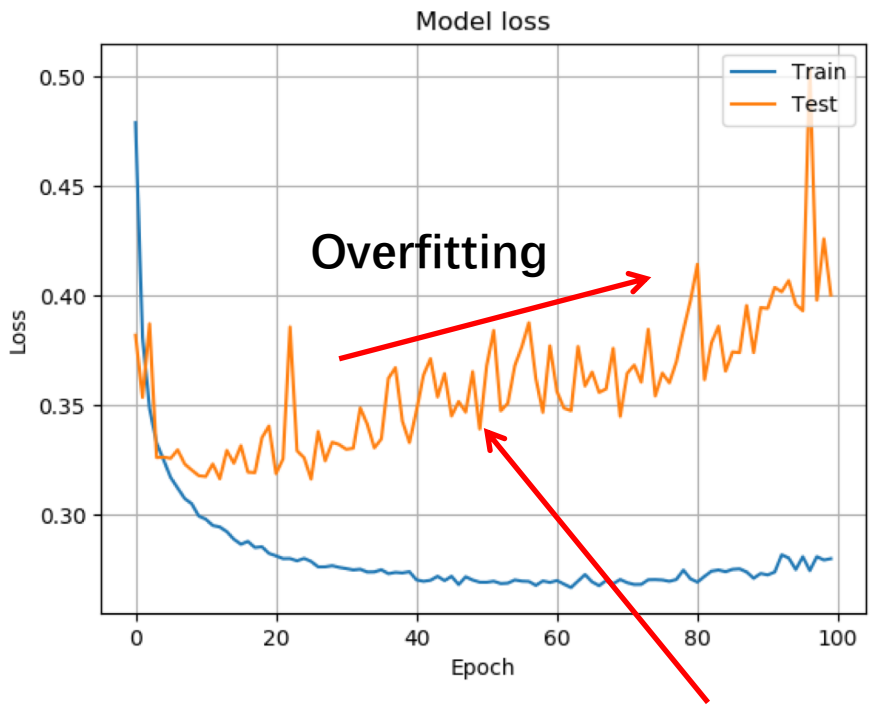


3. Results: Bound Prediction for CTCF

Accuracy and Losses

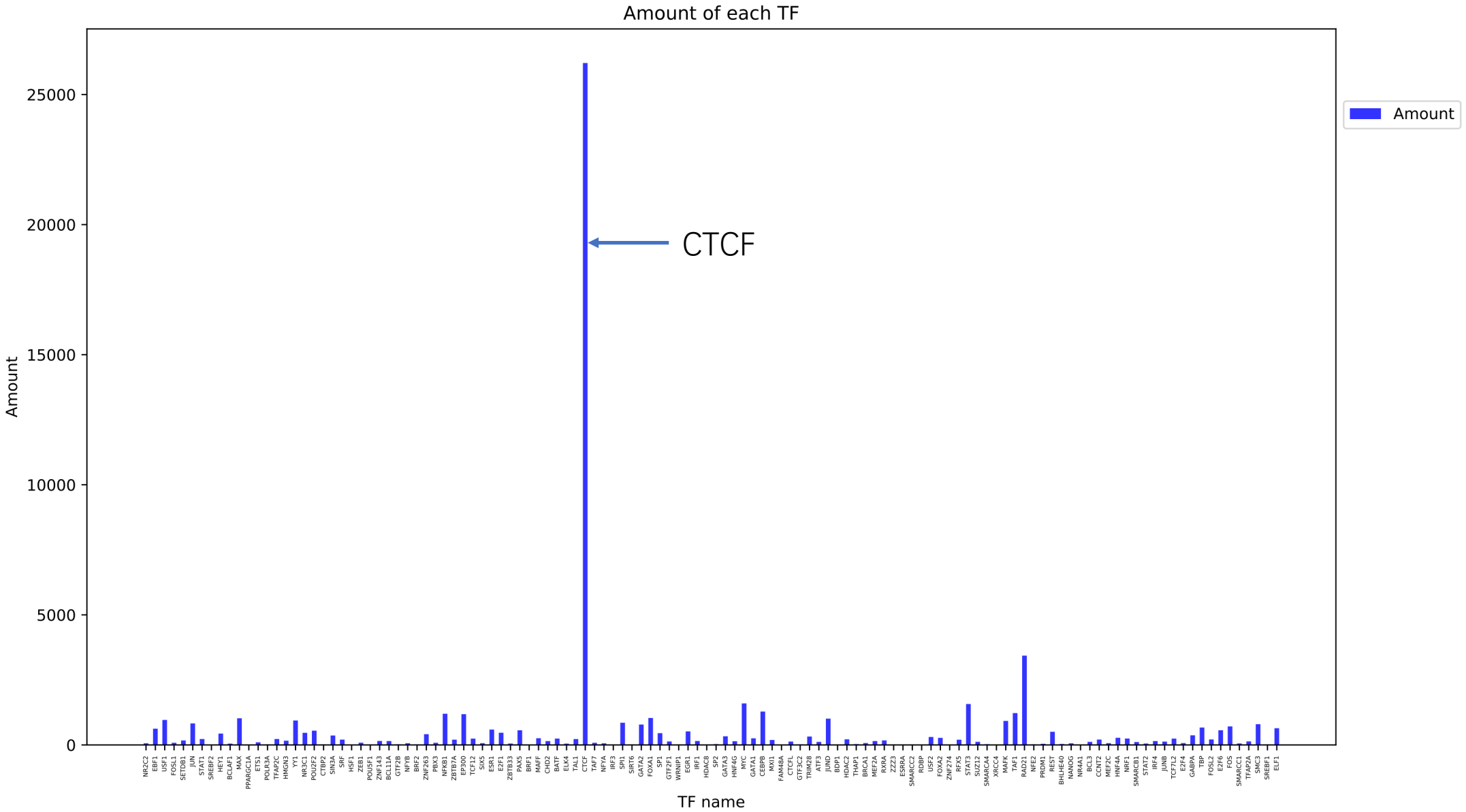
Optimal Accuracy: 87%



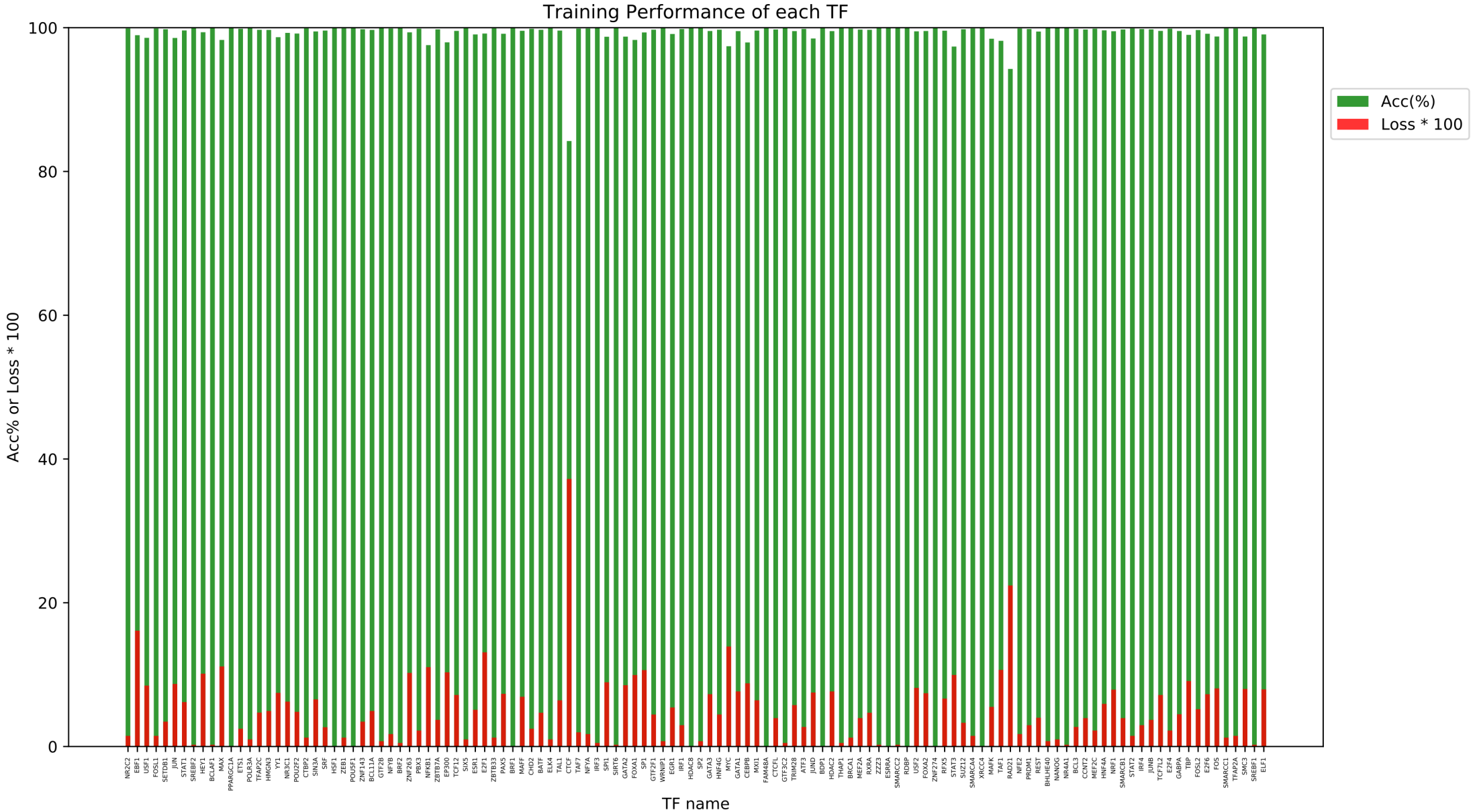


Smoothier After Dropout
But Still Uneven

3. Results: Bound Prediction for All Bound Regions



3. Results: Bound Regions for All Bound Regions



4. Future Goals:

- Data Extension: Expand out datasets to ALL CHROMOSOMES
- Higher Accuracy: Adjust our model till OPTIMAL
- More Features: Comparative Genomics, Epigenomics, Cell Types

5. Outlines:

- 1. Introduction: Motif -> Vector -> Natural Language Processing
- 2. Methods: Word2Vec Algorithm & CNN
- 3. Results: More than 87% Accuracy
- 4. Future Goals: Larger Datasets, Better Model and More Features