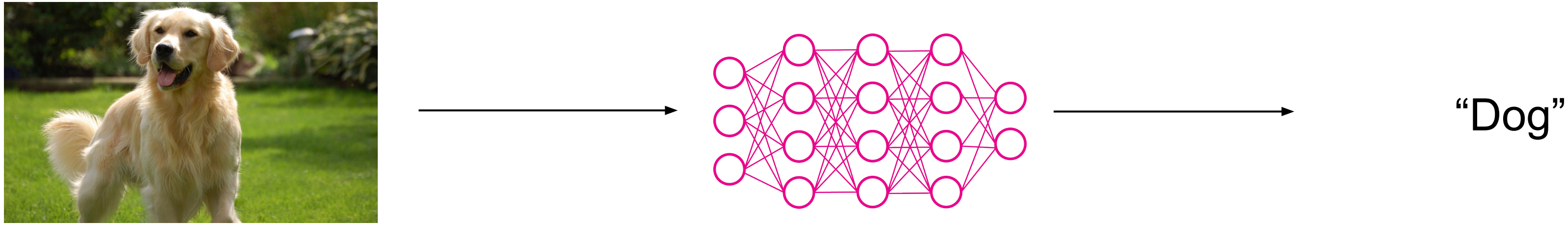
Deep **Mo**tif **Dashboard**: Visualizing and Understanding Genomic Sequences Using Deep Neural Networks

Jack Lanchantin, Ritambhara Singh, Beilun Wang, Yanjun Qi

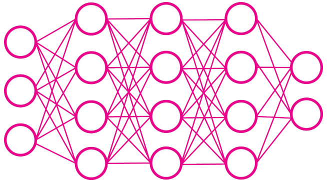
University of Virginia, Department of Computer Science

UNIVERSITY of VIRGINIA

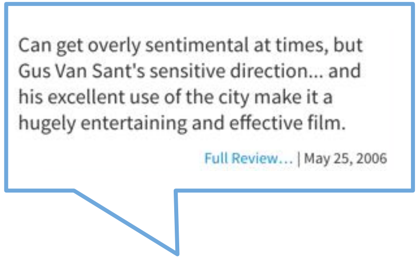
deepmotif.org



“Dog”

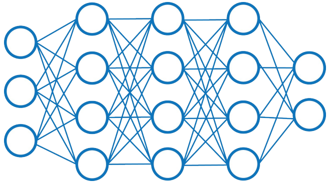


“Dog”

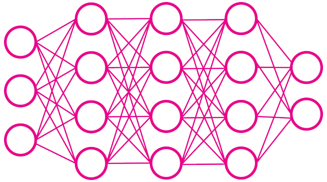


Can get overly sentimental at times, but Gus Van Sant's sensitive direction... and his excellent use of the city make it a hugely entertaining and effective film.

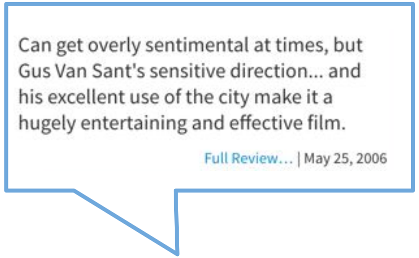
Full Review... | May 25, 2006



👍

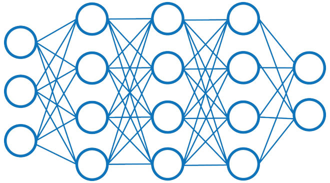


“Dog”



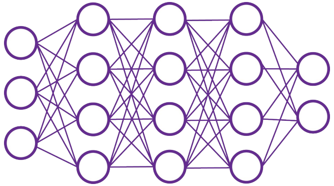
Can get overly sentimental at times, but Gus Van Sant's sensitive direction... and his excellent use of the city make it a hugely entertaining and effective film.

Full Review... | May 25, 2006

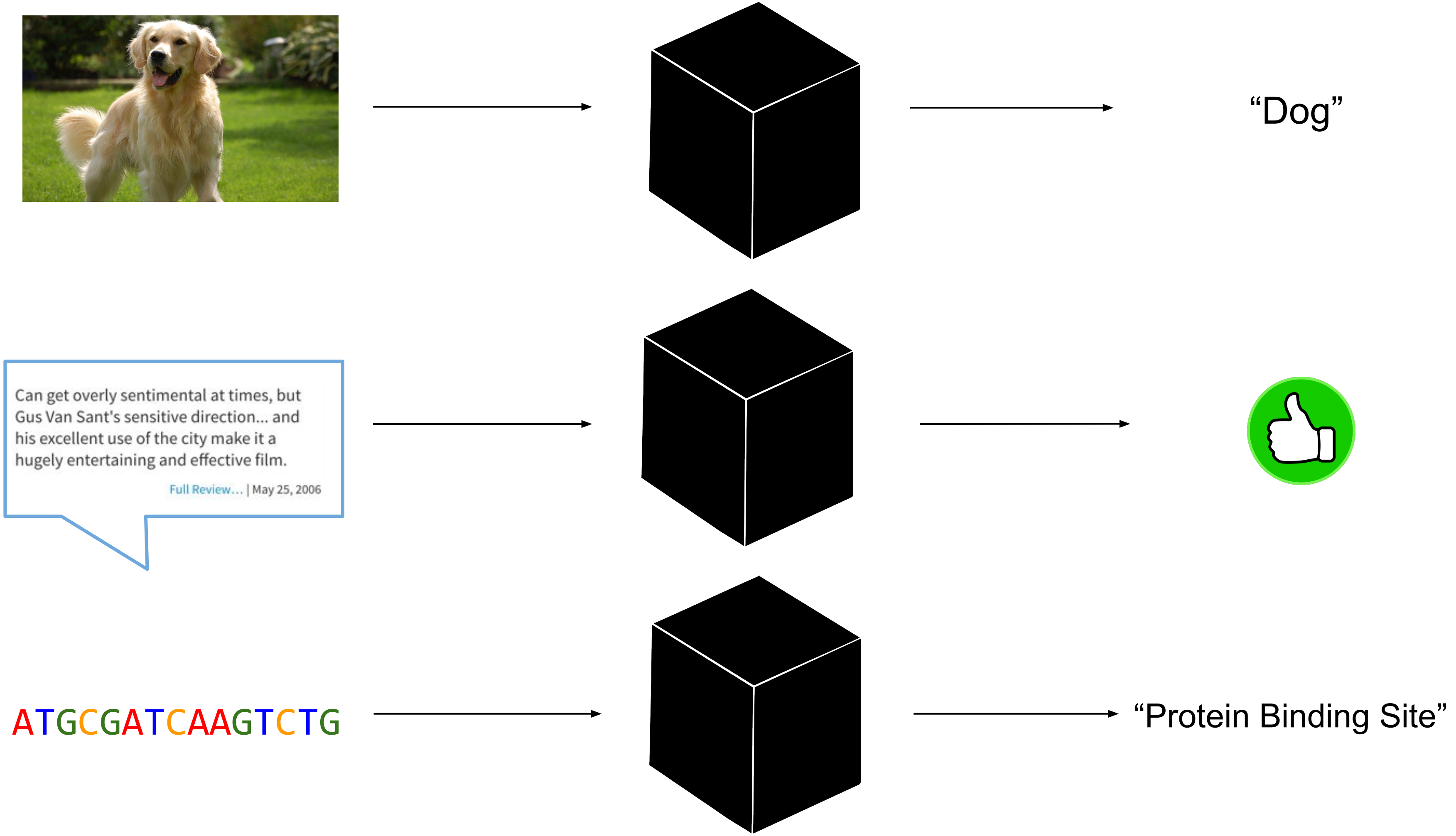


👍

ATGCGATCAAGTCTG



“Protein Binding Site”



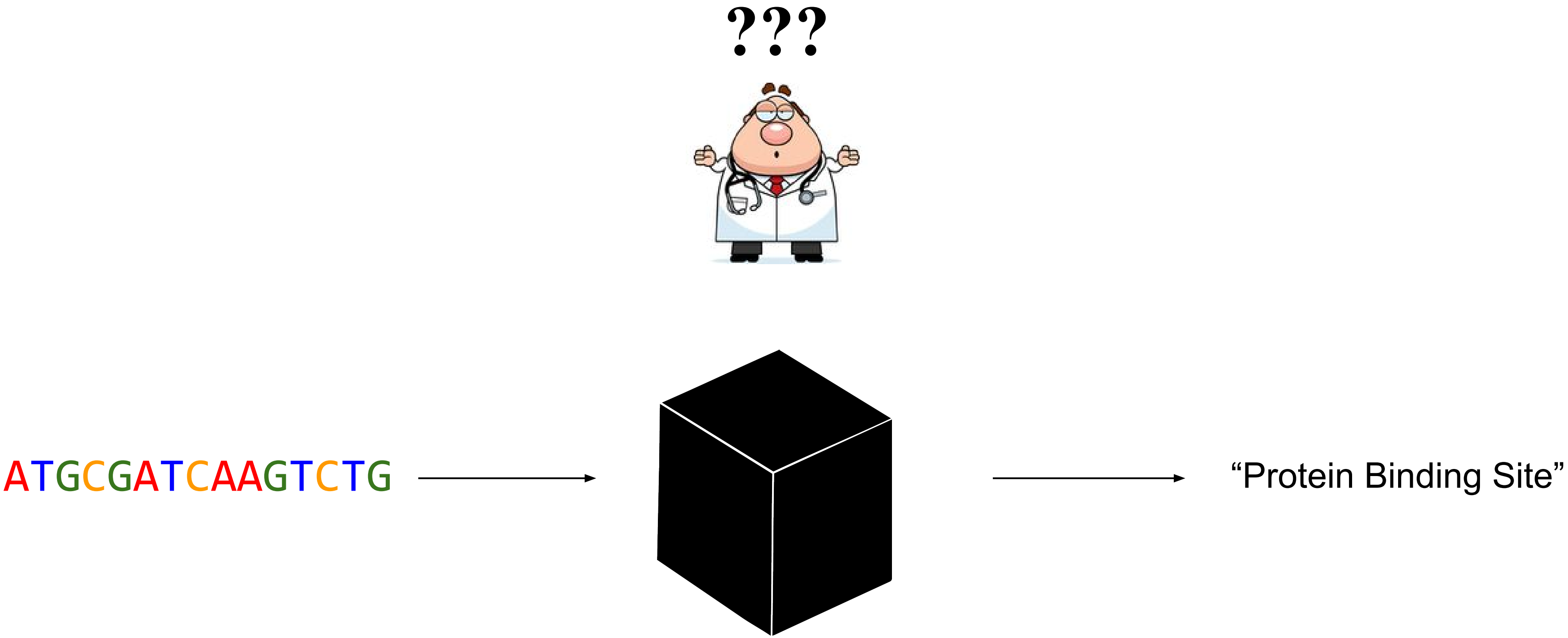
“Dog”

Can get overly sentimental at times, but Gus Van Sant's sensitive direction... and his excellent use of the city make it a hugely entertaining and effective film.

Full Review... | May 25, 2006

ATGCGATCAAGTCTG

“Protein Binding Site”

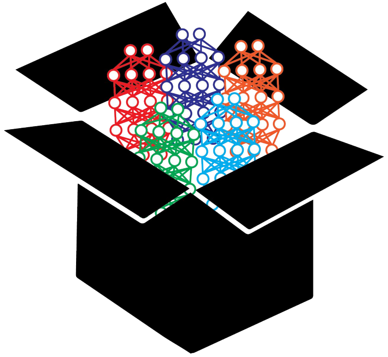


???

ATGCGATCAAGTCTG

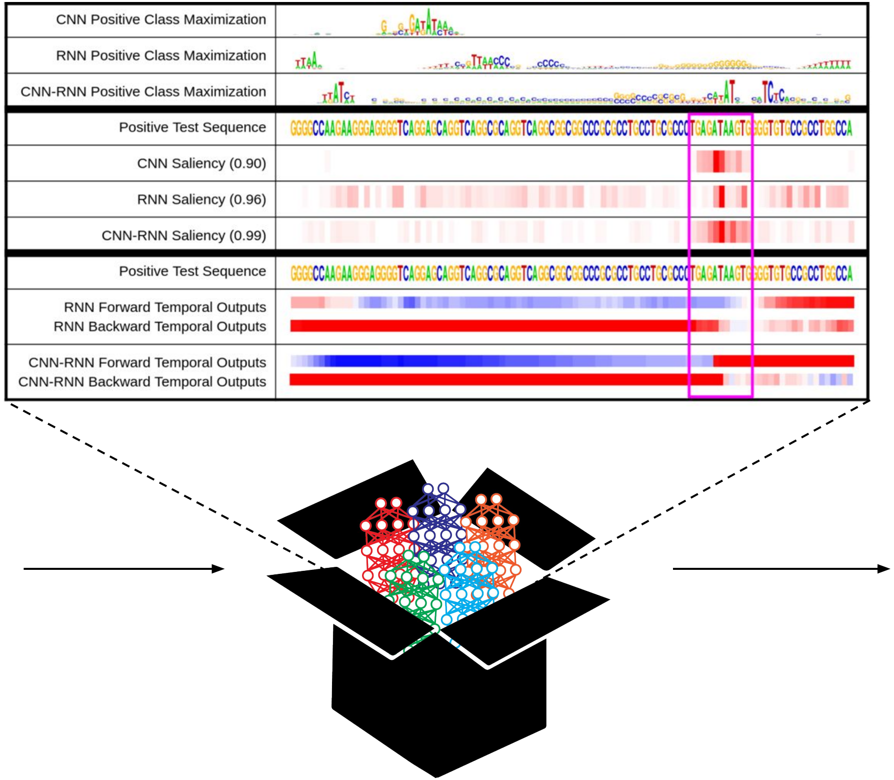
“Protein Binding Site”

Deep **Mo**tif Dashboard: Opening the black box for deep-learning based genomic sequence classifications



ATGCGATCAAGTCTG → → “Protein Binding Site”

Deep **Mo**tif Dashboard: Opening the black box for deep-learning based genomic sequence classifications


CNN Positive Class Maximization
RNN Positive Class Maximization
CNN-RNN Positive Class Maximization
Positive Test Sequence
CNN Saliency (0.90)
RNN Saliency (0.96)
CNN-RNN Saliency (0.99)
Positive Test Sequence
RNN Forward Temporal Outputs
RNN Backward Temporal Outputs
CNN-RNN Forward Temporal Outputs
CNN-RNN Backward Temporal Outputs

ATGCGATCAAGTCTG → → “Protein Binding Site”

Introduction

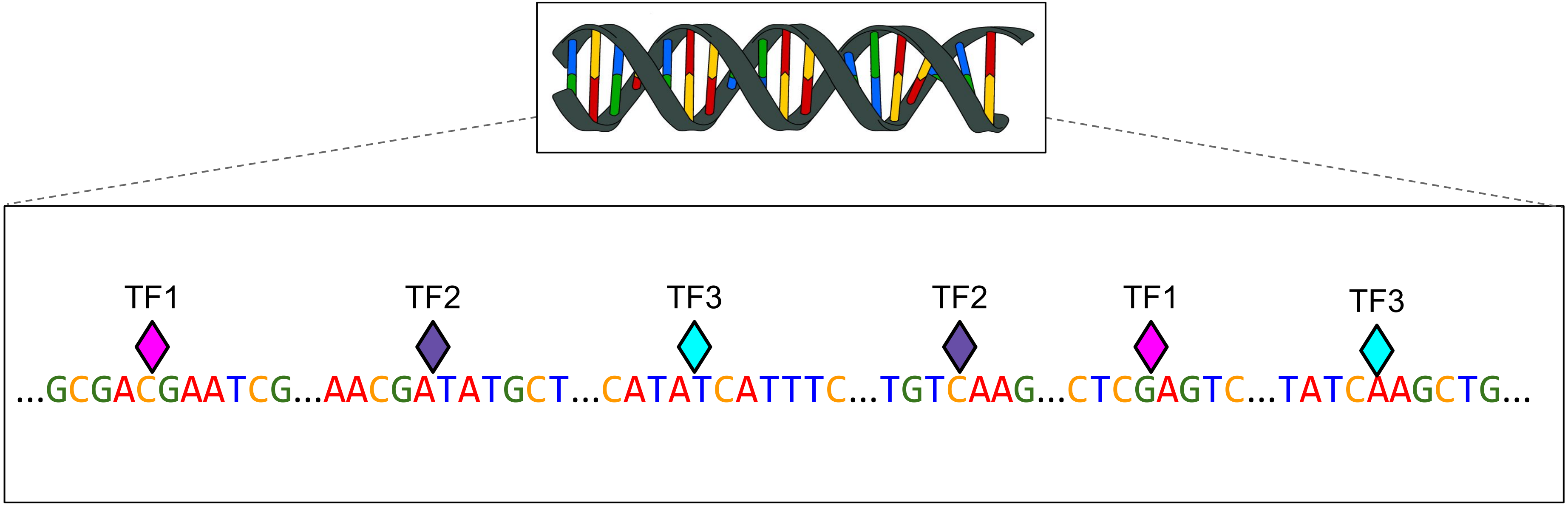
TFBS Classification Task

Neural Models

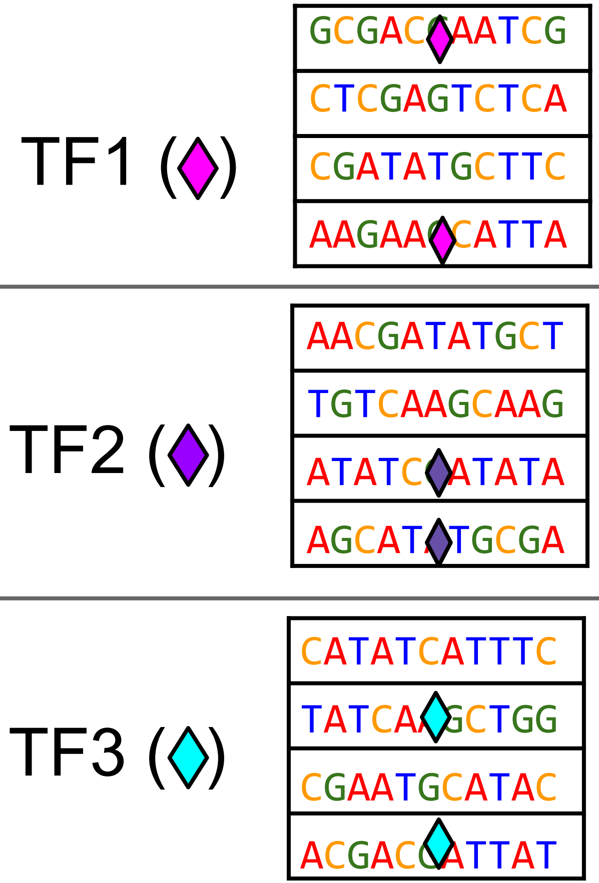
Visualization Methods

Evaluation and Results

Transcription Factor Binding Sites (TFBSs)


TF1
TF2
TF3
TF2
TF1
TF3
...GCGACGAATCG...AACGATATGCT...CATATCATTTC...TGTCAAG...CTCGAGTC...TATCAAGCTG...

TFBS Classification Datasets



TF1 (◆)

GCGACGAATCG

CTCGAGTCTCA

CGATATGCTTC

AAGAAGCATTA

TF2 (◆)

AACGATATGCT

TGTCAAGCAAG

ATATCGATATA

AGCATATGCGA

TF3 (◆)

CATATCATTTC

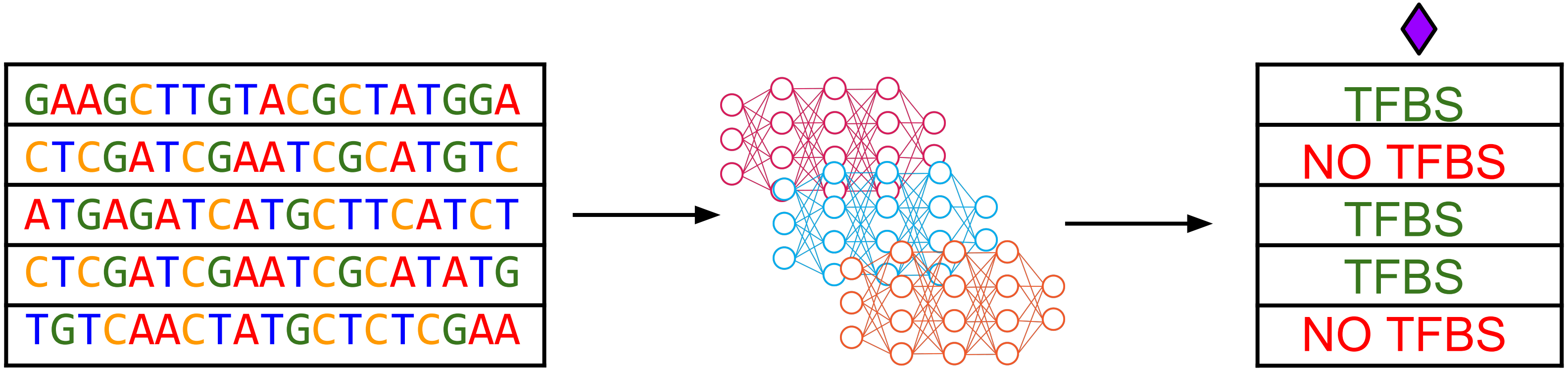
TATCAAGCTGG

CGAATGCATAC

ACGACGATTAT

Deep **Mo**tif (DeMo) Dashboard Approach

1.



GAAGCTTGTACGCTATGGA

CTCGATCGAATCGCATGTC

ATGAGATCATGCTTCATCT

CTCGATCGAATCGCATATG

TGTCAACTATGCTCTCGAA

TFBS

NO TFBS

TFBS

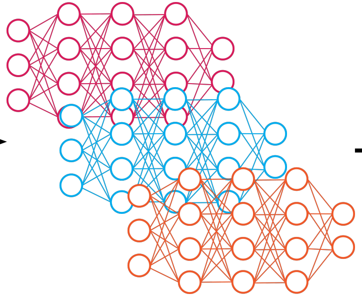
TFBS

NO TFBS

Deep Motif (DeMo) Dashboard Approach

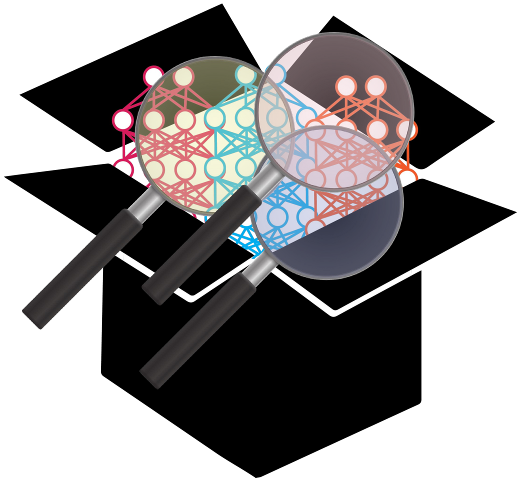
1.

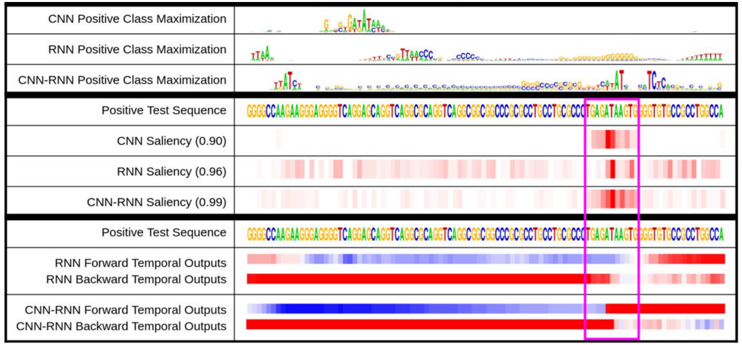
GAAGCTTGTACGCTATGGA
CTCGATCGAATCGCATGTC
ATGAGATCATGCTTCATCT
CTCGATCGAATCGCATATG
TGTCAACTATGCTCTCGAA



TFBS
NO TFBS
TFBS
TFBS
NO TFBS

2.




CNN Positive Class Maximization
RNN Positive Class Maximization
CNN-RNN Positive Class Maximization
Positive Test Sequence
GGGGCCAAGAAGGGAGGGGTCAGGAGCAGGTCAGGCGCAGGTCAGGCGGCGGCCCGCGCCTGCCTGCGCCCTGAGATAAGTGGGGTGTGCCGCCTGGCCA
CNN Saliency (0.90)
RNN Saliency (0.96)
CNN-RNN Saliency (0.99)
Positive Test Sequence
GGGGCCAAGAAGGGAGGGGTCAGGAGCAGGTCAGGCGCAGGTCAGGCGGCGGCCCGCGCCTGCCTGCGCCCTGAGATAAGTGGGGTGTGCCGCCTGGCCA
RNN Forward Temporal Outputs
RNN Backward Temporal Outputs
CNN-RNN Forward Temporal Outputs
CNN-RNN Backward Temporal Outputs

Introduction

TFBS Classification Task

Neural Models

Visualization Methods

Evaluation and Results

Neural Network Models

1. Convolutional (CNN)
2. Recurrent (RNN)
3. Convolutional-Recurrent (CNN-RNN)

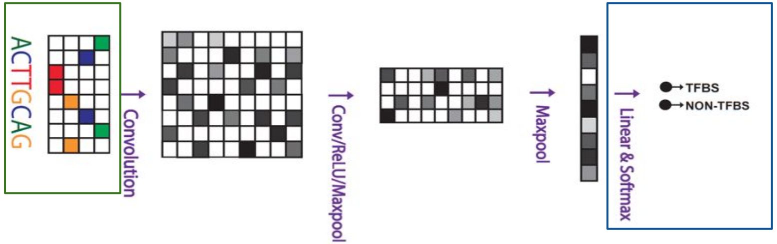
3 Neural Models

Input Sequence

Probability of Binding Site

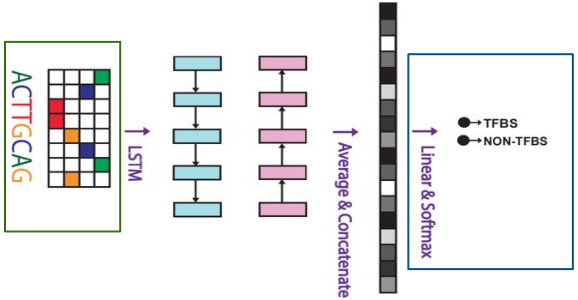
1. Convolutional (CNN)

(short local patterns, or motifs)



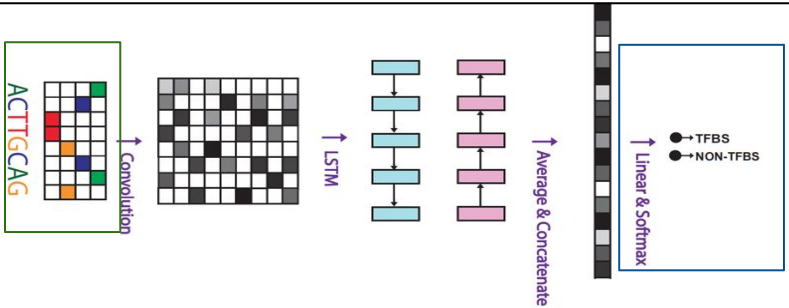
2. Recurrent (RNN)

(long term dependencies)



3. Convolutional-Recurrent (CNN-RNN)

(long term dependencies among motifs)



Introduction

TFBS Classification Task

Neural Models

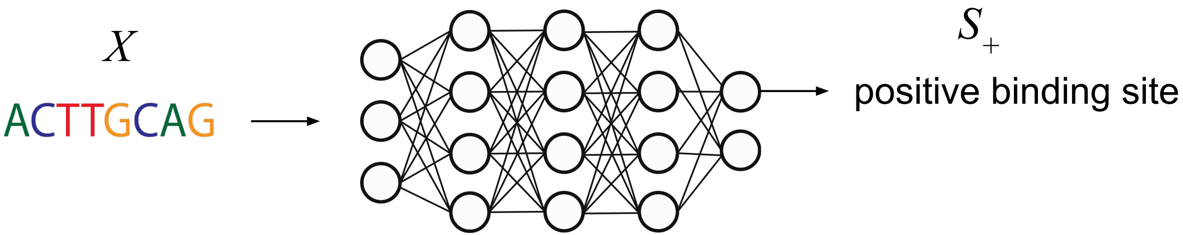
Visualization Methods

Evaluation and Results

Visualization Methods

1. Saliency Maps
2. Temporal Output Values
3. Class Optimization

1. Saliency Map



X

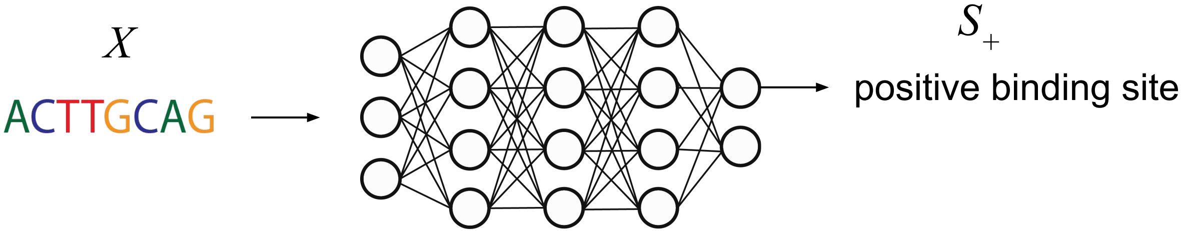
ACTTGCAG

S_+

positive binding site

Which nucleotides are most important for classification?

1. Saliency Map



X

ACTTGCAG

S_+

positive binding site

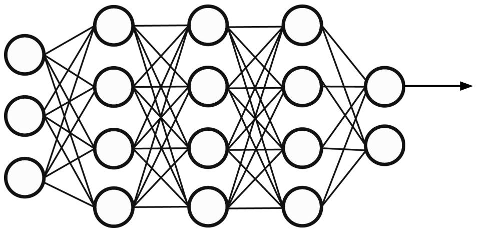
$$S_+(X) \approx w^T X + b = \sum_{i=1}^{|X|} w_i x_i$$

$$w = \left.\frac{\partial S_+}{\partial X}\right|_{X_0} = \text{“\textit{saliency map}”}$$

1. Saliency Map

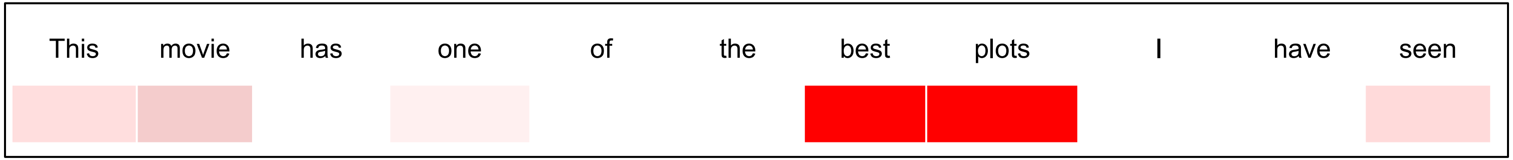
X

This movie has one of the best plots I have seen →



S_+

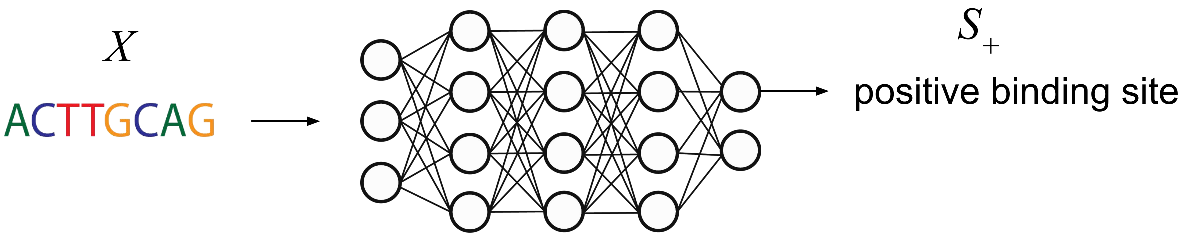
Positive sentiment



This	movie	has	one	of	the	best	plots	I	have	seen

🟥 = important for classification

1. Saliency Map

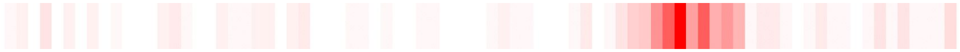


X

ACTTGCAG

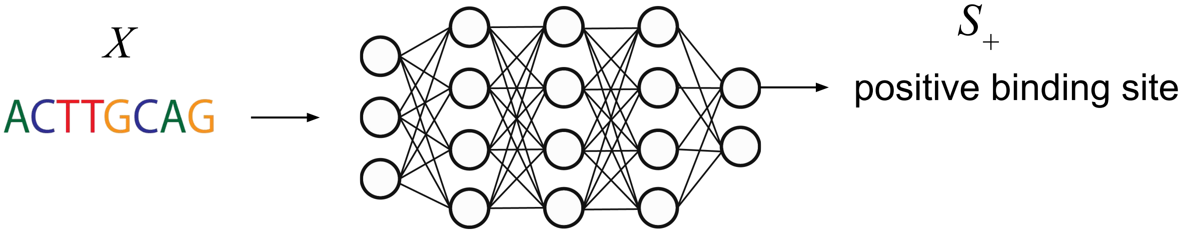
S_+

positive binding site

Positive Test Sequence	TGCTCGCATCCTATTGGCCACGTTAGTCACATGGCCCCACCTGGCTGCAAAGCACGCTGGGAAACGTAGTCTTTCTT
Saliency Map	

🟥 = important nucleotide for prediction

2. Temporal Output Values



X

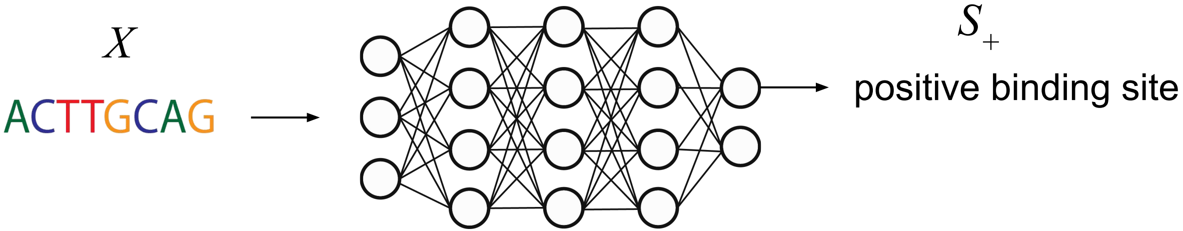
ACTTGCAG

S_+

positive binding site

What are the model's predictions at each timestep of the DNA sequence?

2. Temporal Output Values



X

ACTTGCAG

S_+

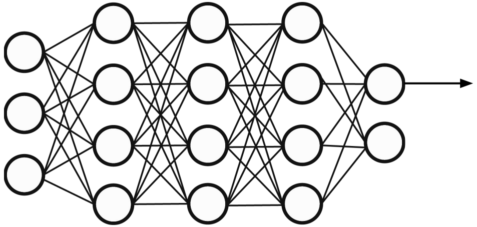
positive binding site

Check the RNN’s prediction scores when we vary the input of the RNN starting from the beginning to the end of a sequence.

2. Temporal Output Values

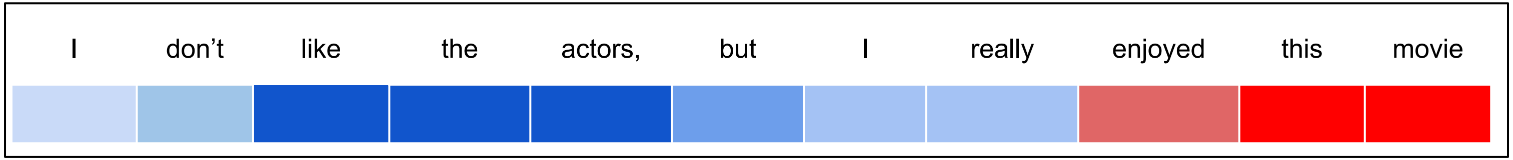
X

I don't like the actors, but I really enjoyed this movie →



S_+

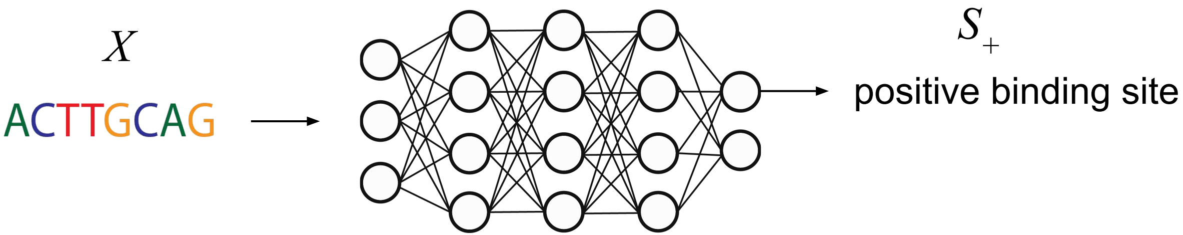
positive sentiment

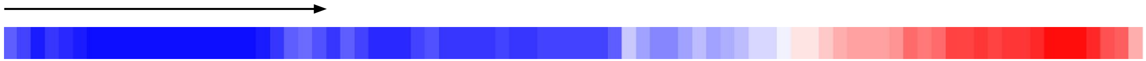
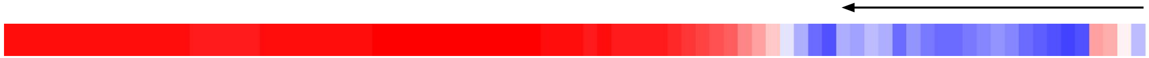


🟦 = negative sentiment

🟥 = positive sentiment

2. Temporal Output Values



Positive Test Sequence	CTTCTGCTCGCATCCTATTGGCCACGTTAGTCACATGGCCCCACCTGGCTGCAAAGCACGCTGGGAAACGTAGTCTTTCTT
RNN Forward Output	
RNN Backward Output	

🟦 = negative binding site prediction 🟥 = positive binding site prediction

Visualization Methods

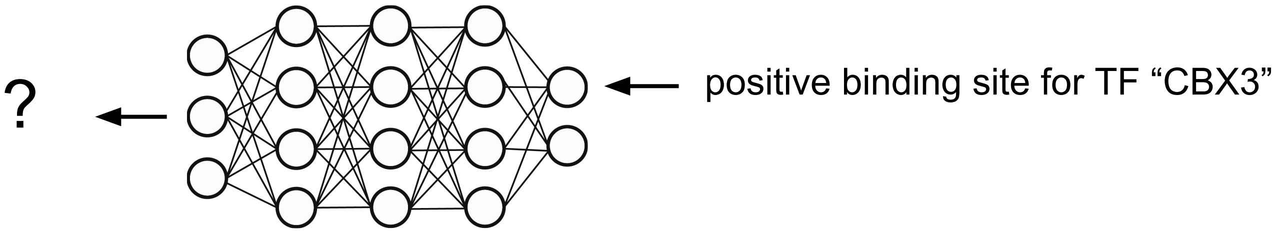
Sequence Specific

1. Saliency Maps
2. Temporal Output Values

TF Specific

3. Class Optimization

3. Class Optimization

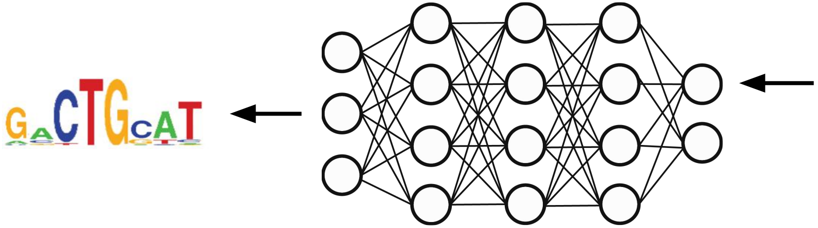


?

positive binding site for TF “CBX3”

For a particular TF, what does the optimal binding site sequence look like?

3. Class Optimization

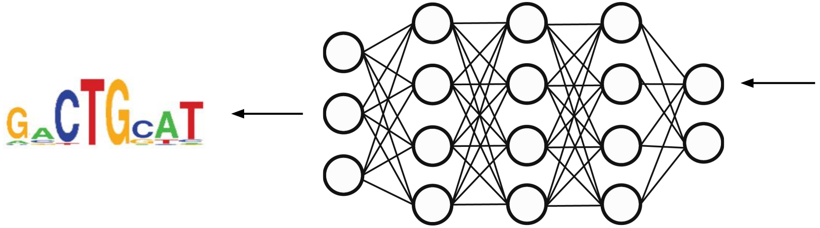


positive binding site for TF “CBX3”

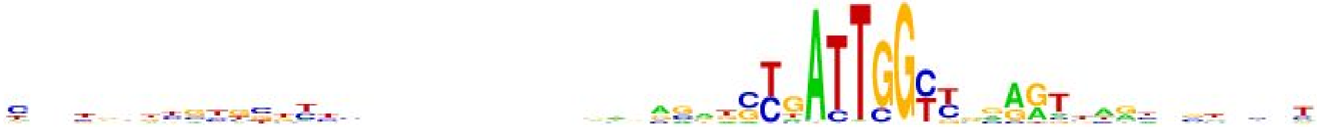
$$\underset{X}{\arg\max}\, S_+(X) + \lambda\|X\|_2^2$$

Where X is the input sequence and the score S_+ is probability of sequence X being a positive binding site

3. Class Optimization



positive binding site for TF “CBX3”

Optimal binding site for TF “CBX3”	

Introduction

TFBS Classification Task

Neural Models

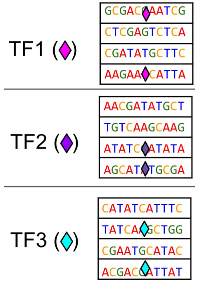
Visualization Methods

Evaluation and Results

Experimental Setup

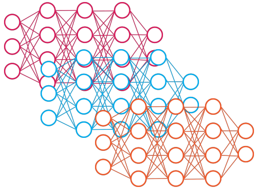
Dataset

- Alipanahi et al. "*Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning*". Nature Biotechnology 2015.
- 108 cancer cell TFs (train separate model for each TF)
- Each sequence is 101-length centered around ChIP-seq peak



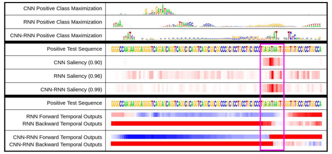
Models

- Test several variations of 3 different models (CNN, RNN, CNN-RNN)

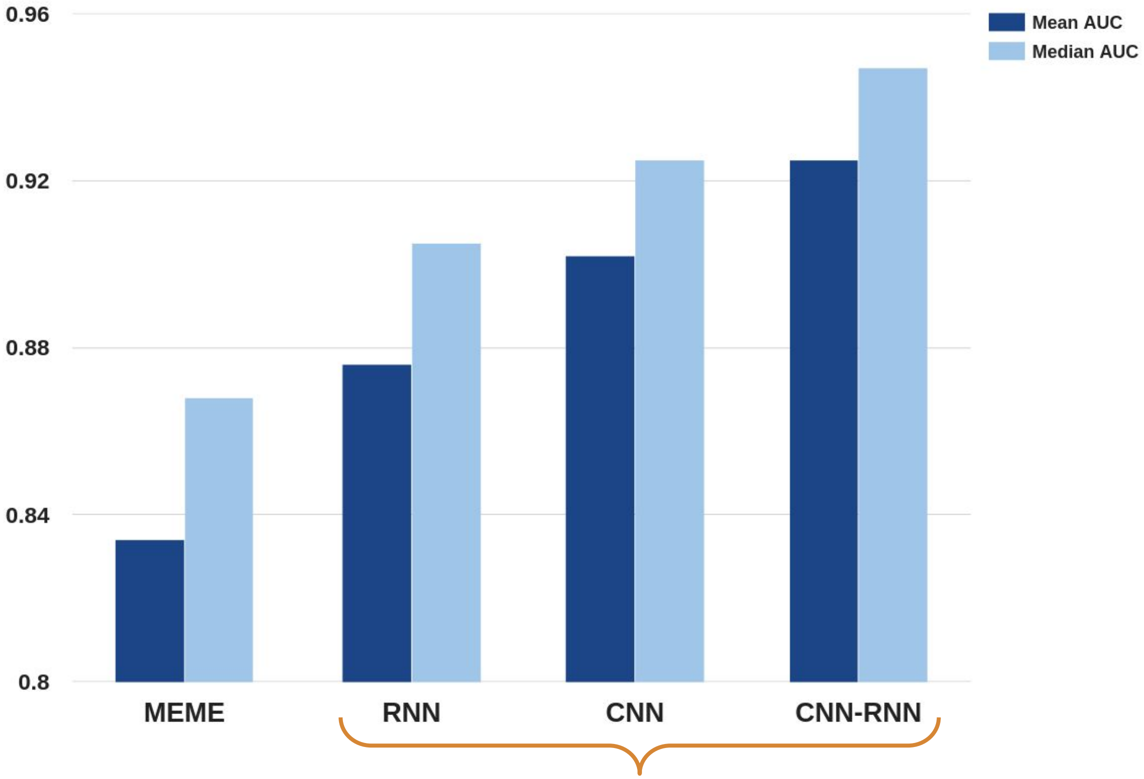


Evaluation

- Compare models using AUC scores on test set
- Evaluate visualization methods manually and by motif matching

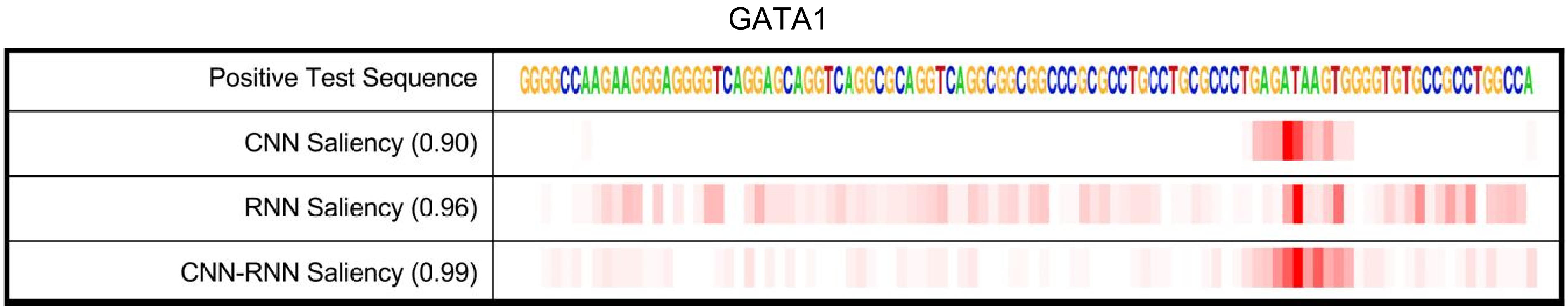


Model Accuracy (AUC Scores)



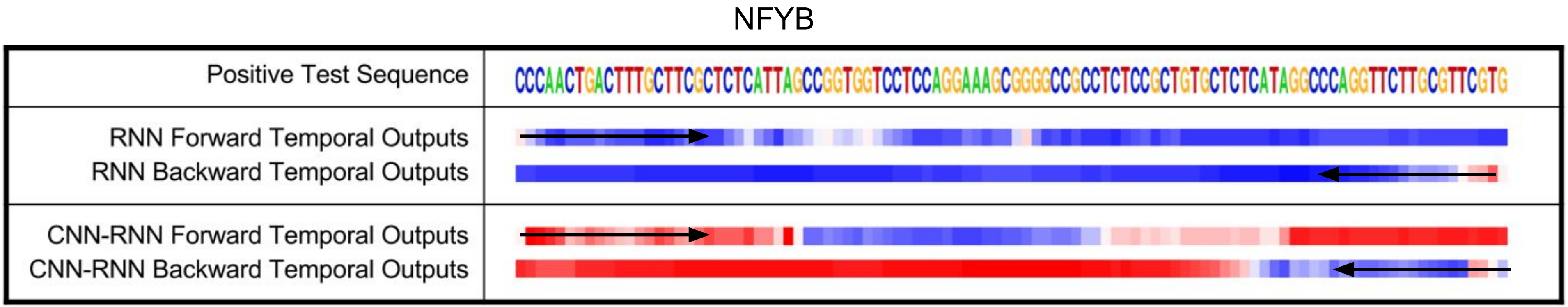
Our Models

1. Saliency Maps



🟥 = important nucleotide for prediction

2. Temporal Output Values

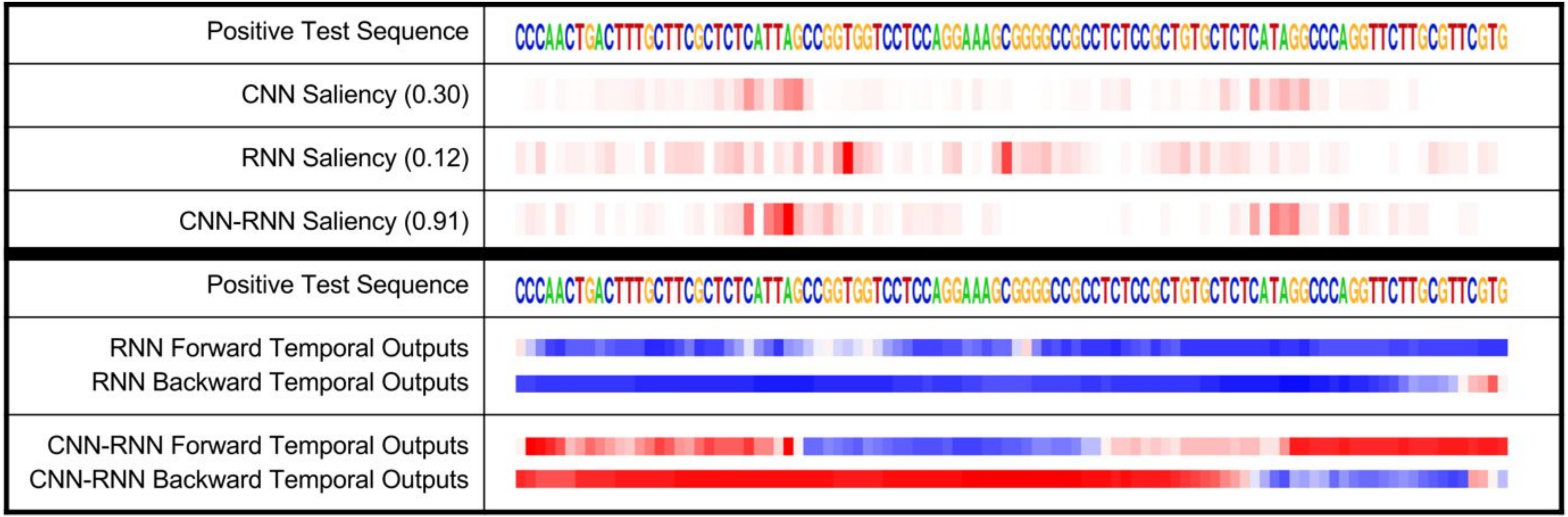


🟦 = negative binding site prediction

🟥 = positive binding site prediction

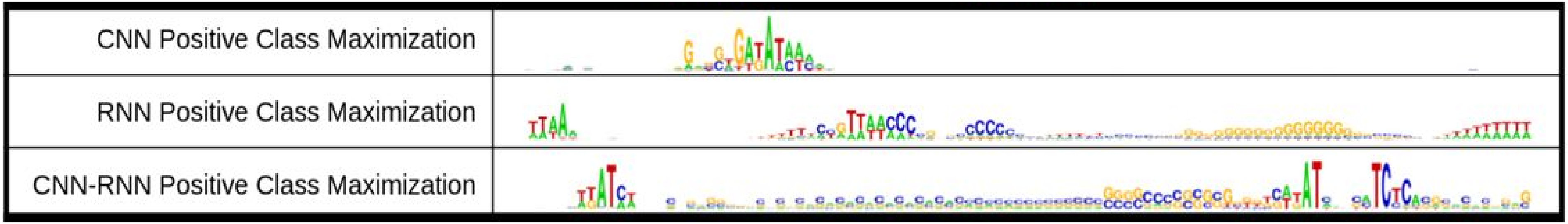
Saliency Map AND Temporal Output Values

NFYB

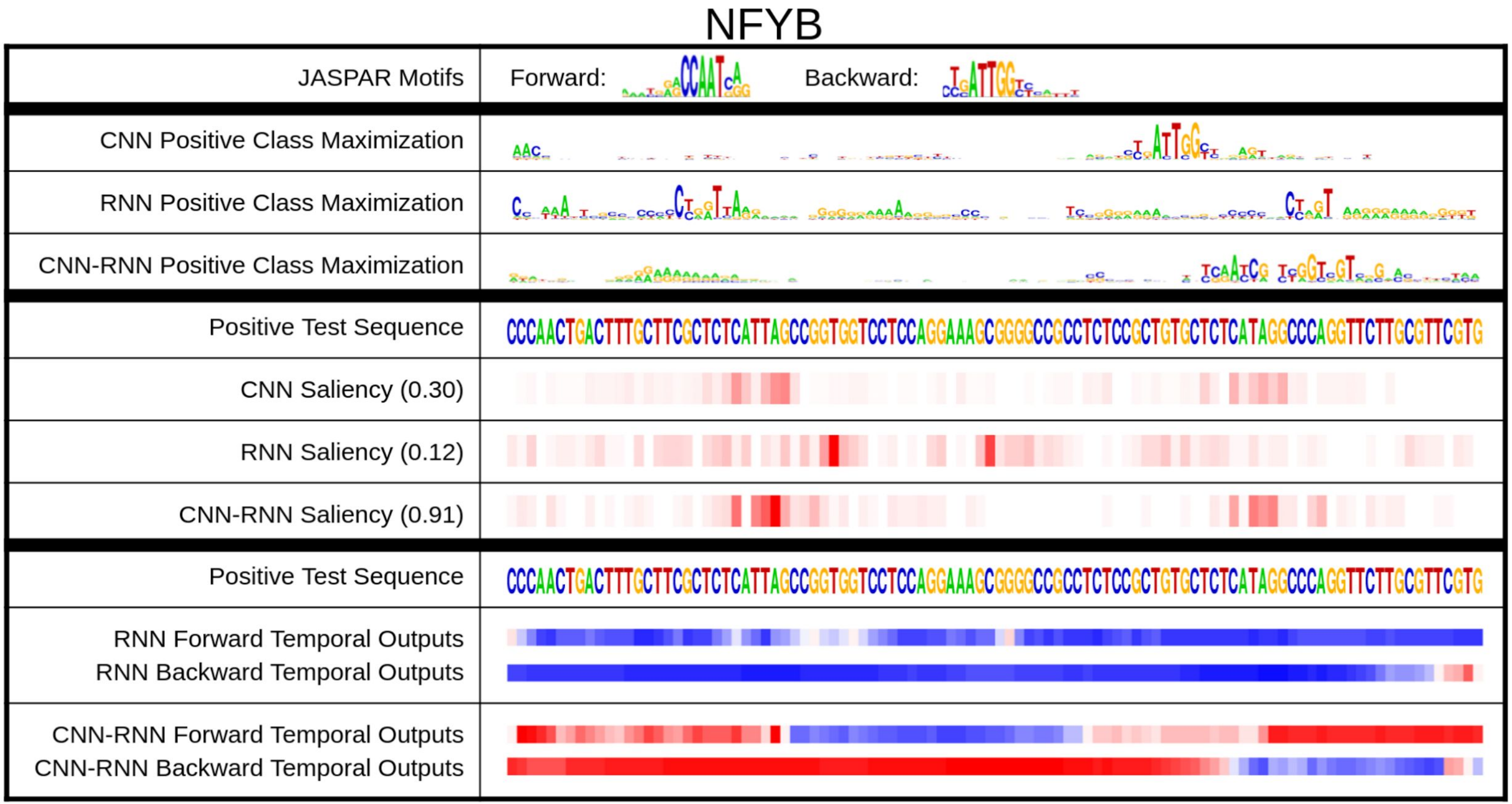


3. Class Optimization

GATA1



DeMo Dashboard



Deep Motif (DeMo) **Dashboard** Contributions and Results

1. Comparative analysis of 3 different neural models on TFBS task
    - CNN-RNNs perform the best

2. Presented 3 different visualization techniques to understand the predictions of neural models
    - Although TFBSs are influenced by motifs, the interactions among motifs are also important

Thank You!

code available at: **deepmotif.org**

Ritambhara Singh

Beilun Wang

Dr. Yanjun Qi

UVA Machine Learning and Biomedicine Group

UNIVERSITY of VIRGINIA