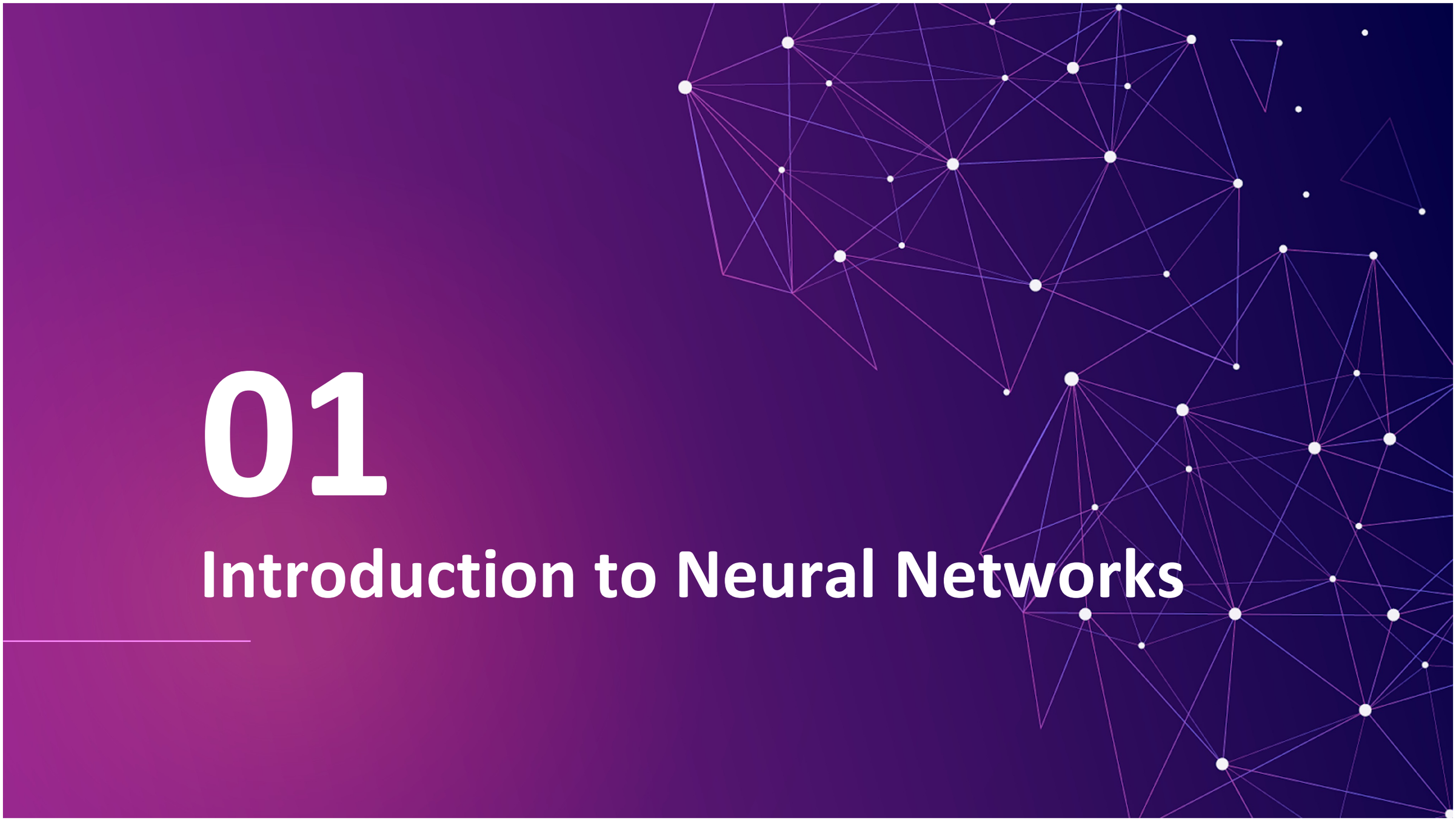
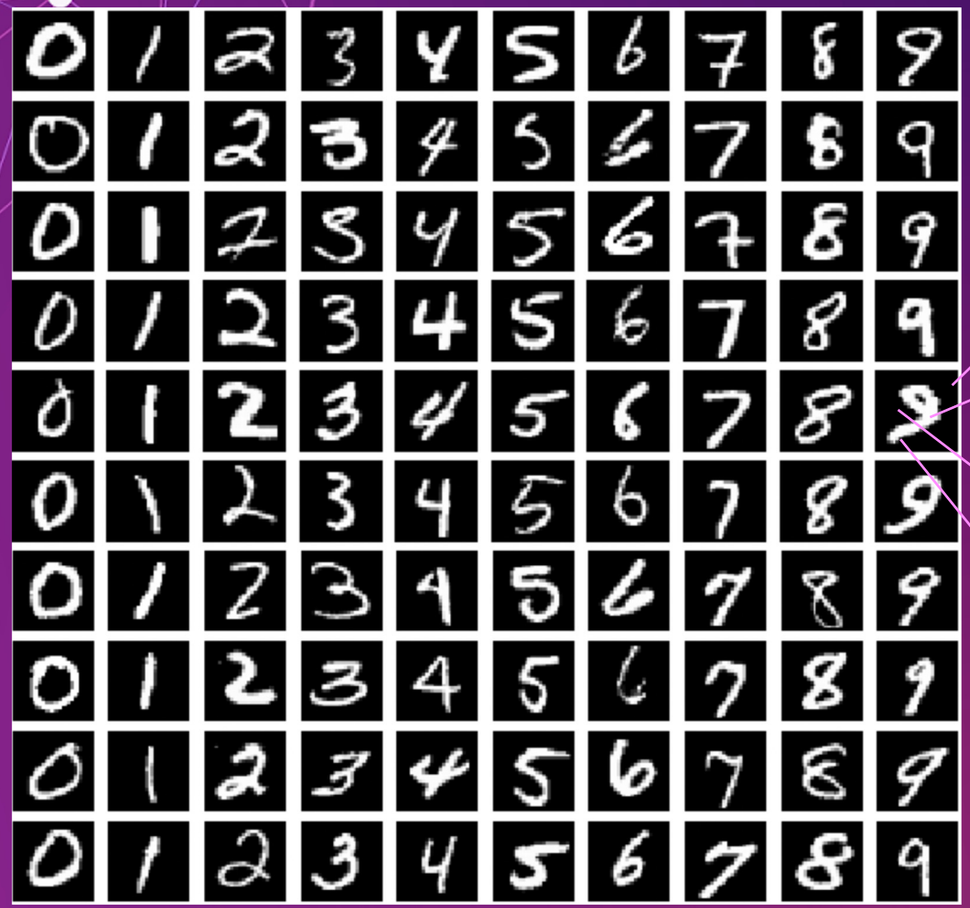
Various Neural Network Architectures for Modeling the Effects of Non-coding DNA

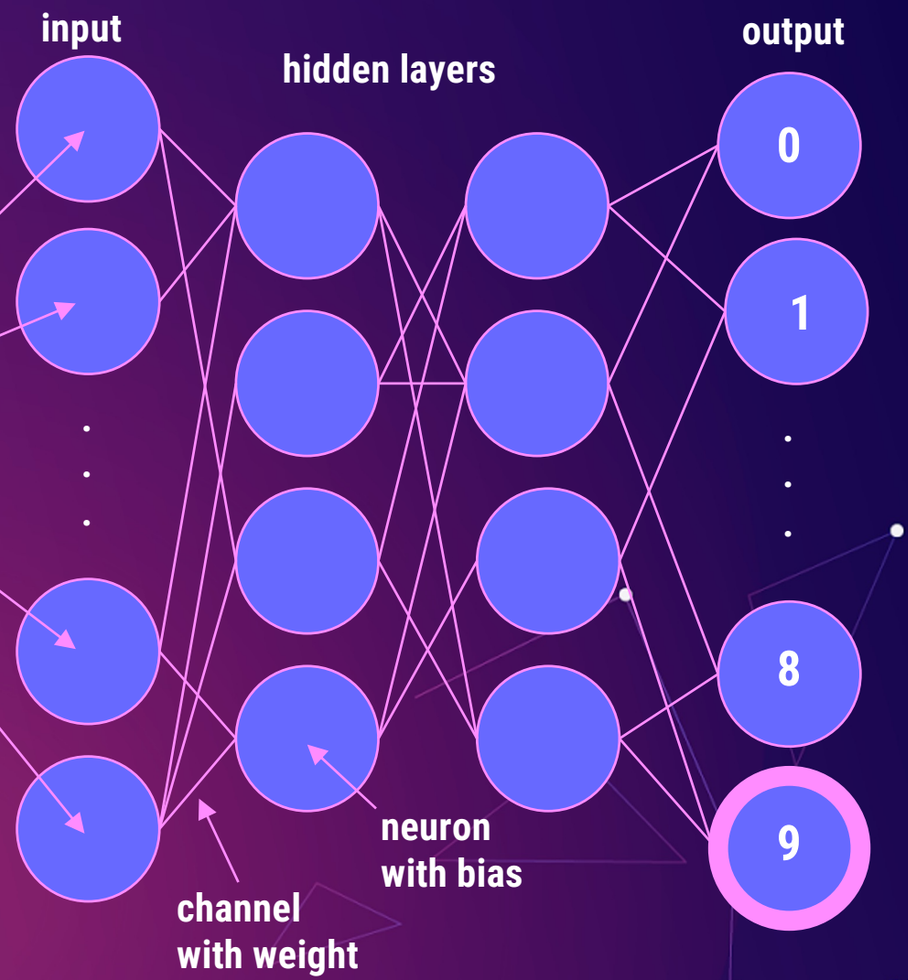
Andria Bao and Sophia Lichterfeld
MIT PRIMES CS Reading Group
December 6, 2022

TABLE OF CONTENTS


01
Introduction to Neural Networks

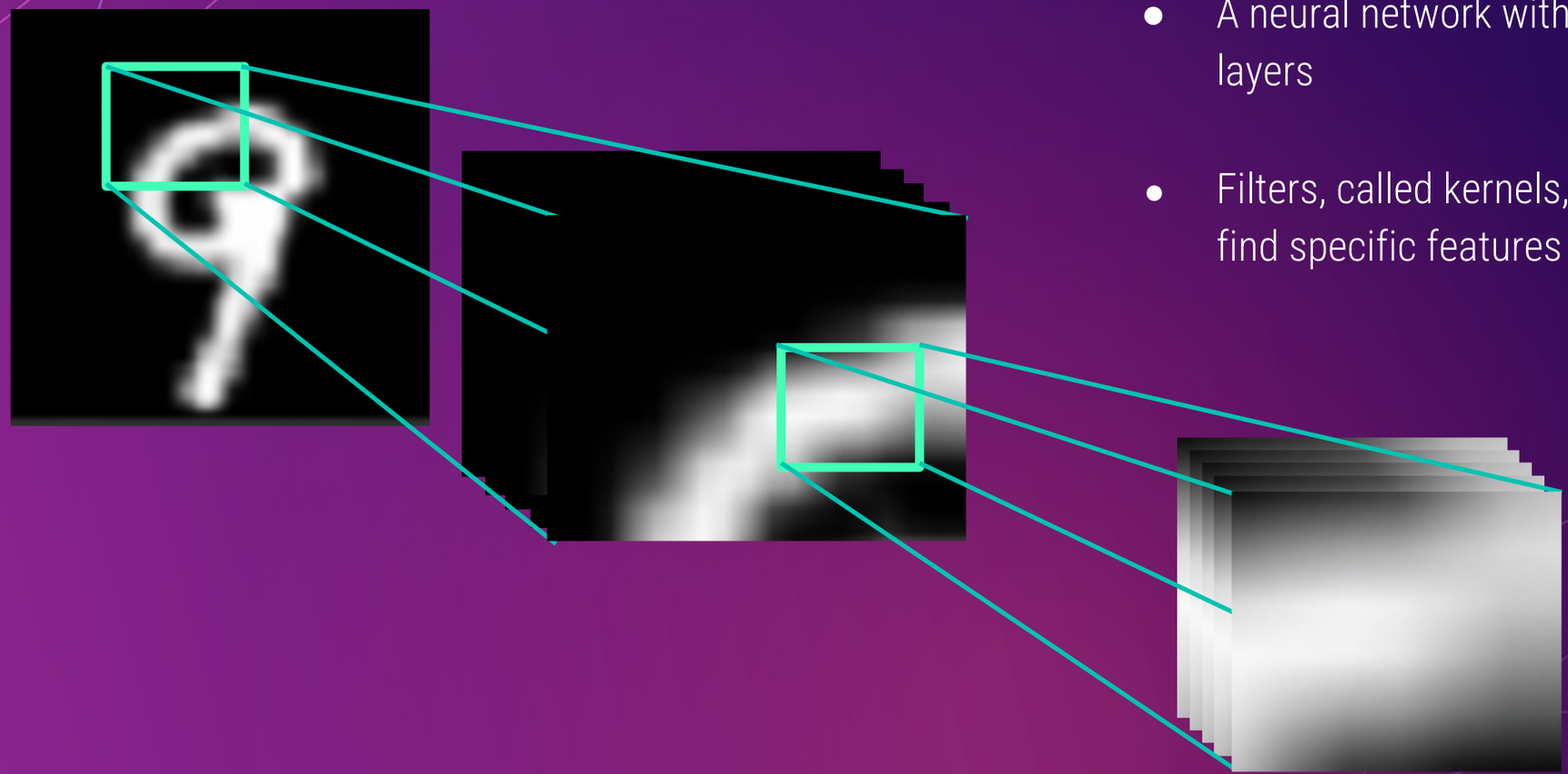
Ex. Recognizing Handwritten Digits





Source: https://www.researchgate.net/figure/Example-images-from-the-MNIST-dataset_fig1_306056875

Convolutional Neural Networks



- A neural network with one or more convolutional layers
- Filters, called kernels, are applied on the data to find specific features in different areas of the image

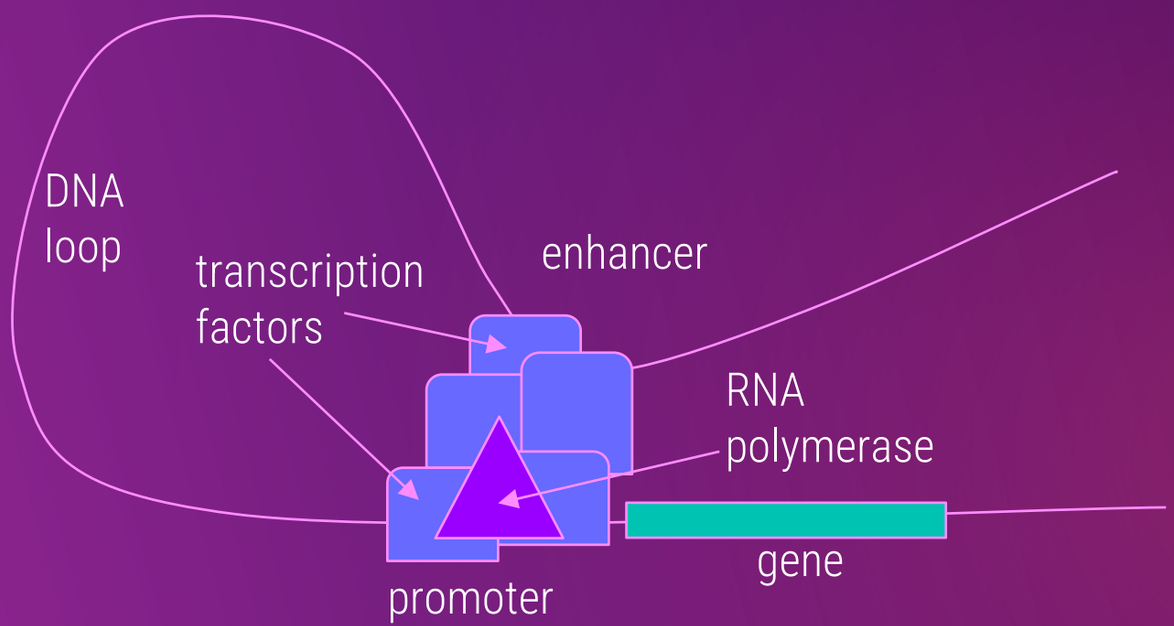
02

Introduction to Non-coding DNA

Key Players in Regulating Transcription

Non-coding DNA

DNA segments that do not encode proteins
→ vs. genes, segments of DNA that are transcribed into mRNA and translated into proteins
→ approximately 98.8% of the human genome is non-coding DNA



Transcription Factors

Proteins that facilitate or hinder the attachment of RNA polymerase.

Promoters

Segments of DNA just upstream of the target gene to which RNA polymerase and general transcription factors bind.

Enhancers

Segments of DNA that can be several thousand base pairs away and either upstream or downstream of the target gene to which gene-specific transcription factors bind.

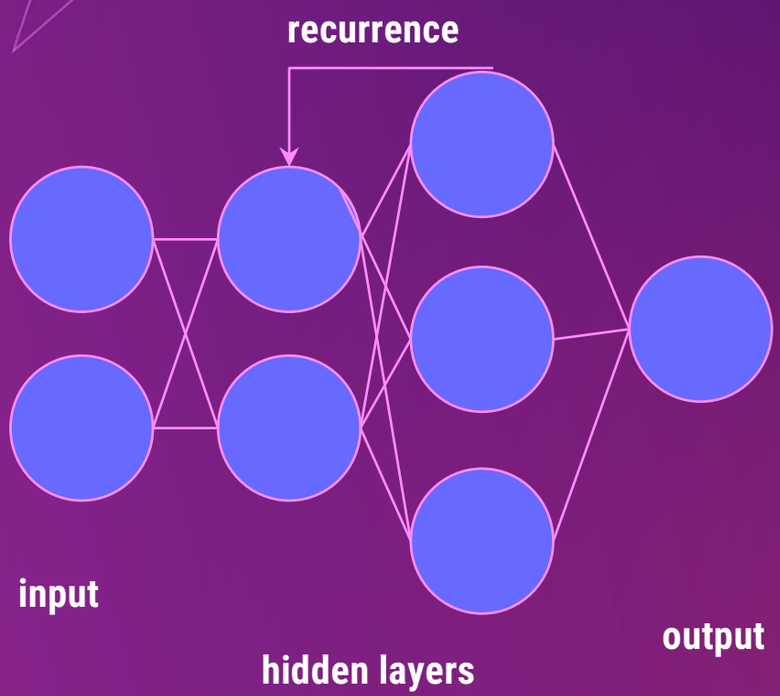
03

DanQ Architecture

Quang D, Xie X. DanQ: a hybrid convolutional and recurrent deep neural network for quantifying the function of DNA sequences. Nucleic Acids Res. 2016 Jun 20;44(11):e107. doi: 10.1093/nar/gkw226. Epub 2016 Apr 15. PMID: 27084946; PMCID: PMC4914104.

RNNs and BLSTMs

Recurrent Neural Networks



Bi-directional Long Short-term Memory Network

- Uses RNN but much more complex
- Remembers past sequences
- Stores long-term and short-term memory

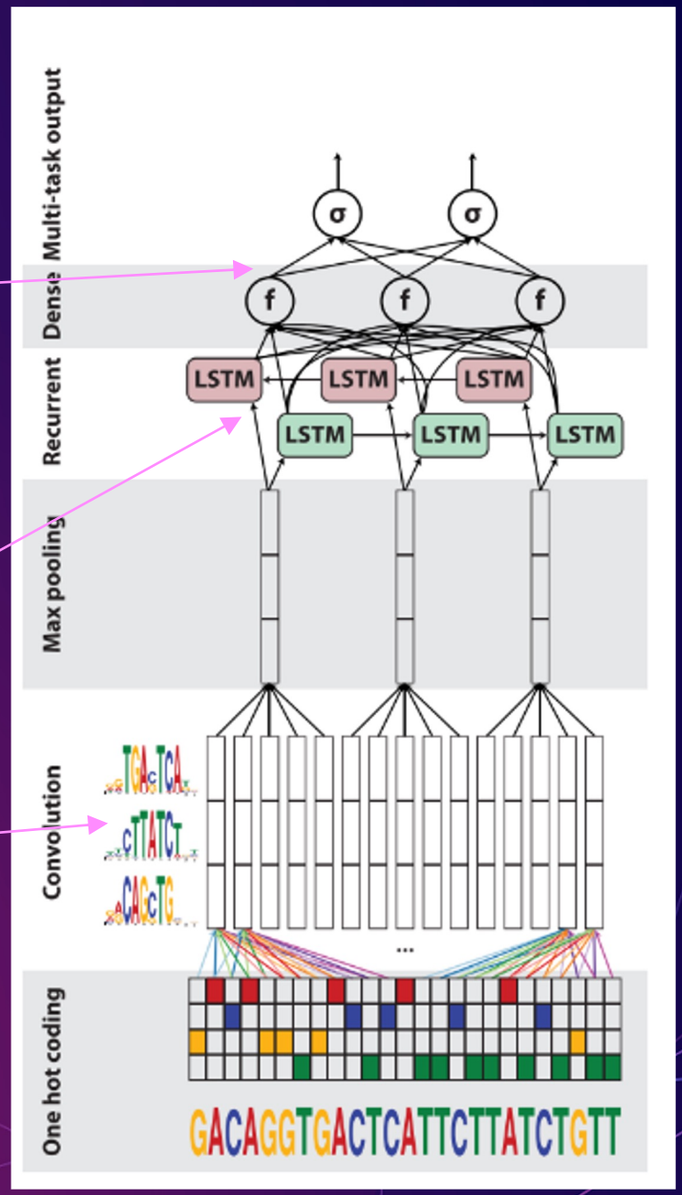
DanQ Architecture

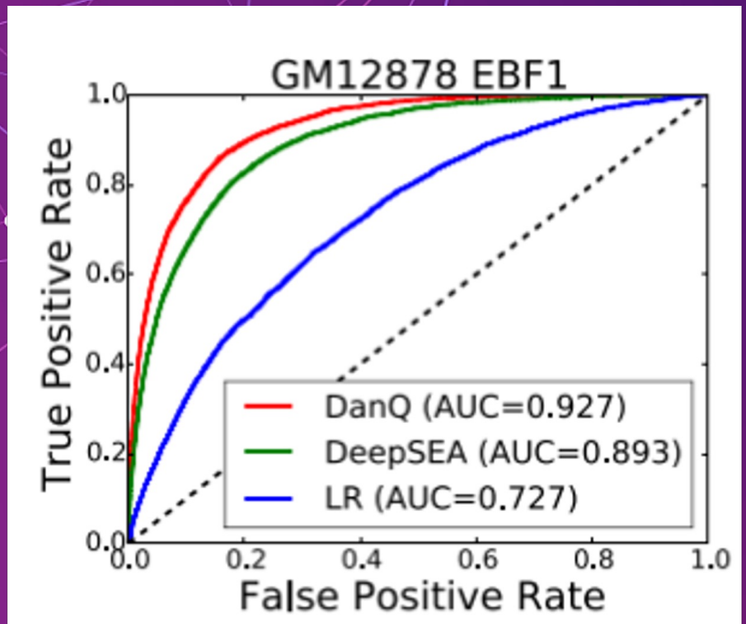
- Based of DeepSEA, a deep learning architecture for predicting gene expression
- DanQ trained with 320 convolution kernals
- DanQ-JASPAR uses JASPAR database and 1024 convolution kernals

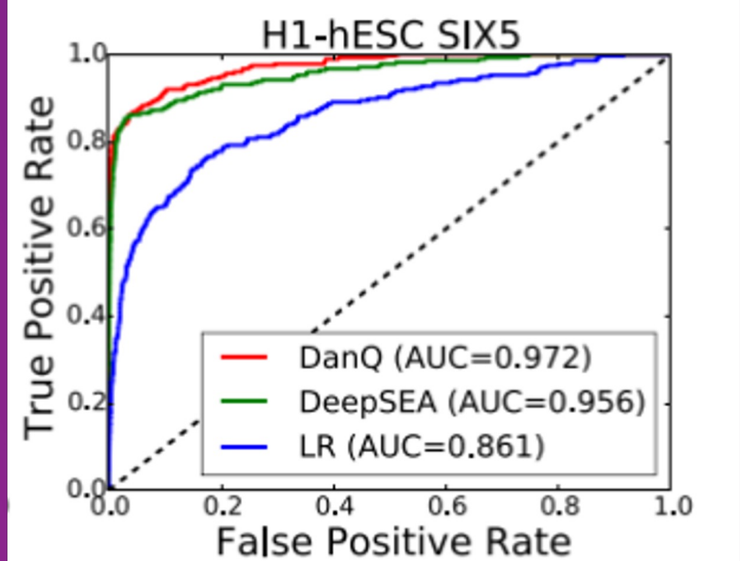
Last two layers prepare loss function & probability predictions

LSTM layers find specific dependencies based on frequency and locations of motifs

Convolution layers find motif sites - long term dependencies between nucleic acide sequences



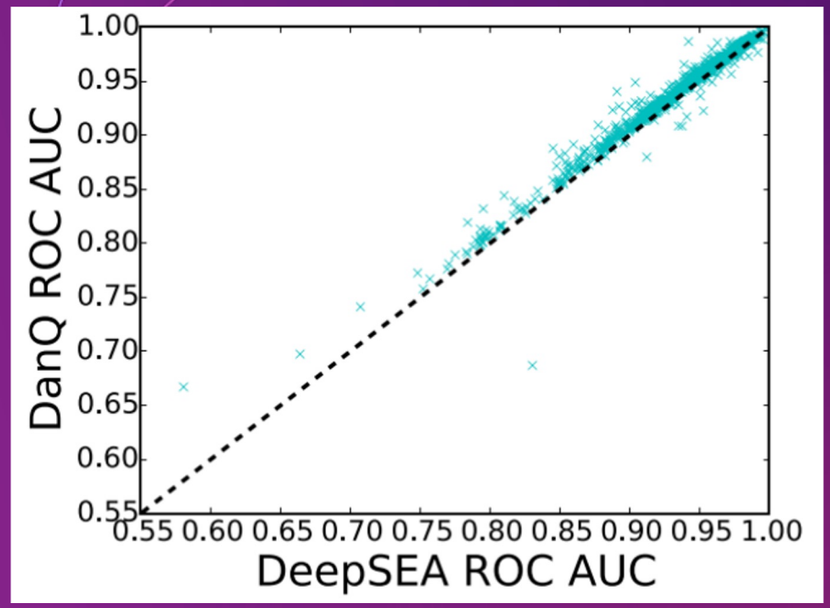




DanQ Results

- AUC ROC: Area under curve of receiver operating characteristics
- DanQ outperforms DeepSEA for 94.1% of targets
- Small improvement of about 1-4%
- Logistic regression has 70% AUC ROC

DanQ Results



- Area under precision recall curve (PR AUC) much more
- LR Baseline PR less than 5%
- Absolute improvement is over 10% and the relative improvement is over 50%
- 97.6% of all DanQ PR AUC scores surpass DeepSEA PR AUC scores

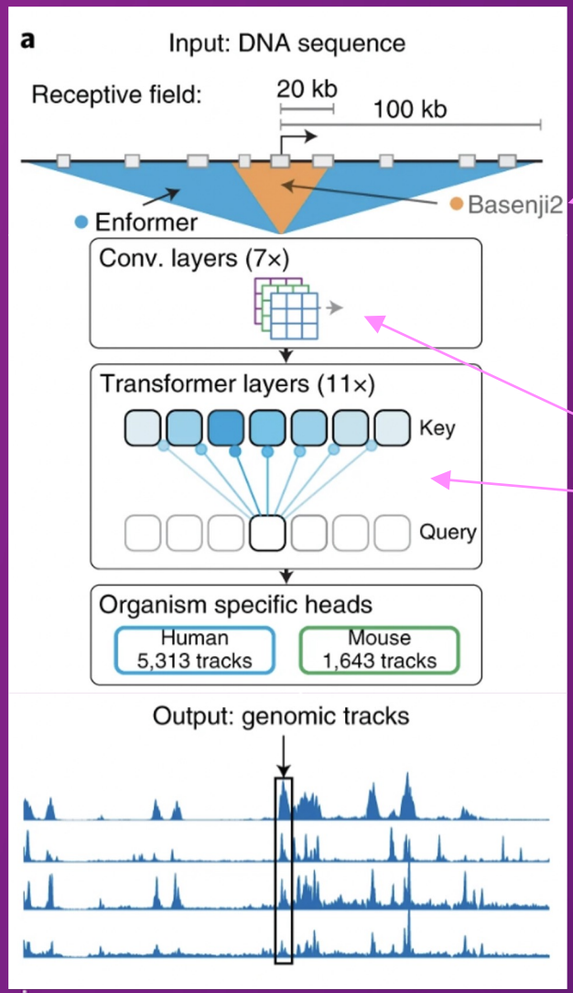
04

Enformer Architecture

Avsec, Ž., Agarwal, V., Visentin, D. *et al.* Effective gene expression prediction from sequence by integrating long-range interactions. *Nat Methods* 18, 1196–1203 (2021). https://doi.org/10.1038/s41592-021-01252-xit

Enformer Architecture

Figure 1a



Basenji2 is a CNN-based architecture that could only successfully detect links within 20kb

The Enformer model consist of 7 convolutional layers followed by 11 transformer layers

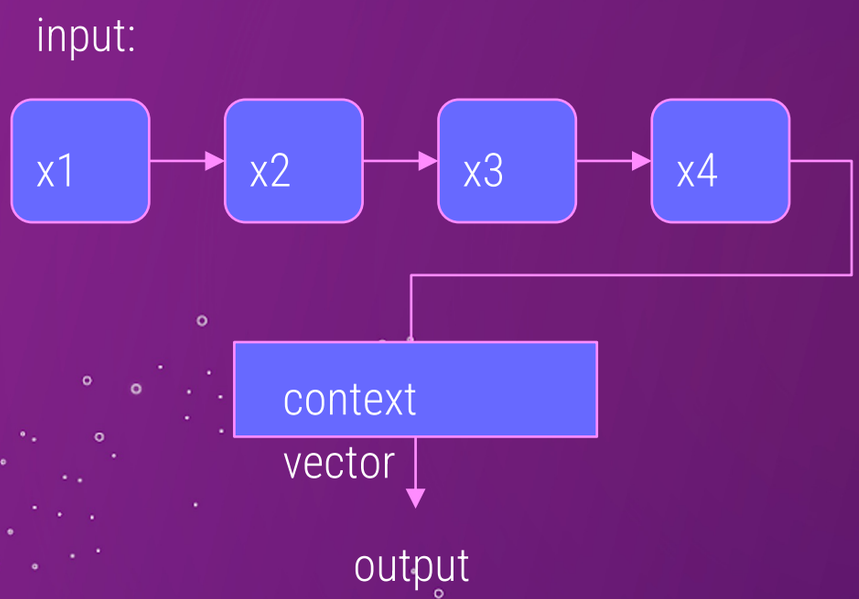
With this construction, Enformer can look for regions of high relevance, specifically enhancer region, up to 100 kb away from the transcription start site

Transformer blocks connect the each position of the input sequence with every other position in the input sequence

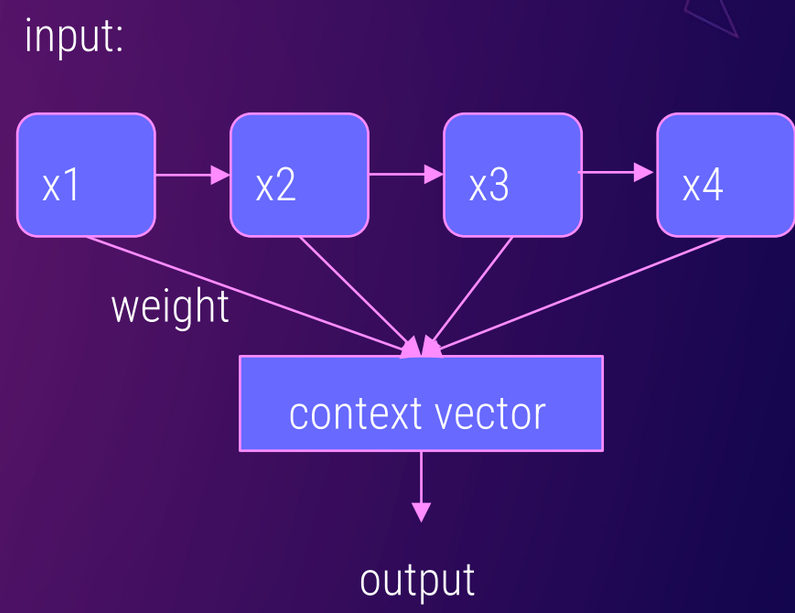
The model looks at which weights show the most relevance to the transcription start site position because this is the area where RNA polymerase must bind to start transcription

Attention Mechanism

Without an attention mechanism, it is difficult to maintain the information from components of the input that are further away

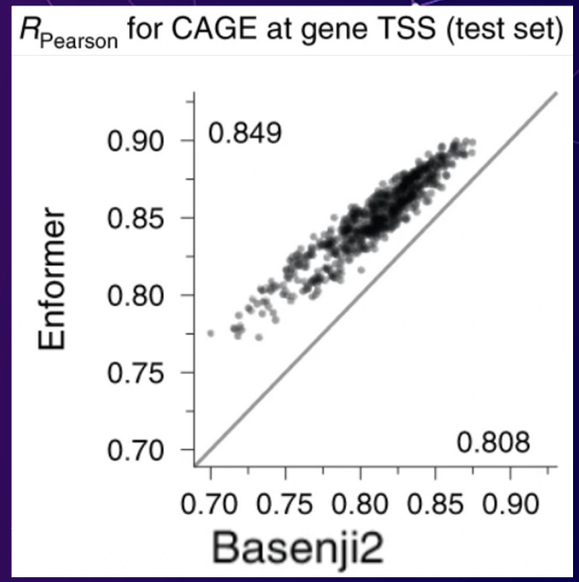


With an attention mechanism, a weighted sum is calculated, which demonstrates higher relevance to the positions with a greater attention weight



Enformer Results

- The Enformer model successfully identifies 84% of enhancer regions for a given gene compared to 47% in previous models that relied more on CNNs
- Compared to Basenji2 (CNN-based), in 100% of tested genes, Enformer had a stronger prediction performance
- When transformer blocks were taken out or the range was purposefully reduced, the modified Enformer could not find distal enhancer regions —> transformer blocks are essential for finding the relationship between DNA regions not located near each other
- The model correctly predicted the effect of a variation in an enhancer region 35kb away from the NLRC5 gene



05

Comparison

Enformer vs. DanQ

Comparison

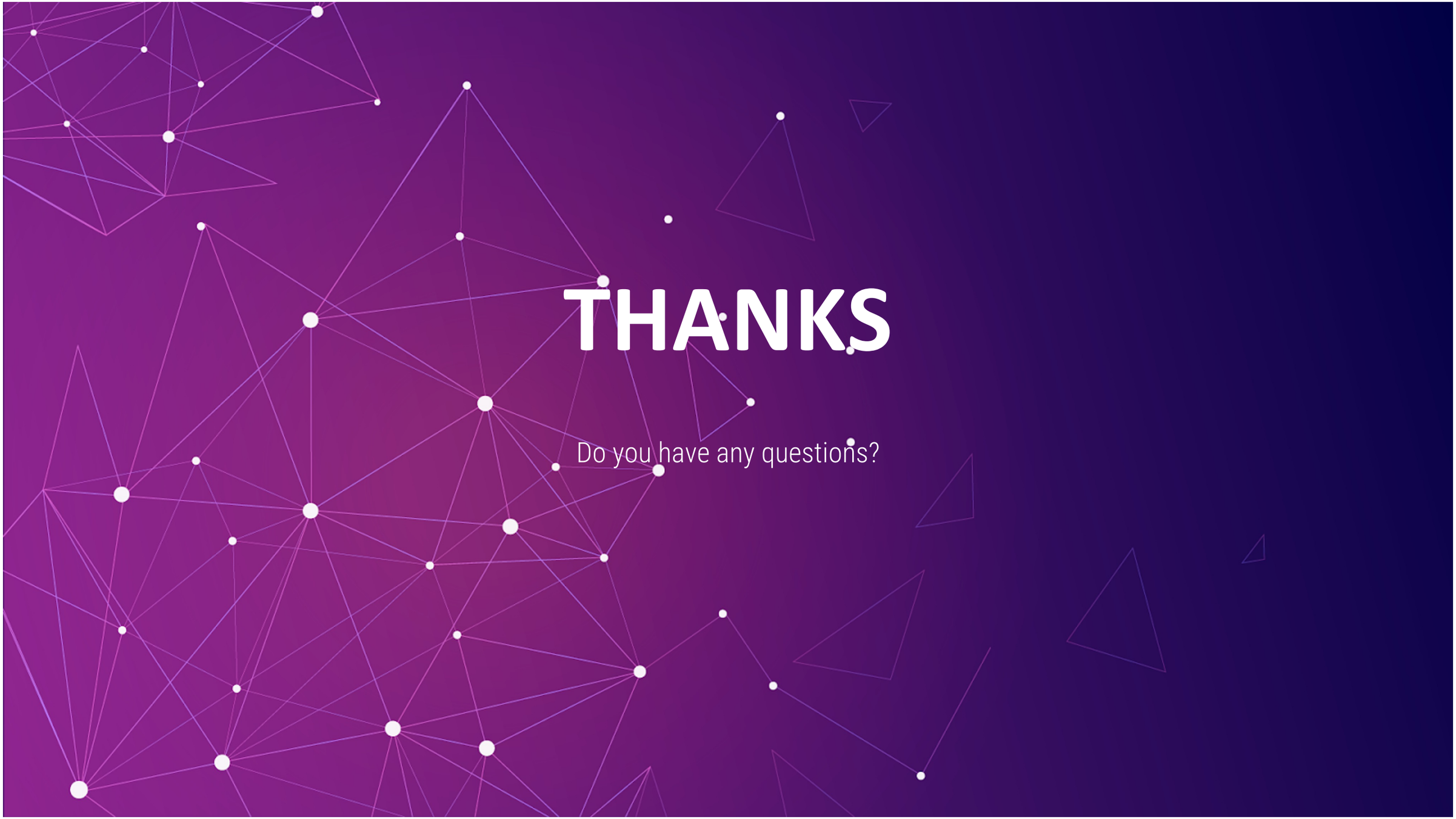
- Both architectures attempt to look at the larger picture
- Both models took in one hot-encoded base pairs as inputs
- Both were significant improvements from previous models

DanQ

- One convolutional layer
- BLSTM layers
- Finds function of DNA sequences
- Published in 2016

Enformer

- 7 convolutional layers
- Transformer layers
- Finds specific enhancer regions that affect gene expression
- Published in 2021


THANKS
Do you have any questions?