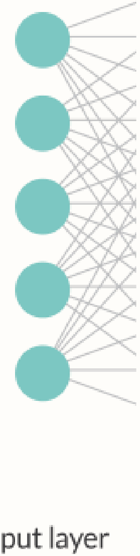
Deep Learning in Genomics:

Models, Interpretability & Motif Discovery

Kimin Nguyen

What is Deep learning?

a subfield of machine learning that uses algorithms called **neural networks** to learn patterns from data

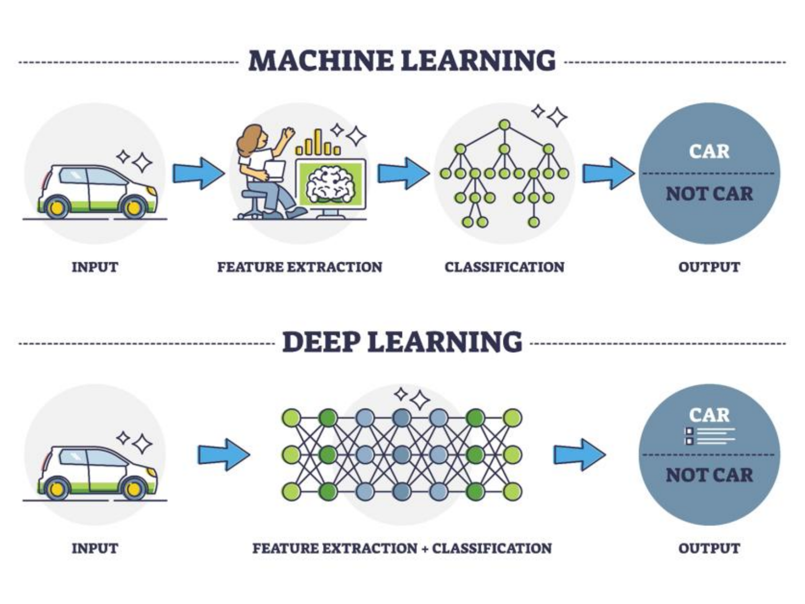


Deep learning vs Machine Learning

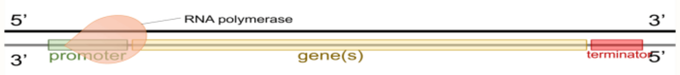
Deep learning can be:

- + Supervised
- + Unsupervised
- + Semi Supervises

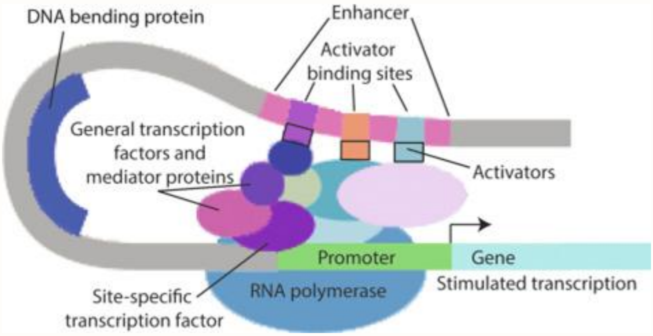
=> Machine Learning requires Data Engineering while Deep Learning learns features automatically from raw data



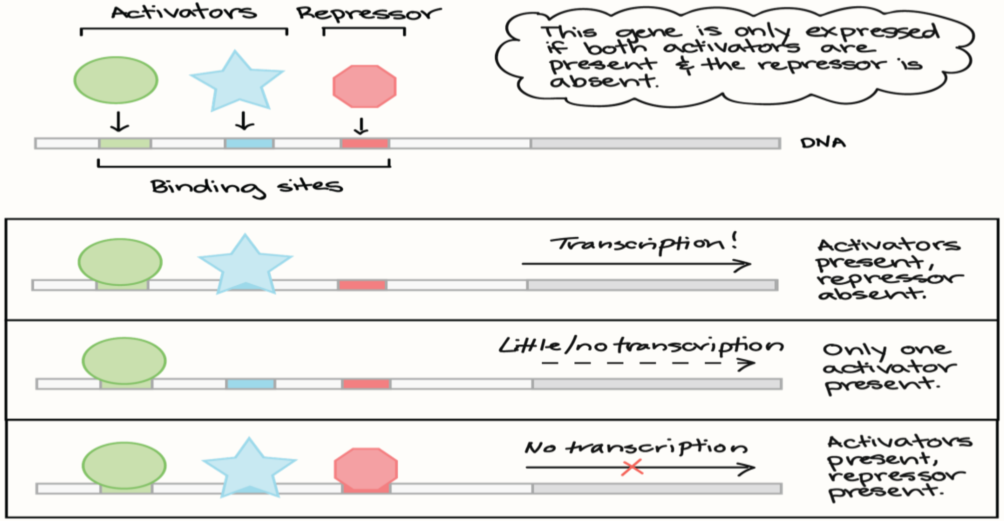
Back to Biology



Transcription factors binding recruited to initiate the process

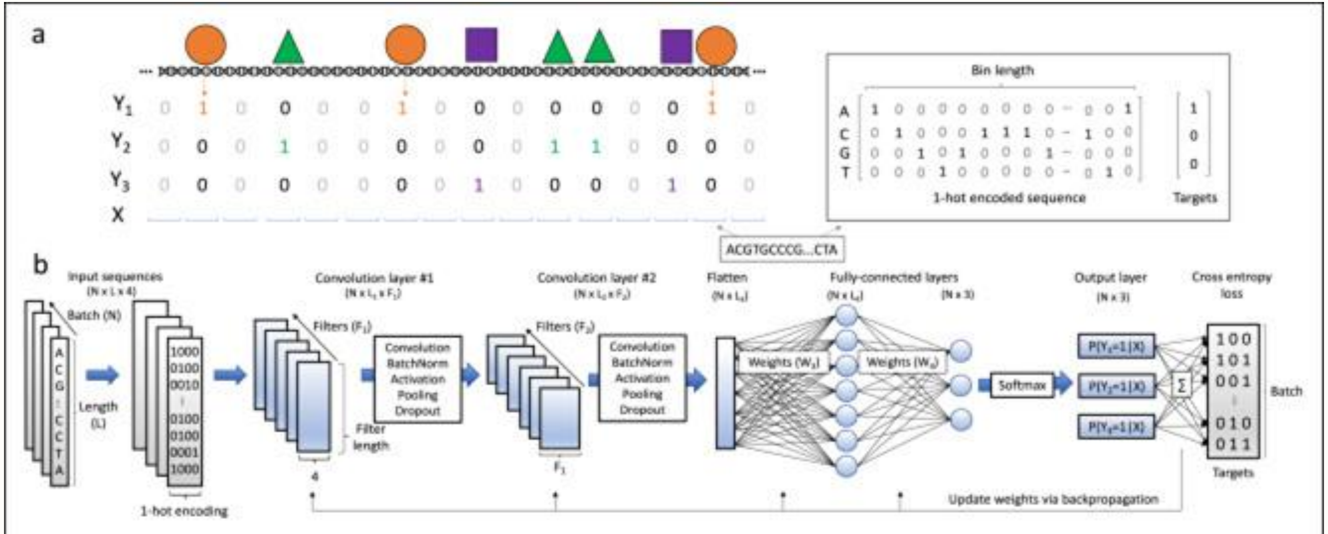


But.........What transcription factor should we consider


Activators
Repressor
This gene is only expressed if both activators are present & the repressor is absent.
DNA
Binding sites
Transcription!
Activators present, repressor absent.
Little/no transcription
Only one activator present.
No transcription
Activators present, repressor present.

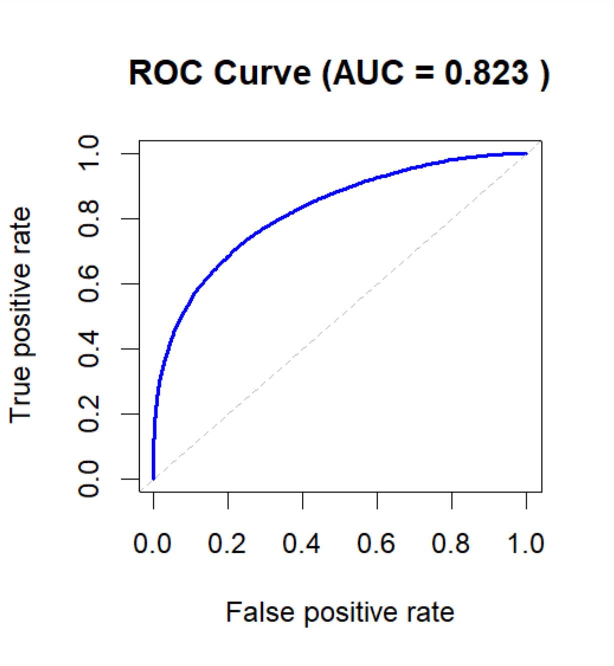
Why CNN?

- treating DNA sequence like 1D “images” where filters (kernels) slide across nucleotide sequences to learn motif-like patterns.
- Model architecture:
  - Convolutional layers: Apply filters to learn local motifs/patterns.
  - Pooling layers: Reduce dimensionality and keep the most important features.
  - Fully connected layers: Combine information for classification or regression.



Current Opinion in Systems Biology

Model evaluation



Parametric evaluation

Accuracy

% of correct predictions (TP + TN / Total) — can be misleading with class imbalance

ROC AUC

How well the model distinguishes between classes (good for balanced data)

PR AUC

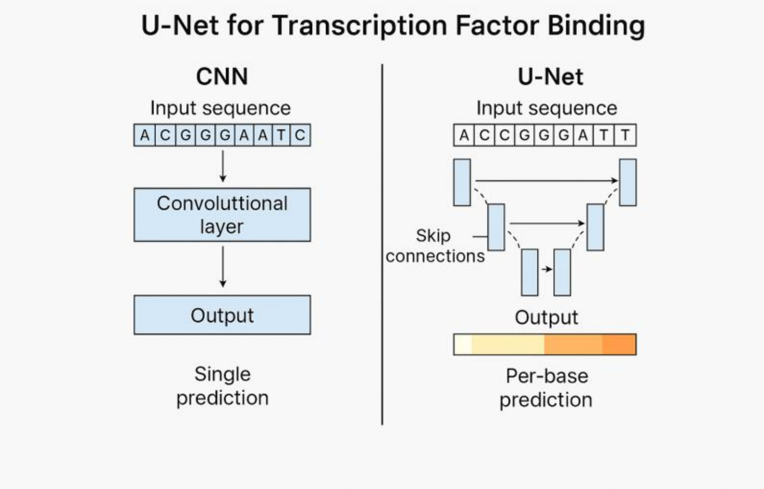
Focuses on **positive class performance** — better for **imbalanced datasets**

F1 Score

Balance between **precision** and **recall** — useful when false positives and negatives matter equally

›››Still CNN but more advance

CNN	***U-NET***
• Outputs 1 label per sequence (e.g., bind or not) • Learns local motifs • Ignores spatial resolution	• Outputs per-base predictions • Learns local + global context • Preserves nucleotide resolution via skip connections



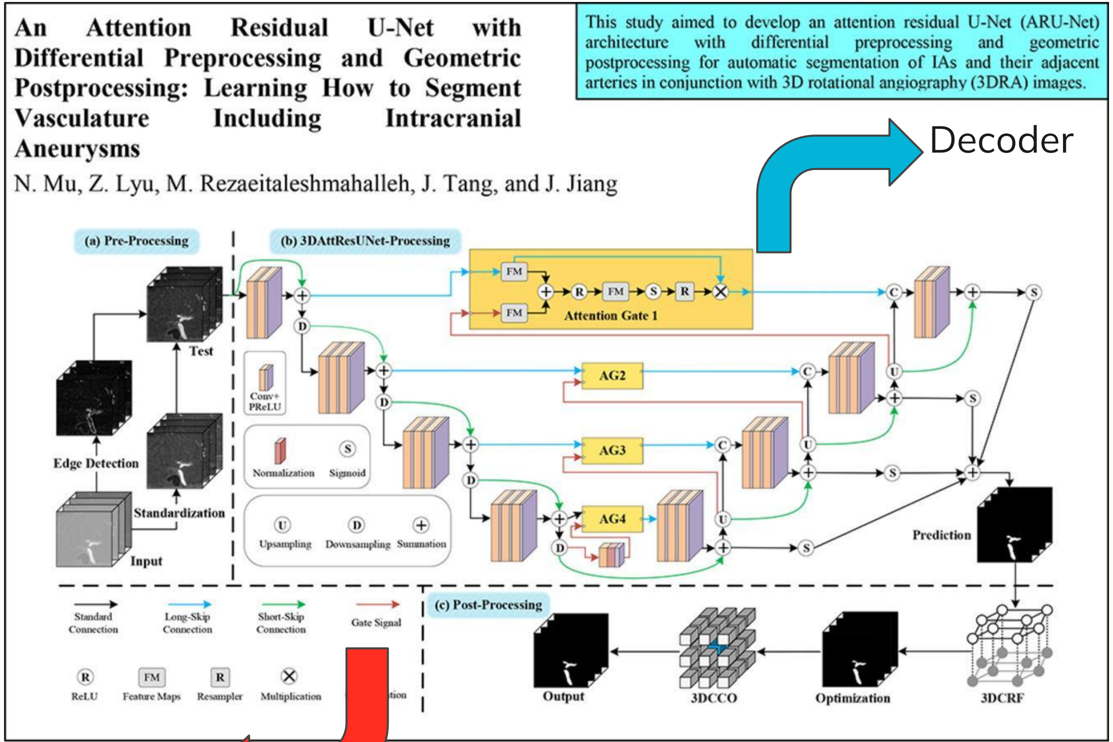
U-Net for Transcription Factor Binding

U-Net for Transcription factor prediction

1. standardize to normalize intensity distributions and edge detection to emphasize vascular boundaries

2. extract hierarchical features using convolutional layers with Parametric ReLU activations
   - Encoder
   - Decoder

3. optimize and 3D Connected Component Optimization (3DCCO) to eliminate small noisy predictions, retaining only biologically plausible connected vessel structures



An Attention Residual U-Net with Differential Preprocessing and Geometric Postprocessing: Learning How to Segment Vasculature Including Intracranial Aneurysms

N. Mu, Z. Lyu, M. Rezaeitaleshmahalleh, J. Tang, and J. Jiang

This study aimed to develop an attention residual U-Net (ARU-Net) architecture with differential preprocessing and geometric postprocessing for automatic segmentation of IAs and their adjacent arteries in conjunction with 3D rotational angiography (3DRA) images.

Encoder

Mu, N., Lyu, Z., Rezaeitaleshmahalleh, M., Tang, J., & Jiang, J. (2023). An attention residual U-net with differential preprocessing and geometric postprocessing: Learning how to segment vasculature including intracranial aneurysms. Current Opinion in Systems Biology. https://doi.org/10.1016/j.coisb.2023.100455

Attribution & Motifs discovery

Highlights important input bases & Finds recurring regulatory patterns
>>> All about the interpreation

1 Gradient-Based Methods

What influences the output?

2 Saliency Maps

Which base positions are biologically relevant

Topics

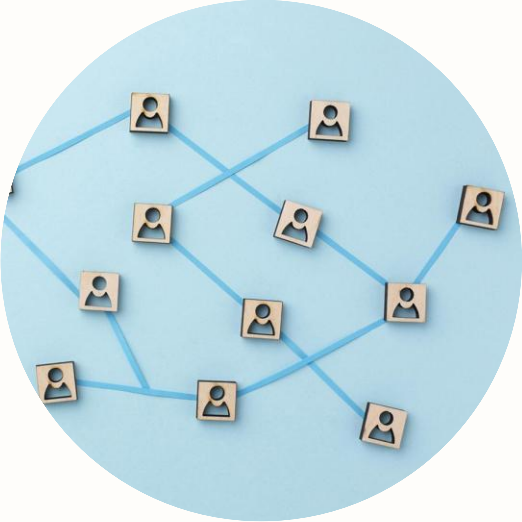
3 TF-MoDISco

- Takes gradient/saliency scores for many sequences.
- Identifies "seqlets"—small high-scoring regions.
- Clusters similar seqlets to form motif patterns.

Advantage & Disadvantage

1. Automatic Feature training
2. High predictive performance

1. Hard to train
2. Interpretability challenge



Thank You