Predicting transcription factor binding site using deep learning

Presenter

Mahboobeh (Mariya) Golchinpour leili

Supervisor

Dr. Fotuhi

Instructors

Dr. Gharaghani and Dr. Karimi-Jafari

Outline

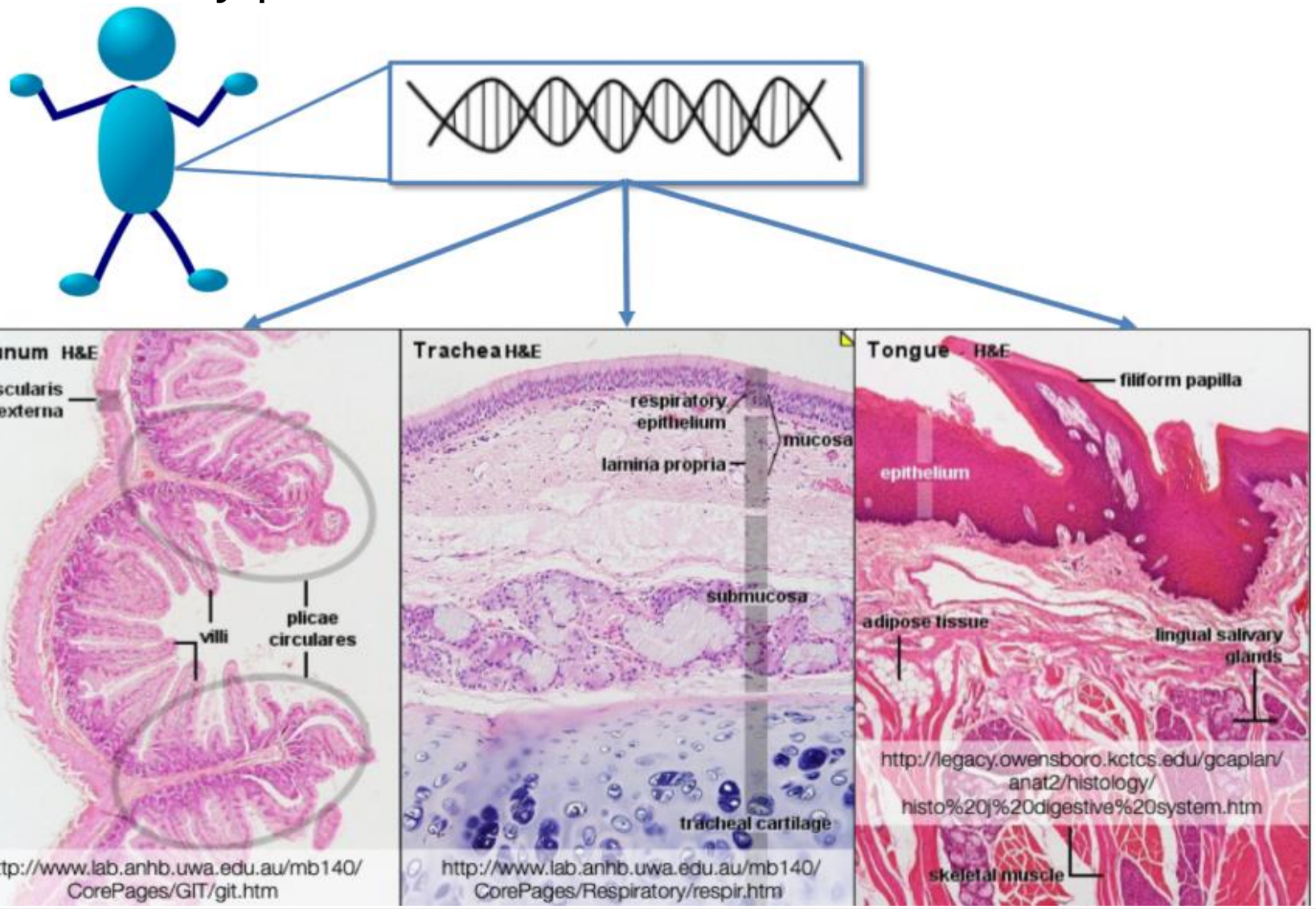
- From DNA to phenotype
- Transcriptional regulation
- Transcription Factors Binding site
- Assays to study Transcription factor binding sites (TFBSs)
- Databases
- Method for representing TFBS
- Evaluation
- Conclusion

Cells have the **same genome**

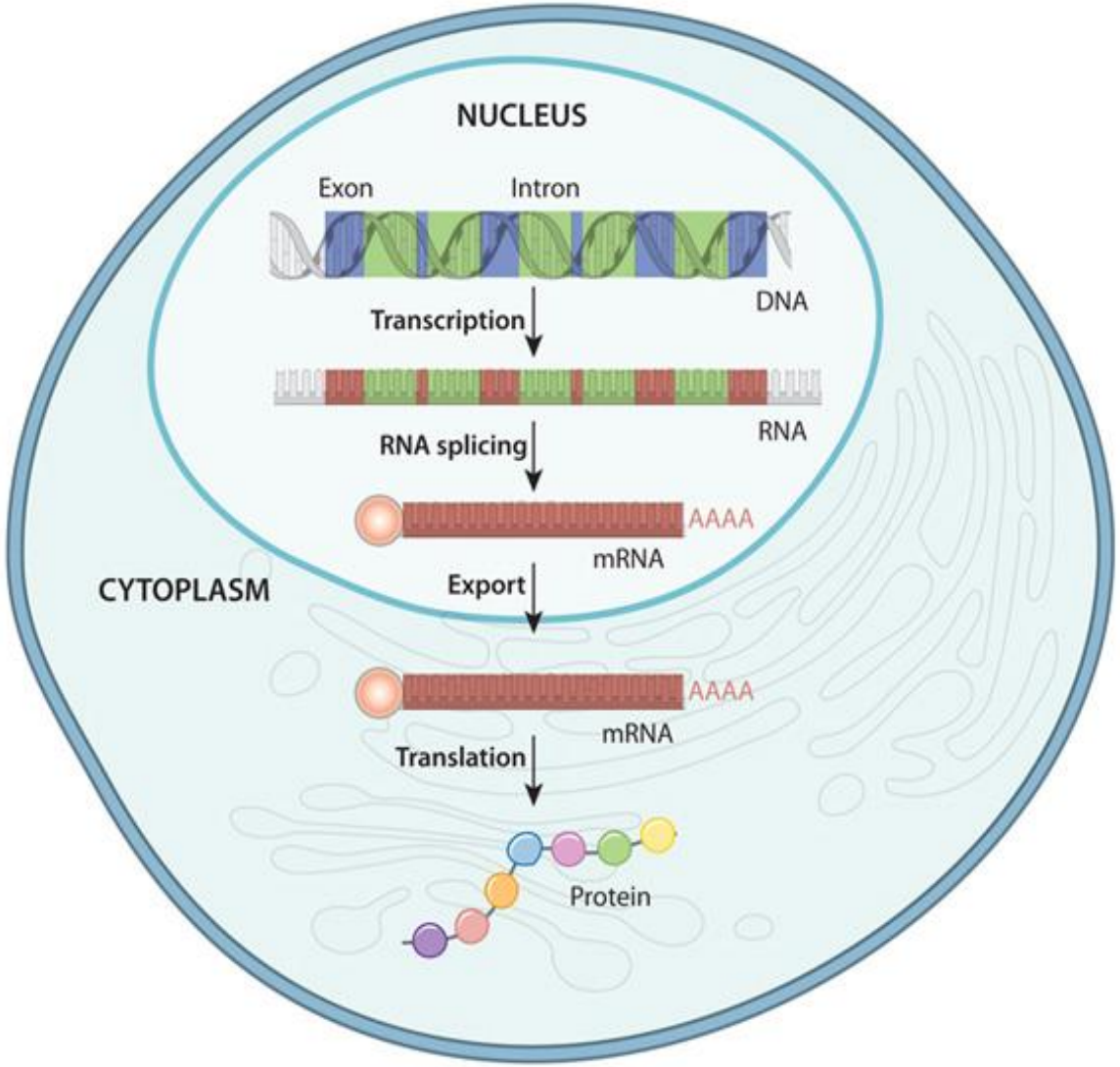
How can they perform **different functions?**

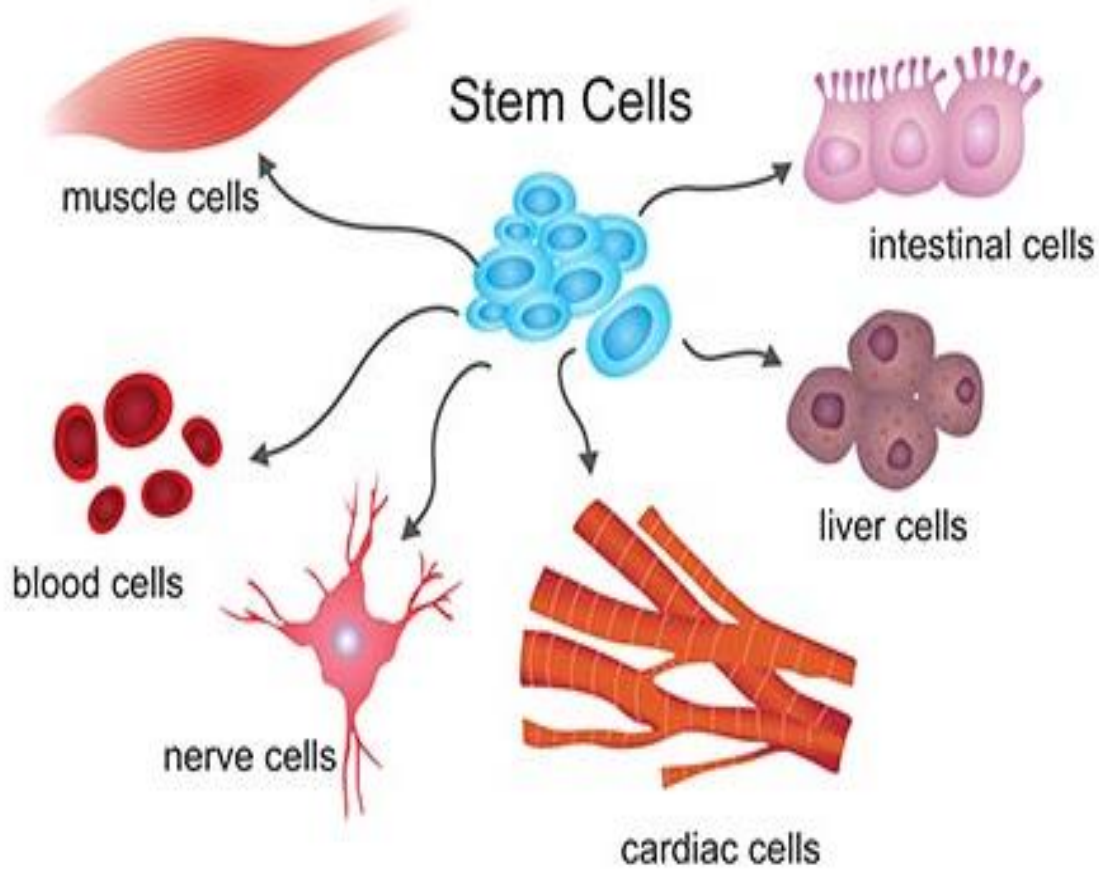
What are cell doing?

How are they doing it?

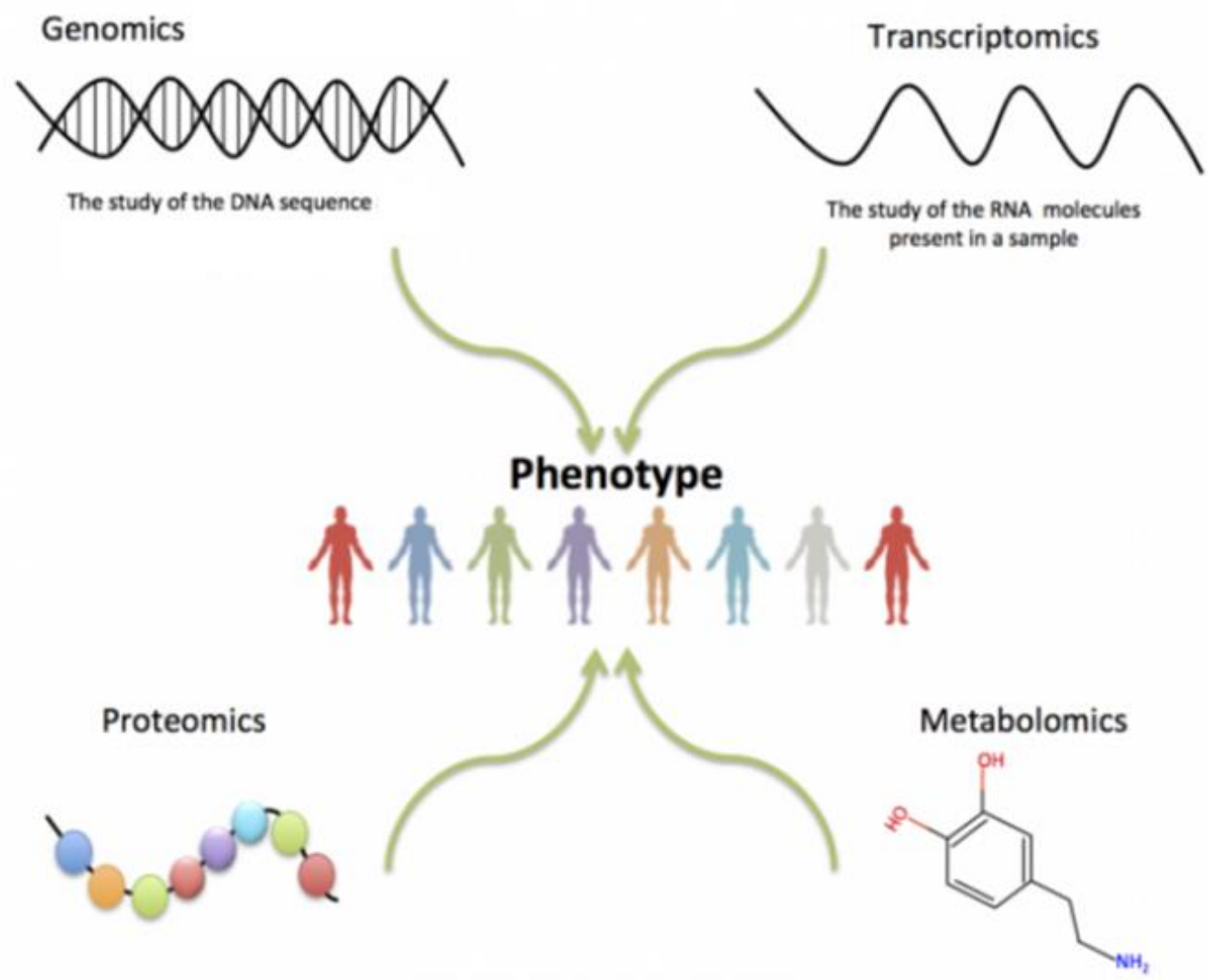


Central Dogma & Cell Differentiation





Genomic regulation contributes to different phenotypes



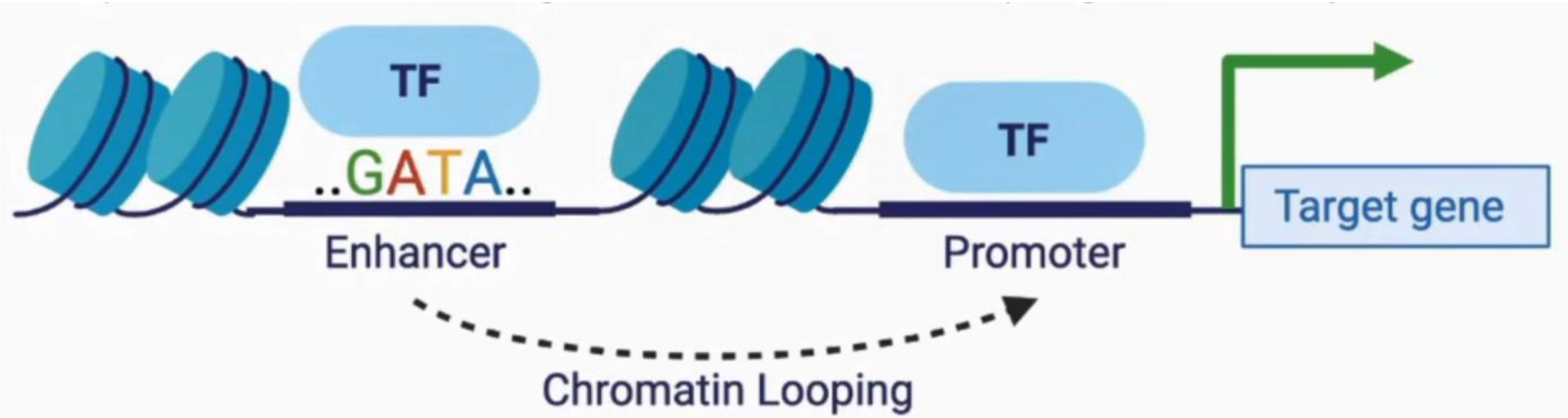
Transcriptional regulation

"Transcription factors (TFs) directly interpret the genome" Lambert, Samuel A., et al (2018)

~1,600 TFs in human
100's expressed at any given time

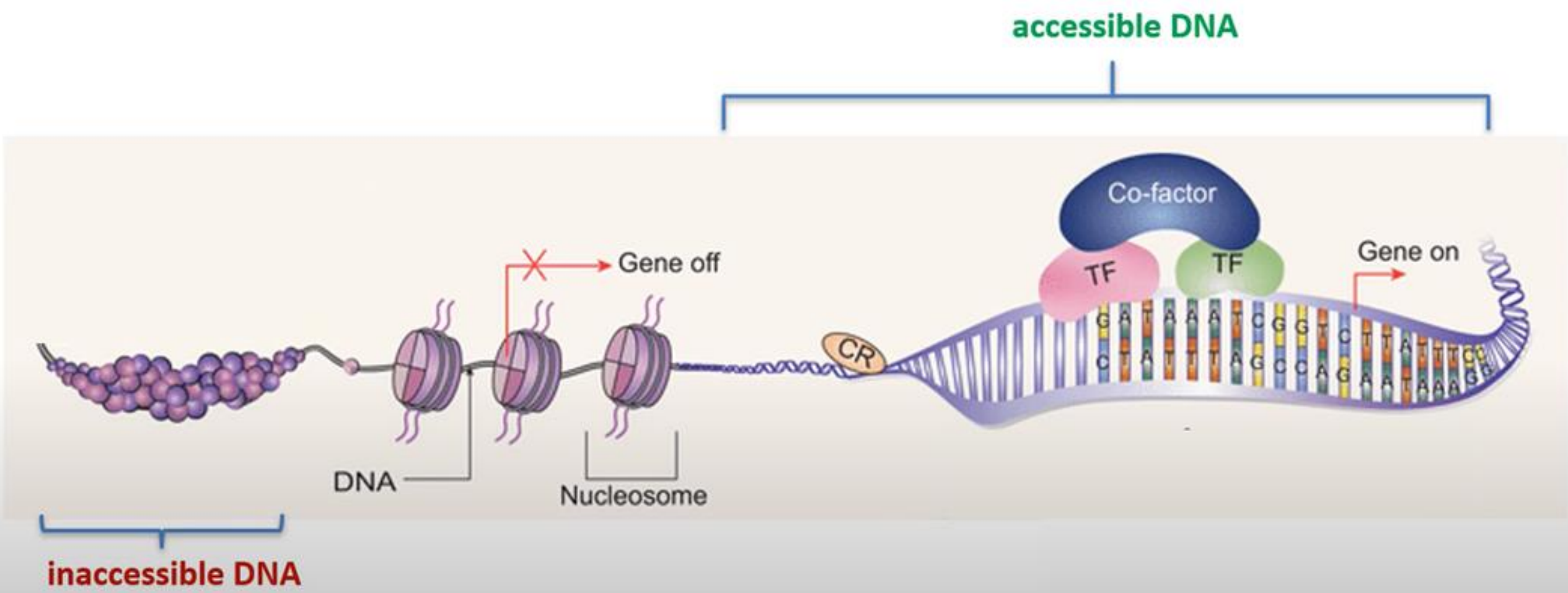
1) TFs have a DNA-binding domain

2) Regulate transcription



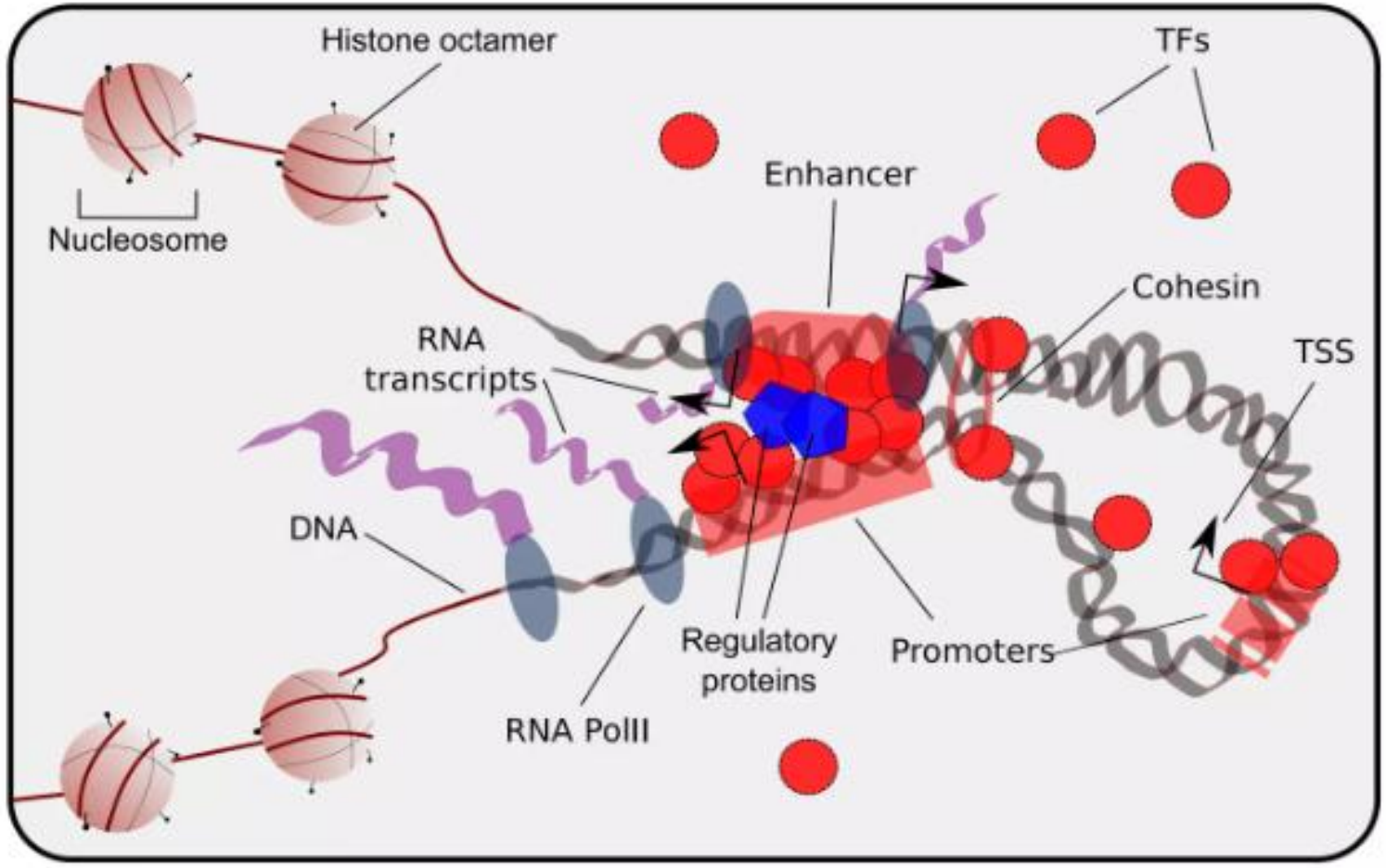
Accessible chromatin and Transcription factor (TF) binding

- TFS binds to DNA at transcription factors binding sites (TFBSs)



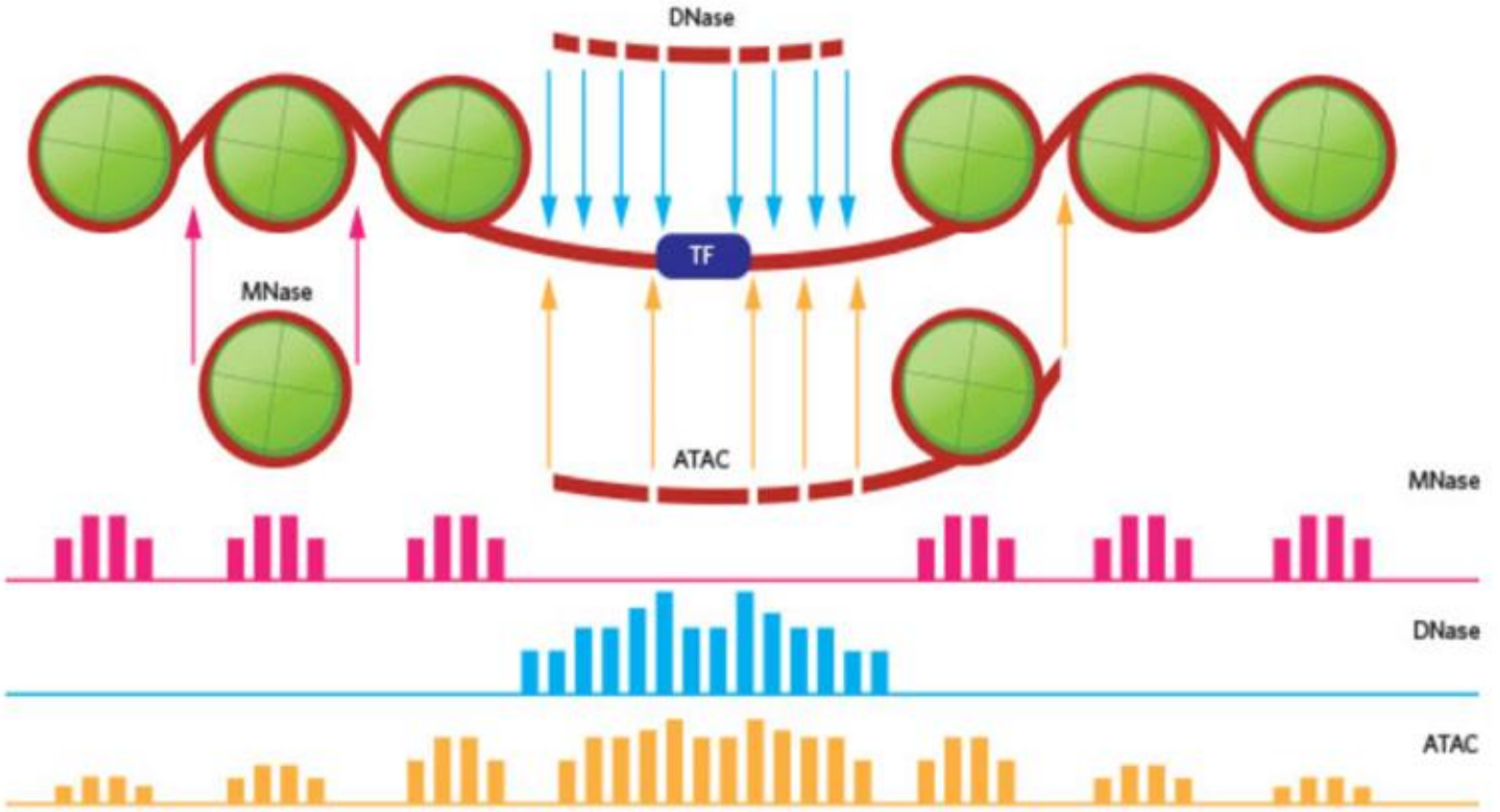
Transcription factor Binding Site

- TFBS are often located in: Gene promoters ,Distal regulatory elements, such as: enhancers, silencers, insulators.



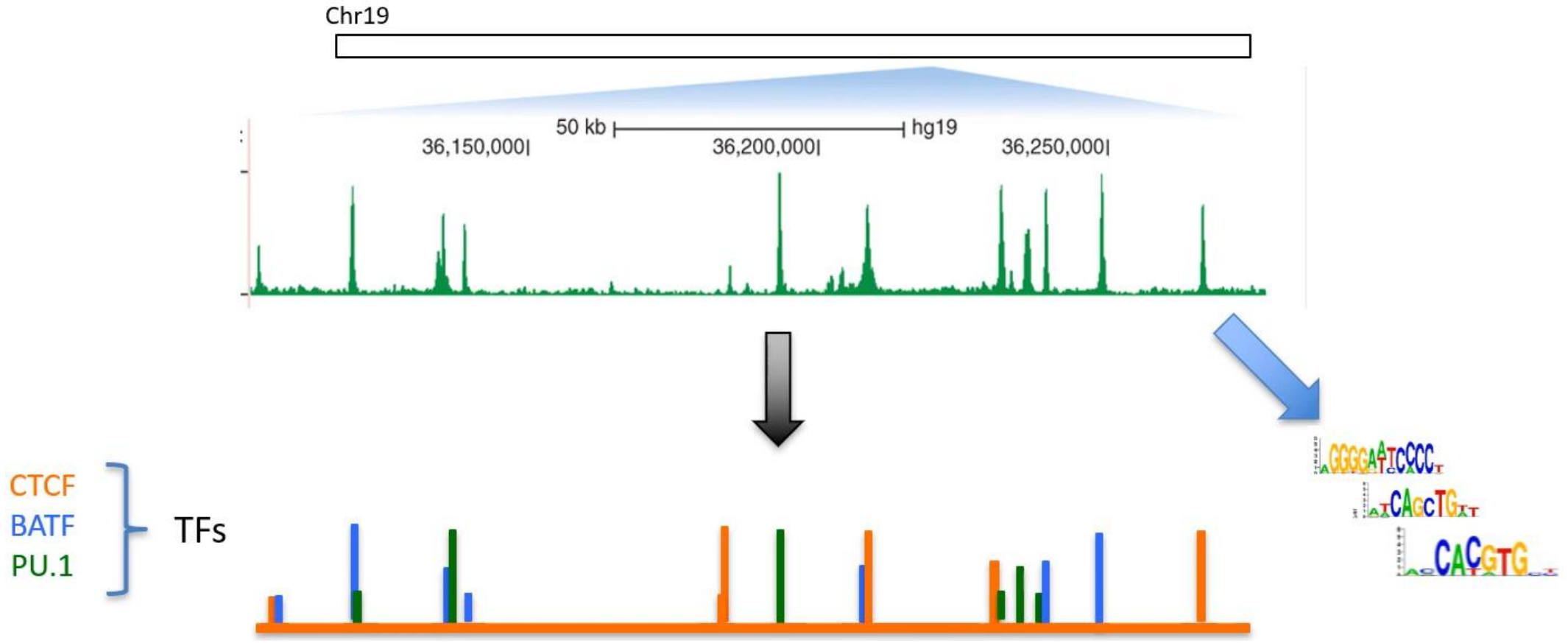
Assays to study Transcription factor binding sites (TFBSs)

- Protein binding microarray (PBM)
- Chip-seq
- ChIP-exo
- DNase-seq
- FAIRE-seq
- MNase-seq
- ATAC-seq



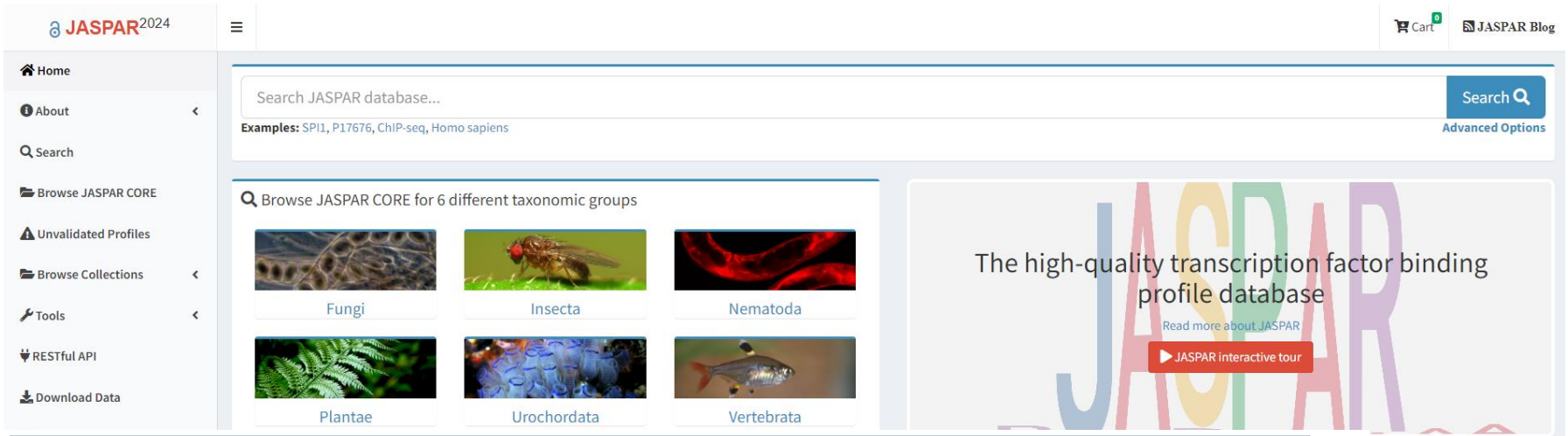
TF Regulatory Grammar

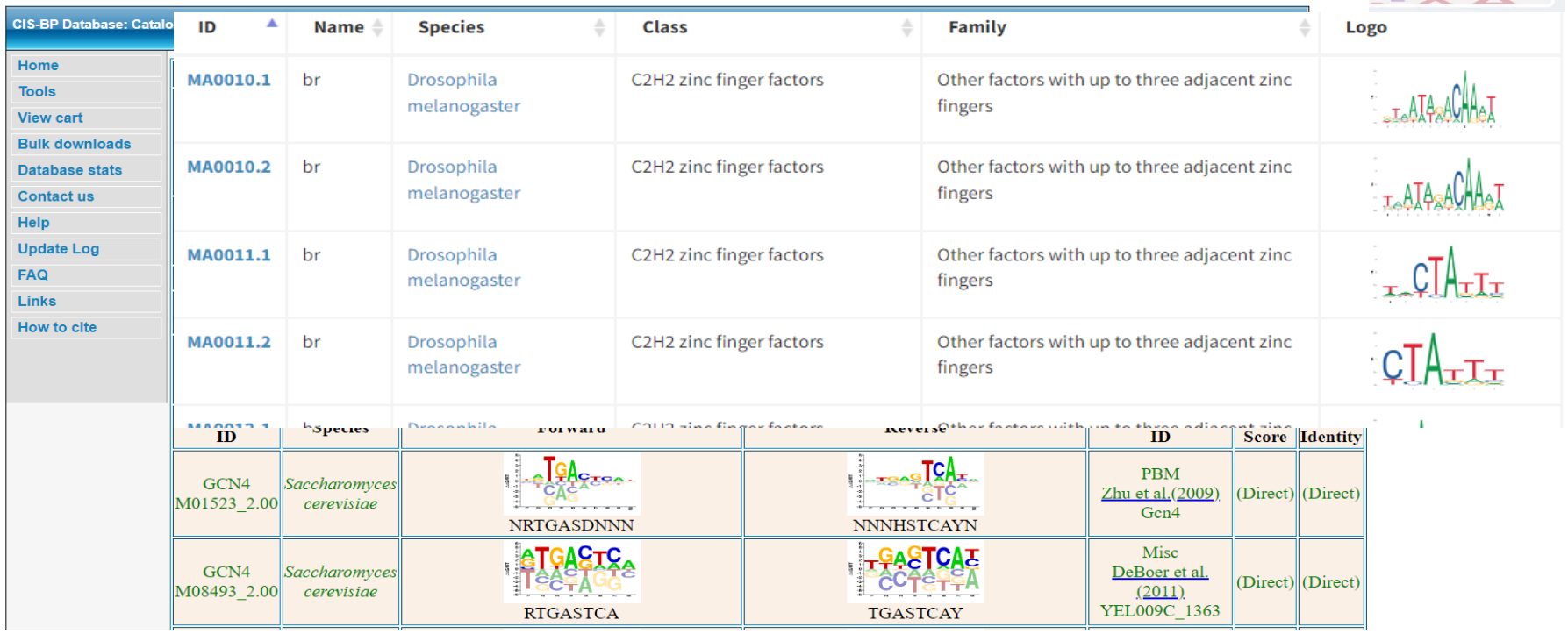
- Given accessible peaks, interpret the TF binding sites
- The TF-specific binding pattern for all TFs



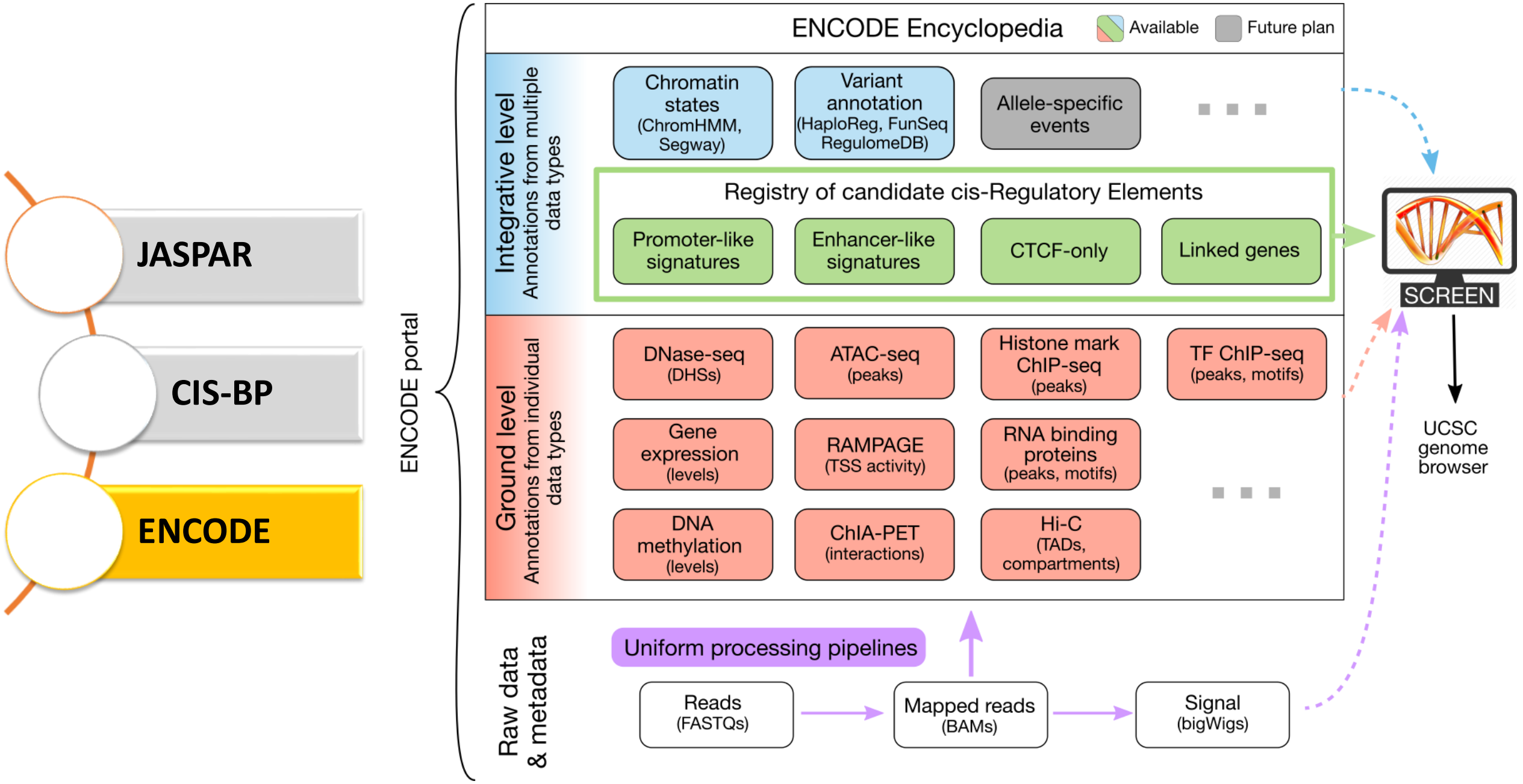
Motif Database

- JASPAR
- CIS-BP
- ENCODE



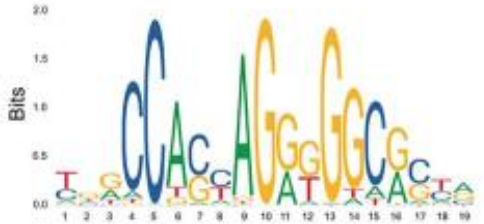


ID	Species	Forward	Reverse	ID	Score	Identity
GCN4 M01523_2.00	*Saccharomyces cerevisiae*	NRTGASDNNN	NNNHSTCAYN	PBM Zhu et al.(2009) Gcn4	(Direct)	(Direct)
GCN4 M08493_2.00	*Saccharomyces cerevisiae*	RTGASTCA	TGASTCAY	Misc DeBoer et al. (2011) YEL009C_1363	(Direct)	(Direct)


JASPAR
CIS-BP
ENCODE
ENCODE portal
ENCODE Encyclopedia
Available
Future plan
Integrative level
Annotations from multiple data types
Chromatin states (ChromHMM, Segway)
Variant annotation (HaploReg, FunSeq RegulomeDB)
Allele-specific events
Registry of candidate cis-Regulatory Elements
Promoter-like signatures
Enhancer-like signatures
CTCF-only
Linked genes
Ground level
Annotations from individual data types
DNase-seq (DHSs)
ATAC-seq (peaks)
Histone mark ChIP-seq (peaks)
TF ChIP-seq (peaks, motifs)
Gene expression (levels)
RAMPAGE (TSS activity)
RNA binding proteins (peaks, motifs)
DNA methylation (levels)
ChIA-PET (interactions)
Hi-C (TADs, compartments)
Raw data & metadata
Uniform processing pipelines
Reads (FASTQs)
Mapped reads (BAMs)
Signal (bigWigs)
SCREEN
UCSC genome browser

Representing TFBS

Position weight matrices (PWMs)



Dinucleotide Weight Matrices (DWM)

Position frequency matrix(PFM)

SVM (kmer-SVM)

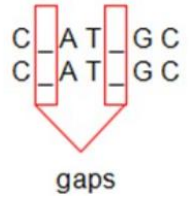
- Gapped K-mer

SVM models

K-mer	Score
CCACCAGGGGGCG	20.0
CCACCAGAGGGCG	19.8
CCACCAGATGGCG	18.0
CCTGCAGAGGGCG	11.5
...	
ACACTAGATGGTG	-9.0
ACTGTAGATGGTG	-15.2
ACTGTGGATGGTT	-16.0
AATCCCGGGGATT	-20.0

C A A T T G C

C T A T G G C



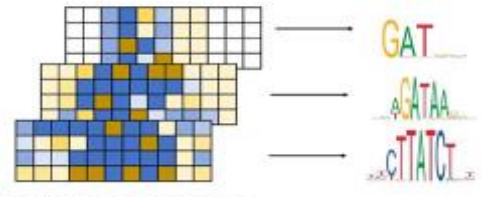
- HINT-ATAC
- TOBIAS
- TAMC
- TRACE

Footprinting

Deep-learning methods

- DeepBind
- DanQ
- BindSpace
- BindVAE

Deep neural networks models



Position Weight matrix(PWM)

- PFM is a 4×L matrix.
- Wαm is the probability of seeing nucleotide α at position m.
- PWMs assuming **nucleotide independence** within TFBSs.

Motifs	T	C	G	G	G	G	g	T	T	T	t	t
	c	C	G	G	t	G	A	c	T	T	a	C
	a	C	G	G	G	G	A	T	T	T	t	C
	T	t	G	G	G	G	A	c	T	T	t	t
	a	a	G	G	G	G	A	c	T	T	C	C
	T	t	G	G	G	G	A	c	T	T	C	C
	T	C	G	G	G	G	A	T	T	c	a	t
	T	C	G	G	G	G	A	T	T	c	C	t
	T	a	G	G	G	G	A	a	c	T	a	C
	T	C	G	G	G	t	A	T	a	a	C	C
SCORE(*Motifs*)	3	+ 4	+ 0	+ 0	+ 1	+ 1	+ 1	+ 5	+ 2	+ 3	+ 6	+ 4 = 30
COUNT(*Motifs*)												
A:	2	2	0	0	0	0	9	1	1	1	3	0
C:	1	6	0	0	0	0	0	4	1	2	4	6
G:	0	0	10	10	9	9	1	0	0	0	0	0
T:	7	2	0	0	1	1	0	5	8	7	3	4
PROFILE(*Motifs*)												
A:	.2	.2	0	0	0	0	.9	.1	.1	.1	.3	0
C:	.1	.6	0	0	0	0	0	.4	.1	.2	.4	.6
G:	0	0	1	1	.9	.9	.1	0	0	0	0	0
T:	.7	.2	0	0	.1	.1	0	.5	.8	.7	.3	.4
CONSENSUS(*Motifs*)	T	C	G	G	G	G	A	T	T	T	C	C

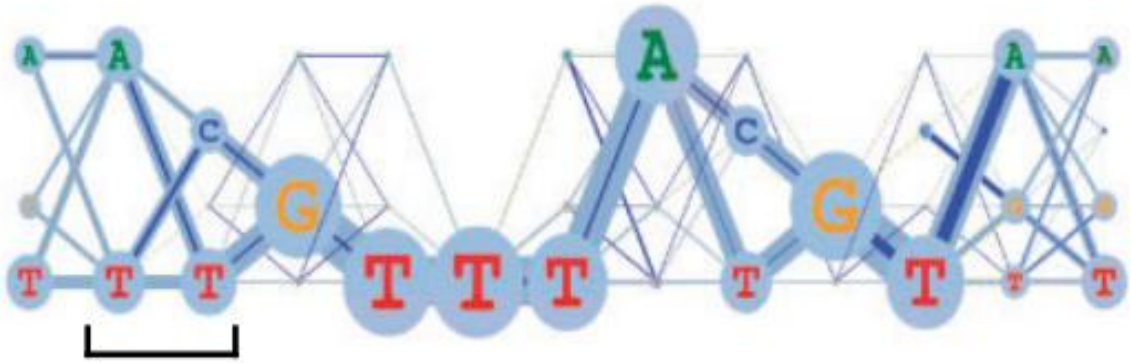
From motif matrix to count matrix to profile matrix to consensus string to motif logo

Dinucleotide Weight Matrices (DWM)

- Dinucleotide interactions between all pairs of positions within the TFBSs.

- $D_{\alpha 1 \alpha 2;m1m2}$: gives the probability of observing each pair of nucleotides $\alpha 1$ and $\alpha 2$ at each pair of positions m1 and m2 in a binding site.

0-order PWMs



1-order PWMs

K-mer

- ***K*-mer** to refer to a substring of length ***k*** in a string
- Define **COUNT(*Text*, *Pattern*)** as the number of times that a ***k-mer*** *Pattern* appears as a substring of ***Text***

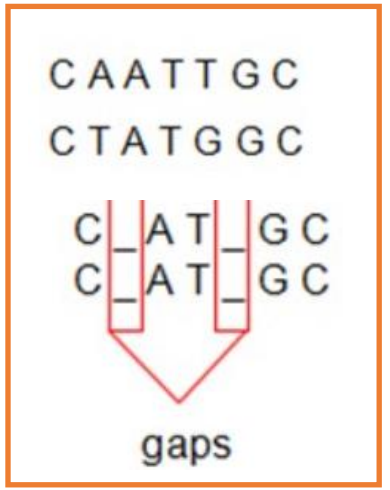
$$\text{COUNT}(\text{ACA}\mathbf{ACTAT}\text{GCAT}\mathbf{ACTAT}\text{CGGGA}\mathbf{ACTAT}\text{CCT}, \mathbf{ACTAT}) = 3$$

Different parameters for k-mers

Length	Window	Tokenized
3	3	ATC GCG TAC GAT CCG
4	4	ATCG CGTA CGAT
5	5	ATCGC GTACG ATCCG
4	2	ATCG CGCG CGTA TACG CGAT ATCC
4	3	ATCG GCGT TACG GATC

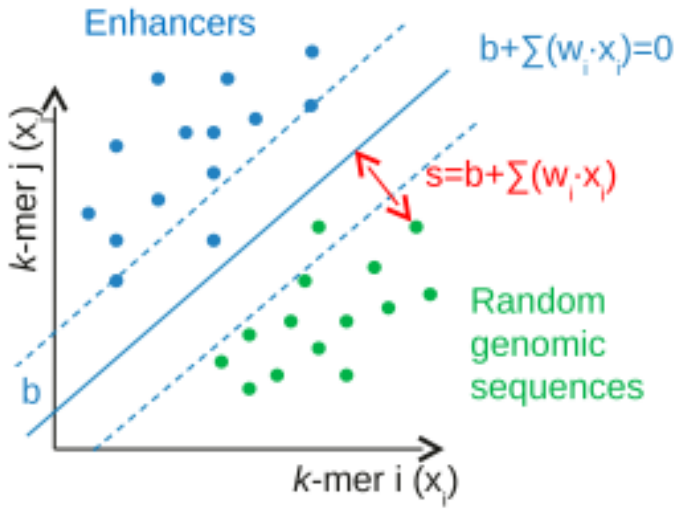
SVM based (kmer-SVM) framework for Enhancer prediction

- The EP300 gene provides instructions for making a protein called p300 (turning on transcription).
- Using the SVM to distinguishes (enhancer) and negative (random genomic) sequence sets.
- Gapped k-mers allow for gaps, providing a more flexible representation of sequence motifs.



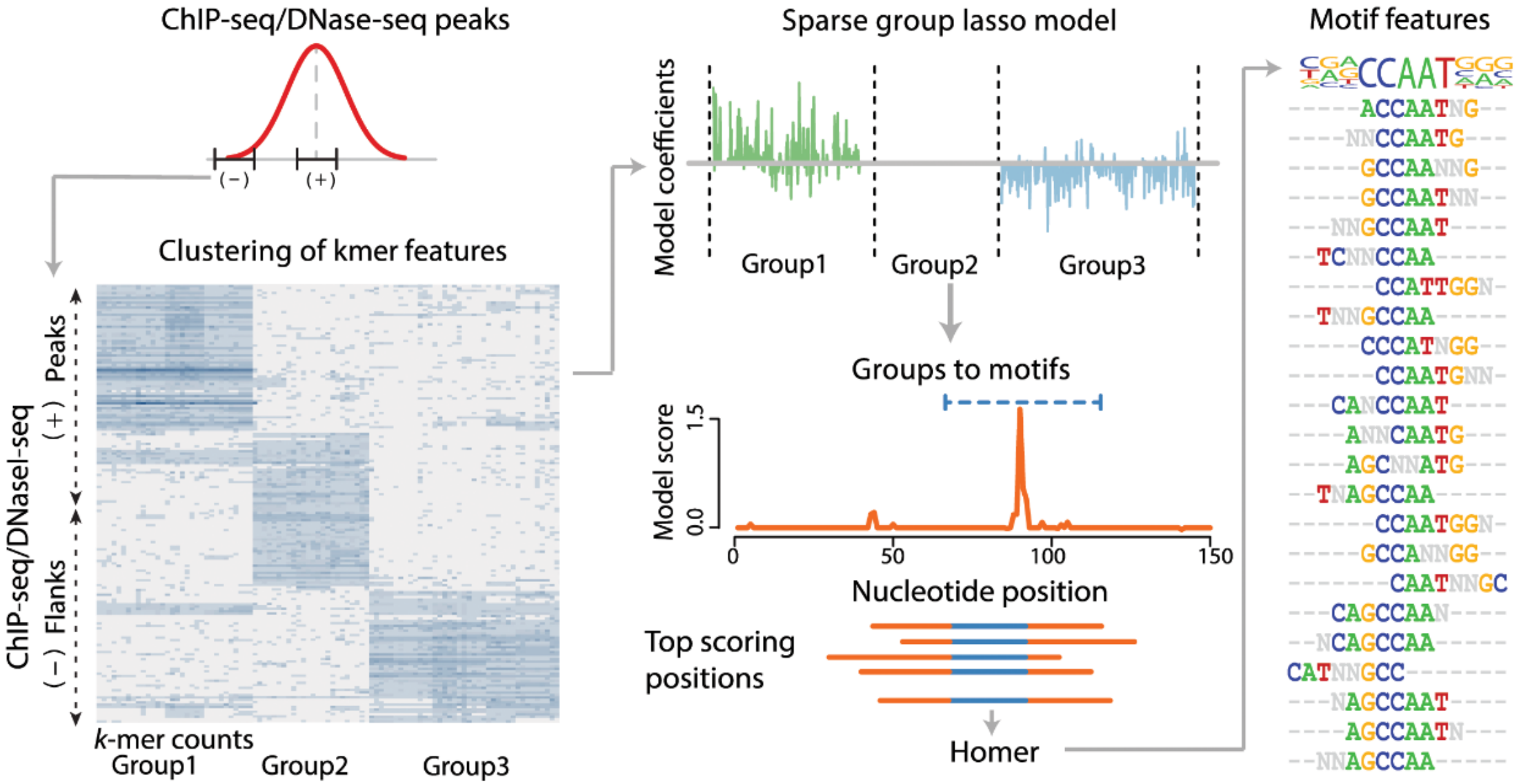
Sequences → *k*-mer frequencies → SVM training → Predictive feature analysis

k-mer	counts
5′-AAAAAA-3′ 3′-TTTTTT-5′	x_1
5′-AAAAAC-3′ 3′-TTTTTG-5′	x_2
5′-AAAAAG-3′ 3′-TTTTTC-5′	x_3
...	...
5′-TTTAAA-3′ 3′-AAATTT-5′	x_n



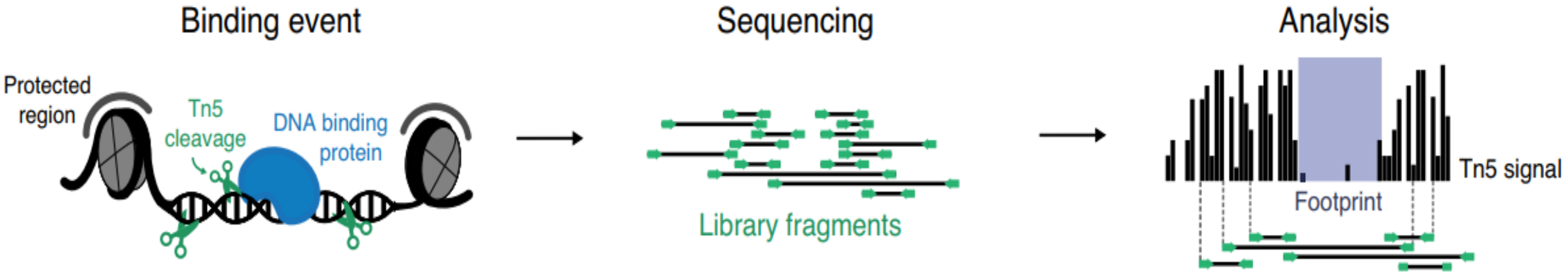
k-mer	weights (w_i)
5′-AATGAG-3′ 3′-TTACTC-5′	+3.94
5′-AATTAG-3′ 3′-TTAATC-5′	+3.84
5′-AGCTGC-3′ 3′-TCGACG-5′	+3.65
...	...
5′-CAGGTA-3′ 3′-GTCCAT-5′	-2.06

Classification task between peaks and flanks



Footprinting

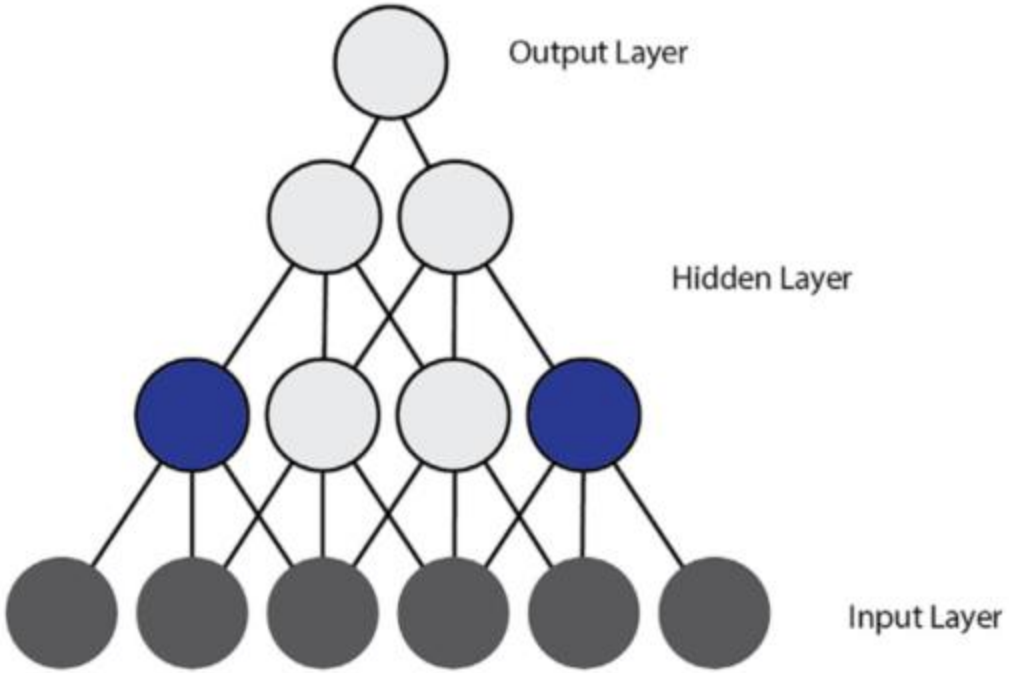
- **DGF** is a computational analysis of chromatin accessibility assays such as **ATAC-seq**
- Employs DNA effector enzymes that only **cut accessible DNA regions**



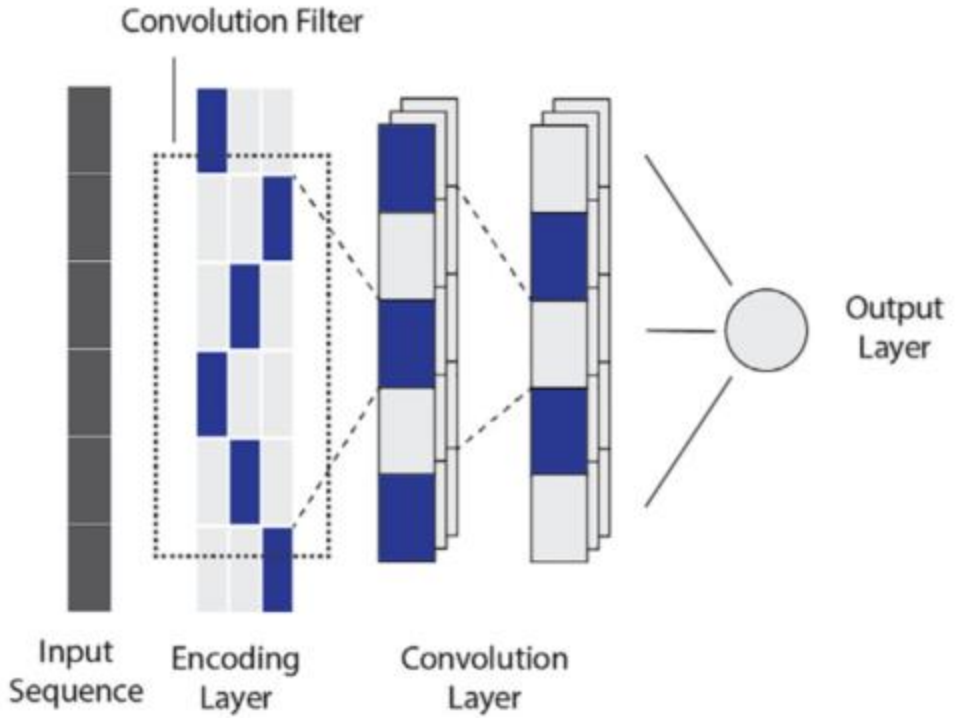
Deep-learning architectures used in TFBS prediction

- Multi-layer perceptron
- Convolutional neural network
- Recurrent neural network
- Bi-LSTM
- Transformer
- Attention
- Autoencoder

MLP & CNN



Multi-layer Perceptron

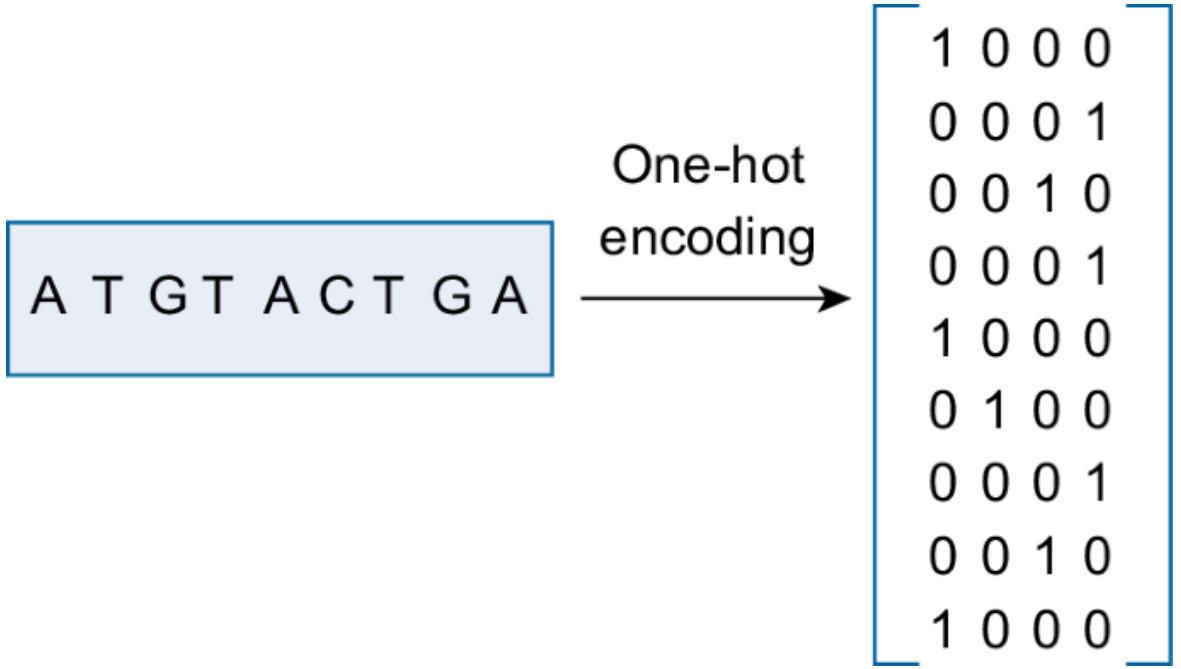


Convolutional Neural Network

One hot-encoding for DNA sequence

- Each nucleotide is represented as a one-hot vector
- RNA sequences can also be encoded similarly by simply changing T to U

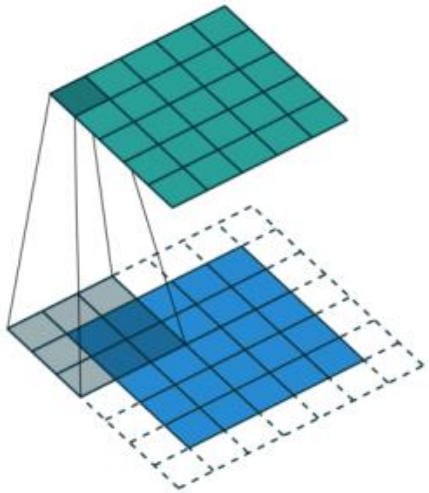
A = (1,0,0,0)
G = (0,1,0,0)
C = (0,0,1,0)
T = (0,0,0,1)



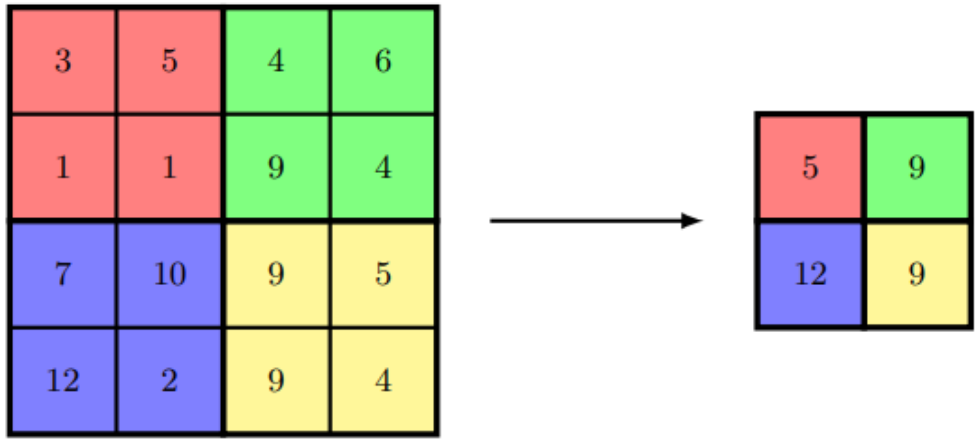
Convolution over one hot-encoding matrix

- CNNs represent genomic sequences as 1D or 2D images with four associated channels (A, C, G, T)

	C	G	A	T	A	A	C	C	G	A	T	A	T
A	0	0	1	0	1	1	0	0	0	1	0	1	0
C	1	0	0	0	0	0	1	1	0	0	0	0	0
G	0	1	0	0	0	0	0	0	1	0	0	0	0
T	0	0	0	1	0	0	0	0	0	0	1	0	1

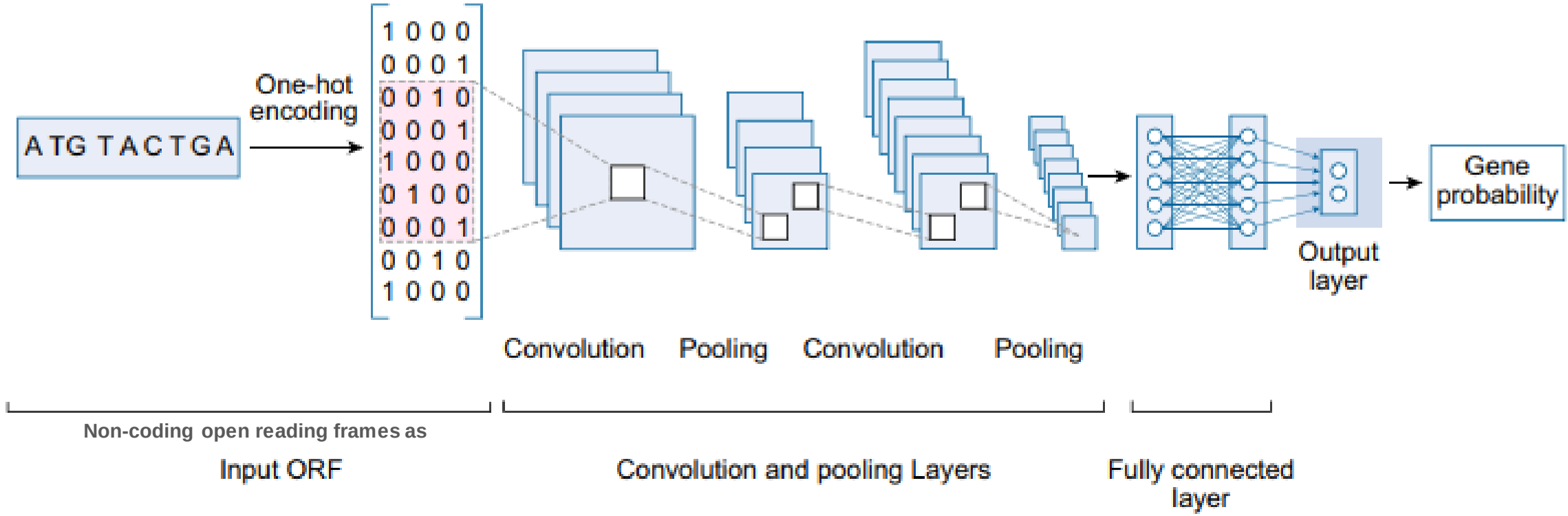


- Pooling layer(Max/average)

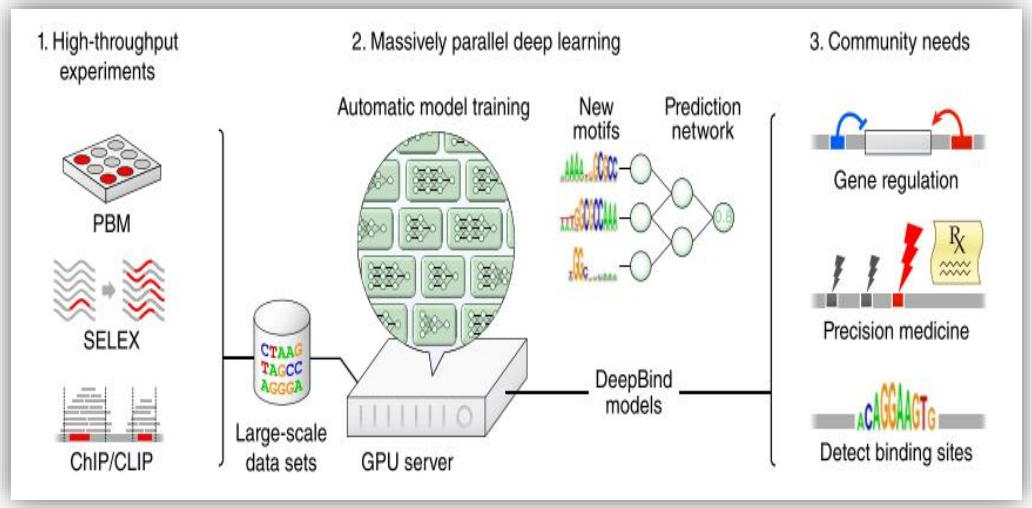


Convolution Neural Network for DNA sequence

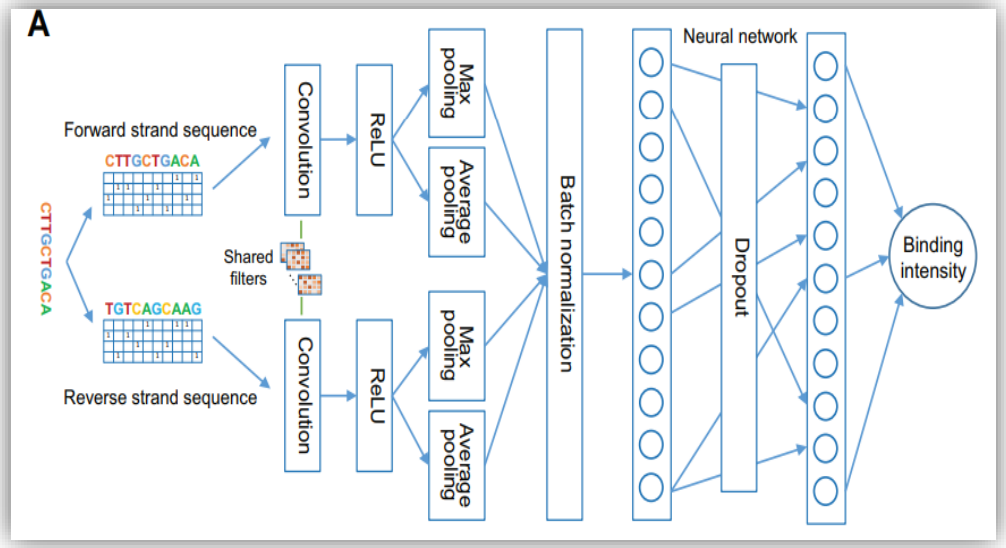
- CNN architectures designed for motif discovery and classification consist of one or more sets of four layers.
  - Convolutional layer
  - Pooling layer(Max/average)
  - Fully connected NN layer
  - Output layer

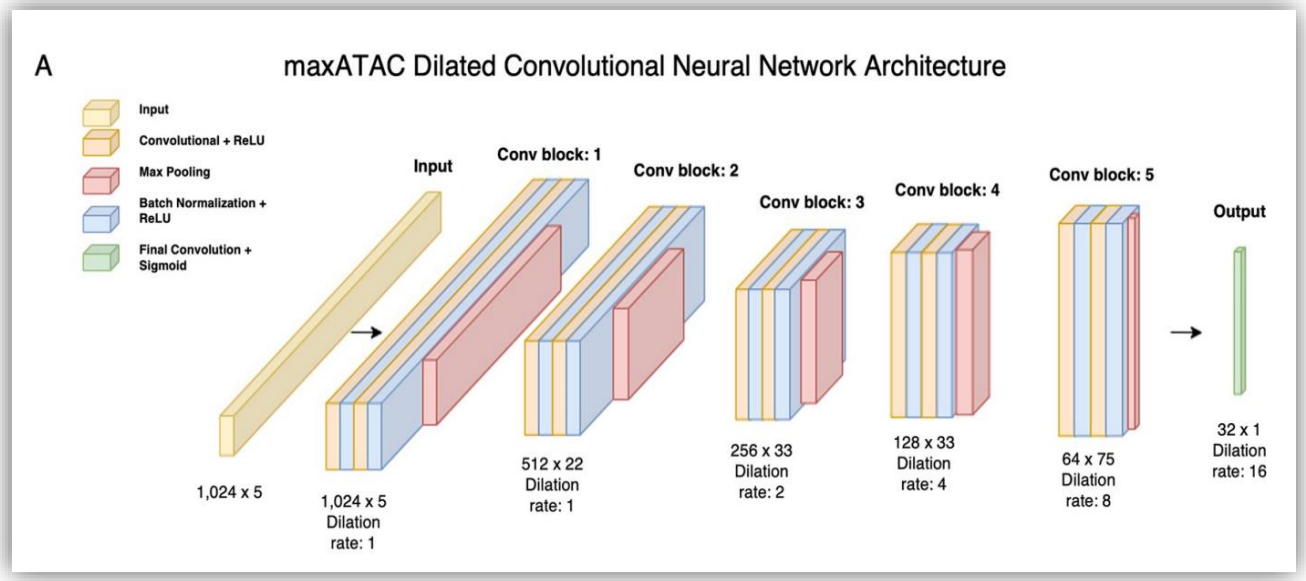


DeepBind: Predicting TF binding sites using CNN

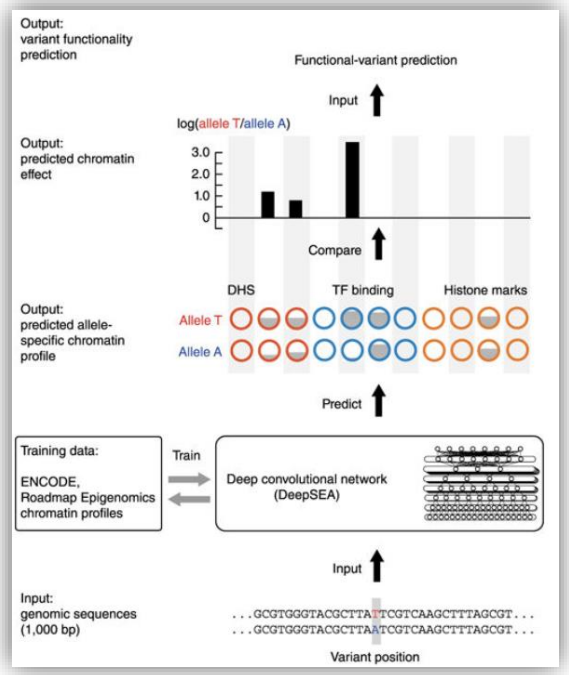


DeFine: predicting TF-DNA binding intensities from DNA sequences





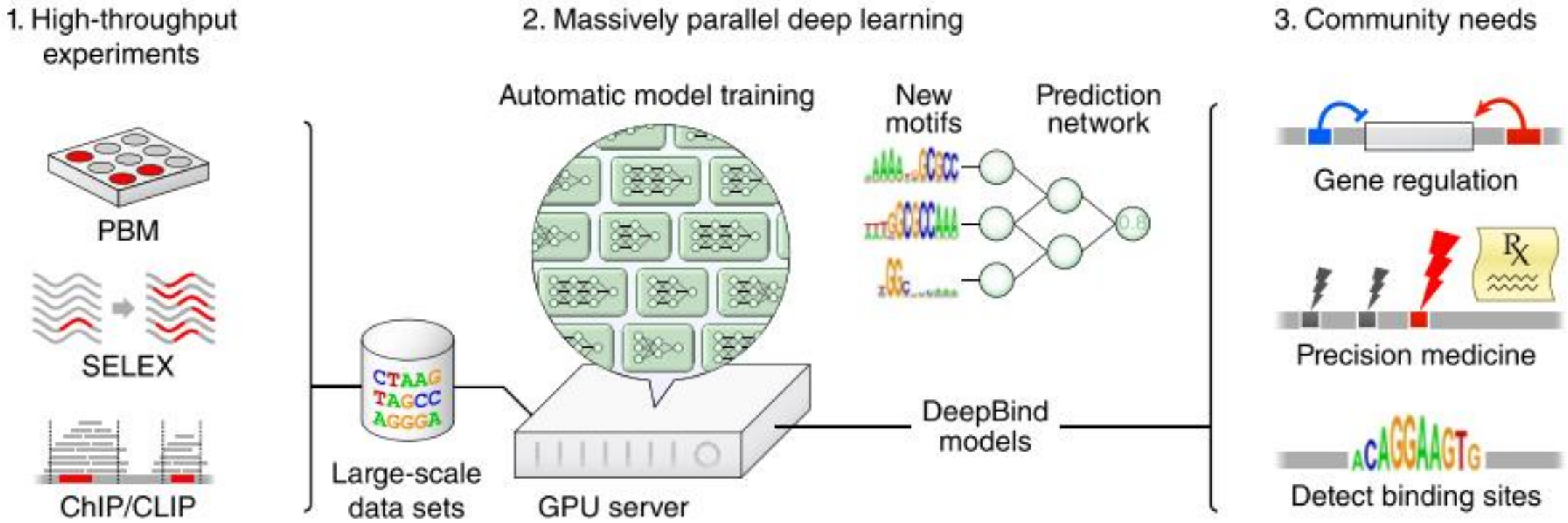
MaxATAC uses CNN to predict TFBS



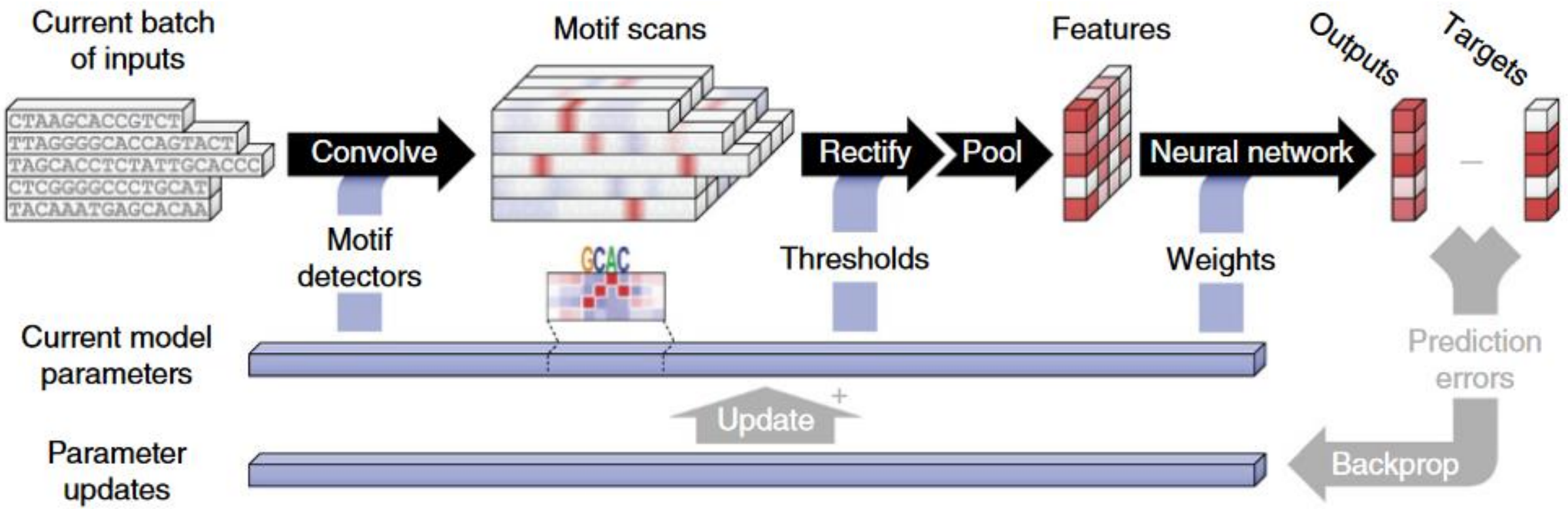
DeepSEA

DeepBind: Predicting TF binding sites using CNN

- Predicting the sequence specificities of DNA- and RNA-binding proteins
- Multiple experimental profiling technologies
  - protein binding microarrays (PBM), ChIP-seq, and HT-SELEX

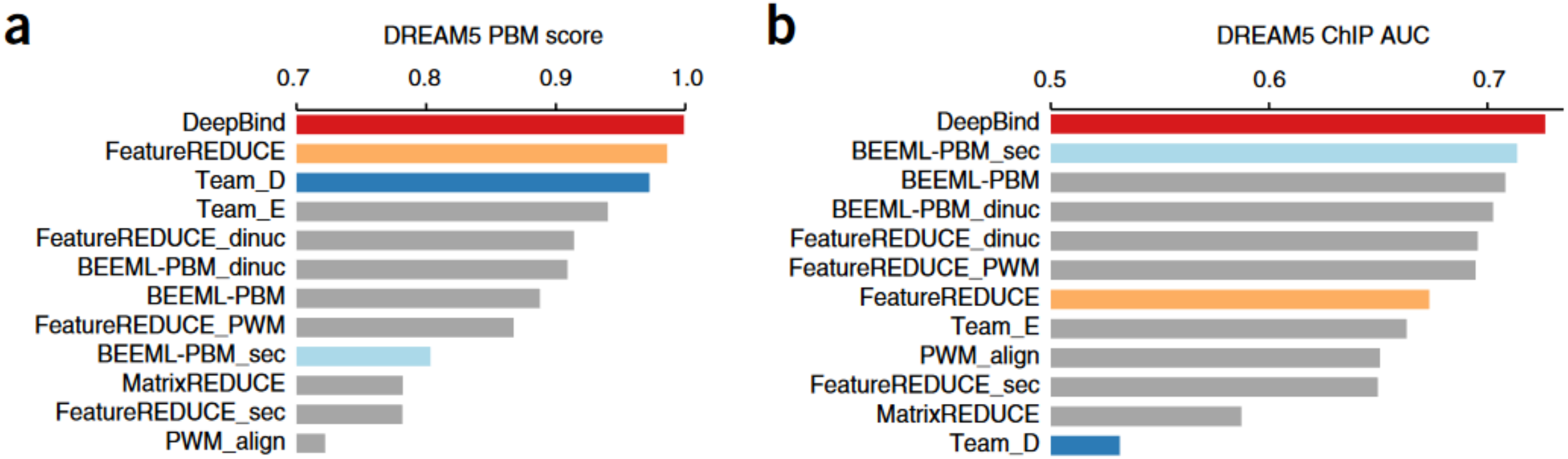


ConvNet for protein-sequence binding



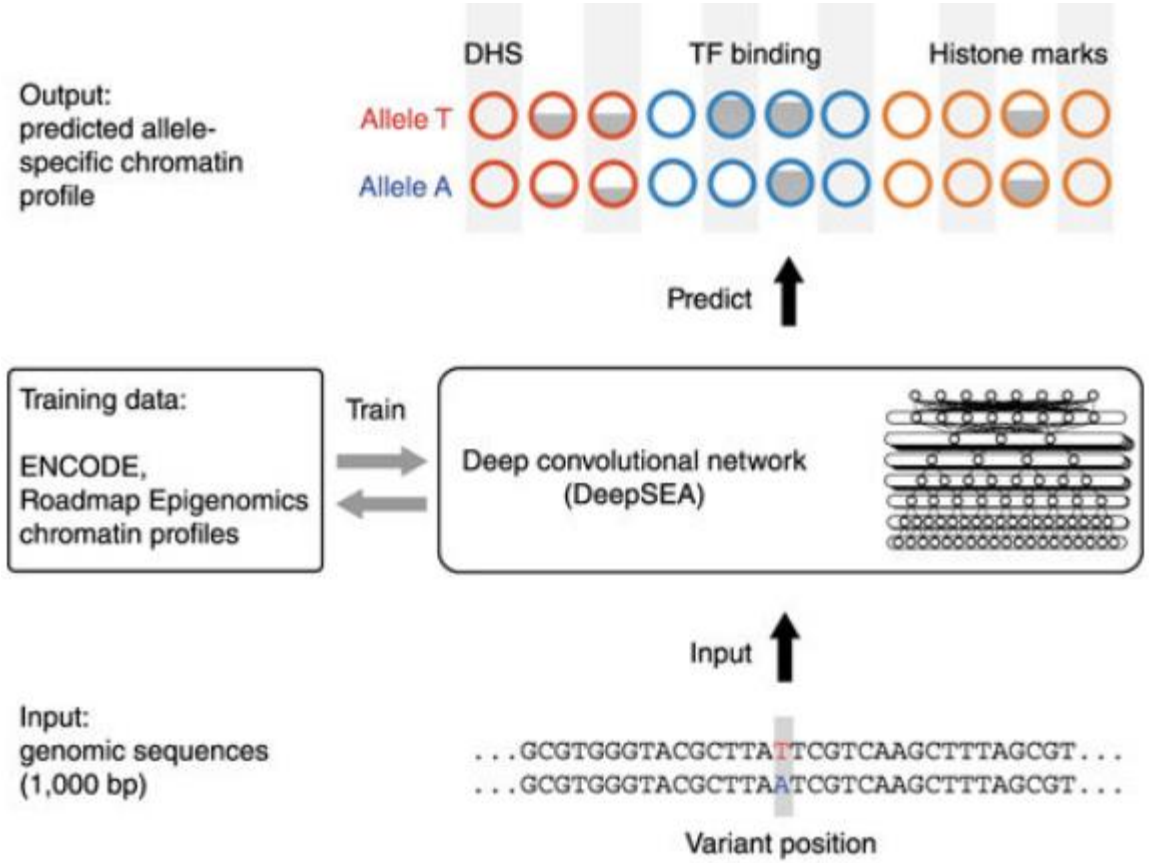
Quantitative comparisons

DNA transcription factors/DREAM5 TF-DNA Motif Recognition Challenge



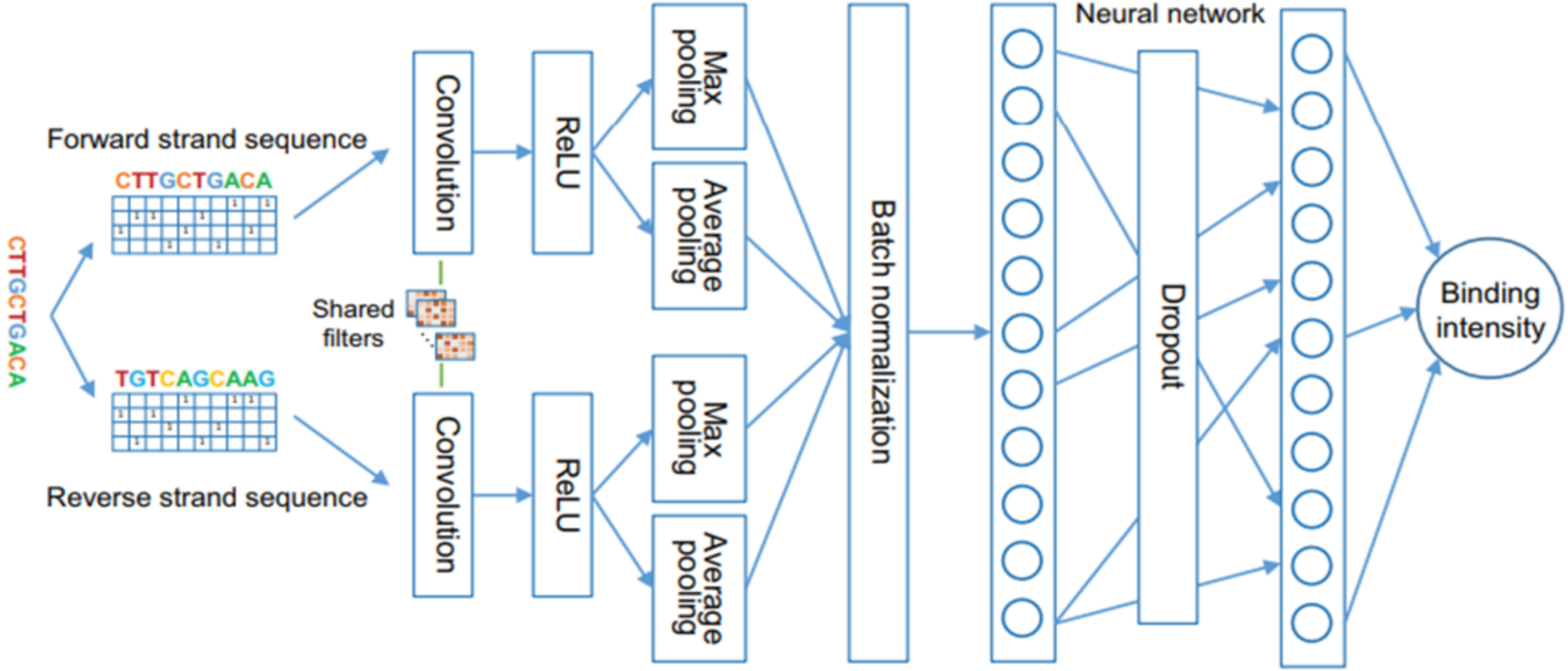
DeepSEA: Predicting effects of noncoding variants with deep learning based sequence model

- A variant in noncoding DNA can turn on a gene and cause a protein to be produced in the wrong place or at the wrong time.
- Predicting chromatin effects of noncoding variants
- Improvement on DeepBind: Add chromatin features and alternate background models
- 3 Convolution layer, 2 Maxpooling layer



DeFine: predicting TF-DNA binding intensities from DNA sequences

- Accurately predict the TF binding intensities to given DNA sequences by leveraging large-scale TF ChIP-seq data.
- Accurately classifying TF-DNA binding or unbinding



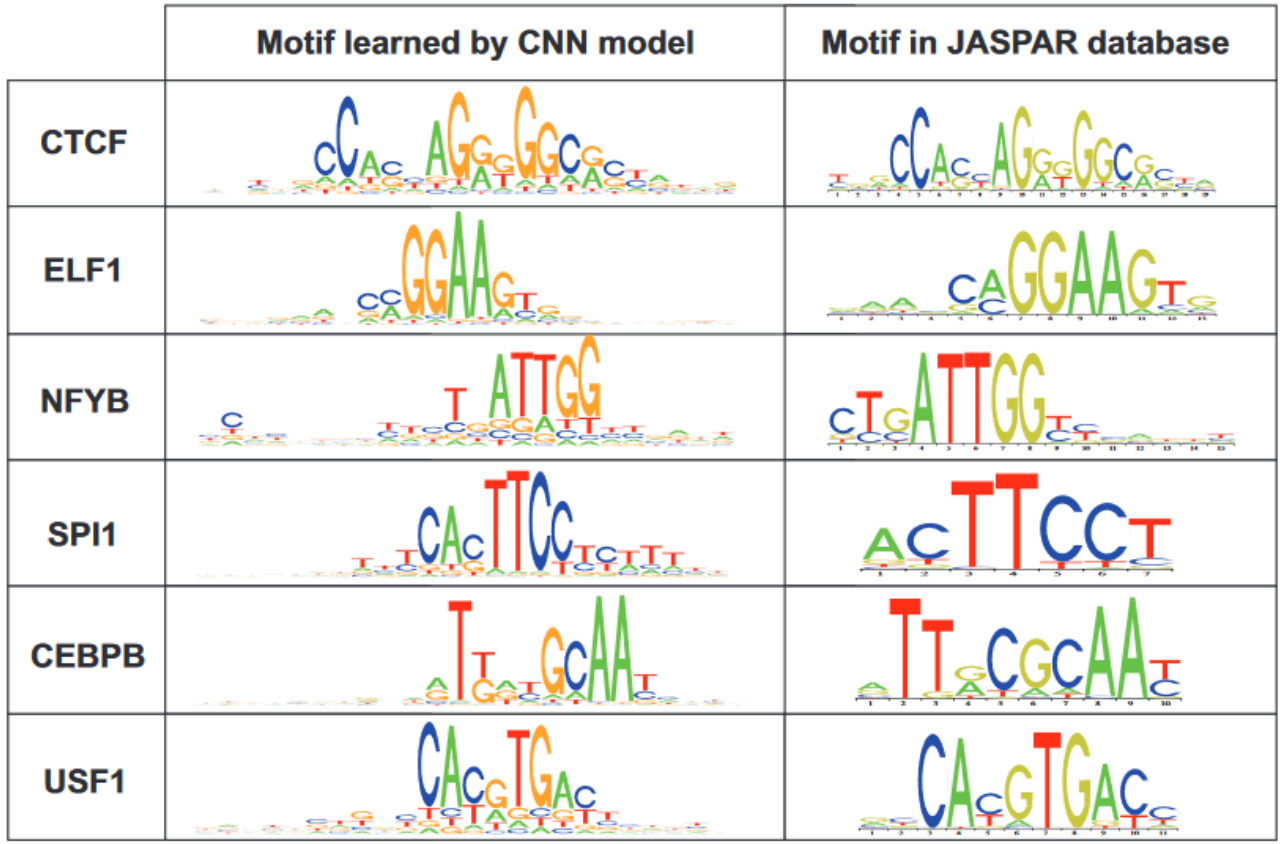
TF binding motifs revealed by models

Examples of TF motifs learned by the deep CNN models and compared with the motifs recorded in JASPAR.

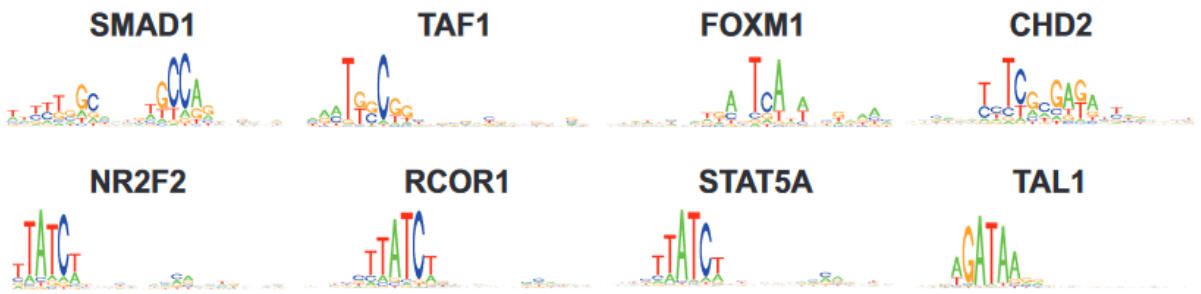
The sequence signatures automatically learned by the deep CNN models of DeFine captured the known TF binding motifs.

Some motifs discovered de novo by the deep CNN models for TFs were not annotated in the JASPAR database.

A

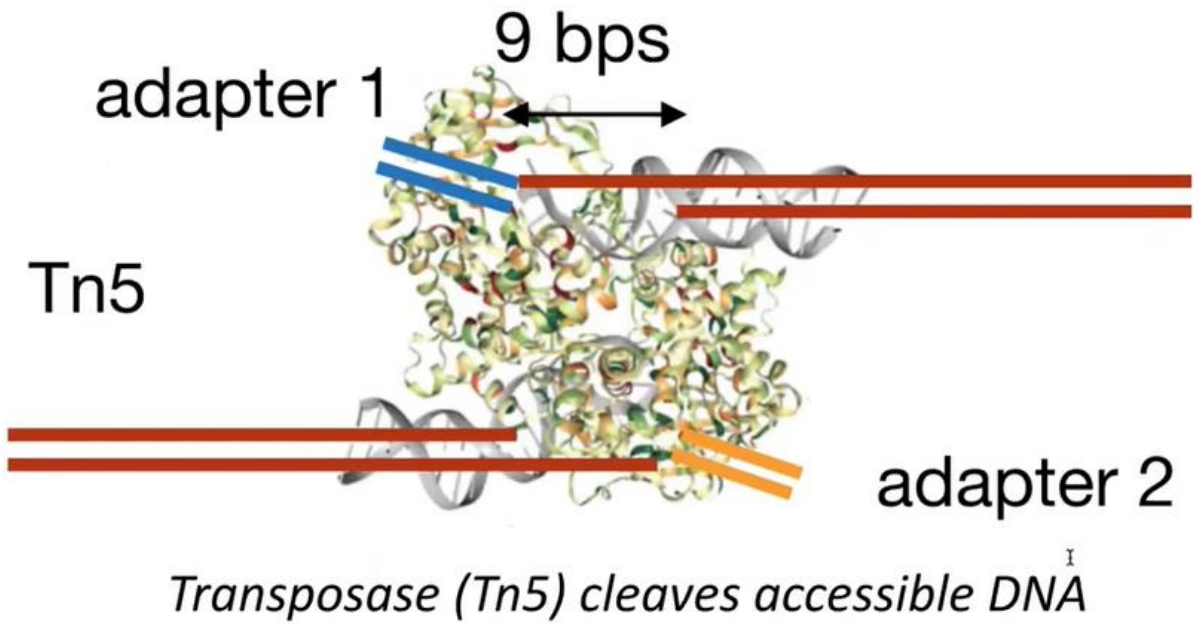


B



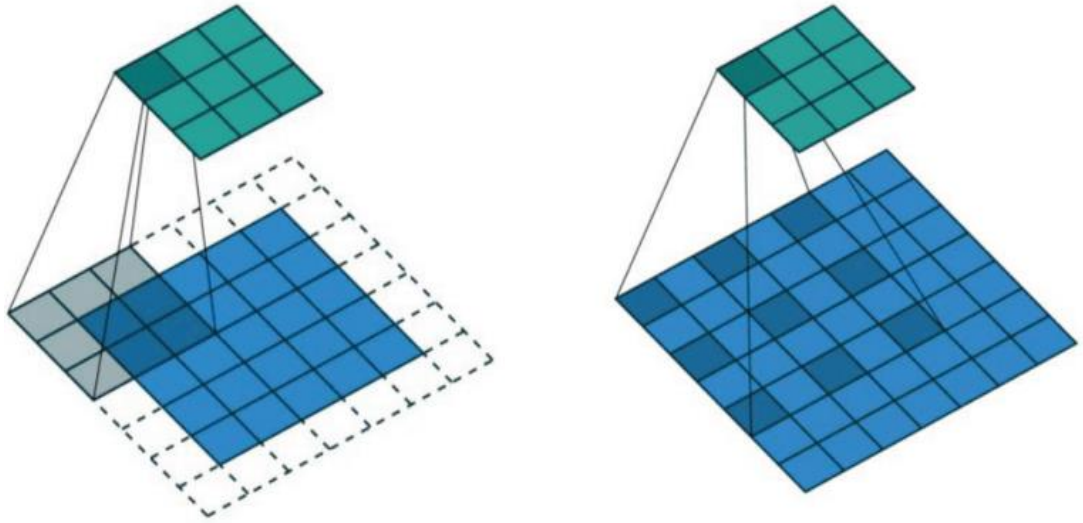
MaxATAC uses DNA sequence and ATAC-seq signal with Dilated CNN to predict TFBS

- ATAC-seq = Assay for Transposase Accessible Chromatin
- Dilated convolutional layers capture spatially distant relationships across the input sequences
- High-resolution TFBS predictions and information-sharing between proximal sequence and accessibility signals.



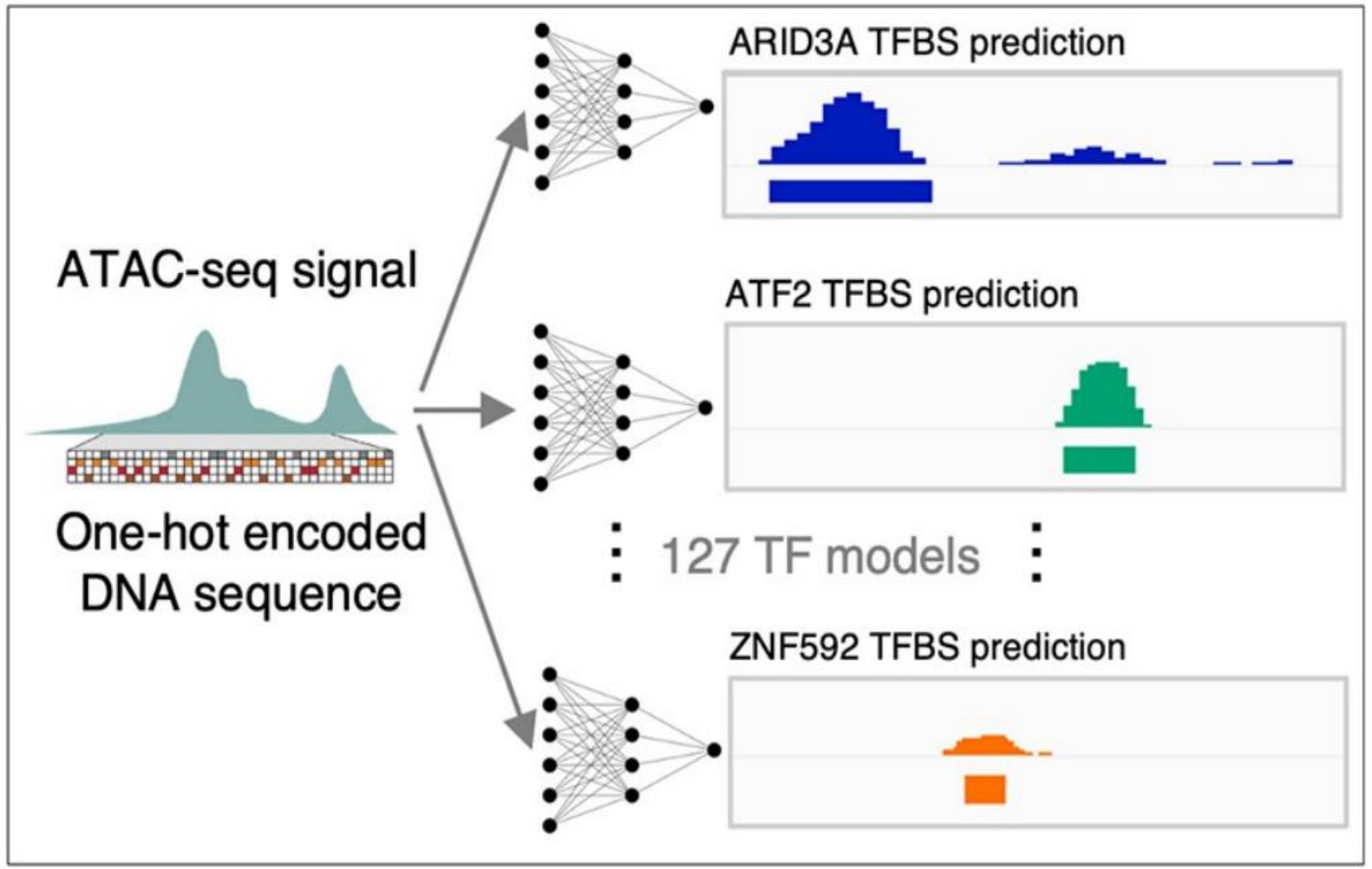
Transposase (Tn5) cleaves accessible DNA

Dilated Convolution

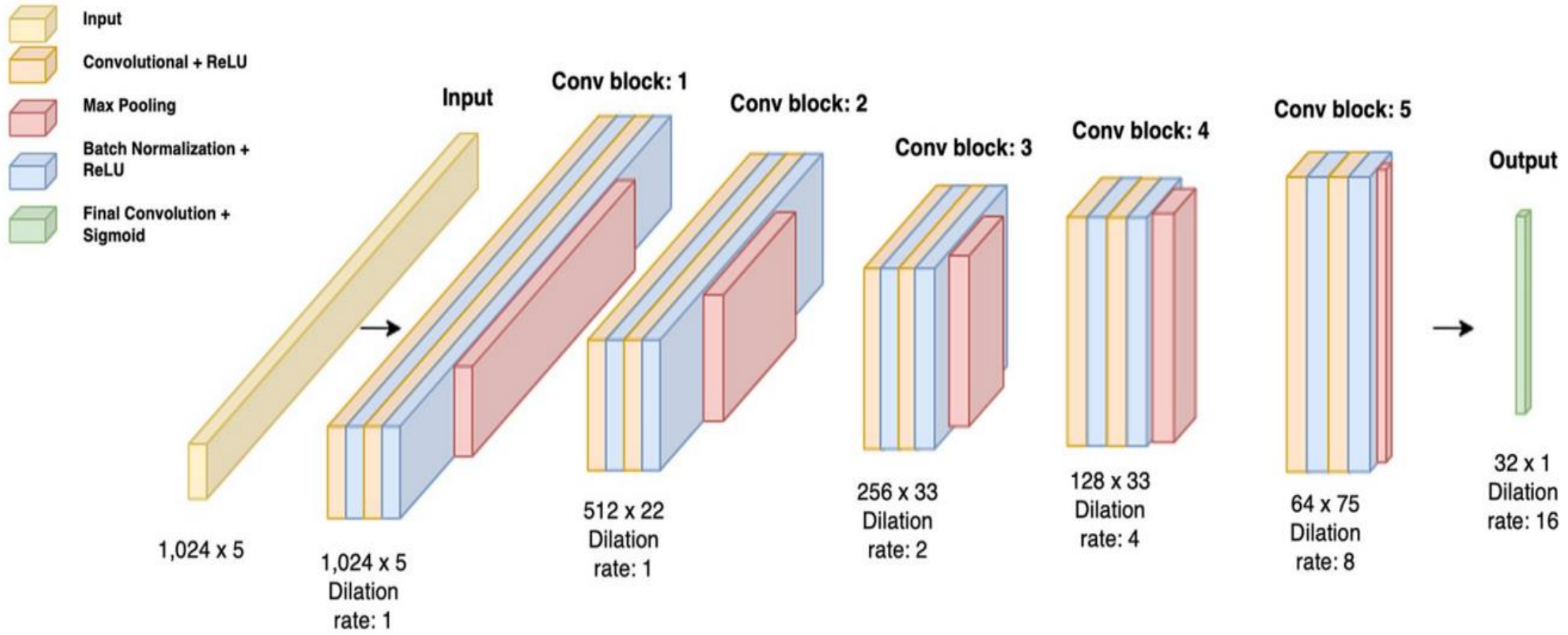


MaxATAC uses Dilated CNN to predict TFBS

- MaxATAC, uses DNA sequence and **ATAC-seq signal** to predict TF binding in a new cell type
- maxATAC Models are available for 127 human TFs

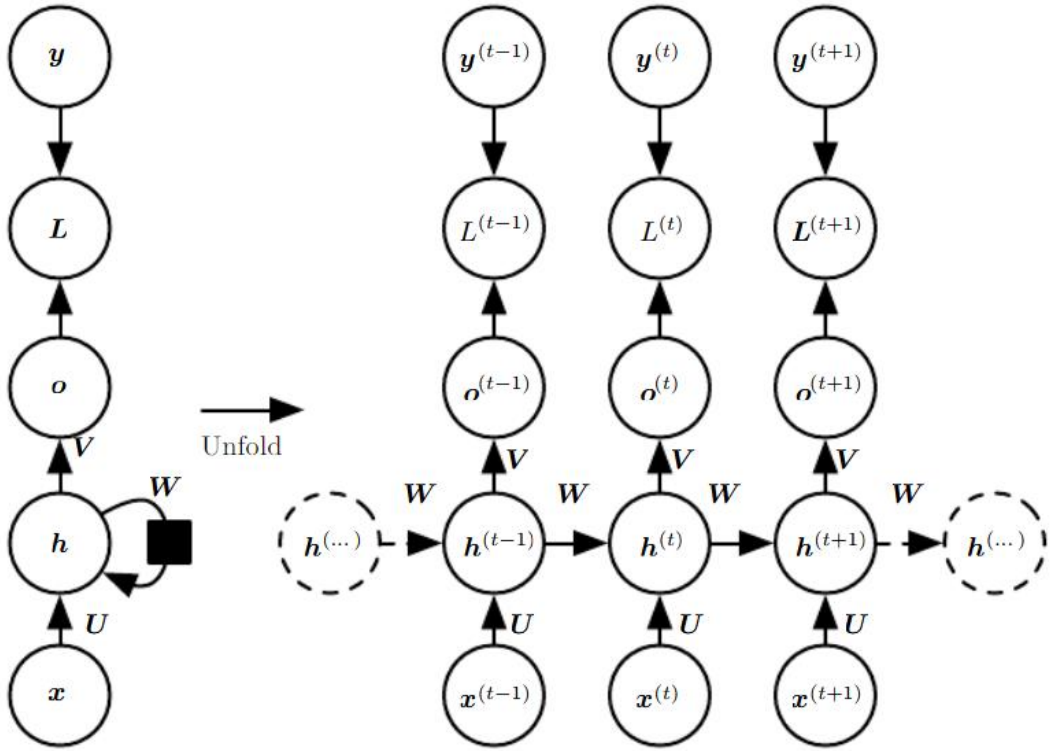


MaxATAC uses Dilated CNN to predict TFBS



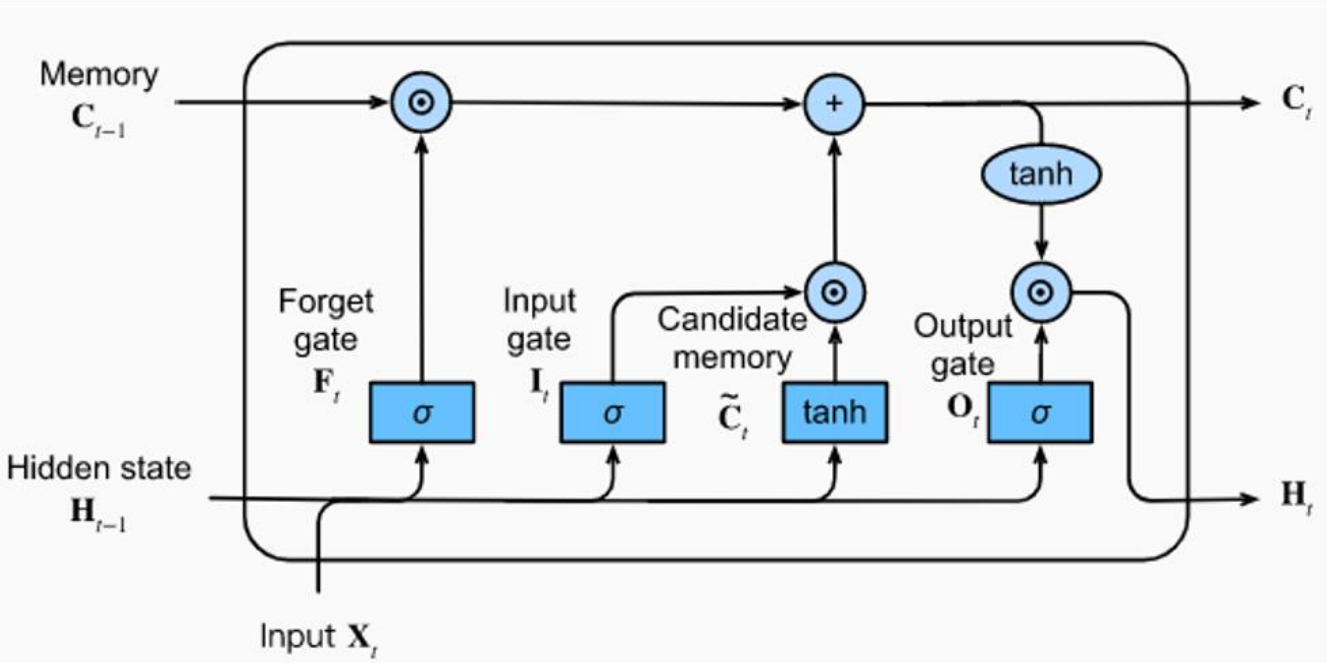
Recurrent neural network (RNN)

- Recurrent networks have recurrent connections between hidden units
  - Allow information to be passed from one step of the sequence to the next.
  - This recurrence allows the network to capture dependencies and patterns in the data that involve temporal relationships.
- RNN architecture is incapable of learning long-term dependencies.



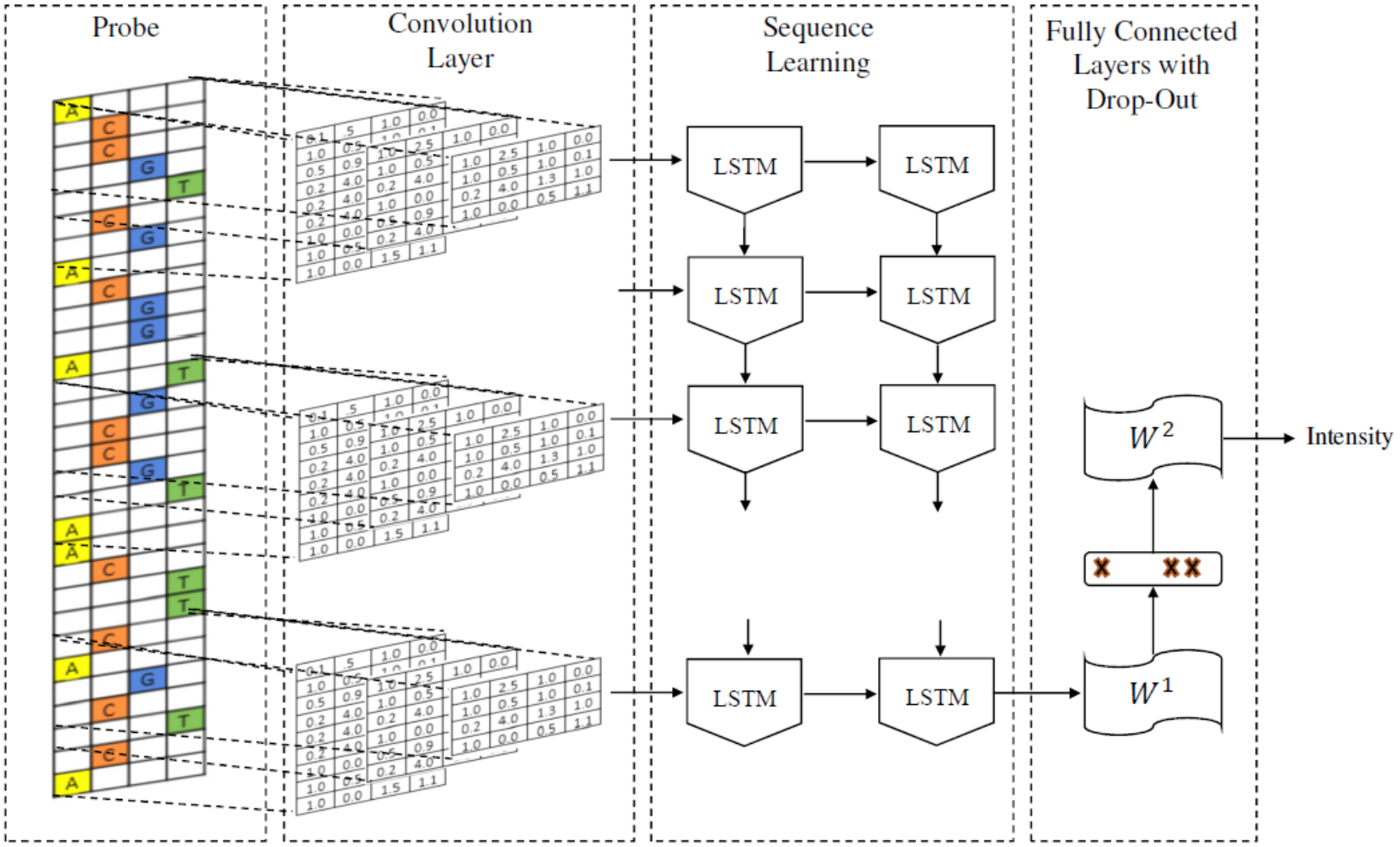
Long Short-Term Memory (LSTM) & Bi-LSTM

- To address issue of RNN, introduced Long Short-Term Memory (LSTM).
- LSTM consists of an input gate, output gate, and forget gate
- forget gate that allows the model to either reflect or forget the impact of input data at each time step.

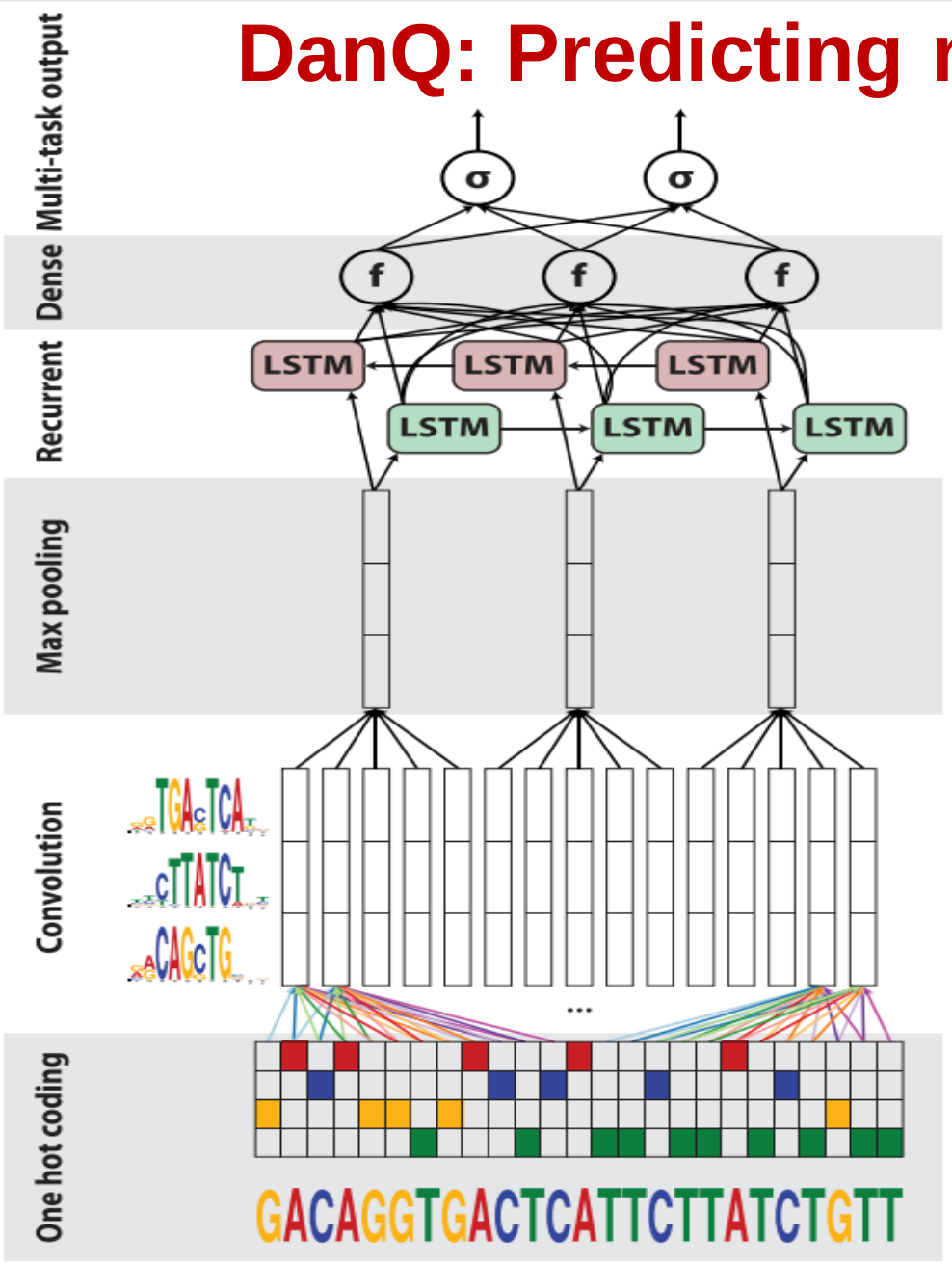


DeeperBind: Enhancing Prediction of Sequence Specificities of DNA Binding Proteins

- DeeperBind is an extension of DeepBind which adds a layer of Deep LSTM to model positional information.
- Predict protein-DNA binding affinity from high-throughput assays that measure the binding affinity.



DanQ: Predicting non-coding function de novo from sequence

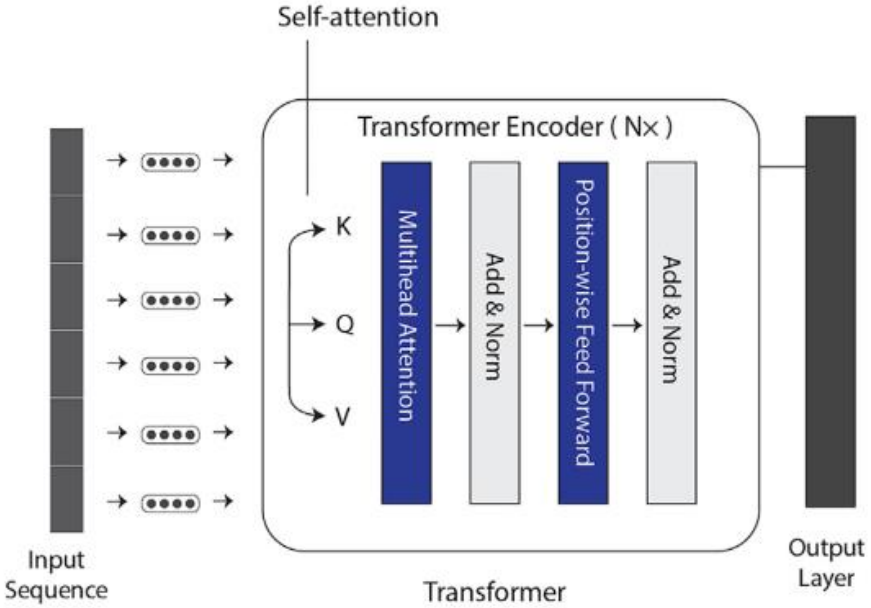


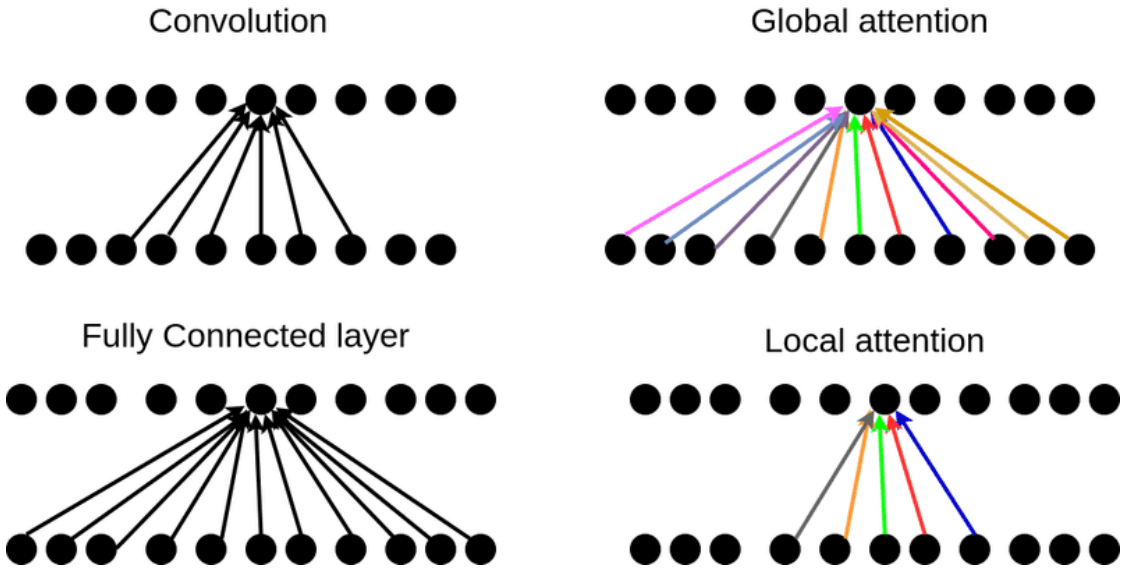
CNN & RNN(Bi-LSTM)

- hybrid convolutional : captures **regulatory motif**
- Bi-LSTM framework: captures **long-term dependencies between the motifs** in order to **learn a regulatory 'grammar'** to improve predictions.
- **Outperformed DeepSEA** even though they were trained exactly on the same dataset
- 1 convolution, 1 maxpool, easier than Deepsea
- Deepsea :3 Convolution layer, 2 Maxpooling layer

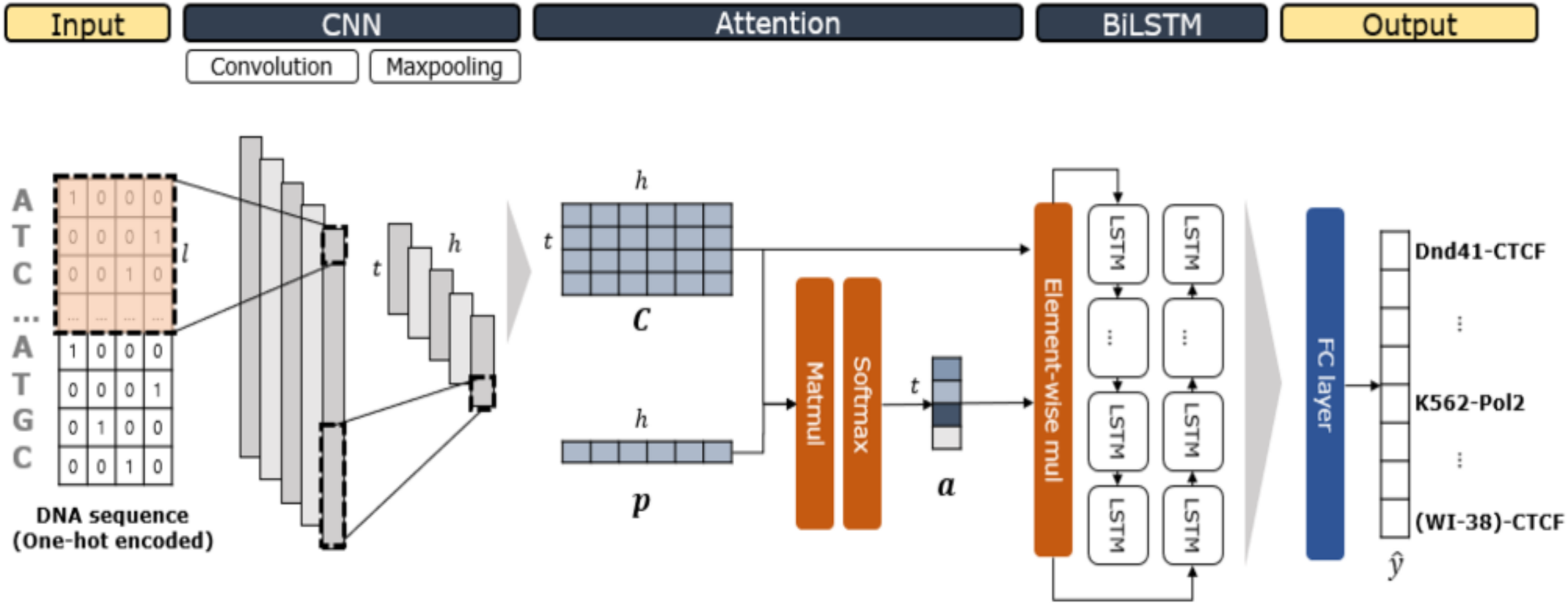
Transformer (Attention)

- Attention mechanism can assign different weight scores to each fragment of an input sequence to focus on more important fragments when generating outputs.



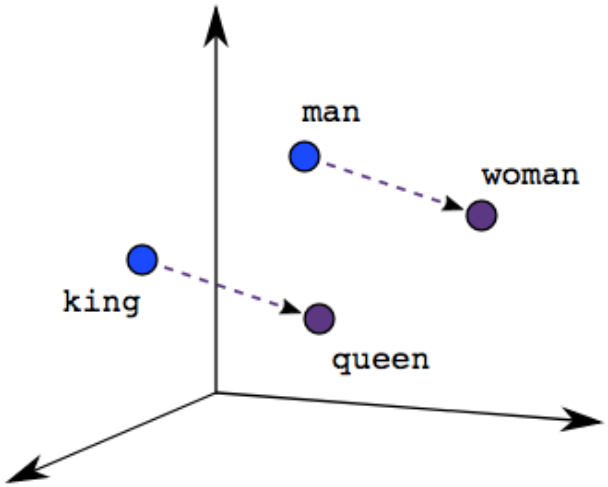


TBiNet
Enhancing the interpretability of transcription factor binding site prediction using attention mechanism

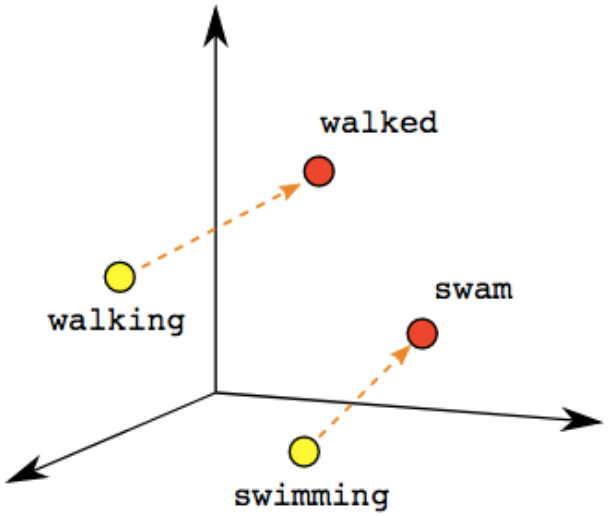


Word Embedding

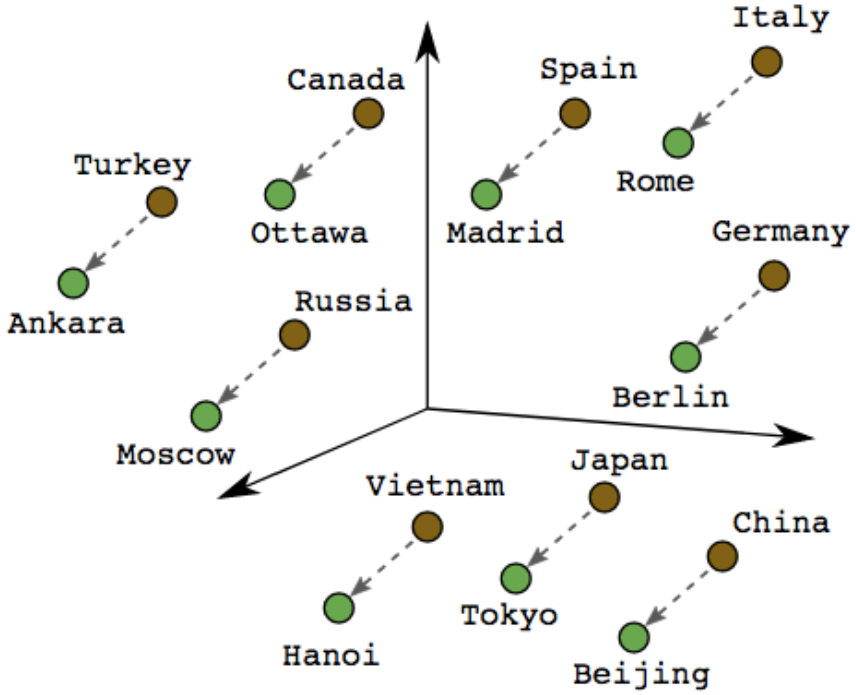
- Word Embeddings in NLP is a technique
- Each word is represented by a real-valued vector with tens or hundreds of dimensions.



Male-Female



Verb Tense

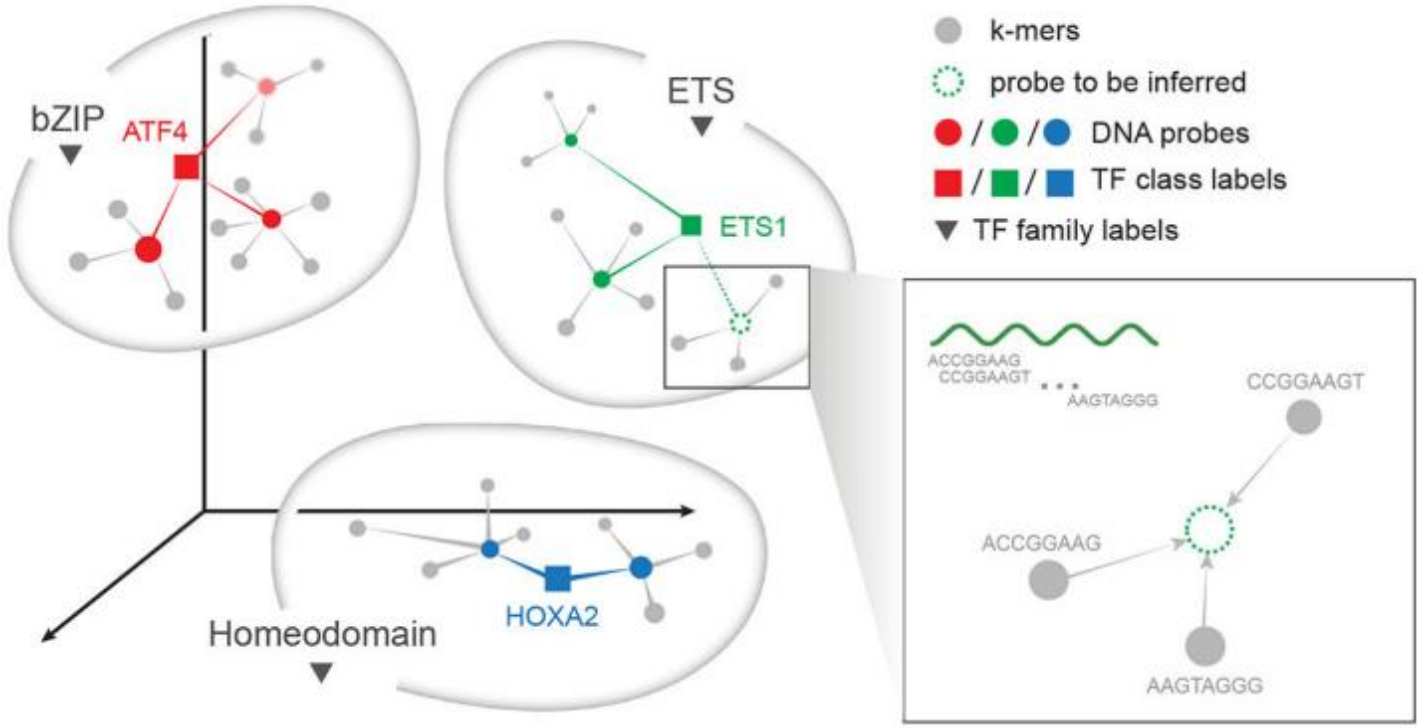


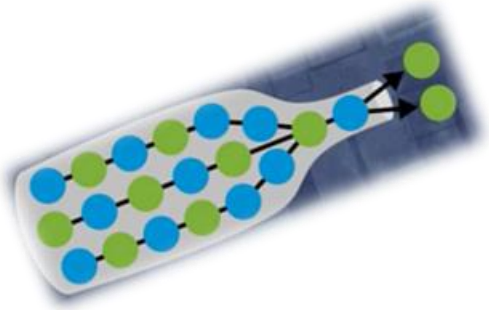
Country-Capital

https://towardsdatascience.com/a-guide-to-word-embeddings-8a23817ab60f

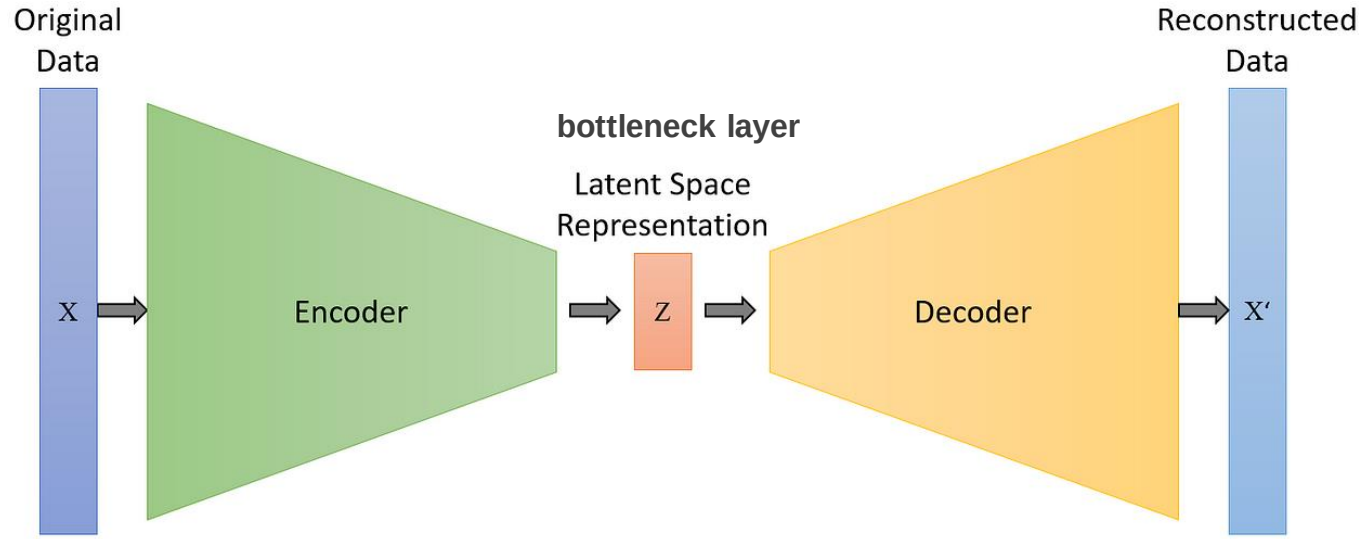
BindSpace: decodes transcription factor binding signals by large-scale sequence embedding

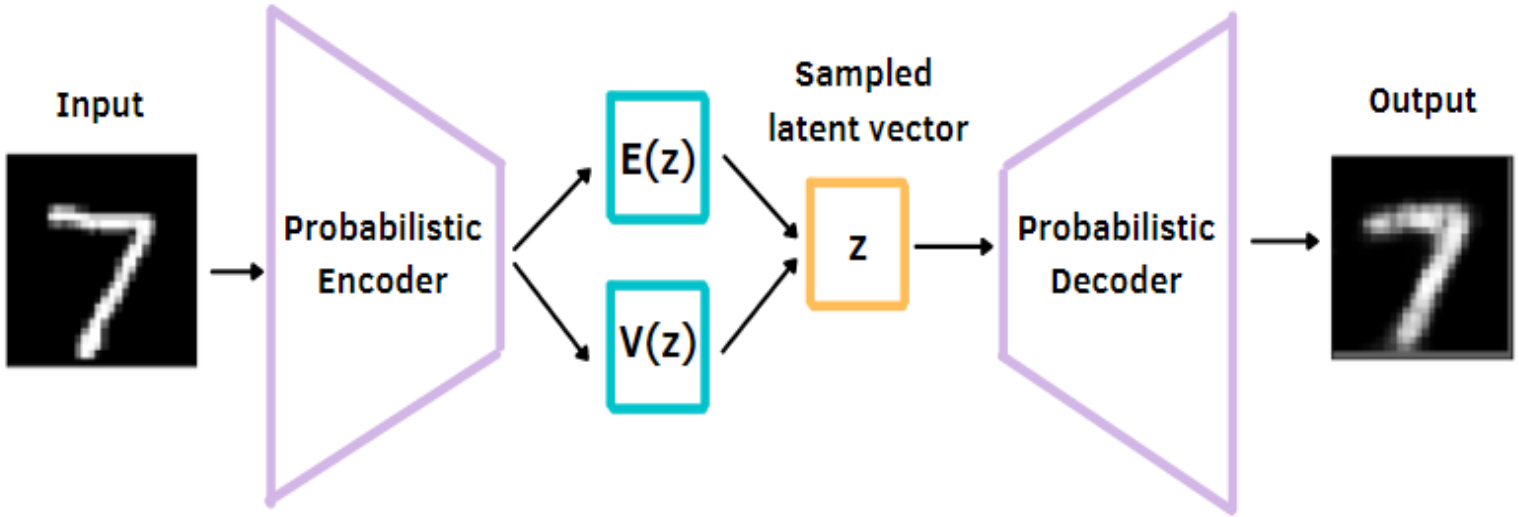
- BindSpace learns k-mer and label embeddings
- Probes embed close to the labels of TFs that bind them and away from other labels





Autoencoder & Variational autoencoder



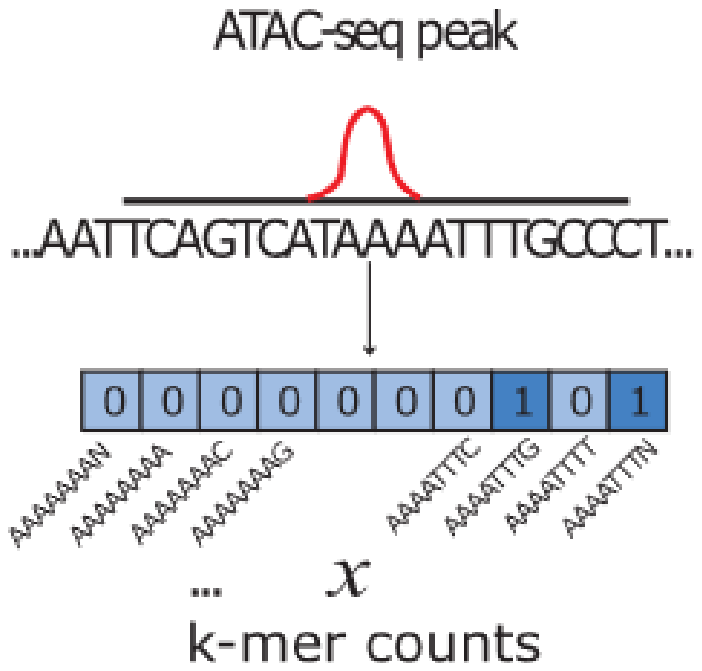


https://miro.medium.com/v2/resize:fit:2000/format:webp/1*pttBvMhZPbxgK9QYDhpNHQ.png

https://medium.com/dataseries/variational-autoencoder-with-pytorch-2d359cbf027b

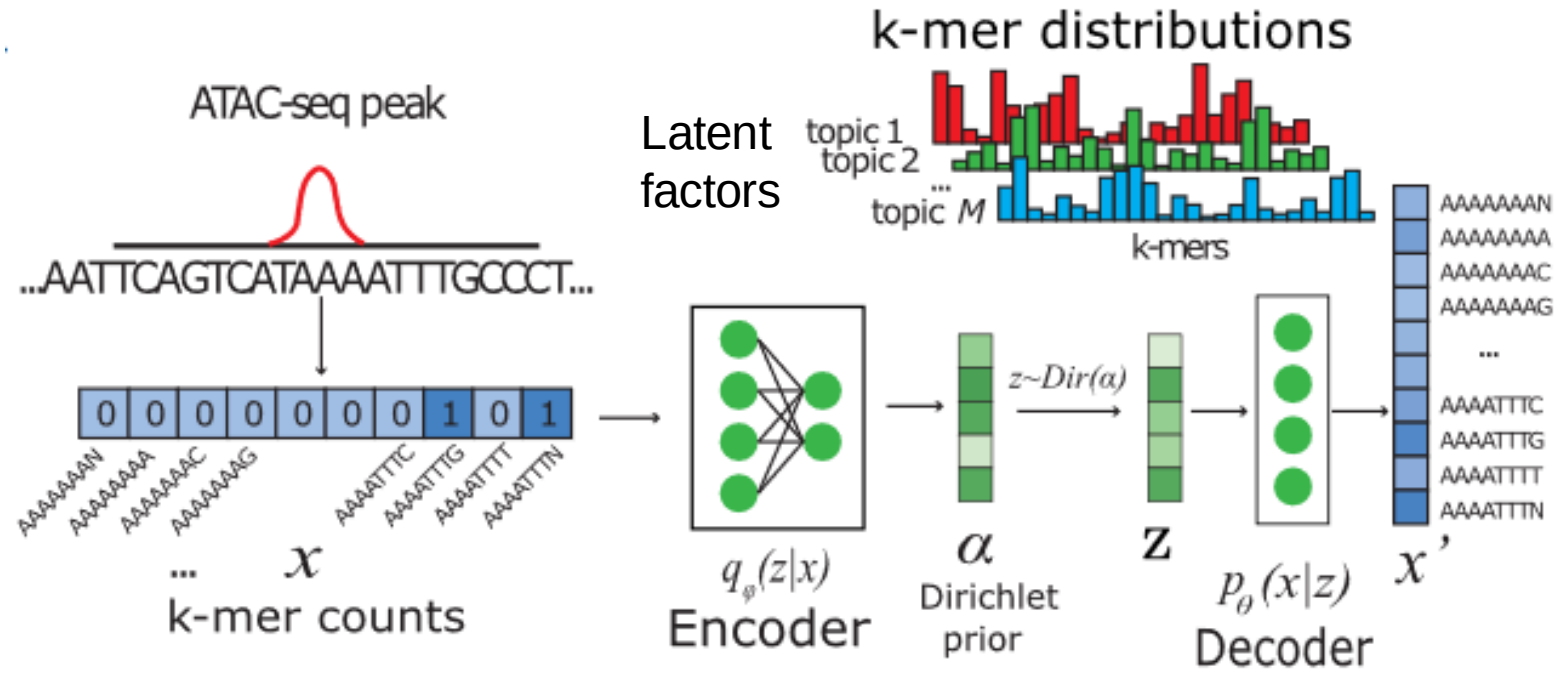
BindVAE
Variational autoencoders for de novo motif discovery from accessible chromatin

- Input: peaks from a cell type (~100k peaks), 200bp length DNA sequences



BindVAE: Variational autoencoders for de novo motif discovery from accessible chromatin

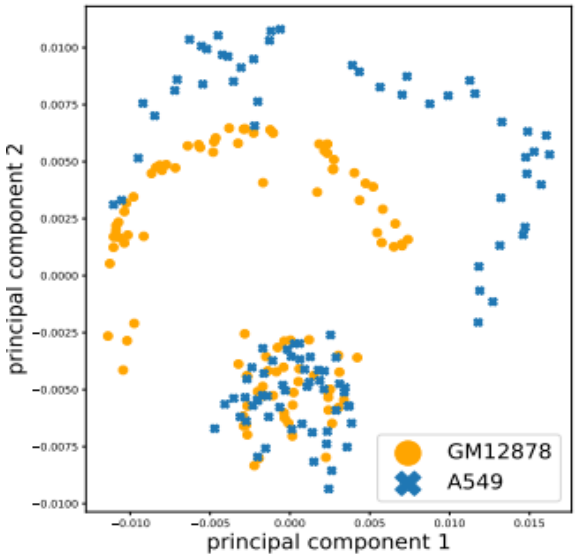
- VAE achieves compression in a probabilistic manner
- Encoder transforms the input x into parameters describing a probability distribution
- The decoder then reconstructs the input from the latent representation z



BindVAE

TF-binding motifs, cell-type specific

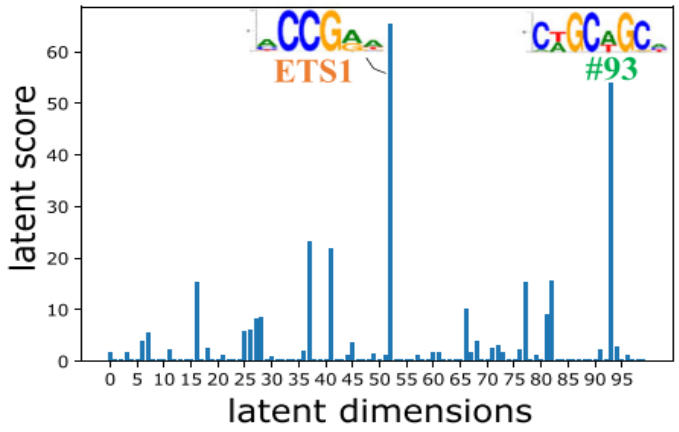
TF	GM12878 motif	A549 motif	CIS-BP motif
HNF4A			
NFIA			
SRY			
ELF5			
OLIG3			



Disentangled output from BindVAE

chr9: 69065430 – 69065460

TTCGGCCCTCTGCAGCCGCCATAGCTCCCCAGCAGAAAC CCGGAAGTGGA



FOXJ3-TBX21 cooperative binding motifs

CAP-SELEX motif
(generated by meme
from enriched probes)

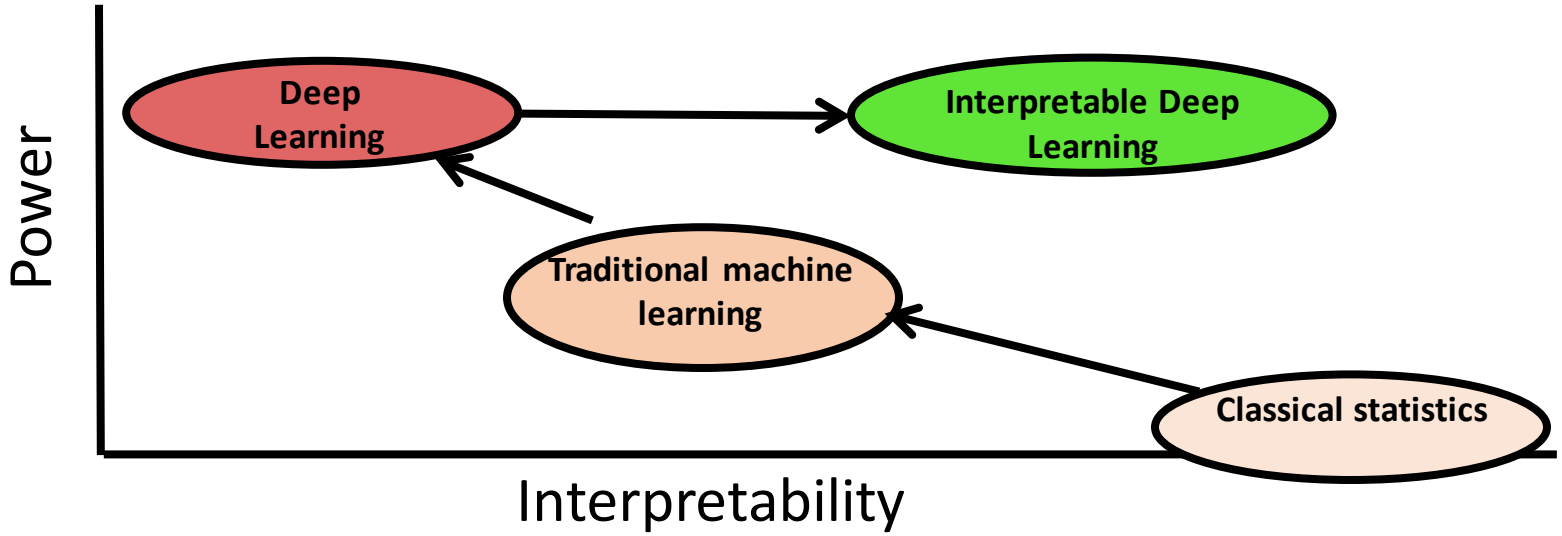
CAP-SELEX motifs
from Jolma et al. (2015)

BindVAE motif

Deep Learning & Poor interpretability

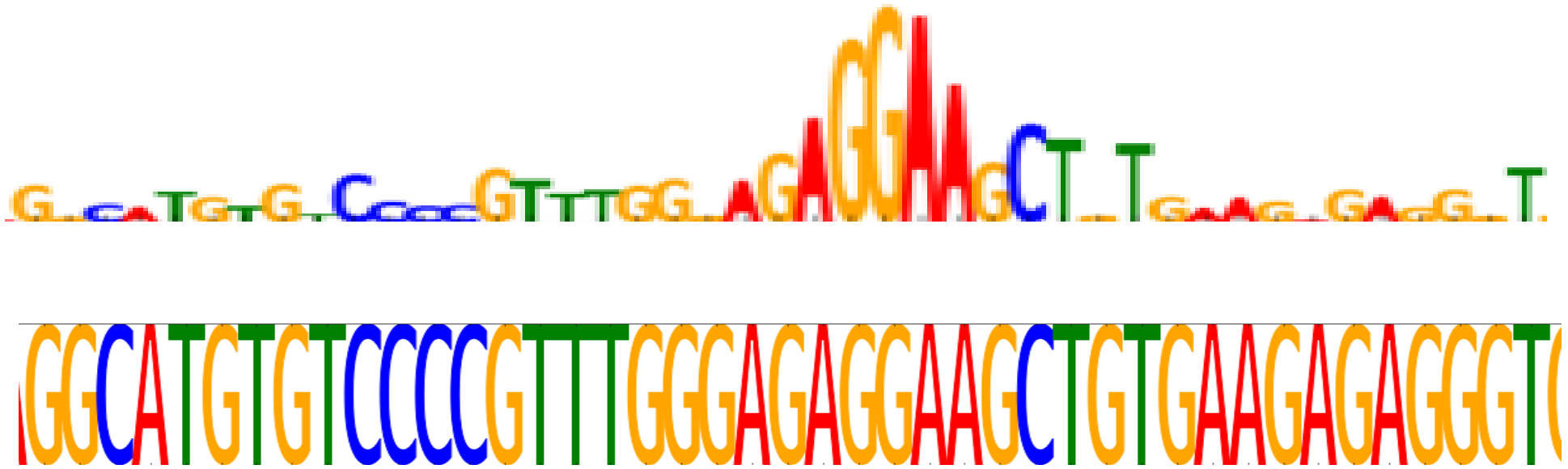
Black box of AI

- Need to understand how machines learn.

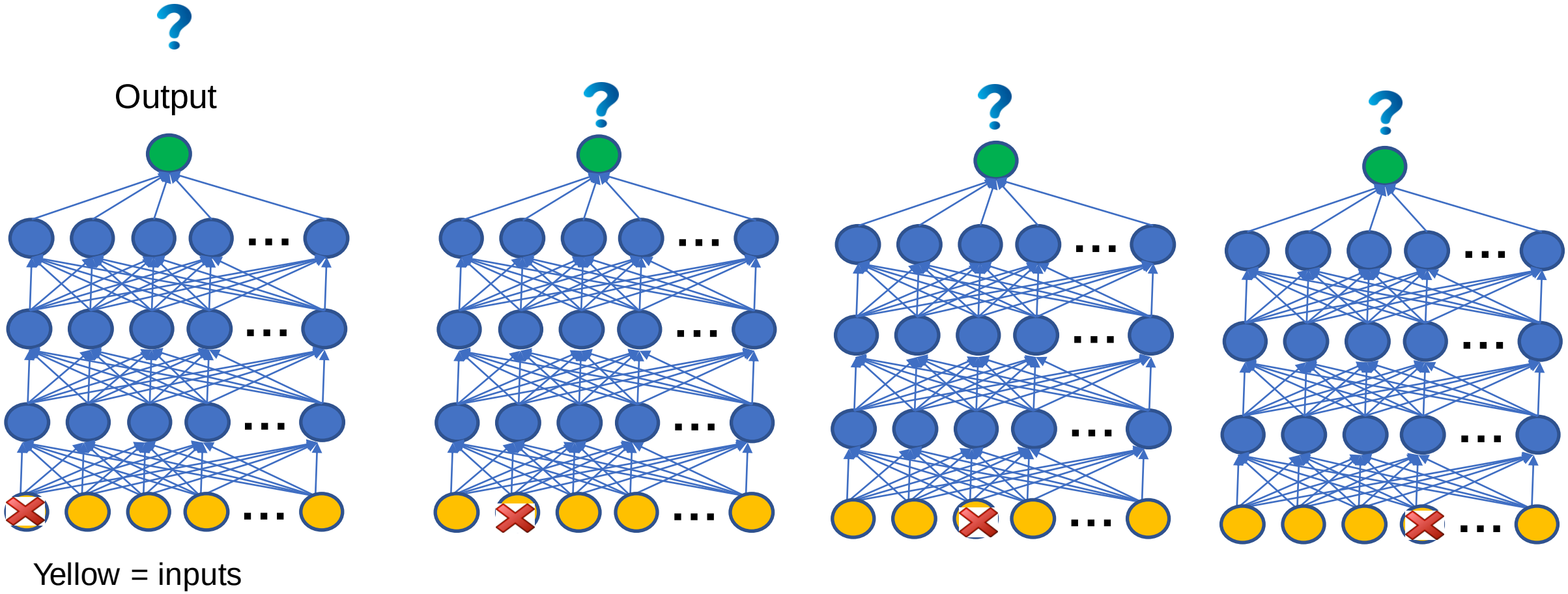


What do we mean by interpretability?

- DeepLIFT: can efficiently reveal important parts of the input for a given prediction
- For a given example and output, give an importance score to individual parts of the input



How can we find the important parts of the input for a given prediction? Perturbation

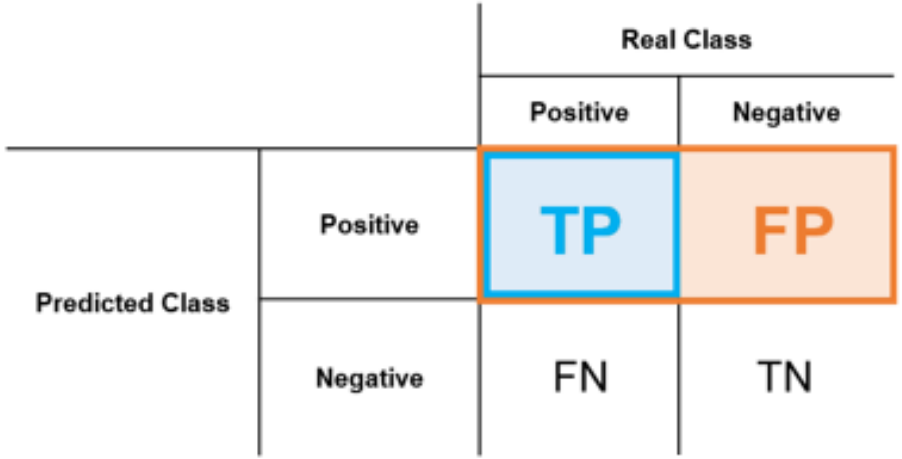


Evaluation

Precision

Of all positive predictions, how many are really positive?

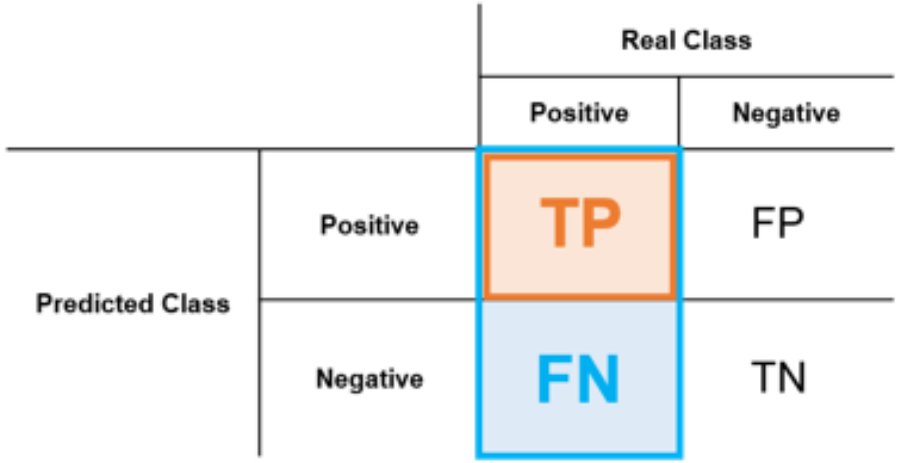
$$\frac{TP}{TP + FP}$$



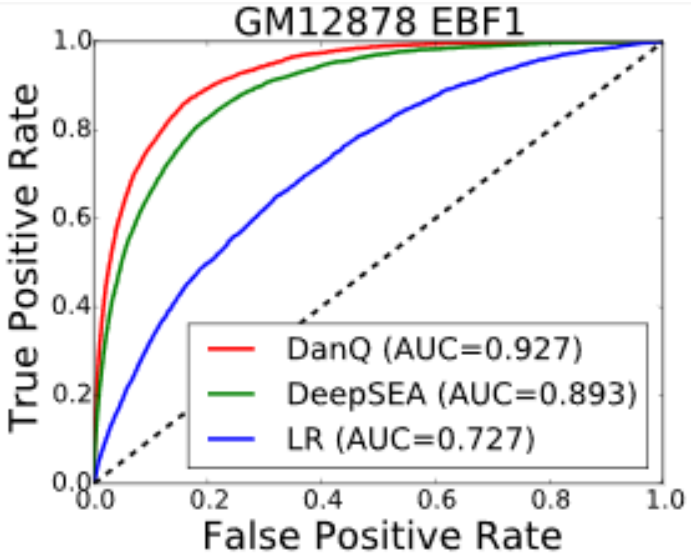
Recall

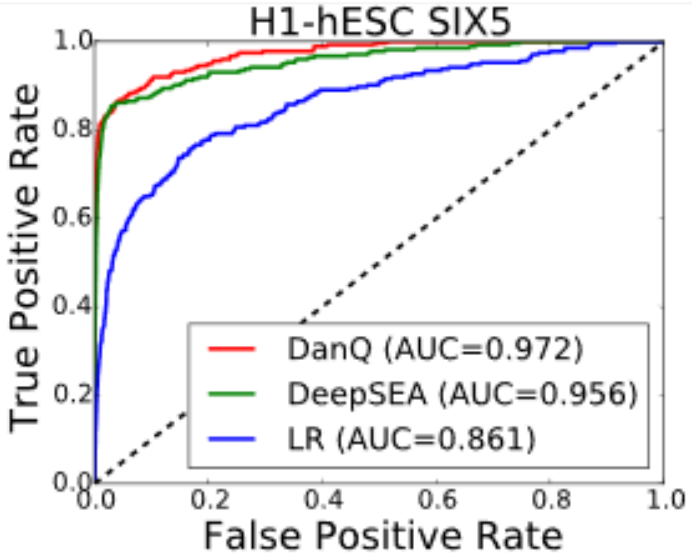
Of all real positive cases, how many are predicted positive?

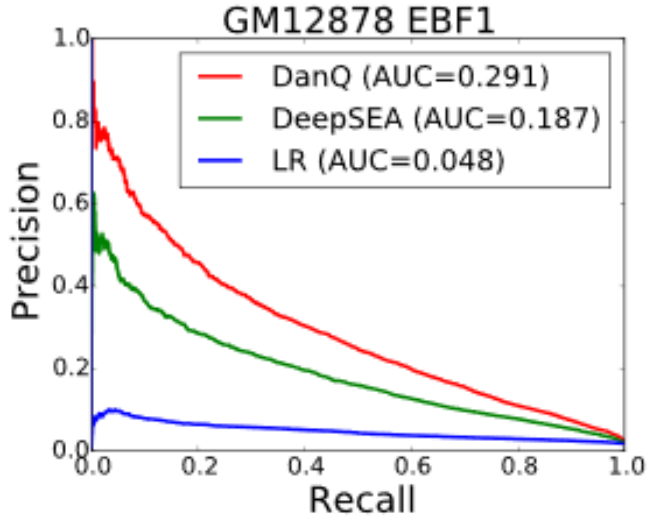
$$\frac{TP}{TP + FN}$$

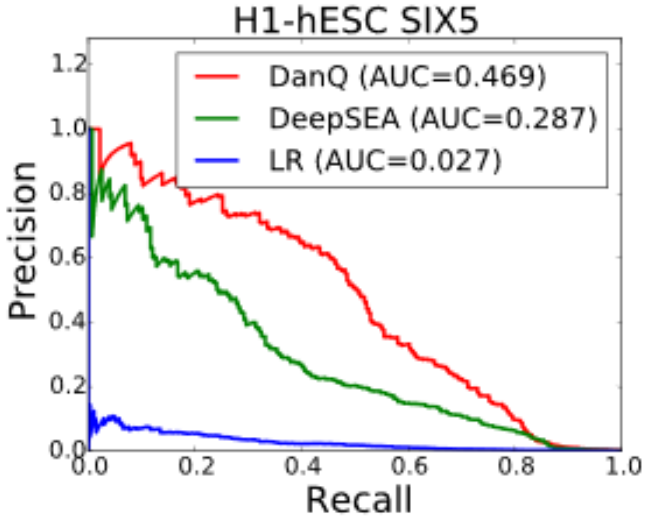


Evaluation

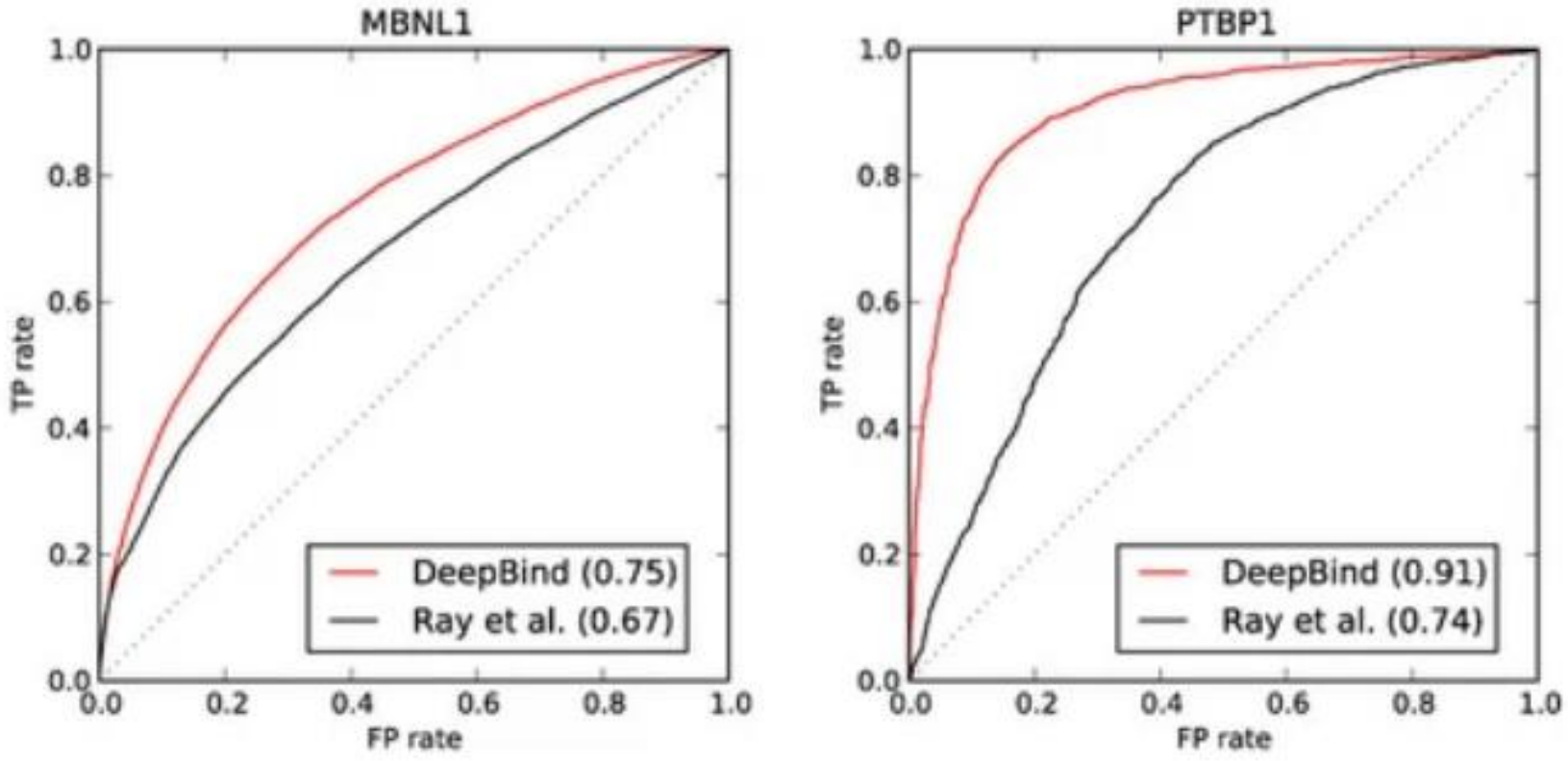

GM12878 EBF1
True Positive Rate
False Positive Rate
DanQ (AUC=0.927)
DeepSEA (AUC=0.893)
LR (AUC=0.727)


H1-hESC SIX5
True Positive Rate
False Positive Rate
DanQ (AUC=0.972)
DeepSEA (AUC=0.956)
LR (AUC=0.861)


GM12878 EBF1
Precision
Recall
DanQ (AUC=0.291)
DeepSEA (AUC=0.187)
LR (AUC=0.048)

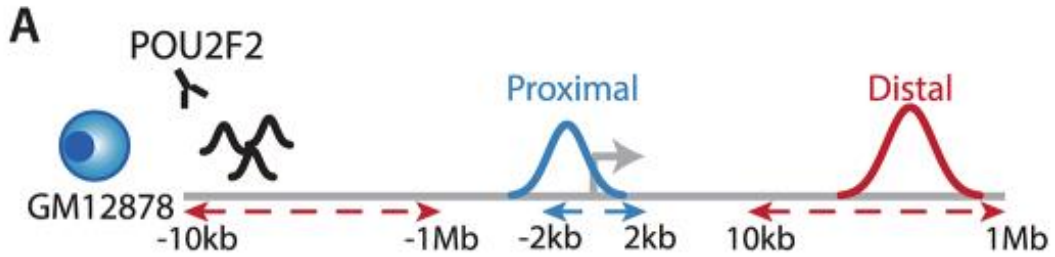

H1-hESC SIX5
Precision
Recall
DanQ (AUC=0.469)
DeepSEA (AUC=0.287)
LR (AUC=0.027)

Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning

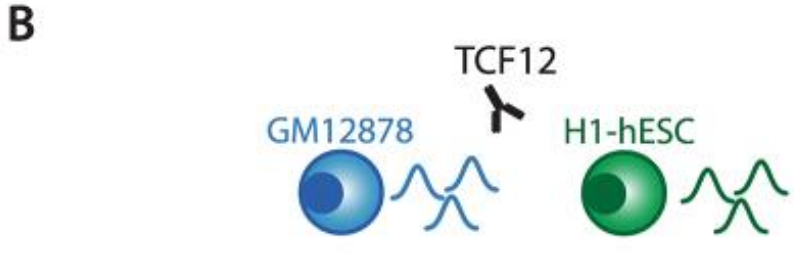


Evaluation

- POU2F2 (POU Class 2 Homeobox 2) is a Protein Coding gene. Diseases associated with POU2F2 include B-Cell Lymphoma and Papilloma.
- TCF12 functioned as a transcriptional repressor.



Common	Proximal	Distal
Oct	YY1	BATF
ETS	ZNF	TCF



Common	GM12878	H1-hESC
TCF	BATF	TEAD
E2A-PU.1	RUNX	PRDM

Conclusion

2006- MEME: discovering and analyzing DNA and protein sequence motifs- Timothy L Bailey

2008-Amadeus -Transcription factor and microRNA motif discovery: the Amadeus platform and a compendium of metazoan target sets

2011-MEME-Chip: motif analysis of large DNA datasets

2011-DREME: motif discovery in transcription factor ChIP-seq data

2011- STEME: efficient EM to find motifs in large data sets

1995- The value of prior knowledge in discovering motifs with MEM- TL Bailey, C Elkan

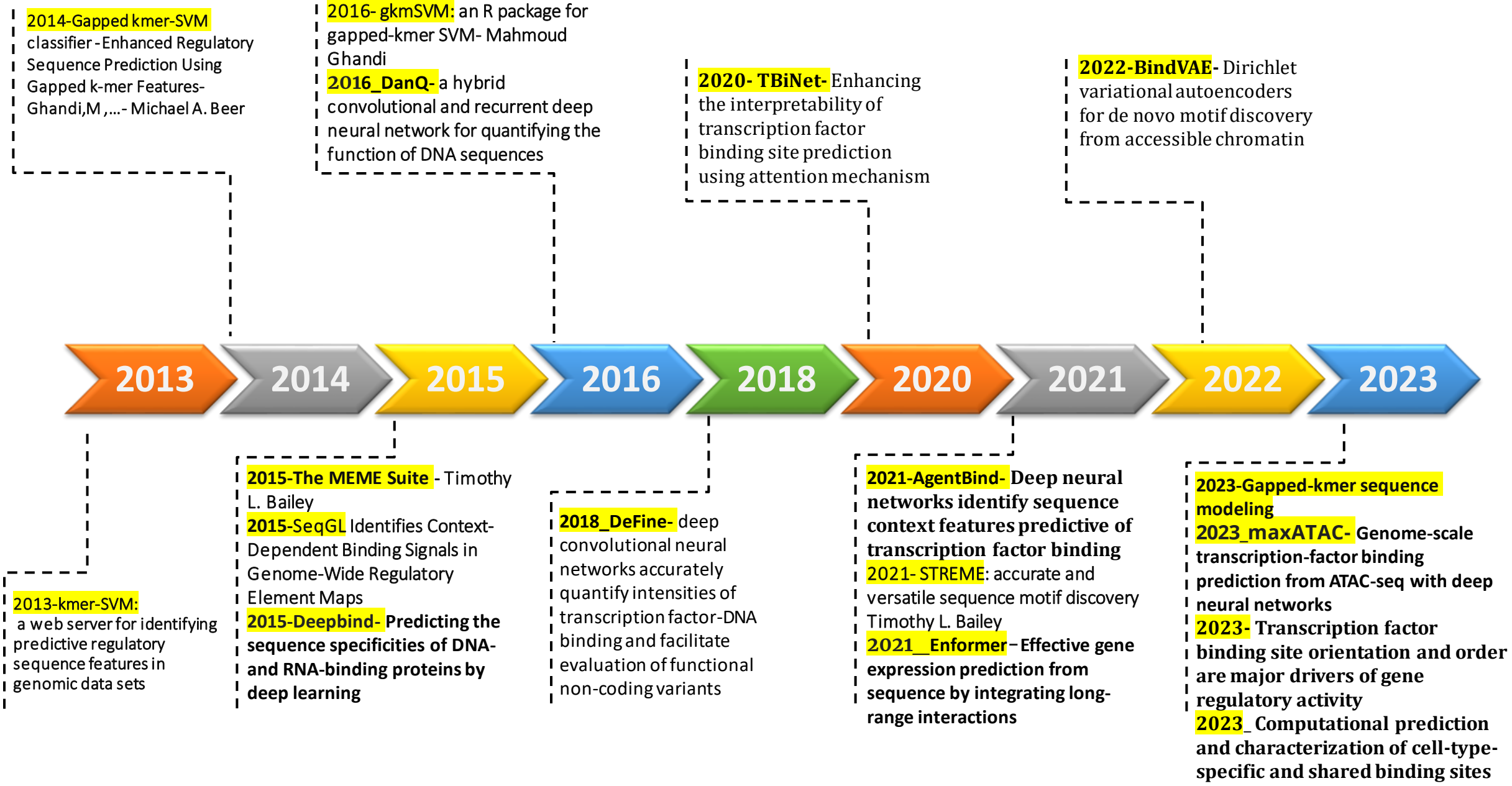
2007-Tomtom -Quantifying similarity between motifs

2007-Trawler: de novo regulatory motif discovery pipeline for chromatin immunoprecipitation

2010-HOMER -Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities Sven Heinz(Heinz et al., 2010),

2012- RSAT peak-motifs: motif analysis in full-size ChIP-seq datasets- Morgane Thomas-Chollier

2012- A complete workflow for the analysis of full-size ChIP-seq (and similar) data sets using peak-motifs- Morgane Thomas-Chollier


2014-Gapped kmer-SVM classifier - Enhanced Regulatory Sequence Prediction Using Gapped k-mer Features- Ghandi,M ,...- Michael A. Beer
2016- gkmSVM: an R package for gapped-kmer SVM- Mahmoud Ghandi
2016_DanQ- a hybrid convolutional and recurrent deep neural network for quantifying the function of DNA sequences
2020- TBiNet- Enhancing the interpretability of transcription factor binding site prediction using attention mechanism
2022-BindVAE- Dirichlet variational autoencoders for de novo motif discovery from accessible chromatin
2013
2014
2015
2016
2018
2020
2021
2022
2023
2013-kmer-SVM: a web server for identifying predictive regulatory sequence features in genomic data sets
2015-The MEME Suite - Timothy L. Bailey
2015-SeqGL Identifies Context-Dependent Binding Signals in Genome-Wide Regulatory Element Maps
2015-Deepbind- Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning
2018_DeFine- deep convolutional neural networks accurately quantify intensities of transcription factor-DNA binding and facilitate evaluation of functional non-coding variants
2021-AgentBind- Deep neural networks identify sequence context features predictive of transcription factor binding
2021- STREME: accurate and versatile sequence motif discovery Timothy L. Bailey
2021__Enformer – Effective gene expression prediction from sequence by integrating long-range interactions
2023-Gapped-kmer sequence modeling
2023_maxATAC- Genome-scale transcription-factor binding prediction from ATAC-seq with deep neural networks
2023- Transcription factor binding site orientation and order are major drivers of gene regulatory activity
2023_ Computational prediction and characterization of cell-type-specific and shared binding sites

References

- Vaswani, Ashish, et al. "Attention is all you need." *Advances in neural information processing systems* 30 (2017).
- Quang, Daniel, and Xiaohui Xie. "DanQ: a hybrid convolutional and recurrent deep neural network for quantifying the function of DNA sequences." *Nucleic acids research* 44.11 (2016): e107-e107.
- Zhou, Jian, and Olga G. Troyanskaya. "Predicting effects of noncoding variants with deep learning–based sequence model." *Nature methods* 12.10 (2015): 931-934.
- Kshirsagar, Meghana, et al. "BindVAE: Dirichlet variational autoencoders for de novo motif discovery from accessible chromatin." *Genome Biology* 23.1 (2022): 174.
- Yuan, Han, et al. "BindSpace decodes transcription factor binding signals by large-scale sequence embedding." *Nature methods* 16.9 (2019): 858-861.
- Park, Sungjoon, et al. "Enhancing the interpretability of transcription factor binding site prediction using attention mechanism." *Scientific reports* 10.1 (2020): 13413.
- Cazares, Tareian A., et al. "maxATAC: Genome-scale transcription-factor binding prediction from ATAC-seq with deep neural networks." *PLOS Computational Biology* 19.1 (2023): e1010863.
- Wang, Meng, et al. "DeFine: deep convolutional neural networks accurately quantify intensities of transcription factor-DNA binding and facilitate evaluation of functional non-coding variants." *Nucleic acids research* 46.11 (2018): e69-e69.
- Setty, Manu, and Christina S. Leslie. "SeqGL identifies context-dependent binding signals in genome-wide regulatory element maps." *PLoS computational biology* 11.5 (2015): e1004271.
- Korhonen, Janne H., et al. "Fast motif matching revisited: high-order PWMs, SNPs and indels." Bioinformatics 33.4 (2017): 514-521.

Thanks