Predicting effects of noncoding variants with deep learning-based sequence model

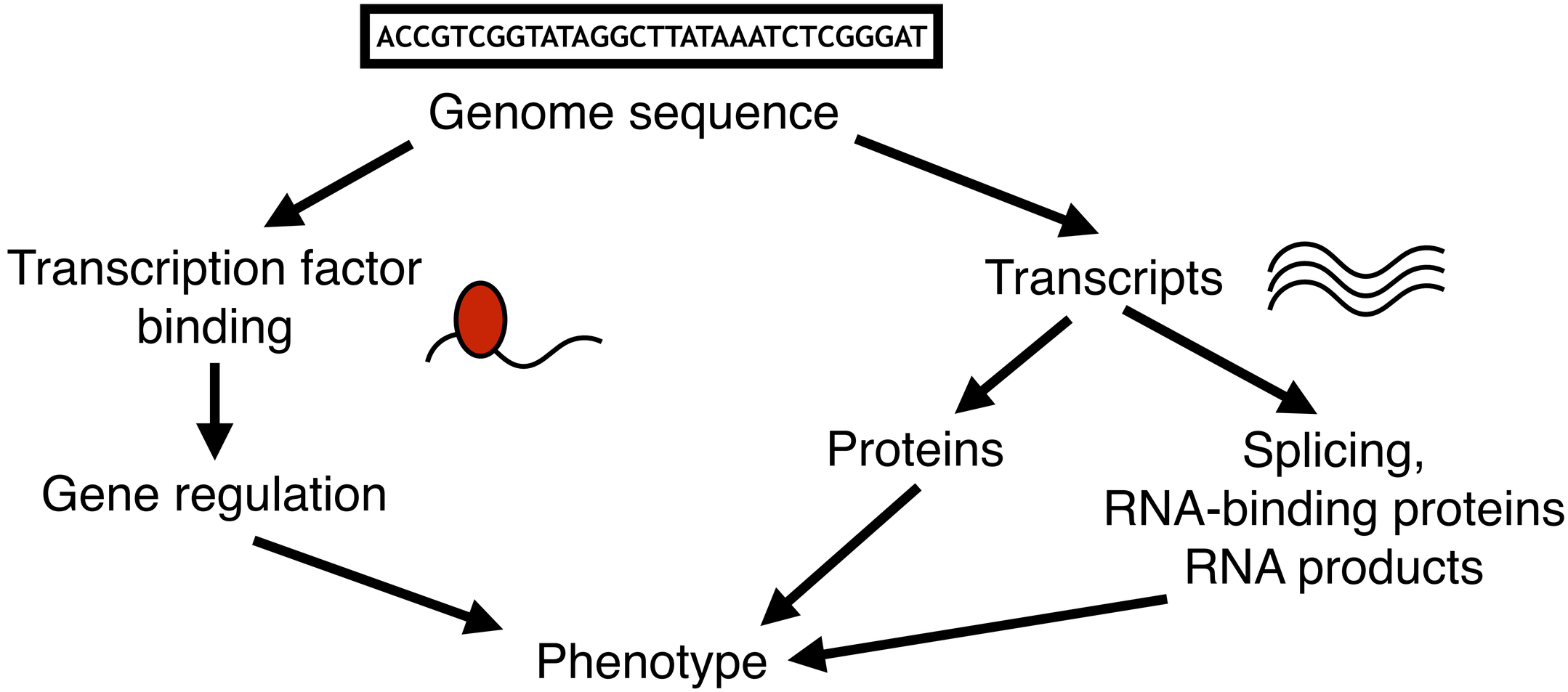
Jian Zhou & Olga G Troyanskaya

How does genotype affect phenotype?

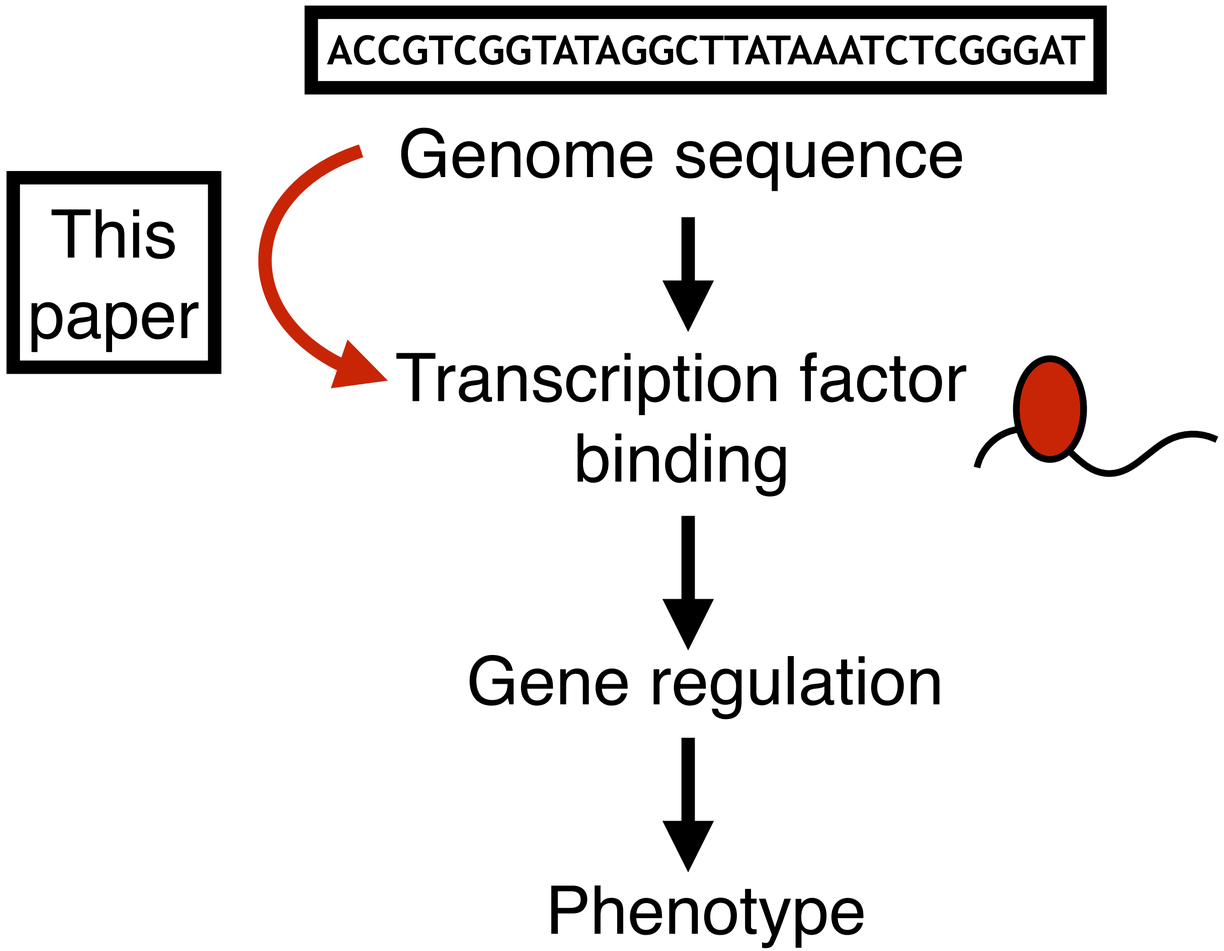
ACCGTCGGTATAGGCTTATAAATC**A**TCGGGATCCTATTAATGAGGAAAA

Genetic traits
Disease
Evolutionary fitness

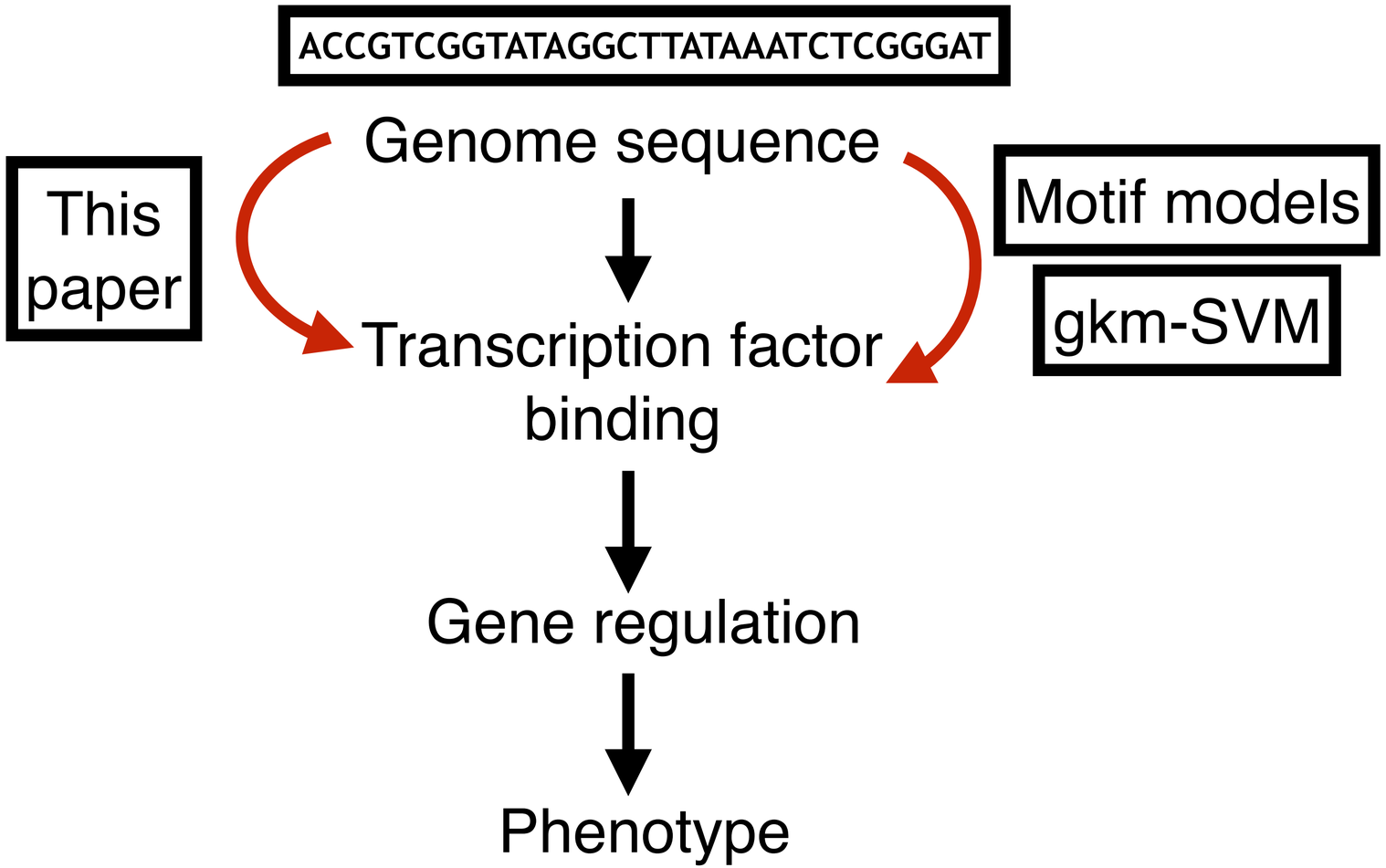
Genotype usually determines phenotype either through (1) protein-coding sequence or (2) transcription factor binding



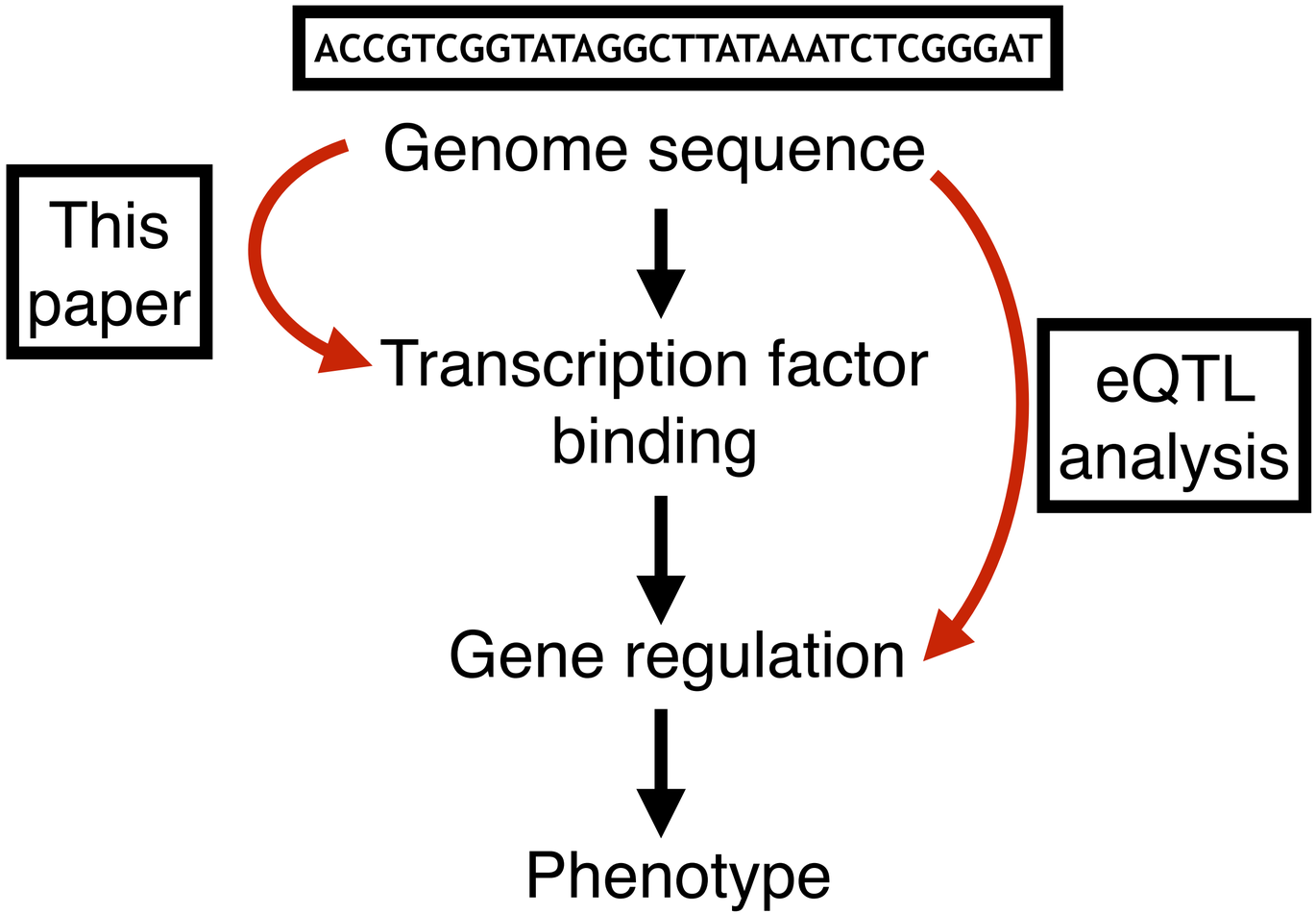
This paper focuses on predicting the effect of sequence on TF binding



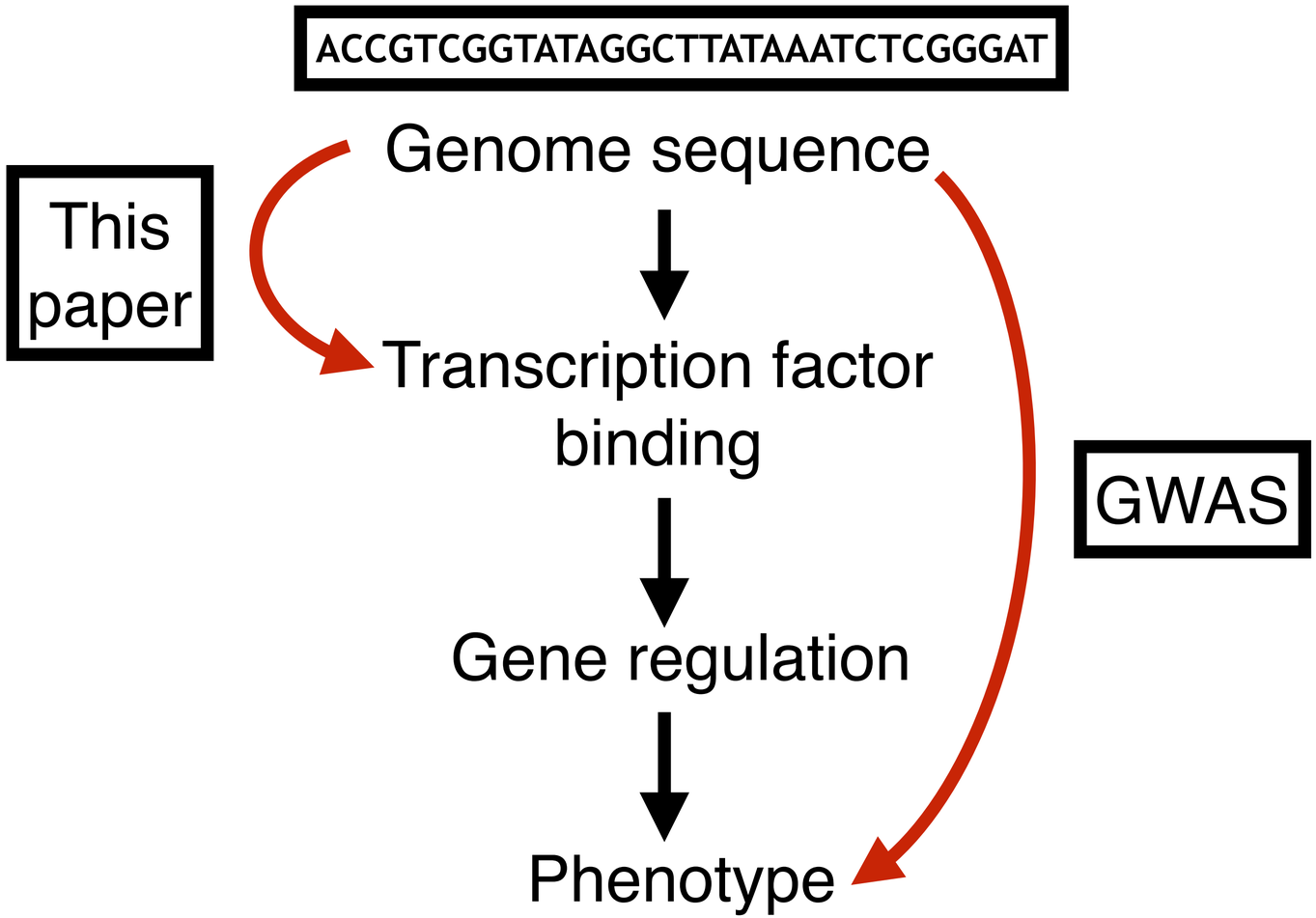
This paper focuses on predicting the effect of sequence on TF binding



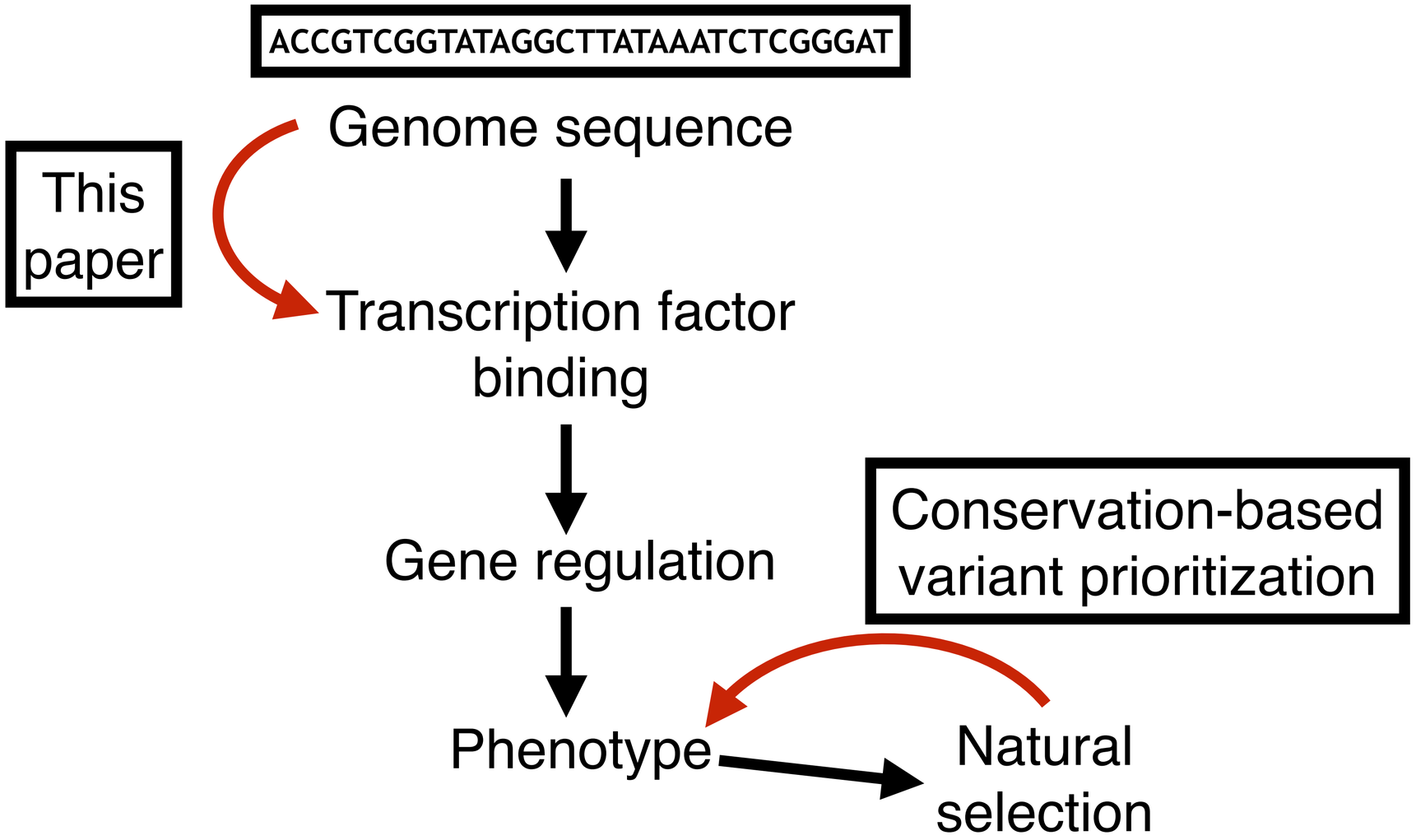
This paper focuses on predicting the effect of sequence on TF binding



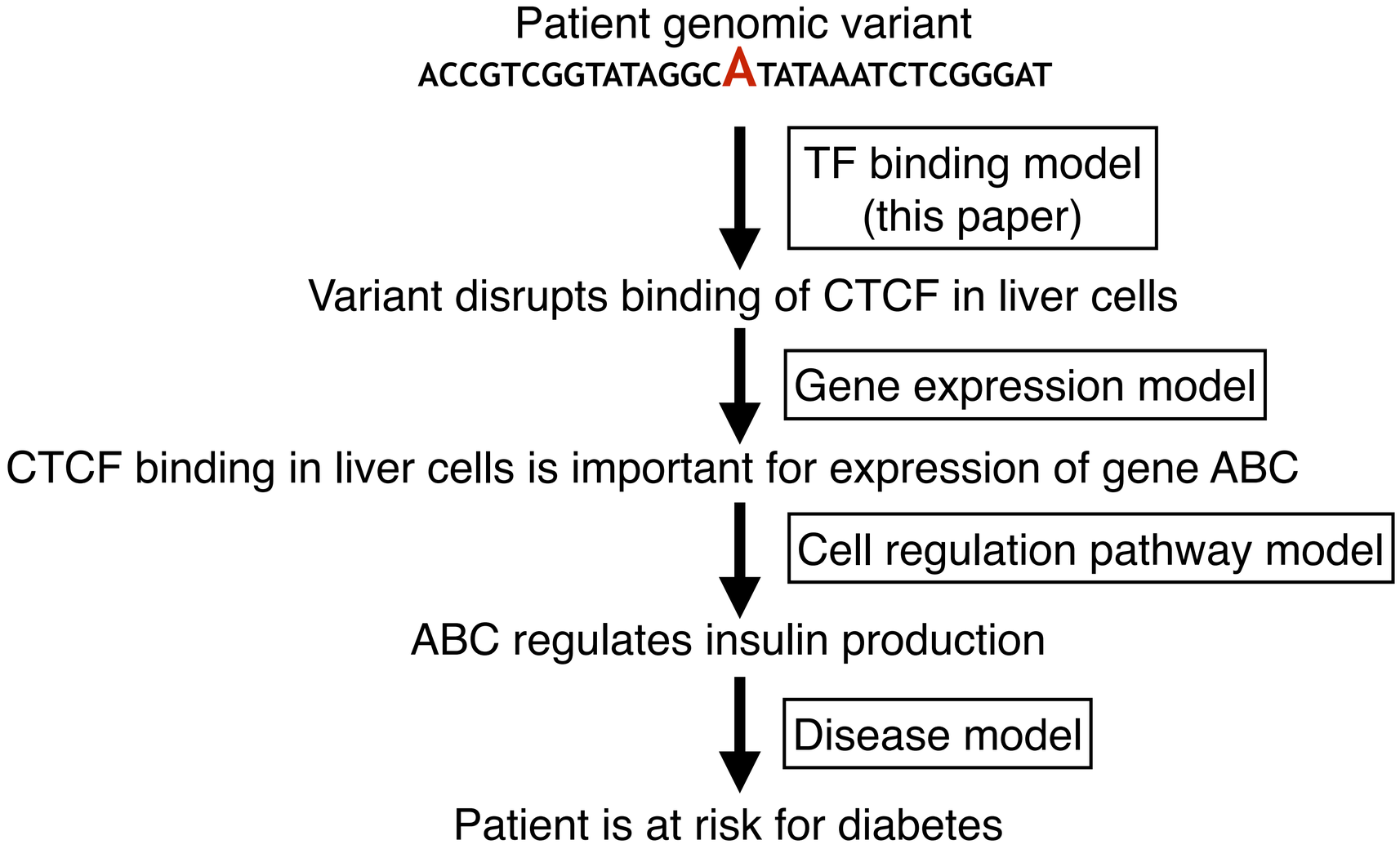
This paper focuses on predicting the effect of sequence on TF binding



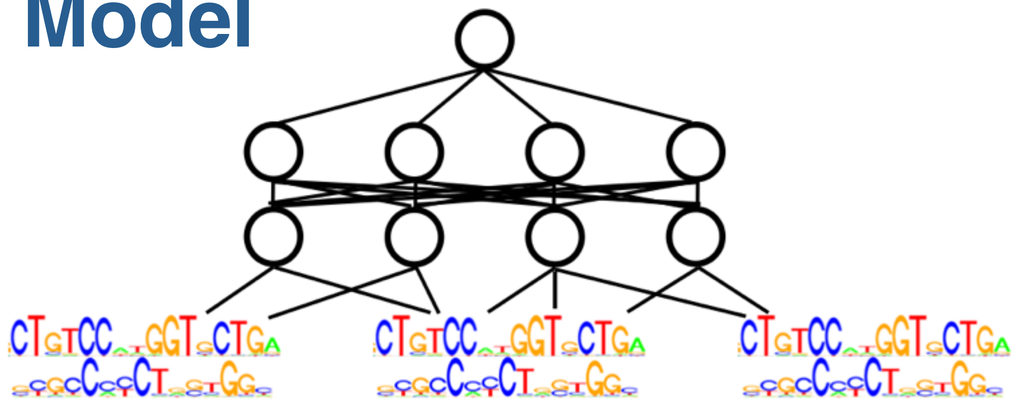
This paper focuses on predicting the effect of sequence on TF binding

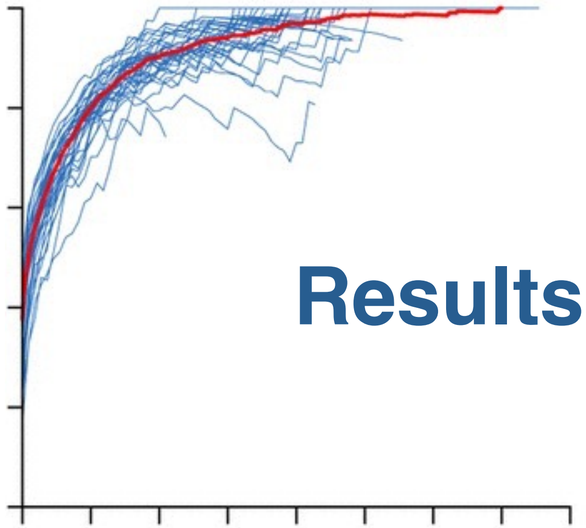


A TF binding predictor is an important step in a variant effect predictor



Model

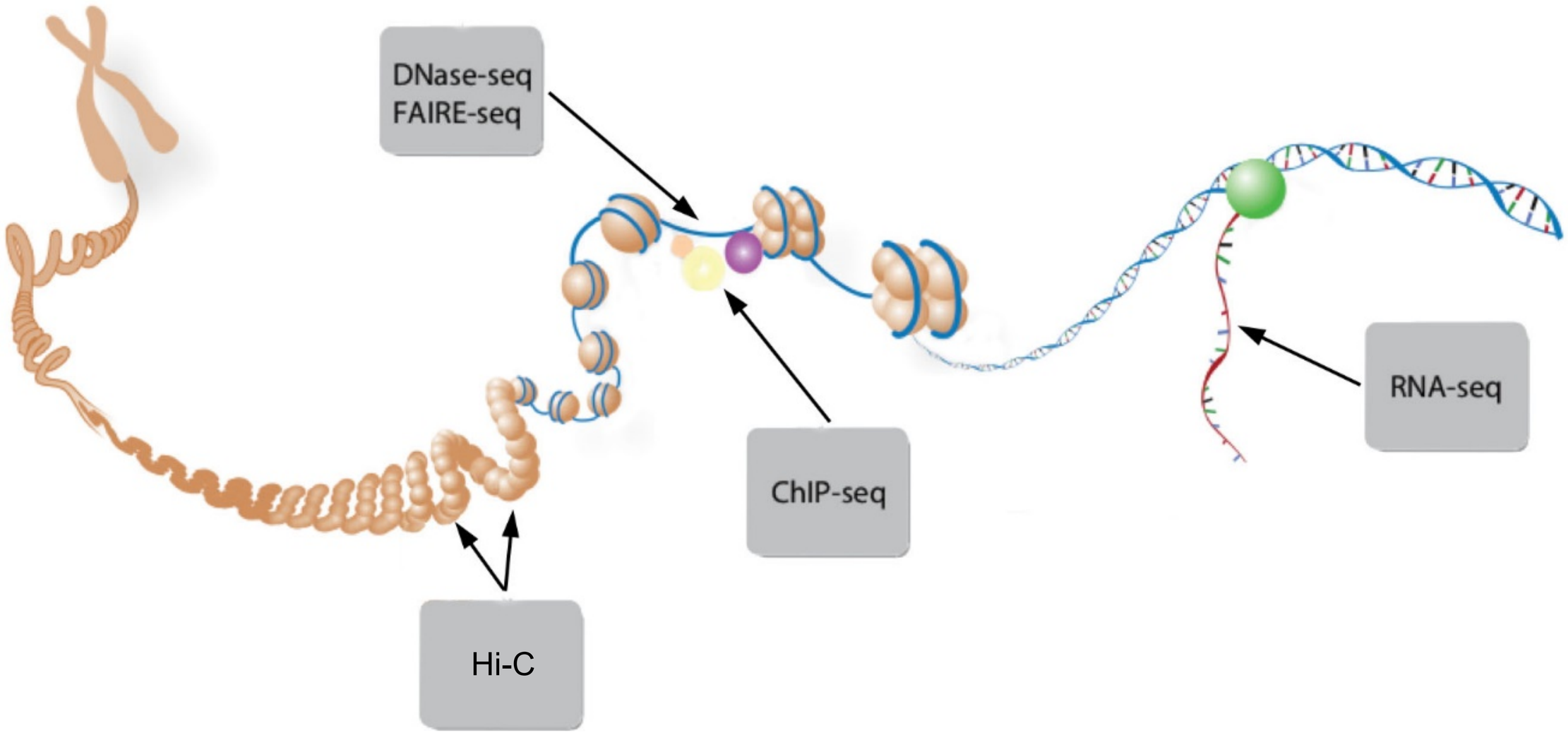




Predicting effects of noncoding variants with deep learning-based sequence model

Jian Zhou & Olga G Troyanskaya

Sequencing-based genomics assays measure many types of genomic activity



ENCODE Project Consortium 2011. *PLoS Biol* 9:e1001046

ChIP-seq measures where a given protein binds along the genome

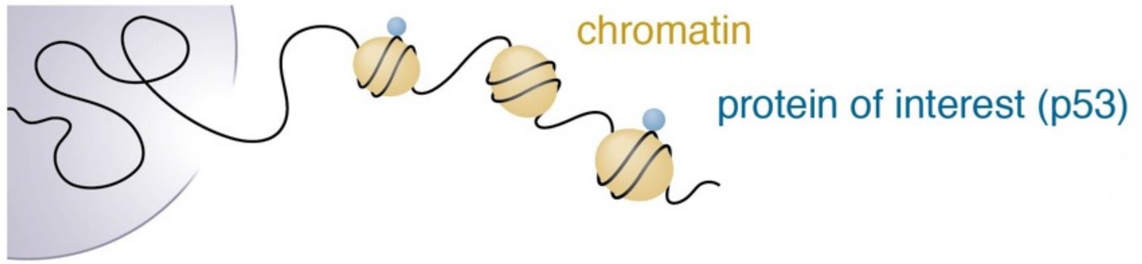


Figure derived from Brookhaven National Laboratory Newsroom 2011-11-30

ChIP-seq measures where a given protein binds along the genome

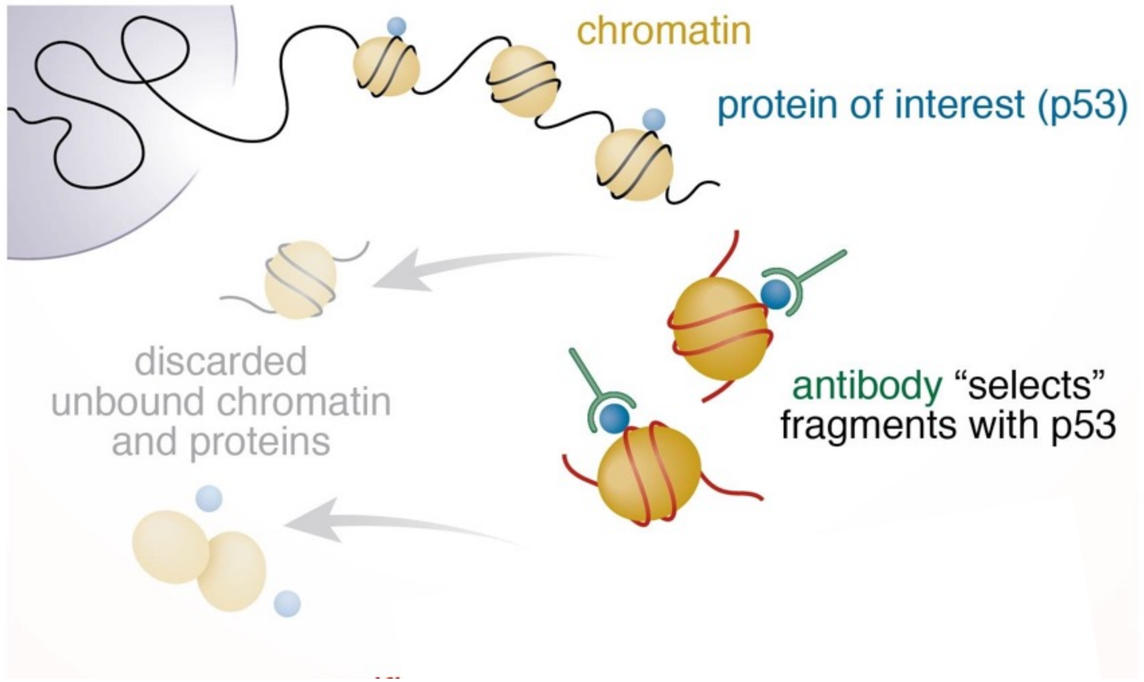


Figure derived from Brookhaven National Laboratory Newsroom 2011-11-30

ChIP-seq measures where a given protein binds along the genome

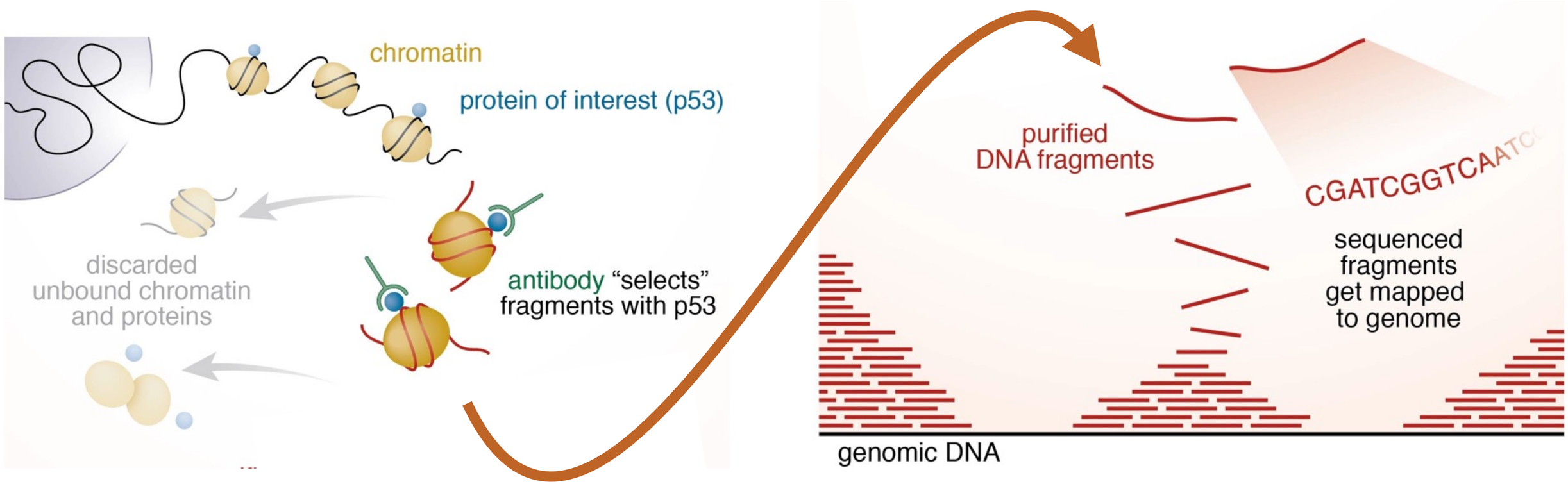


Figure derived from Brookhaven National Laboratory Newsroom 2011-11-30

ChIP-seq measures where a given protein binds along the genome

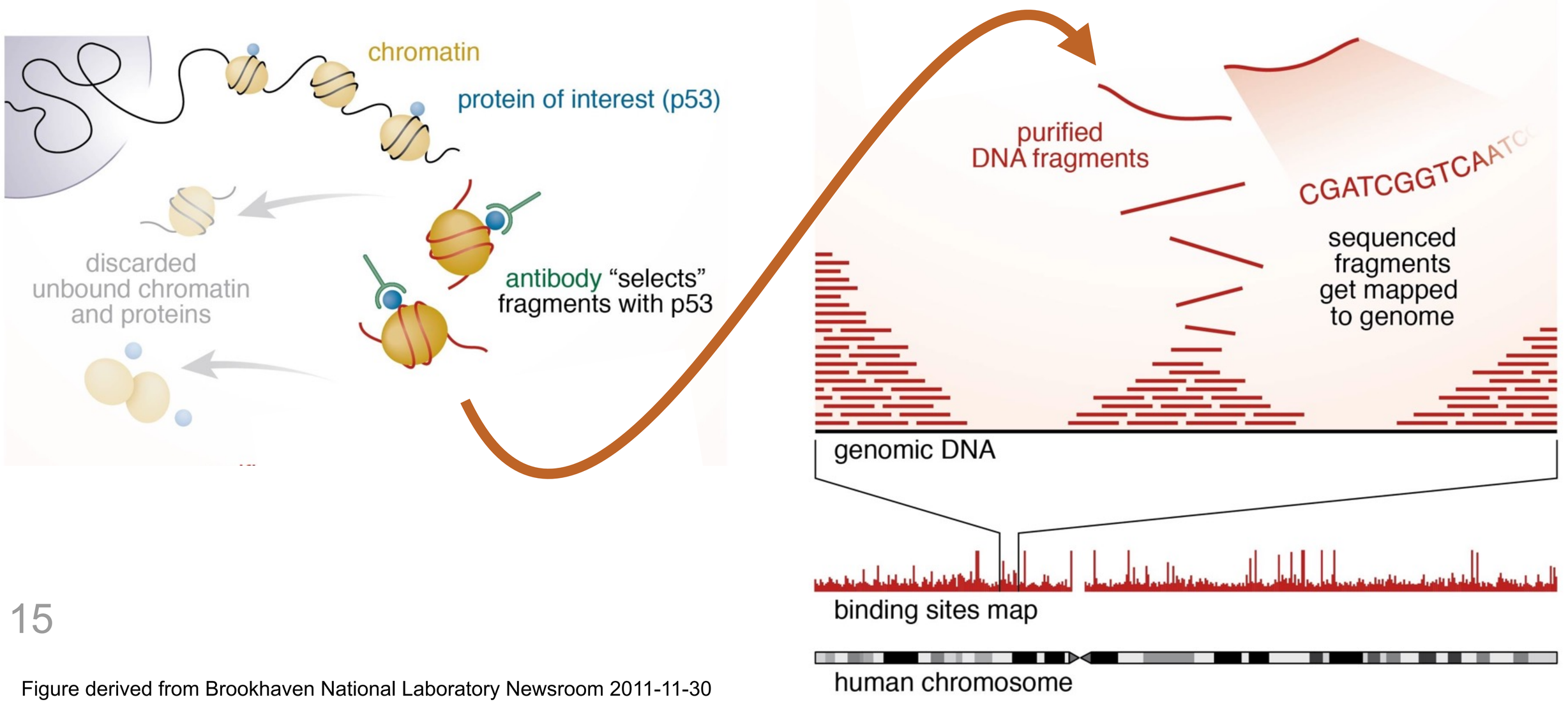
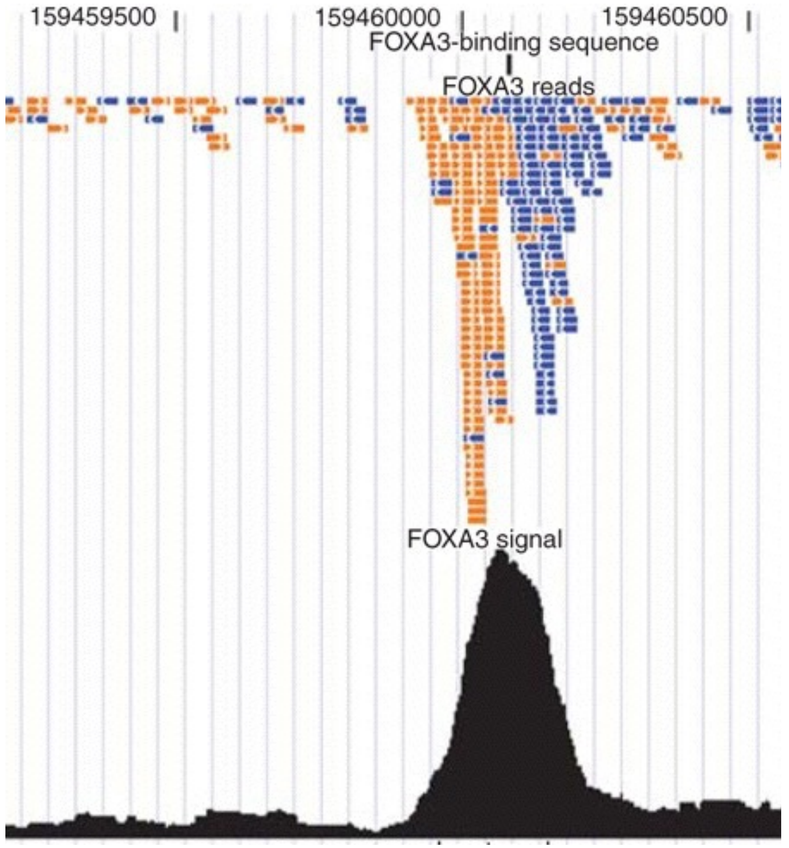


Figure derived from Brookhaven National Laboratory Newsroom 2011-11-30

ChIP-seq peak calls indicate confident transcription factor binding sites

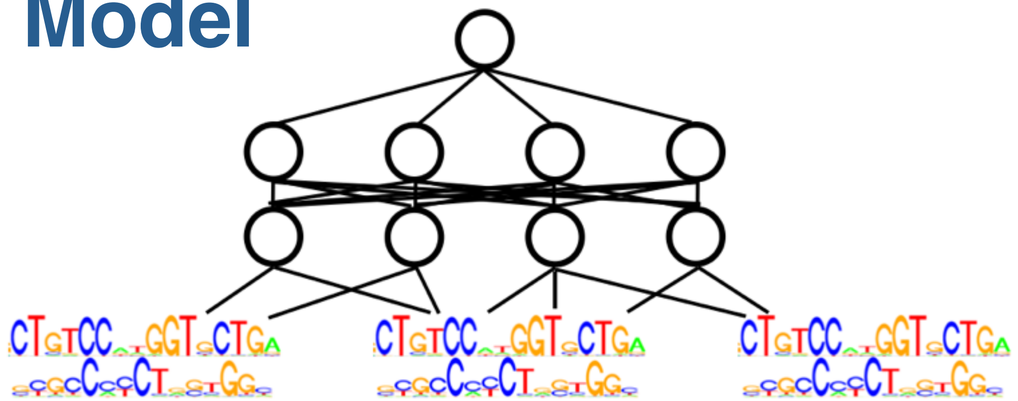
- Peak calling: Stack up the reads in the genome; choose the tall stacks.

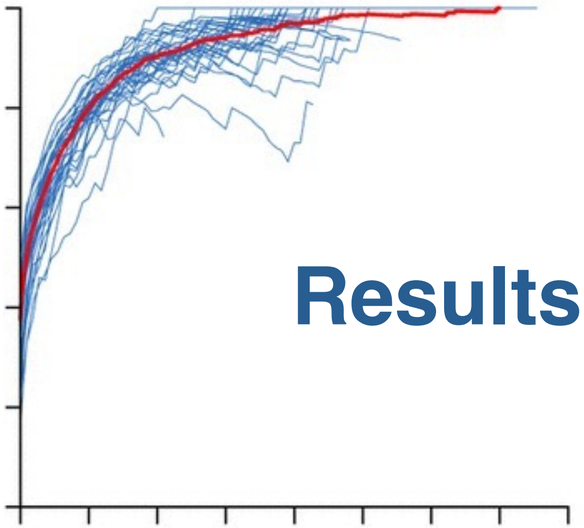
- **Issues to consider**:
  - Sequencing fragment lengths
  - Sequencing read lengths
  - Experimental biases
  - Mappability
  - GC bias
  - How to pick a threshold and assign statistical confidence



Data

Model





Results

Predicting effects of noncoding variants with deep learning-based sequence model

Jian Zhou & Olga G Troyanskaya

Problem setup

Reference sequence

ACCGTCGGTATAGGCTATAAATCTCGGGAT

CTCF binding in liver cells (ChIP-seq)

Deep neural network classifier

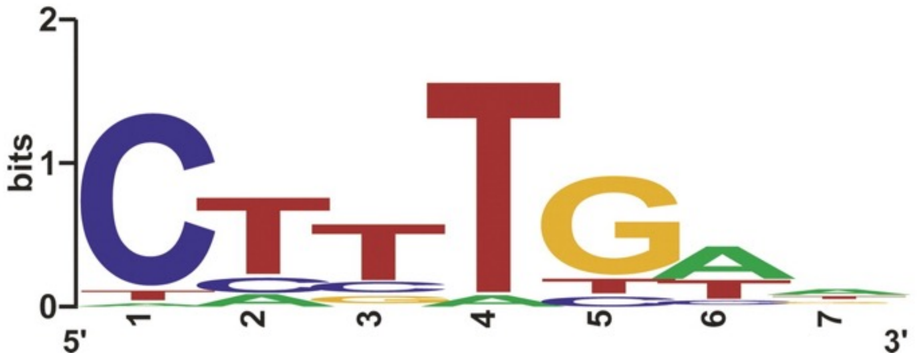
Observed variant

ACCGTCGGTATAGGC**A**TATAAATCTCGGGAT

Does variant affect CTCF binding in liver?

The traditional model for understanding transcription factor binding is the position-weight matrix (PWM)

	1	2	3	4	5	6	7
A	1	4	1	2	0	17	13
C	28	5	5	0	3	3	2
G	0	0	4	0	25	1	7
T	2	22	21	29	4	10	9

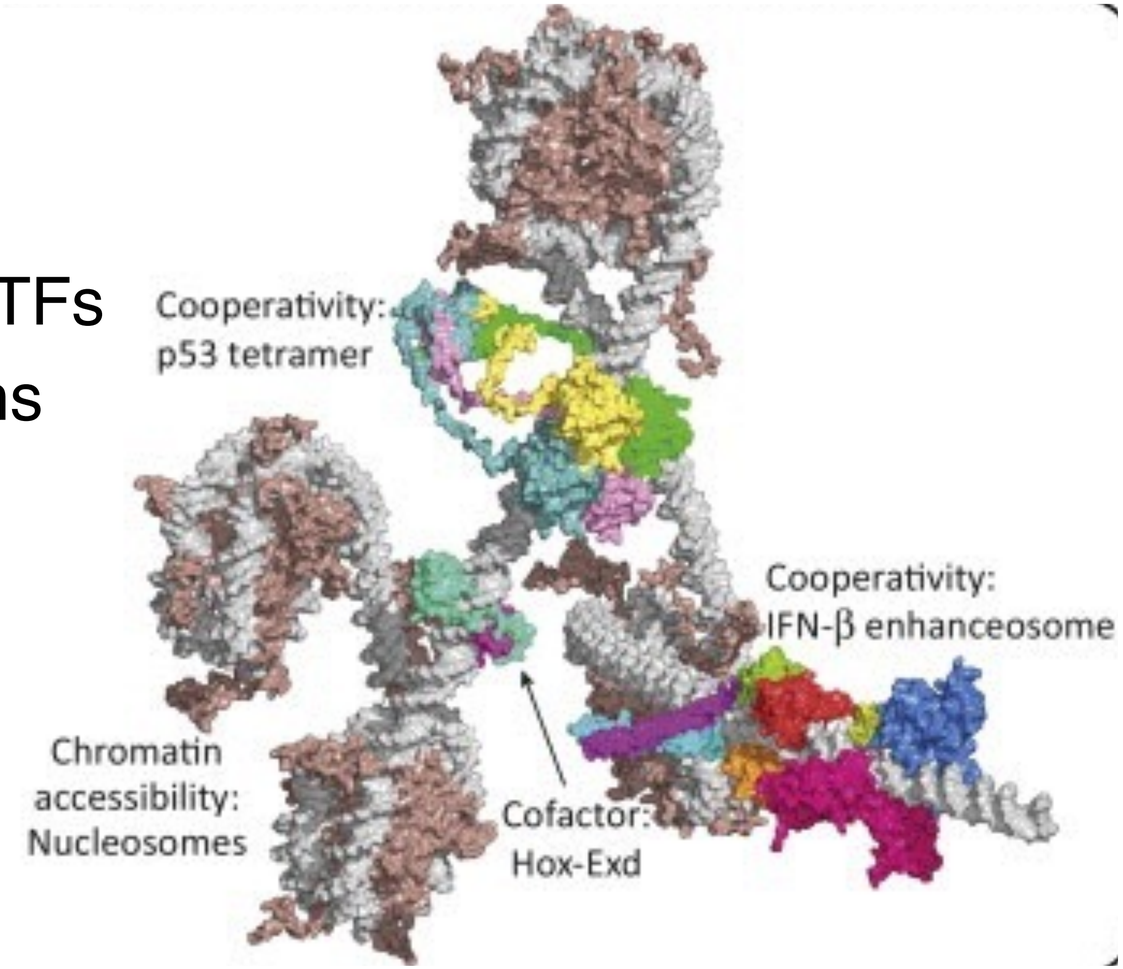


Binds?

ACCGTCGGTATAGGCTTATAAATCTCGGGAT

How can we get a better model than sequence motifs?

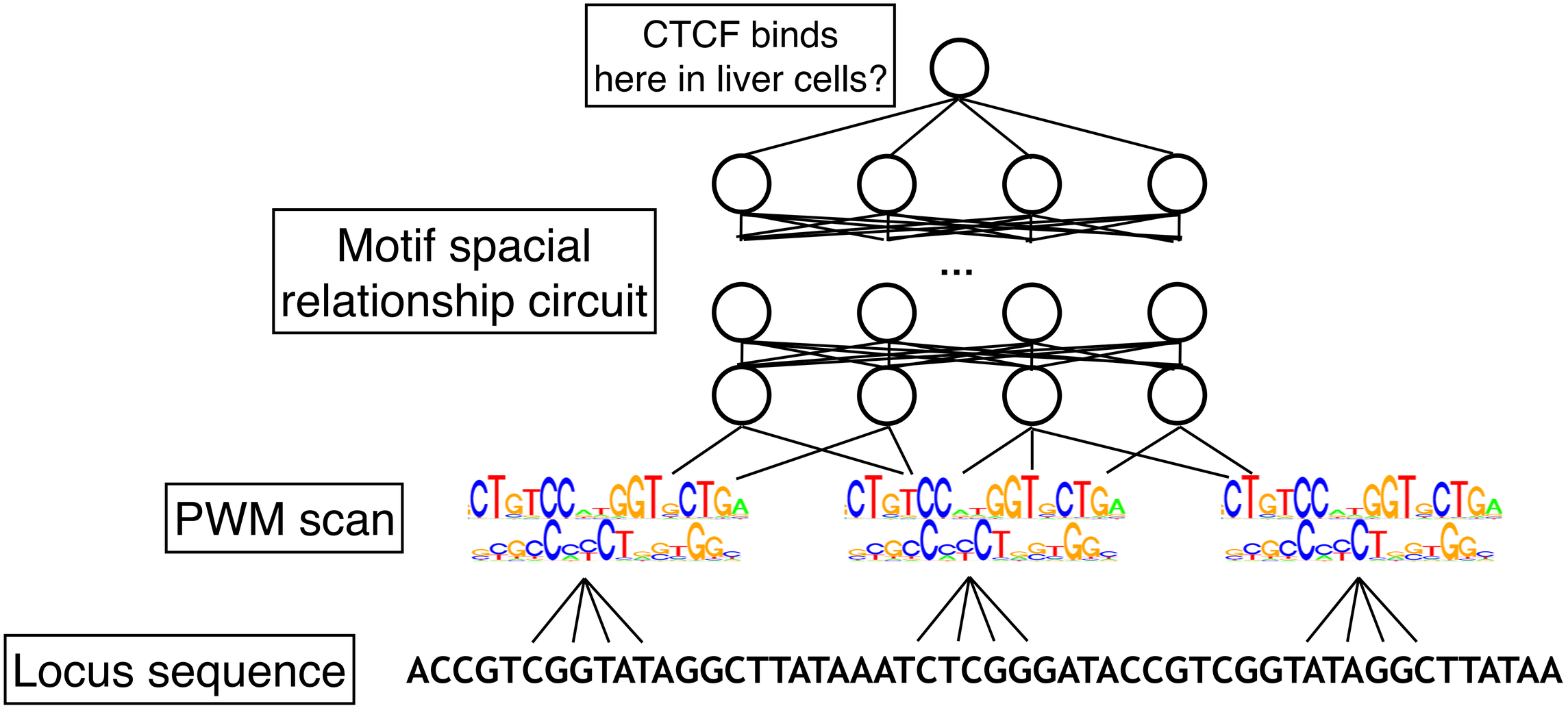
- DNA physical shape
- Variable gaps
- Cooperativity between TFs
- Nucleosome interactions



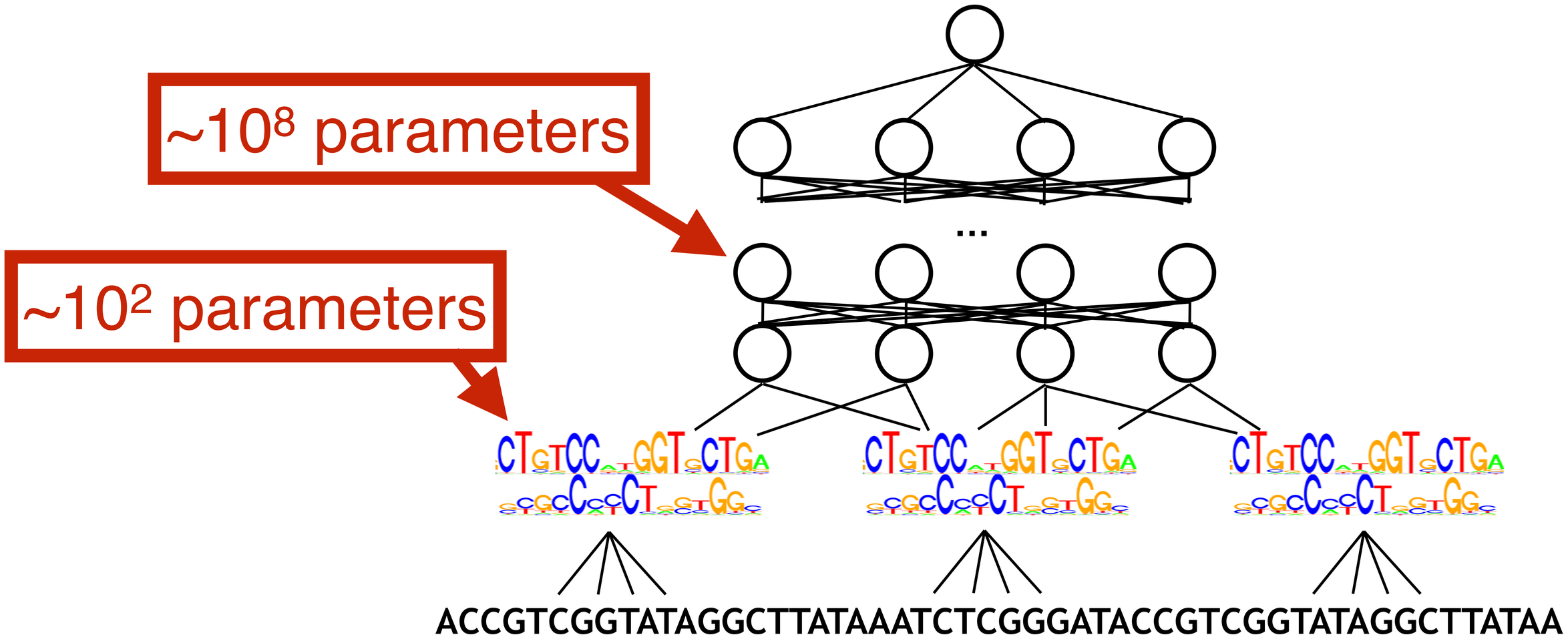
Why deep neural networks?

Deep learning is best when you have more data than sense.
— Jacob Schreiber

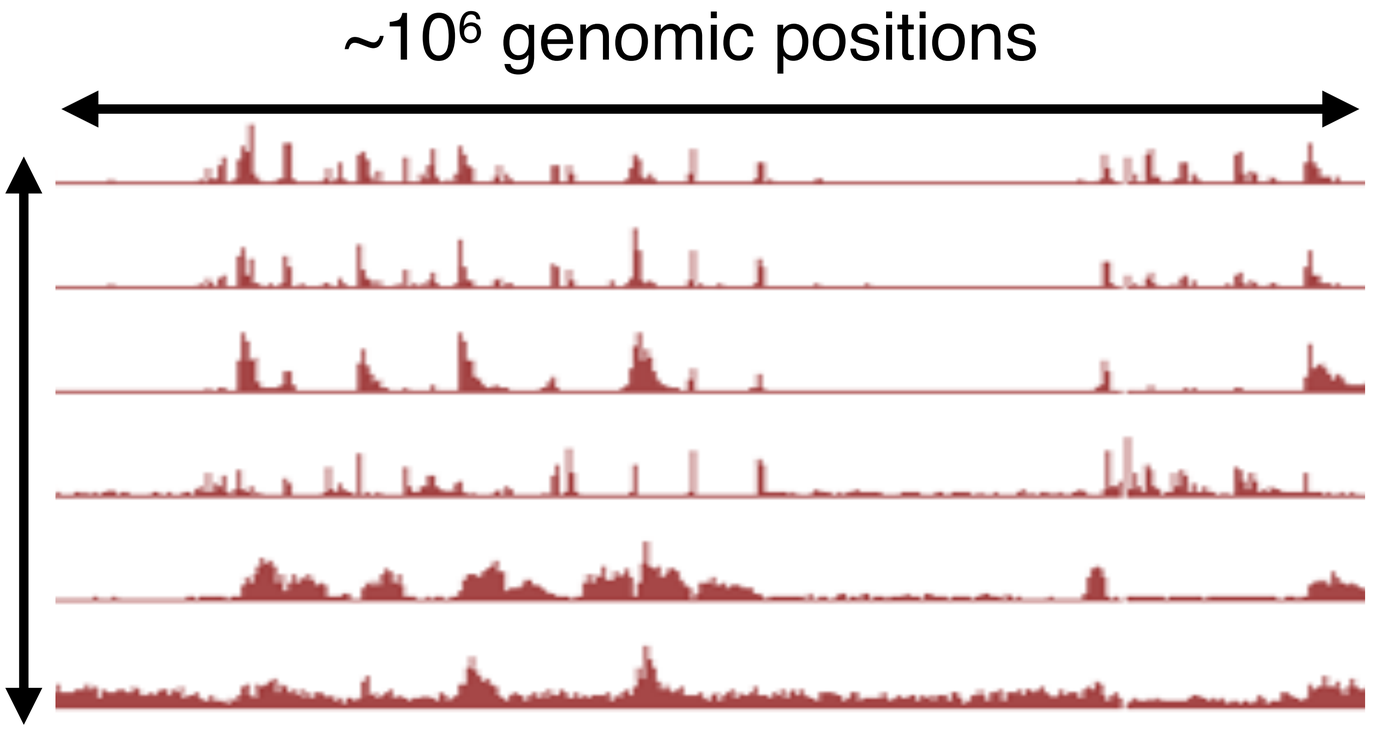
The deep neural network captures complex patterns of motif occurrence



Concern: Deep neural networks have a lot of parameters to train

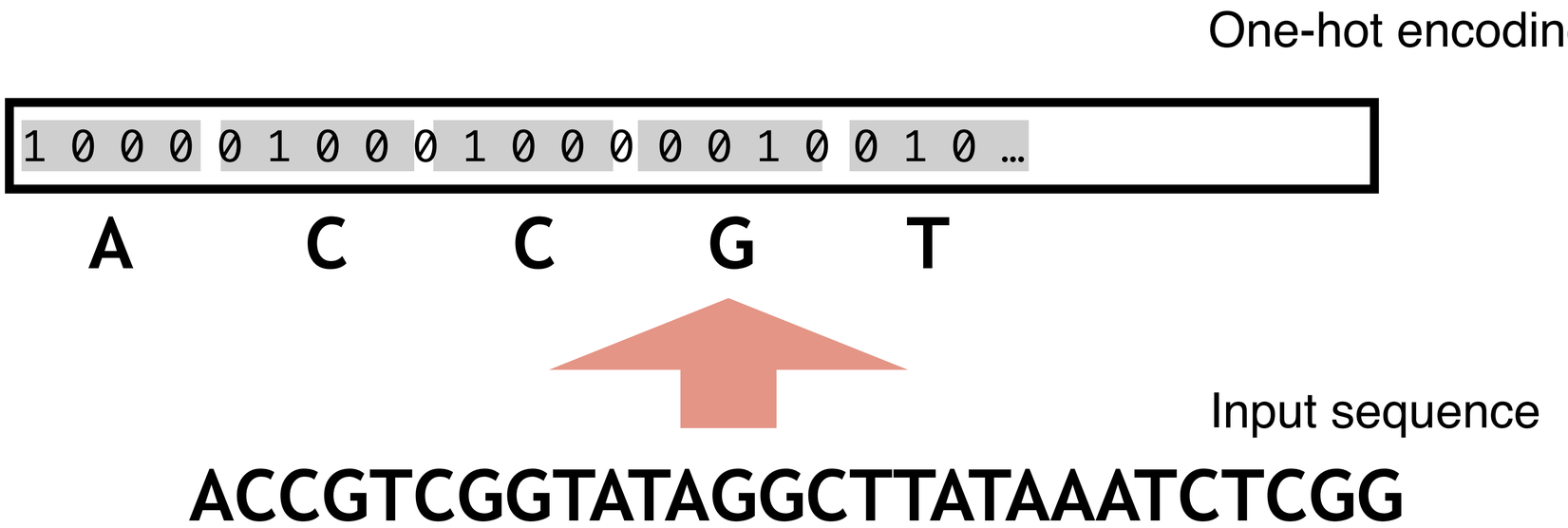


We have plenty of data to train a deep model

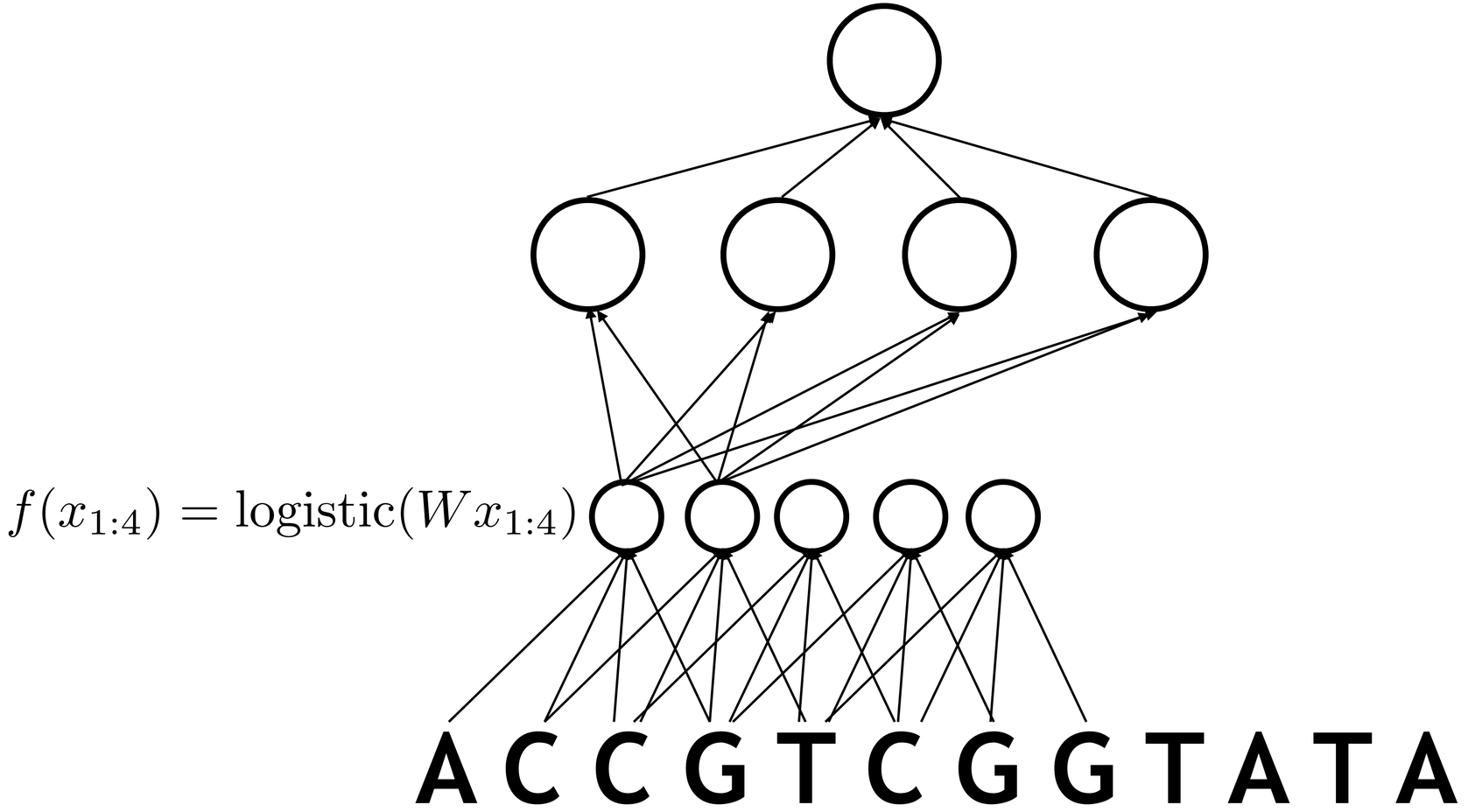


~10^9 training data points > ~10^8 neural network parameters

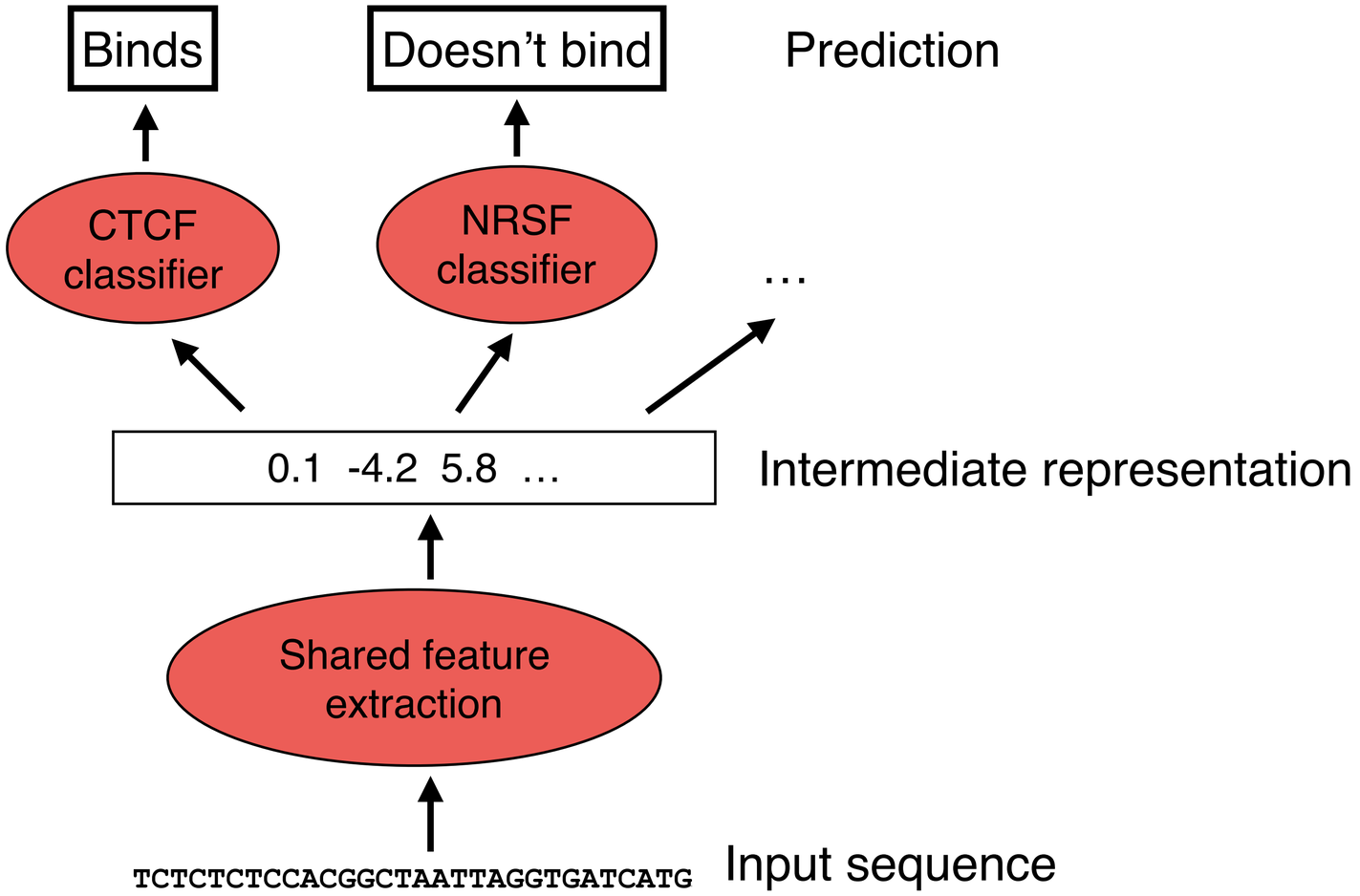
Sequence representation: one-hot encoding



A *convolutional* network reduces parameters by applying the same function across each portion of the input

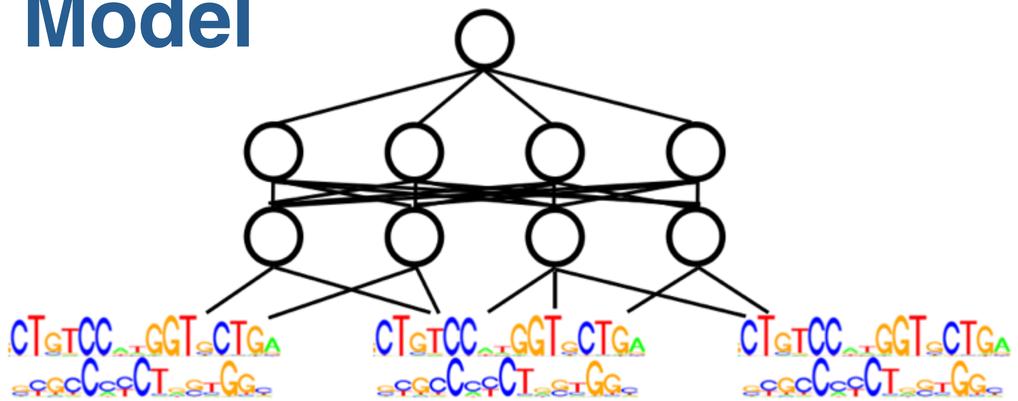


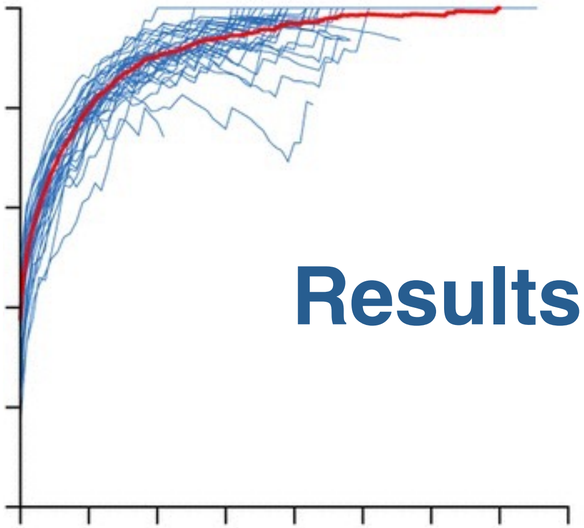
A multi-task approach shares representations between factors



Data

Model





Results

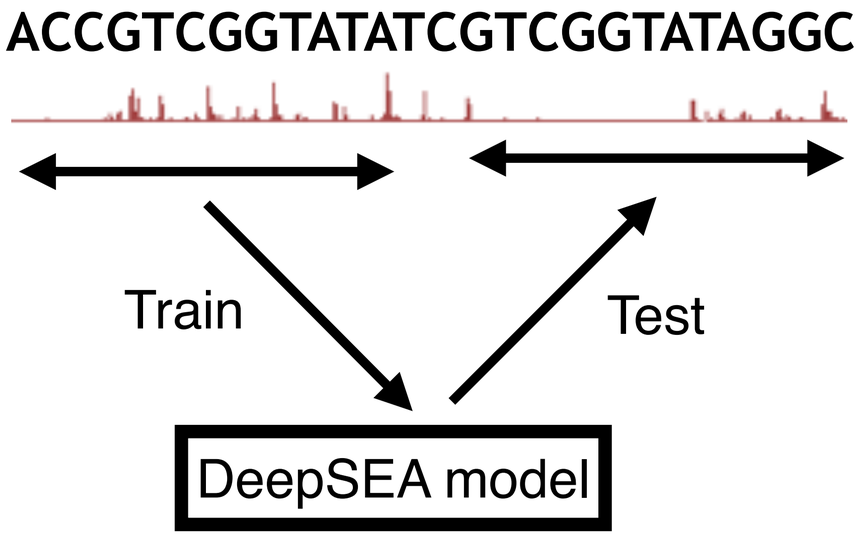
Predicting effects of noncoding variants with deep learning-based sequence model

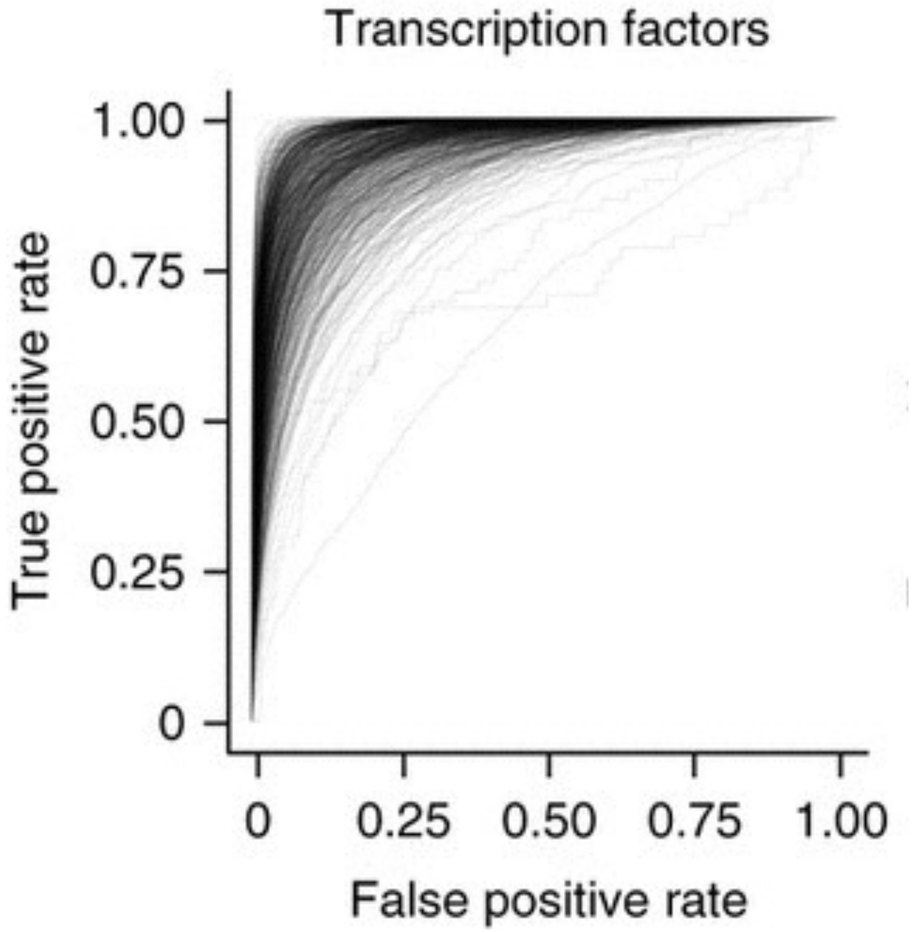
Jian Zhou & Olga G Troyanskaya

DeepSEA accurately predicts TF binding and DNase hypersensitivity

Reference sequence

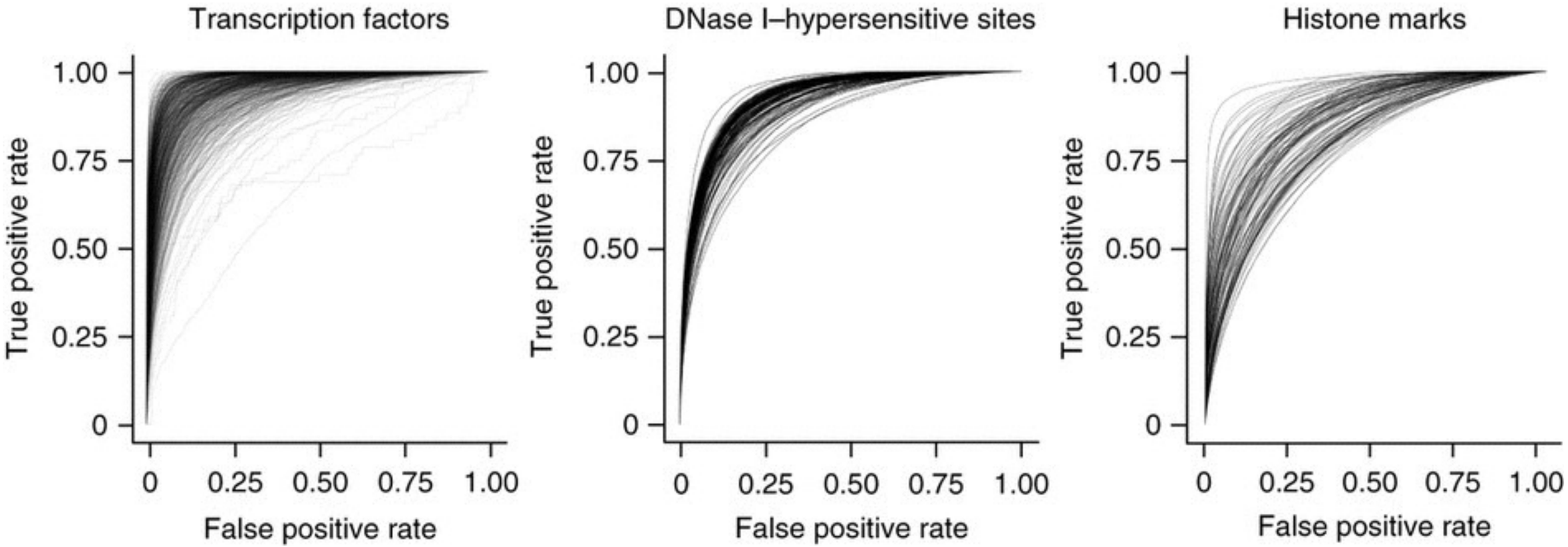
CTCF binding in liver cells (ChIP-seq)



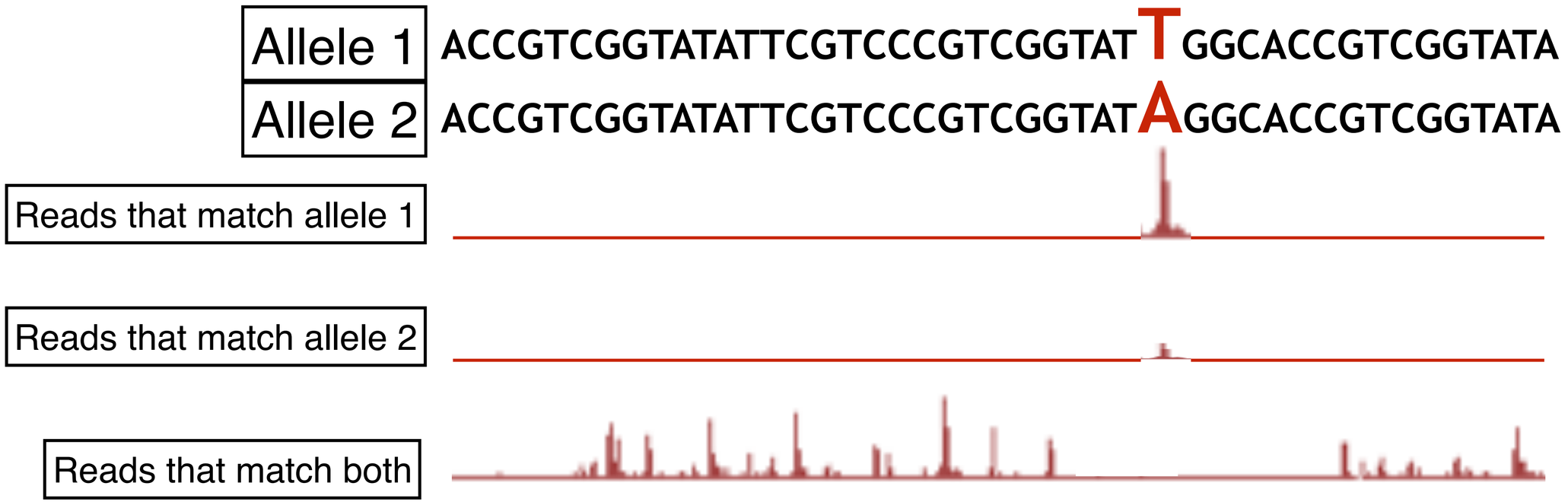


Mean AUC: 0.958

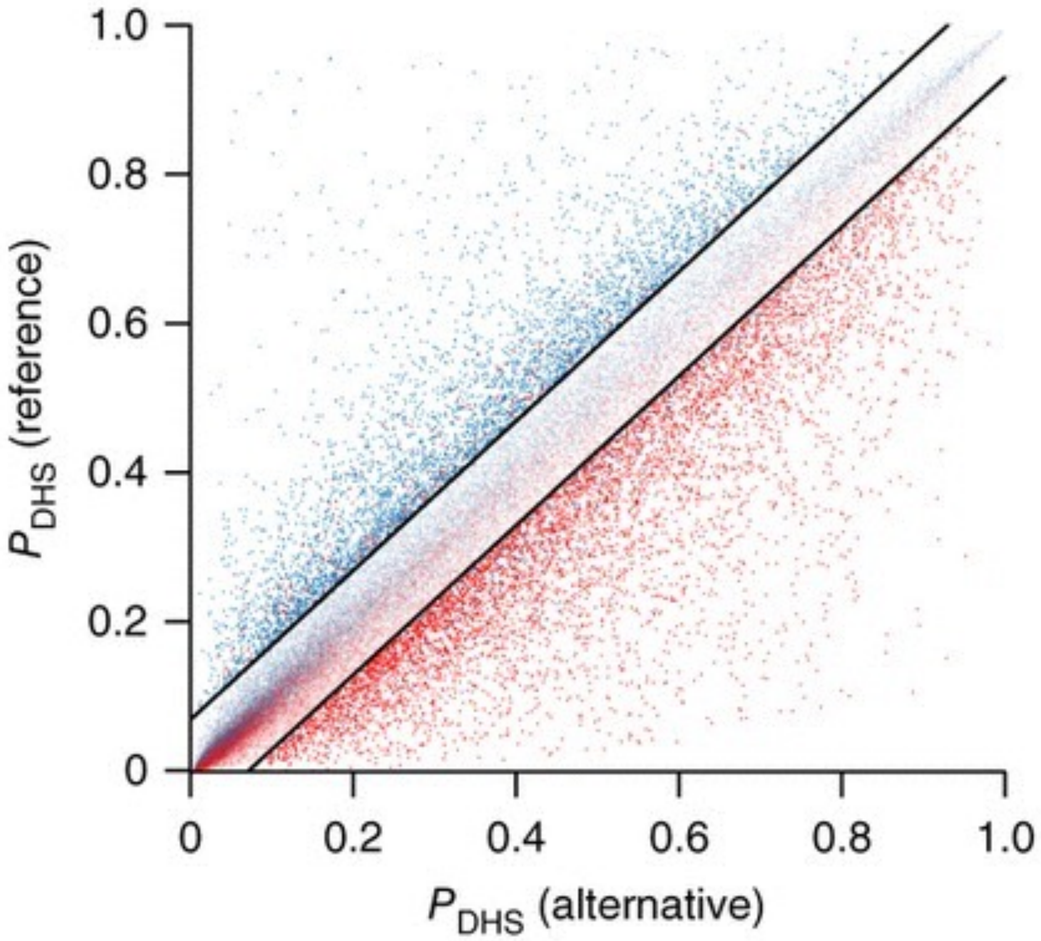
DeepSEA accurately predicts TF binding and DNase hypersensitivity



Can DeepSEA predict allelic imbalance in DNase experiments?



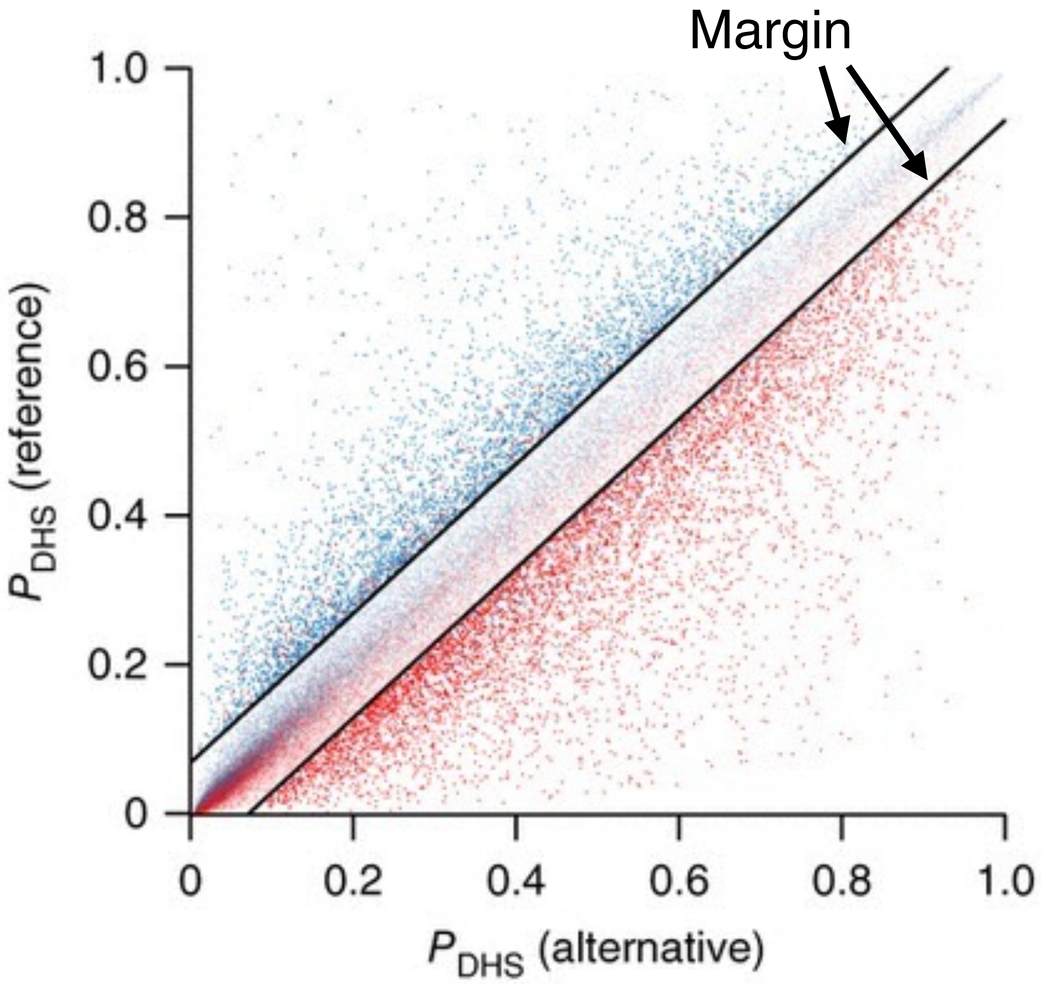
DeepSEA accurately predicts allelic imbalance in DNase experiments

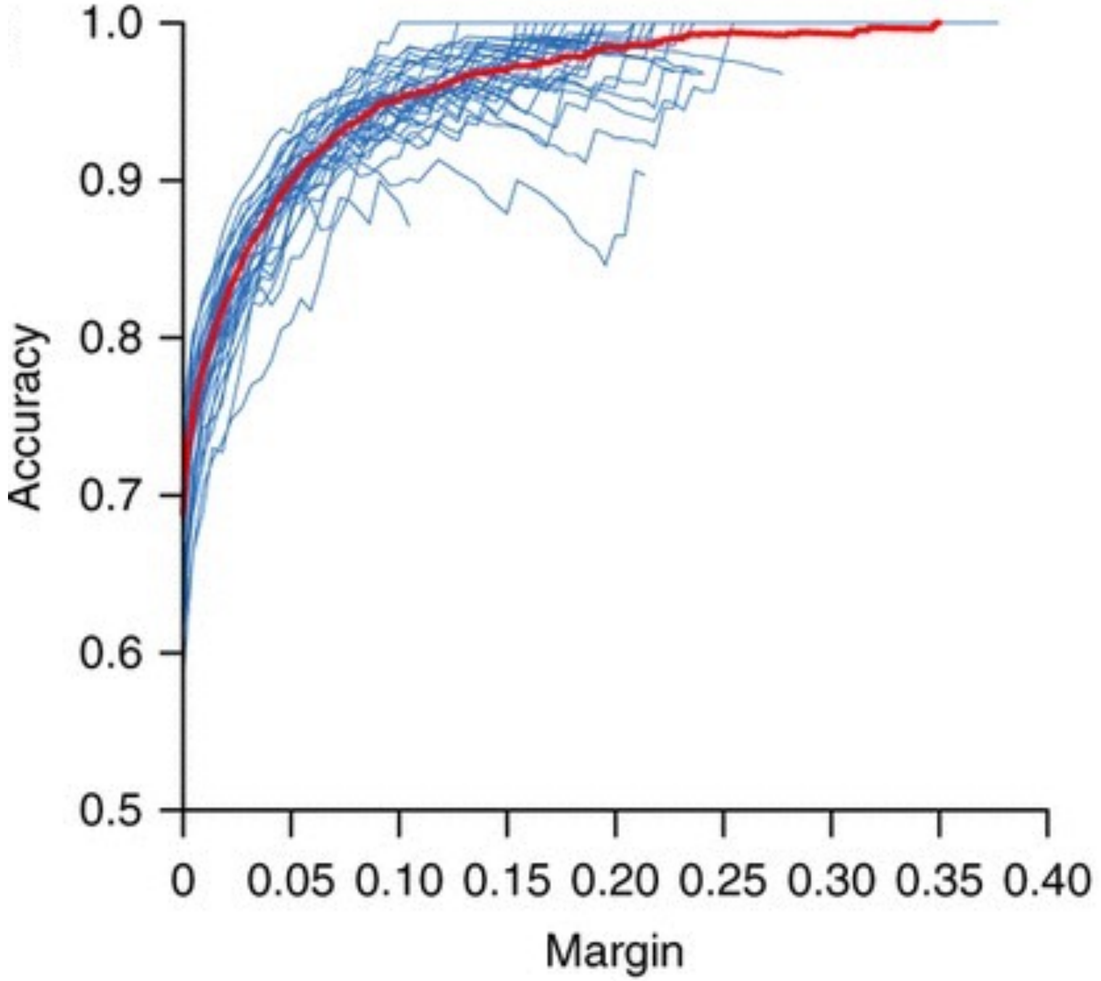


Observed bias for alternative

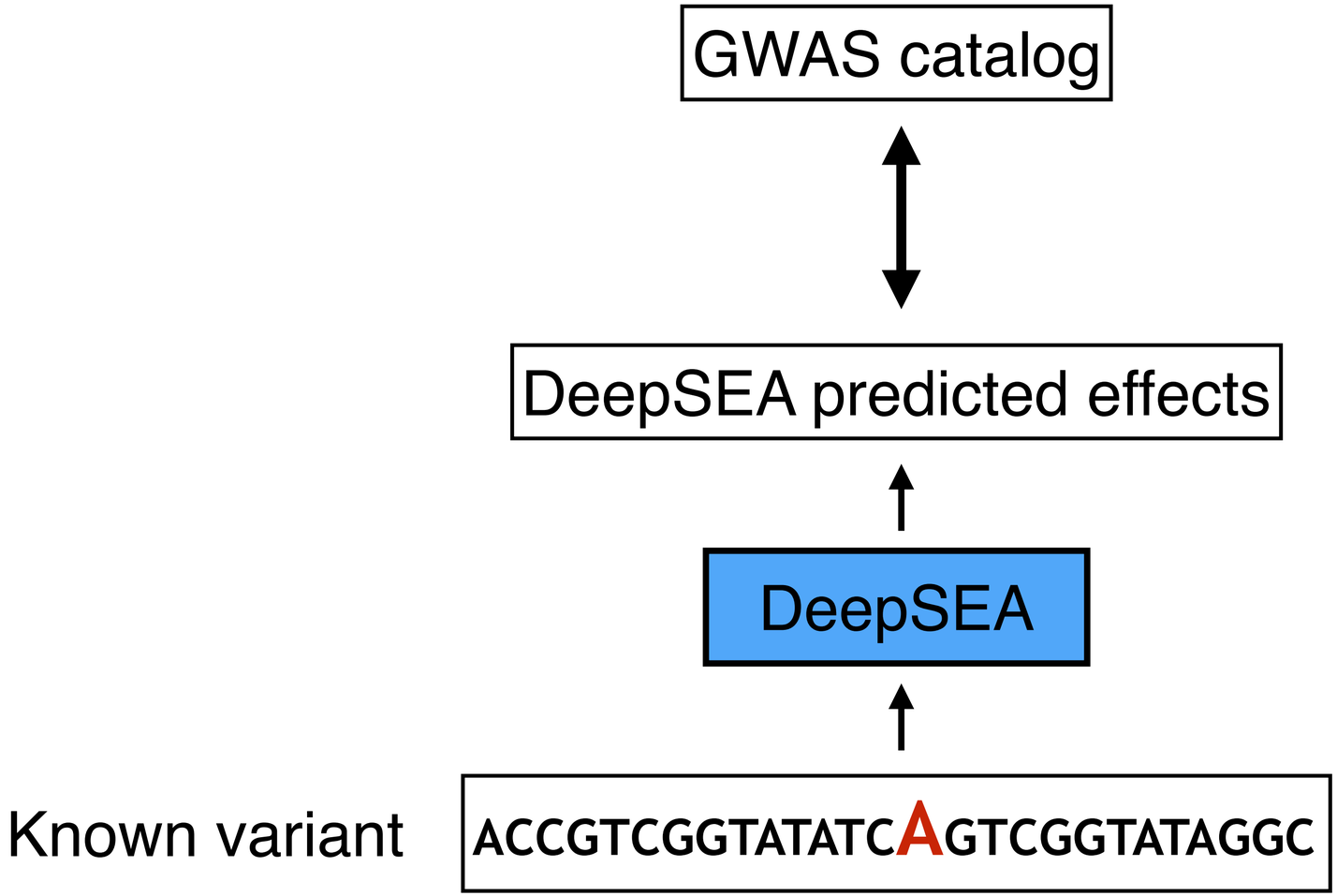
Observed bias for reference

DeepSEA accurately predicts allelic imbalance in DNase experiments

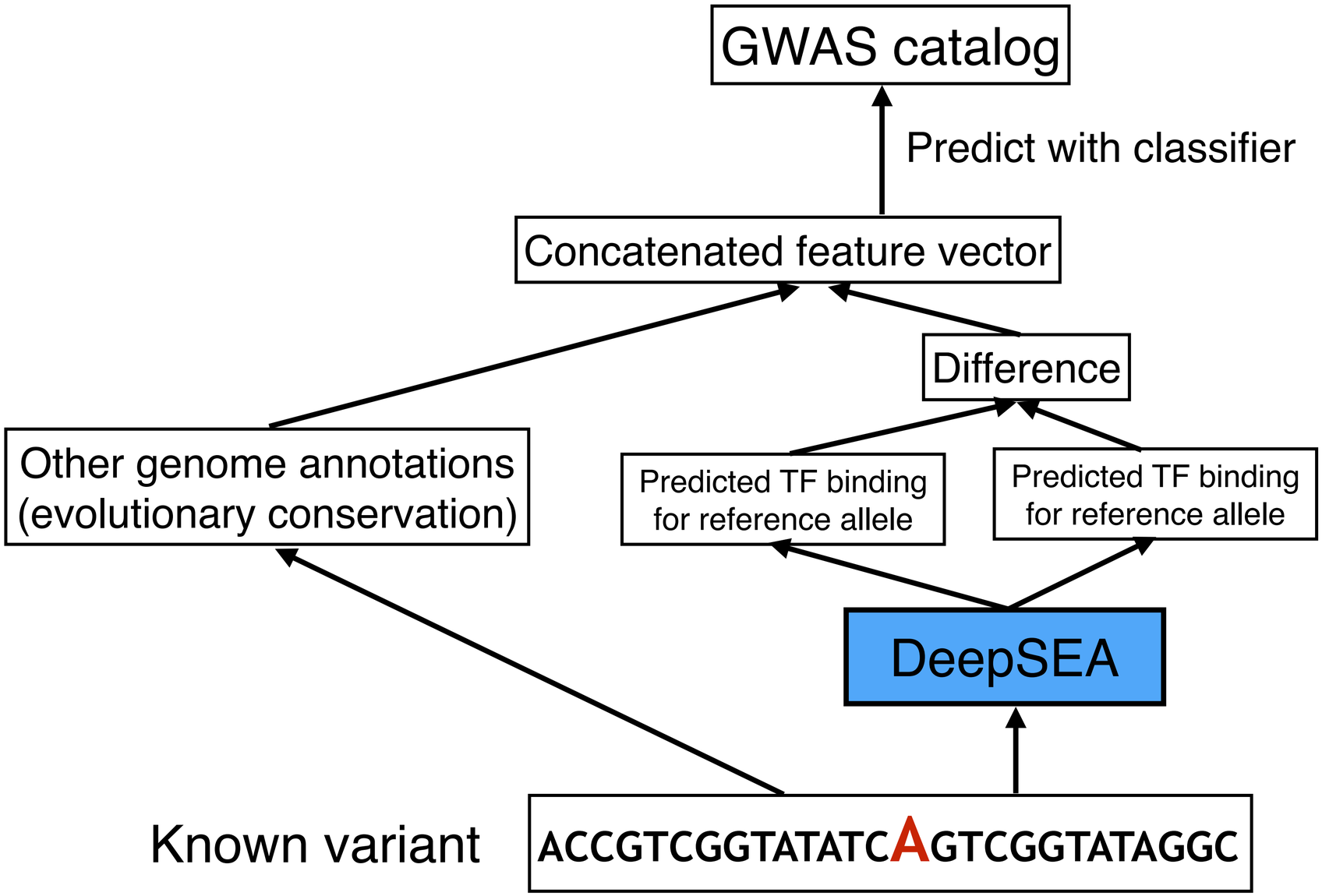




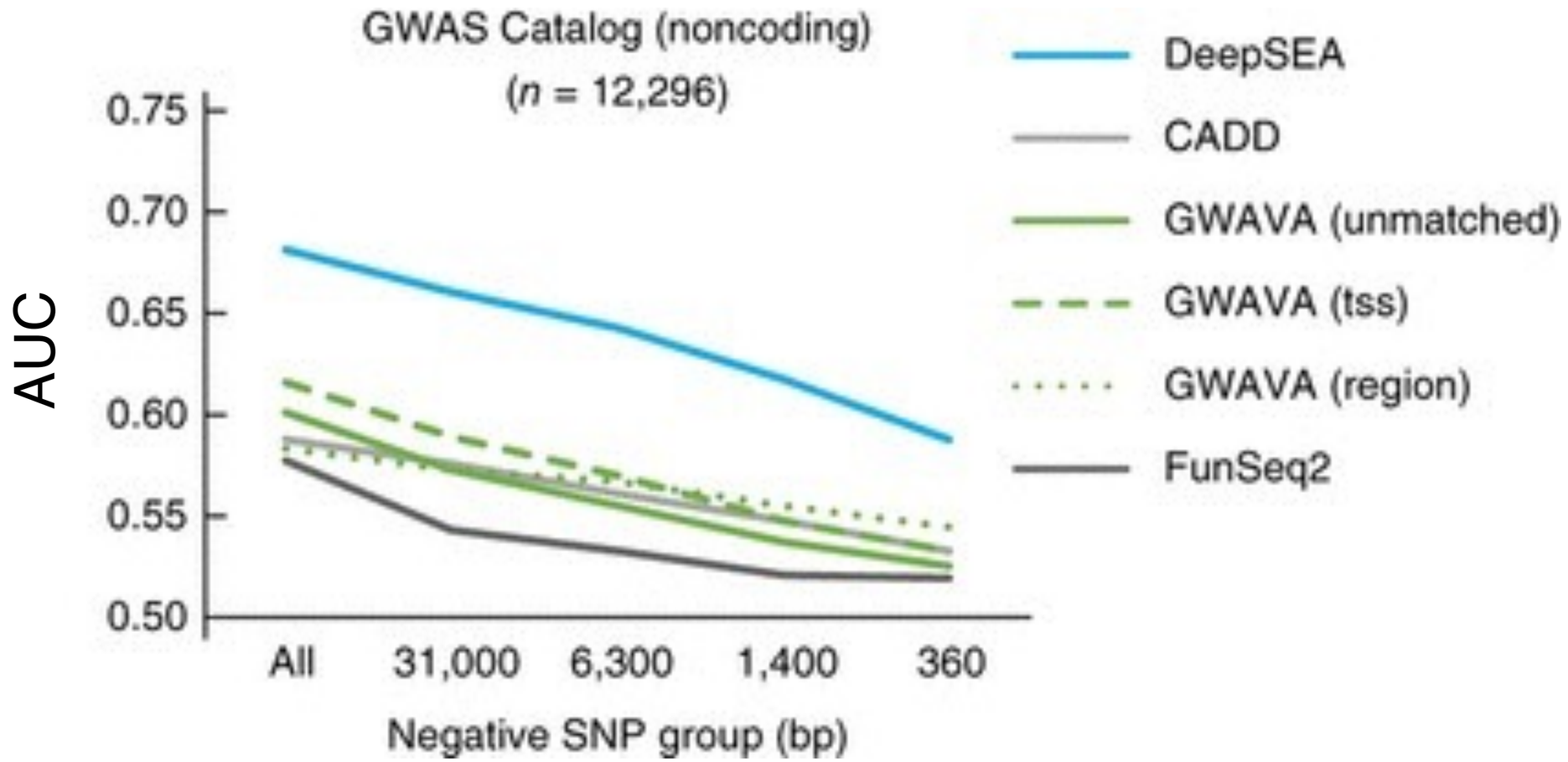
Can DeepSEA predict known regulatory variants?



Can DeepSEA predict known regulatory variants?

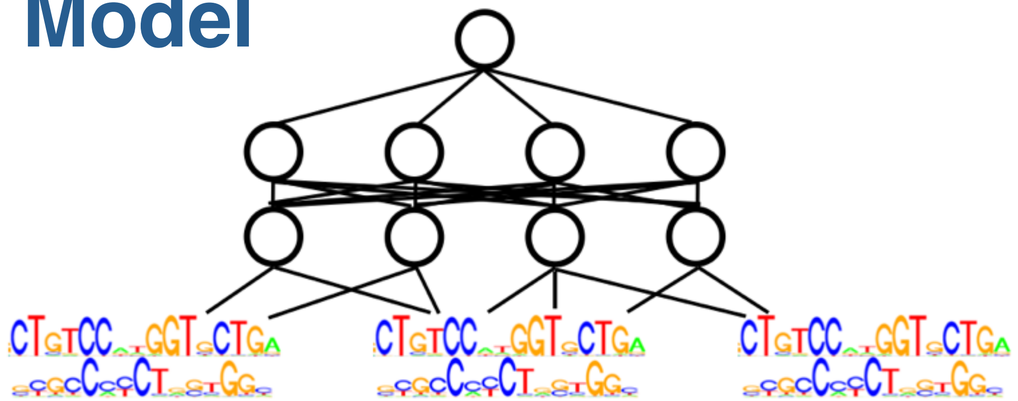


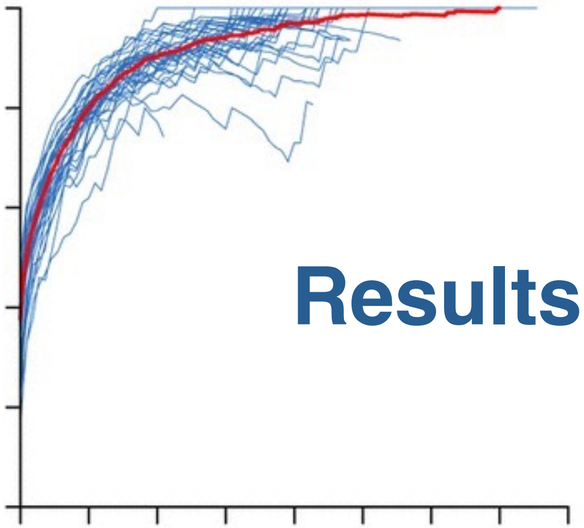
DeepSEA accurately predicts known regulatory variants



Data

Model





Results

Predicting effects of noncoding variants with deep learning-based sequence model

Jian Zhou & Olga G Troyanskaya

Supplementary Note. DeepSEA model configuration

Model Architecture:

1. Convolution layer (320 kernels. Window size: 8. Step size: 1.)
2. Pooling layer (Window size: 4. Step size: 4.)
3. Convolution layer (480 kernels. Window size: 8. Step size: 1.)
4. Pooling layer (Window size: 4. Step size: 4.)
5. Convolution layer (960 kernels. Window size: 8. Step size: 1.)
6. Fully connected layer (925 neurons)
7. Sigmoid output layer

Regularization Parameters:

Dropout proportion (proportion of outputs randomly set to 0):

- Layer 2: 20%
- Layer 4: 20%
- Layer 5: 50%
- All other layers: 0%

L2 regularization (λ_1): 5e-07
L1 sparsity (λ_2): 1e-08
Max kernel norm (λ_3): 0.9