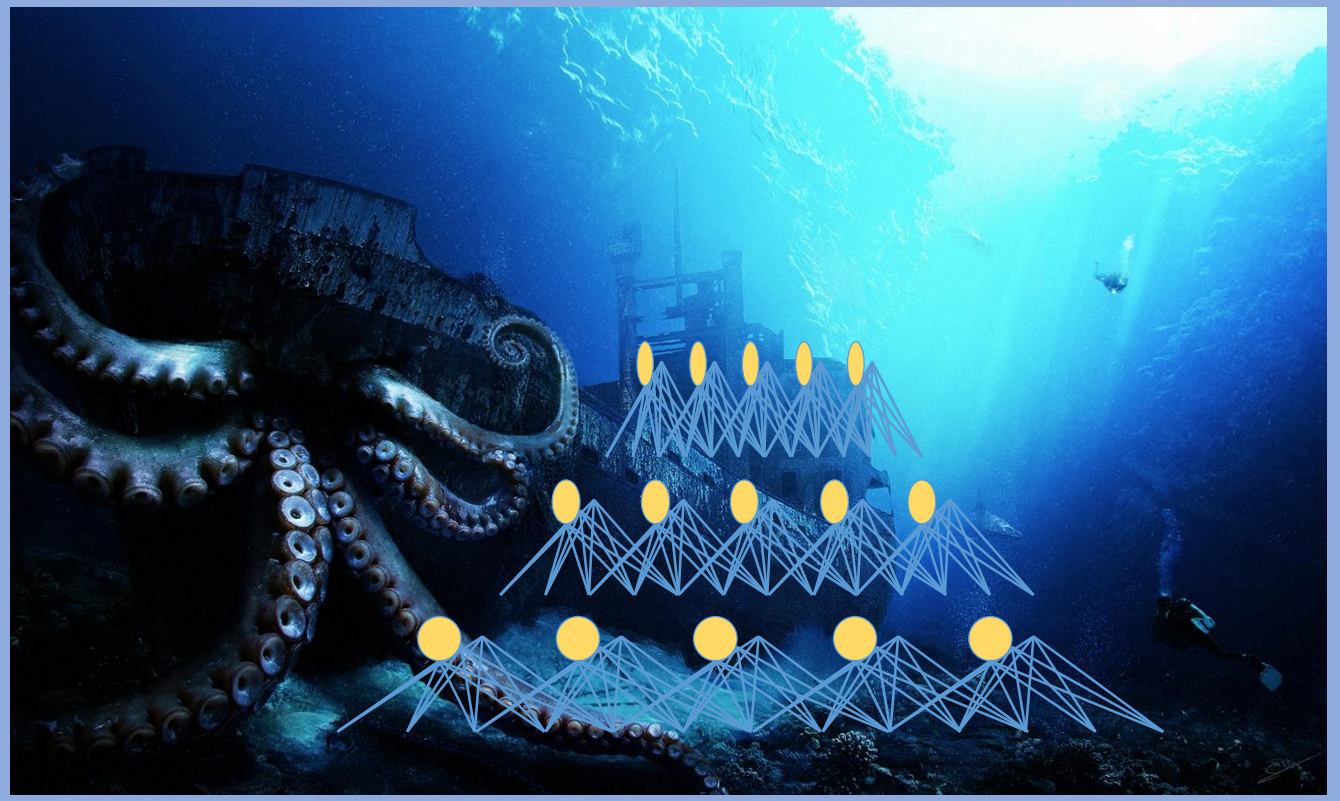
DeepSEA – Zhou & Troyanskaya

predicting the effects of non coding variants on chromatin and regulation

CS231B journal club



Arbel Harpak & Ziyue Gao

DeepSEA – Zhou & Troyanskaya

predicting the effects of non coding variants on chromatin and regulation

Single noncoding SNP

↓

Regulatory effect

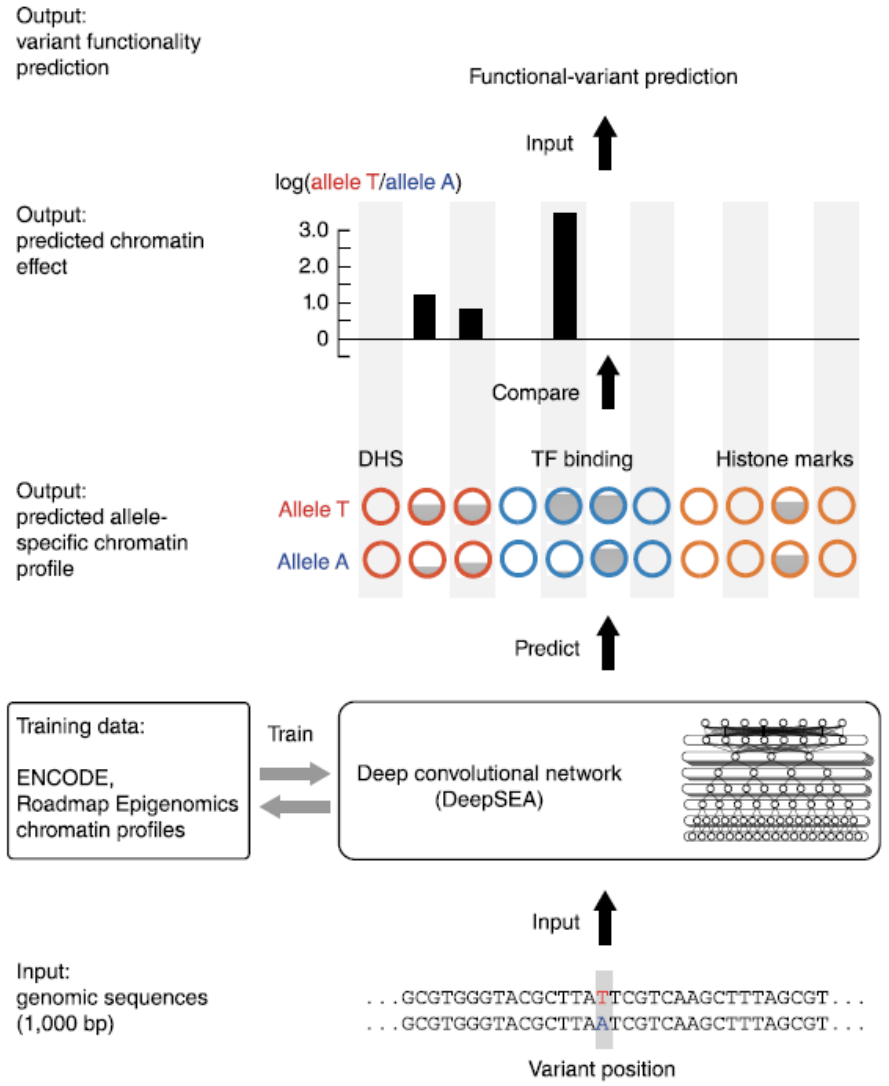
↓

Function

Setup: multitask prediction of regulatory effects

First wave of genomics CNN papers, Introducing:

- Integrating sequence from wide context
- Learn at multiple spatial scale with hierarchical architecture
- Joint learning of diverse chromatin factors sharing predictive features
  (690 TF profiles, 125 DHS, 104 histone mark)

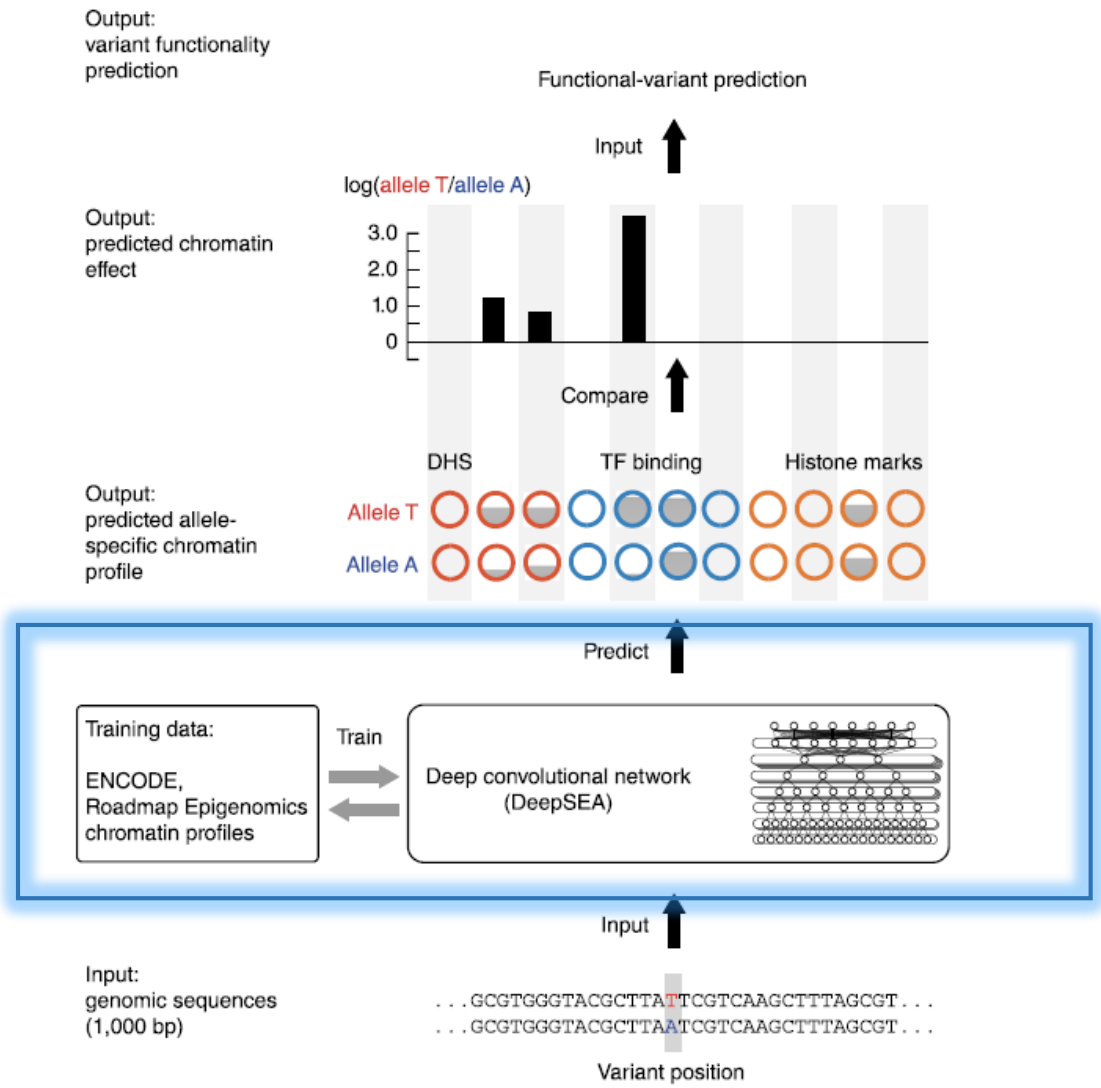


Part of first wave of CNN for genomics

Introducing:

- Integration of sequence from wide context
- Learning at multiple spatial scale with hierarchical architecture
- Joint learning of diverse chromatin factors sharing predictive features

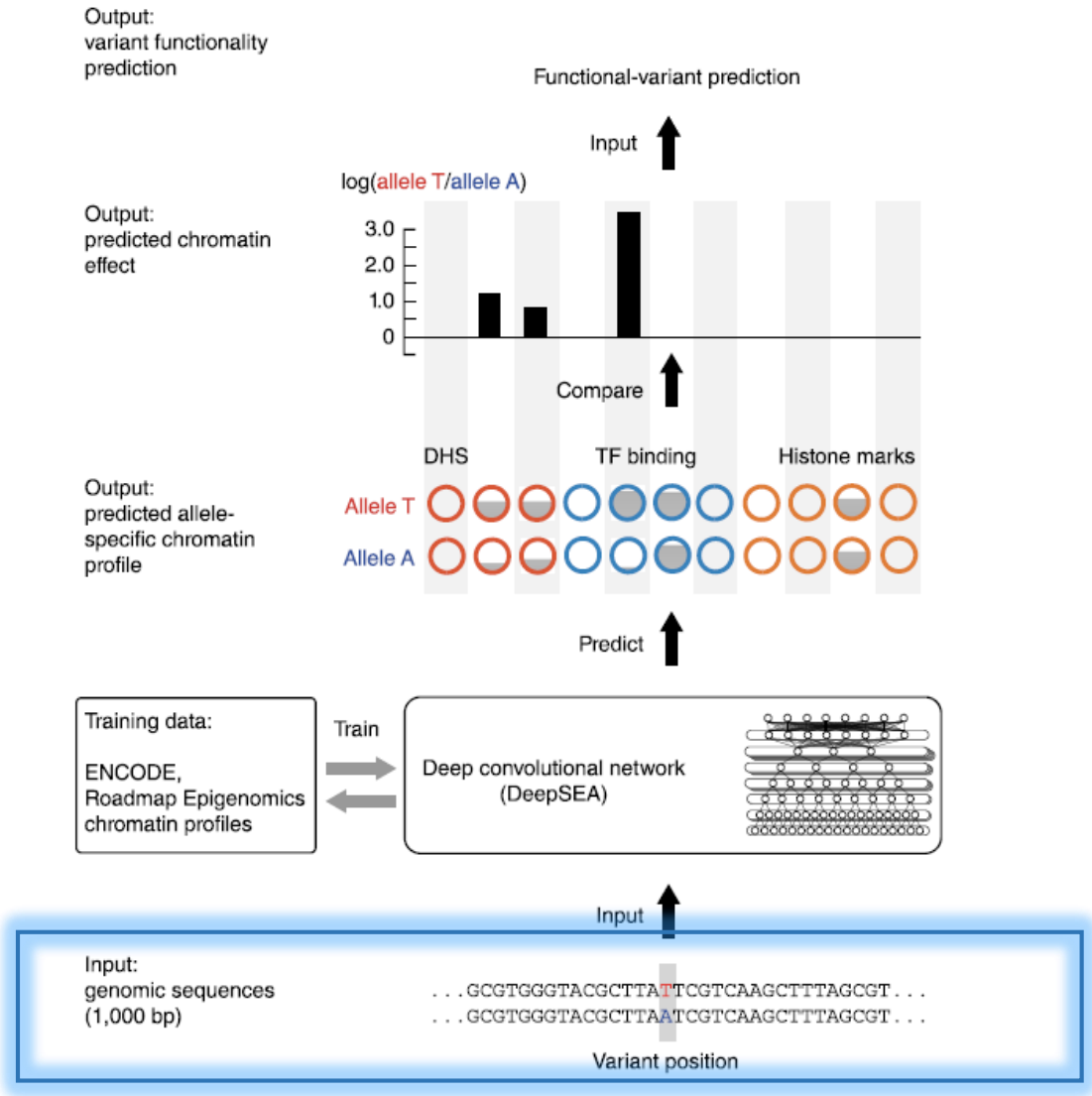
Is multitasking a good idea?



Sequence-only → factor peaks → Overall functional effect

Data:

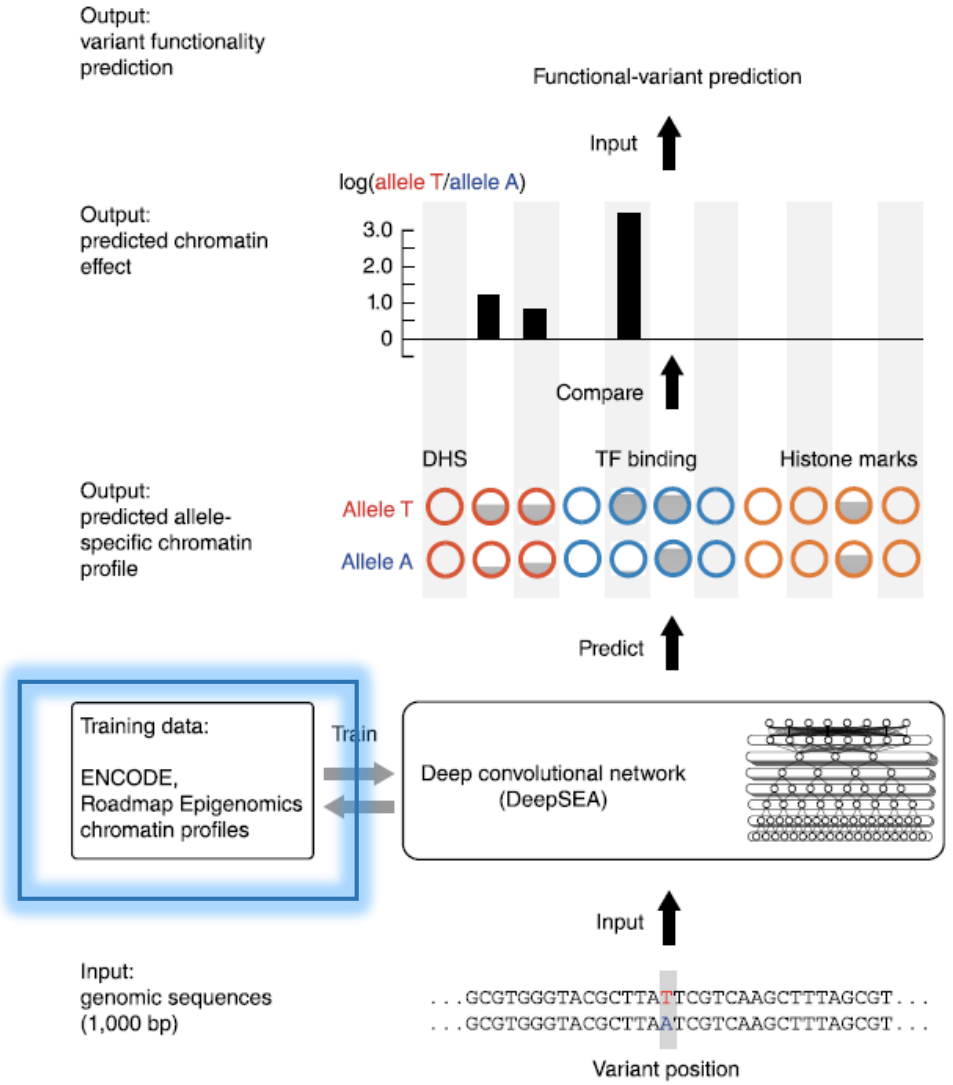
- Samples: stride of 1bp along the genome
- Input features: 1000bp one-hot, reference genome



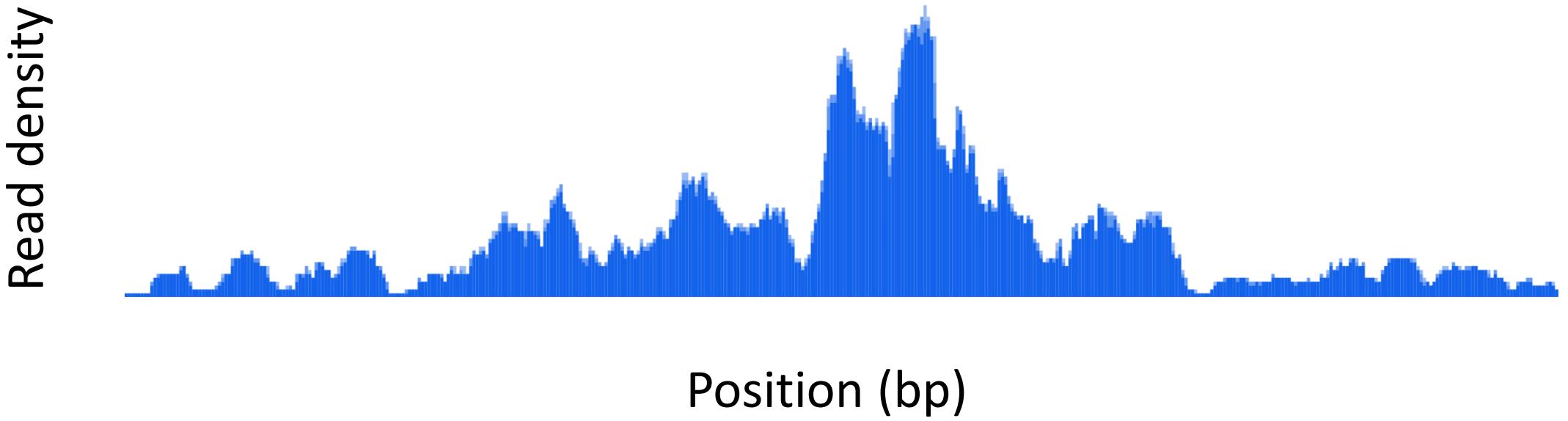
Sequence-only → factor peaks → Overall functional effect

Data:

- Samples: stride of 1bp along the genome
- Input features: 1000bp one-hot, reference genome
- Response (output): 0/1 for each chromatin factor (based on previously called peaks)



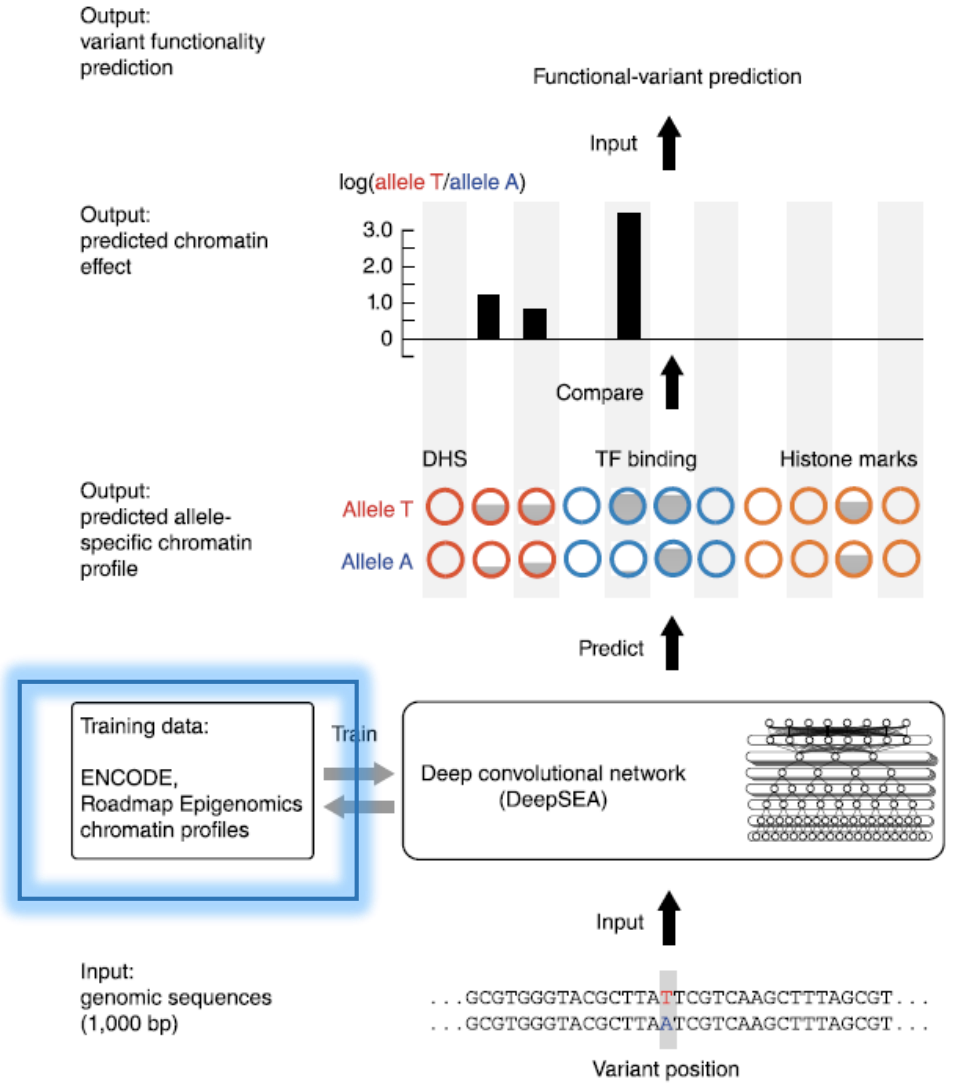
Response: discretized peaks



Sequence-only → factor peaks → Overall functional effect

Data:

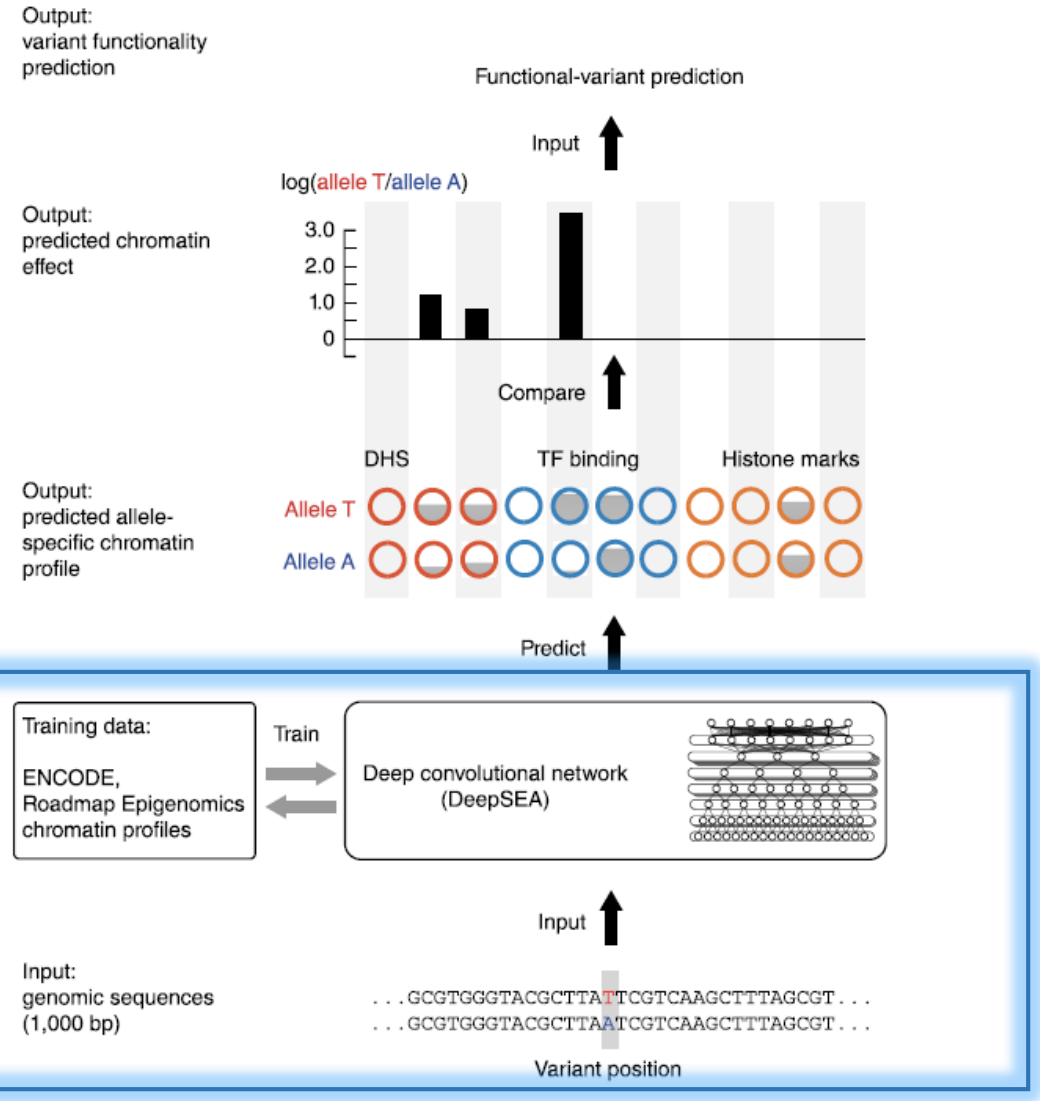
- Samples: stride of 1bp along the genome
- Input features: 1000bp one-hot, reference genome
- Response (output): 0/1 for each chromatin factor (based on previously called peaks)
- Train (only samples with>1TF), validate (only 4000 samples), test (2 chromosomes)



Architecture – your standard deep CNN

Architecture:

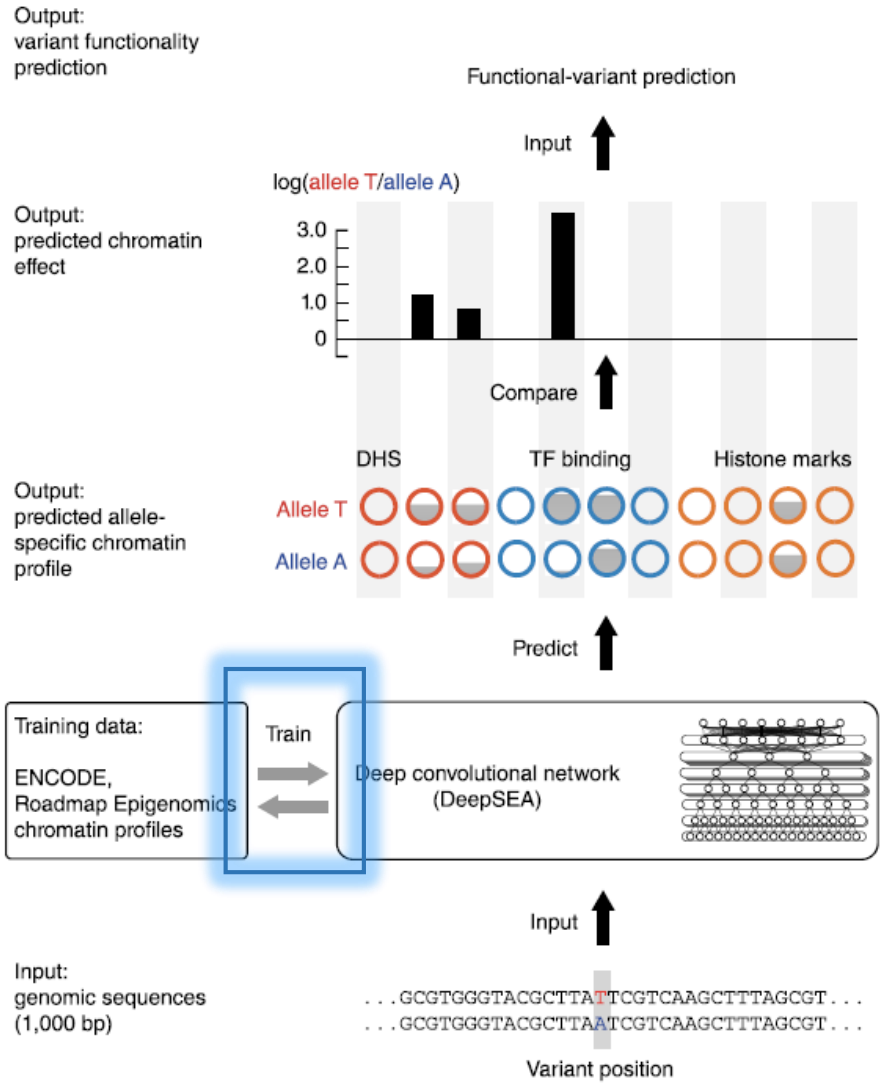
- 3 convolutional layers with ReLU activation + max pooling
- # Kernels is 240, 480, 960 respectively
- Followed by a fully-connected layer with ReLU(WX)
- Last layer (919 outputs) is logistic, represents probability of peak



Objective function: sum of Negative Log Likelihood

$$\text{objective} = \text{NLL} + \text{Regularization}$$

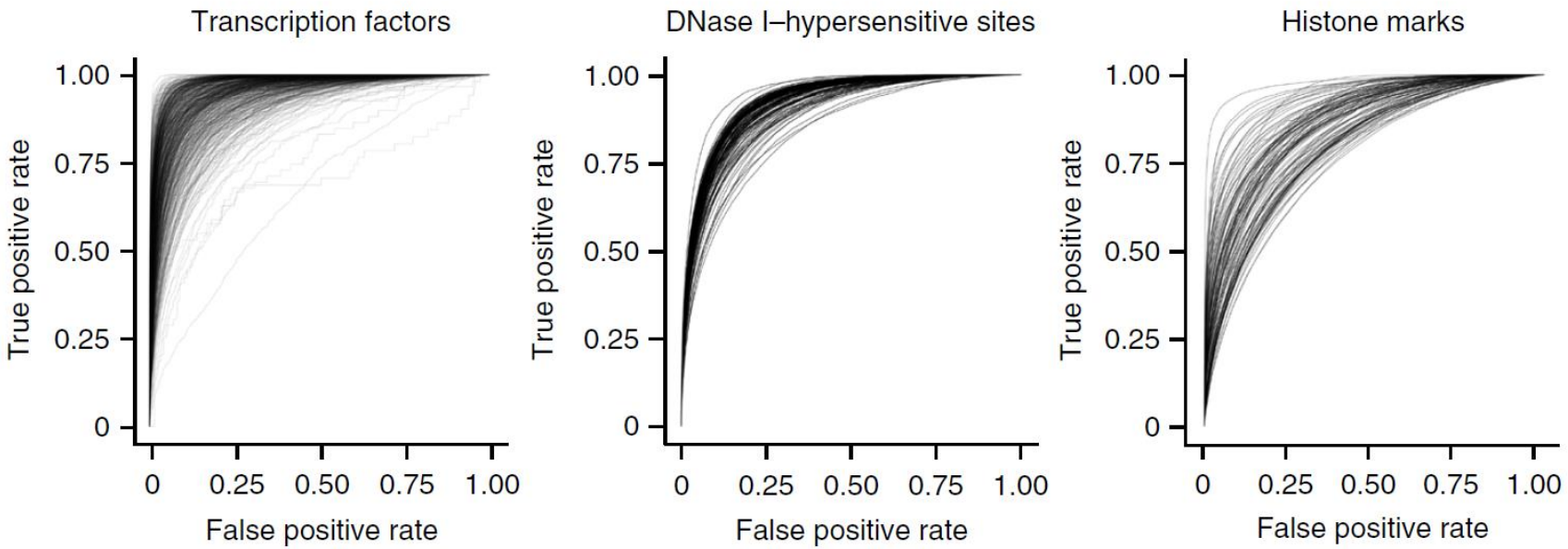
$$\text{NLL} = -\sum_s \sum_t \log(Y_t^s f_t(X^s) + (1 - Y_t^s)(1 - f_t(X^s)))$$



Objective function overweights transcription factors?

$$\text{objective} = \text{NLL} + \text{Regularization}$$

$$\text{NLL} = -\sum_s \sum_t \log(Y_t^s f_t(X^s) + (1 - Y_t^s)(1 - f_t(X^s)))$$



Regularization—I just can't get enough

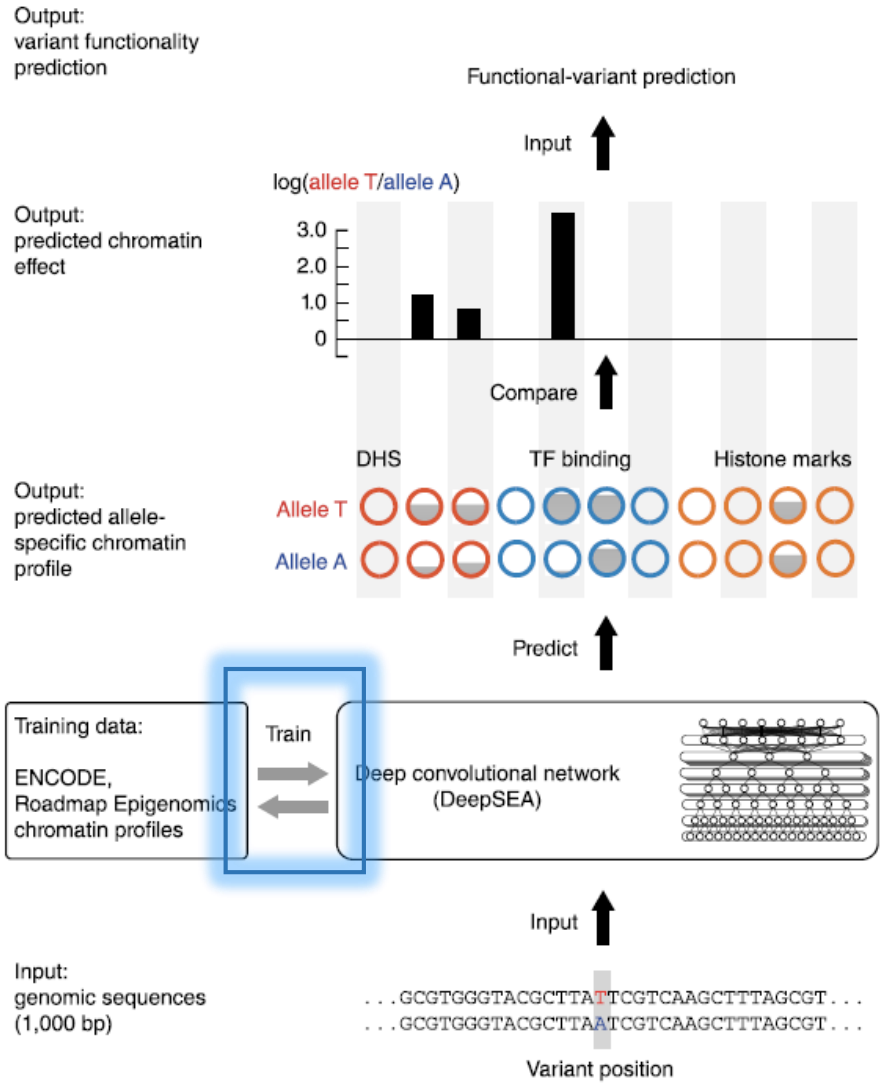
$$\text{objective} = \text{NLL} + \lambda_1 \,||\, W \,||_2^2 + \lambda_2 \,||\, H^{-1} \,||_1$$

Also:

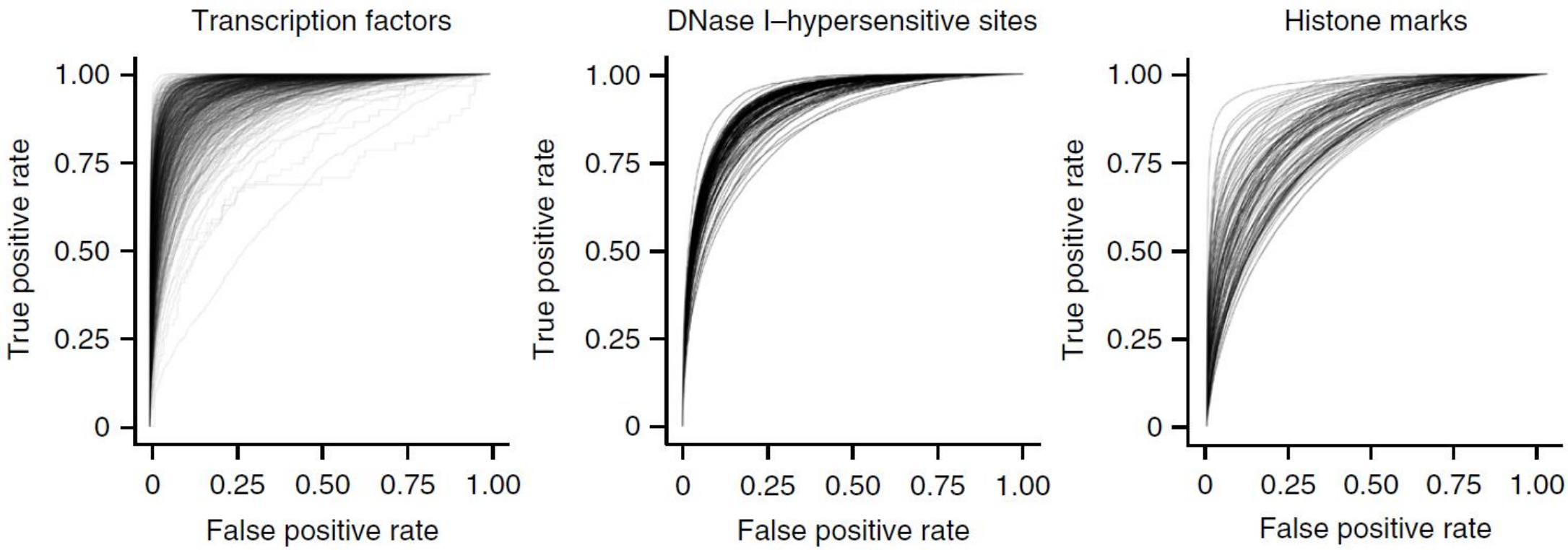
λ_3 - (shared) regularization on weight matrix for each neuron

"λ_4" – dropout training

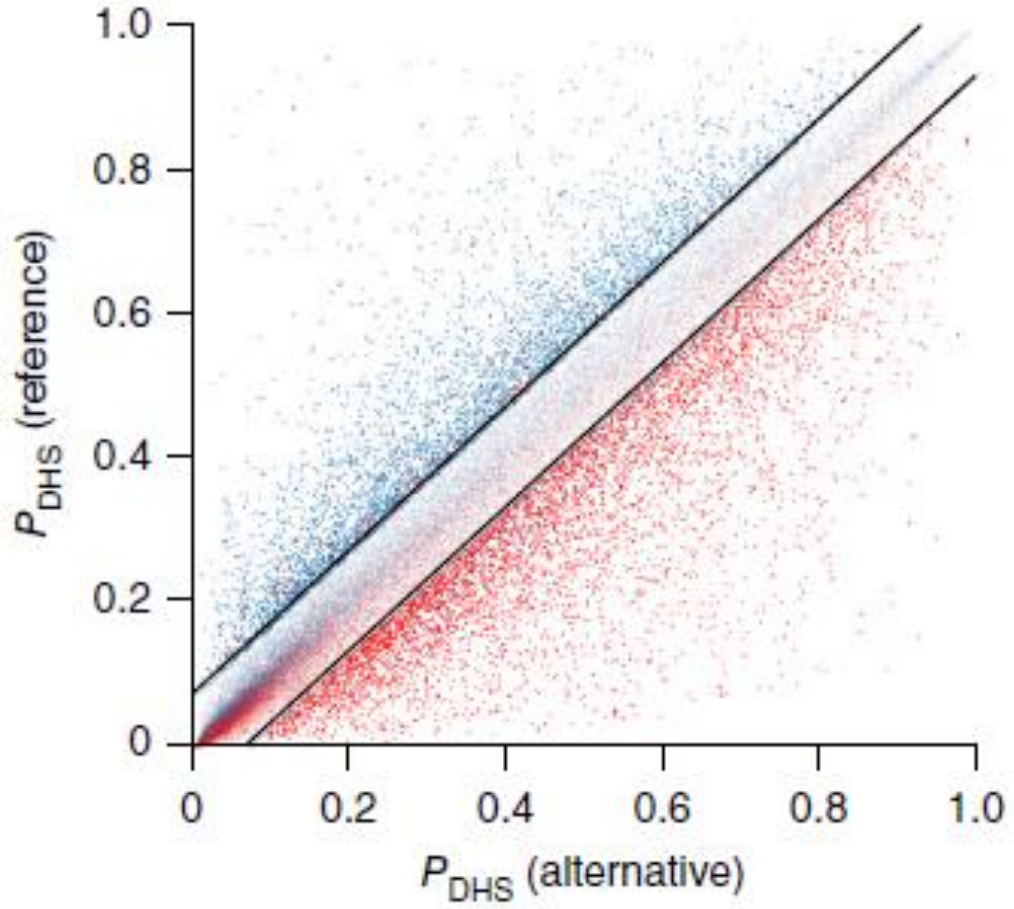
"λ_5" – multi-task prediction



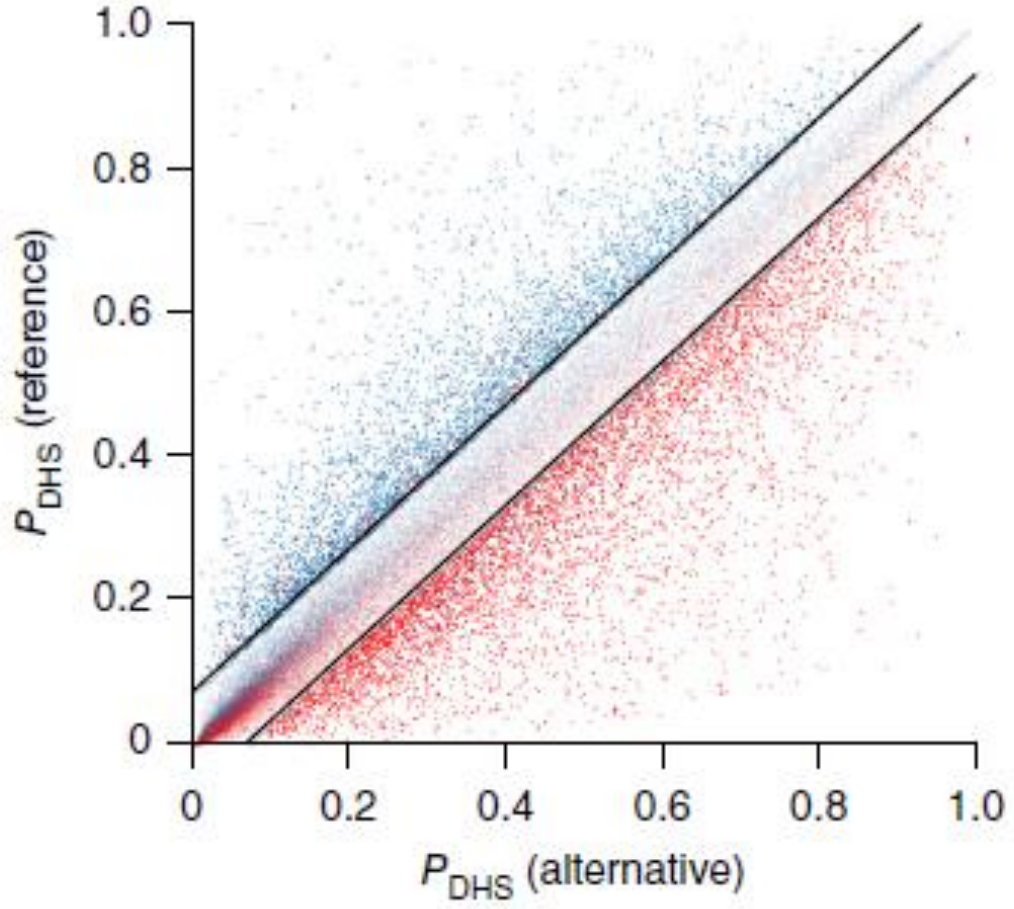
Test performance metrics: The infamous ROC AUC



Test performance of importance scoring: Allelic imbalance



Validation of importance scoring: Allelic imbalance



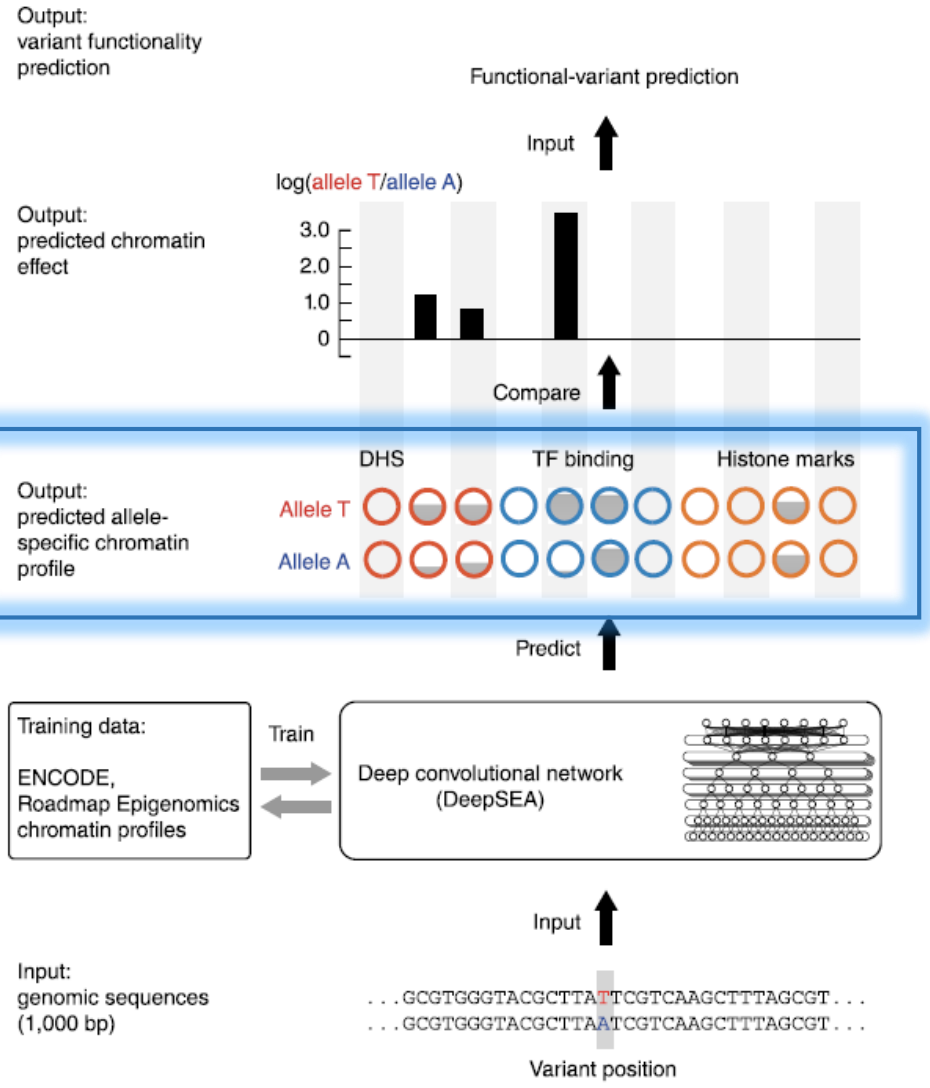
Axes = DeepSEA estimates
Color = Allelic imbalance

Importance scoring: in-silico mutagenesis

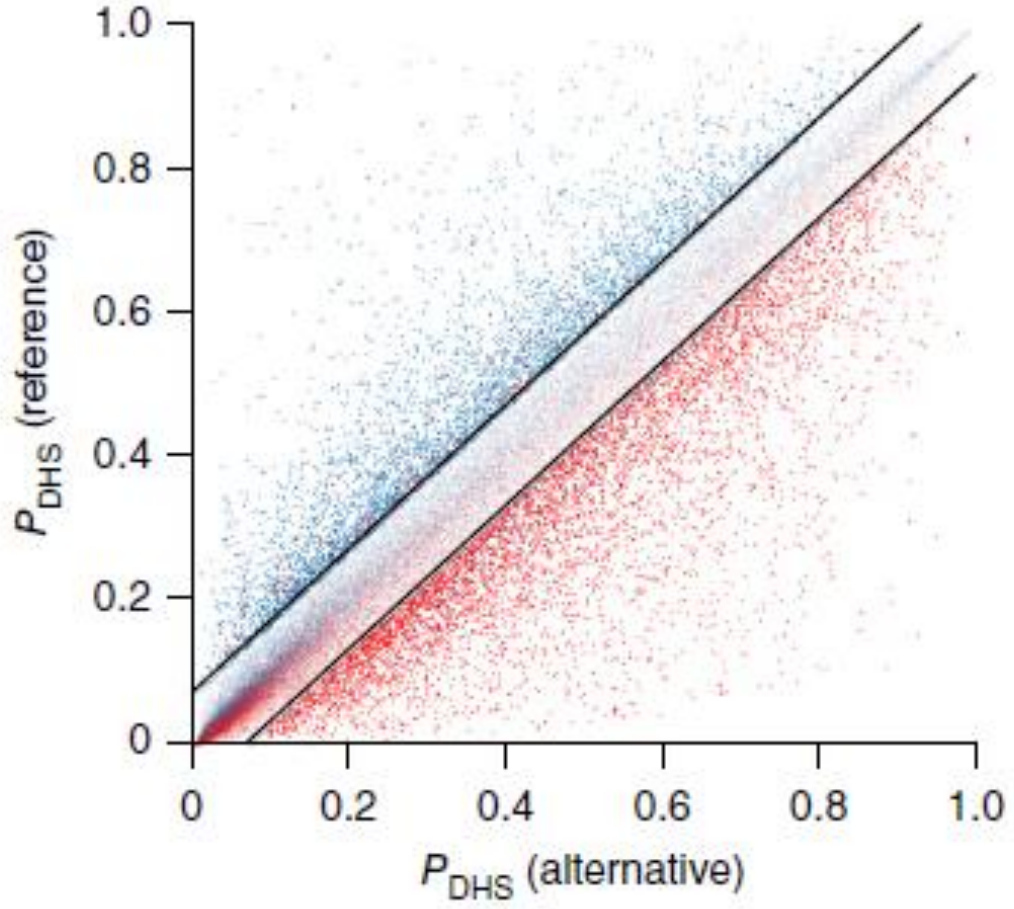
Probability of 1 (peak) with reference allele

$$\log_2\left(\frac{P_0}{1-P_0}\right) - \log_2\left(\frac{P_1}{1-P_1}\right)$$

Probability of 1 (peak) with alternative allele

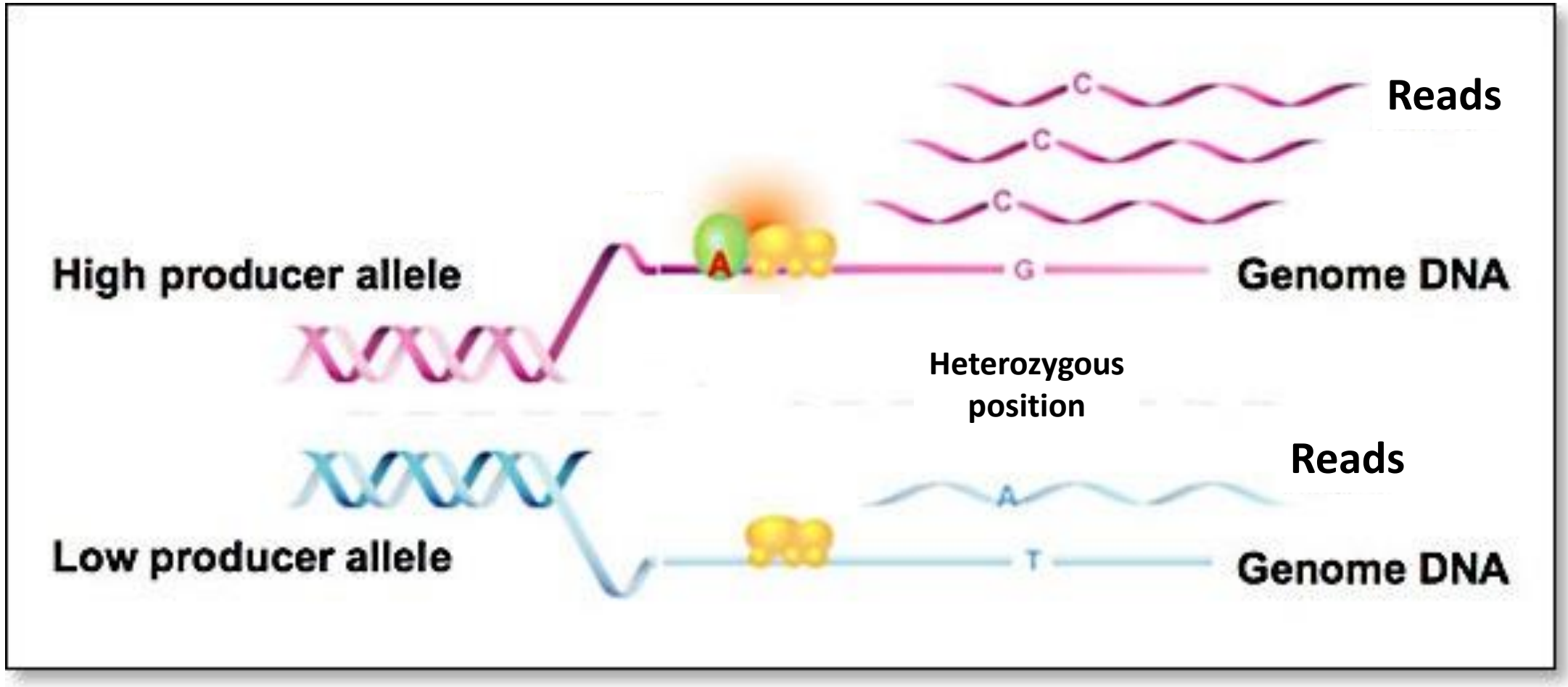


Validation of importance scoring: Allelic imbalance

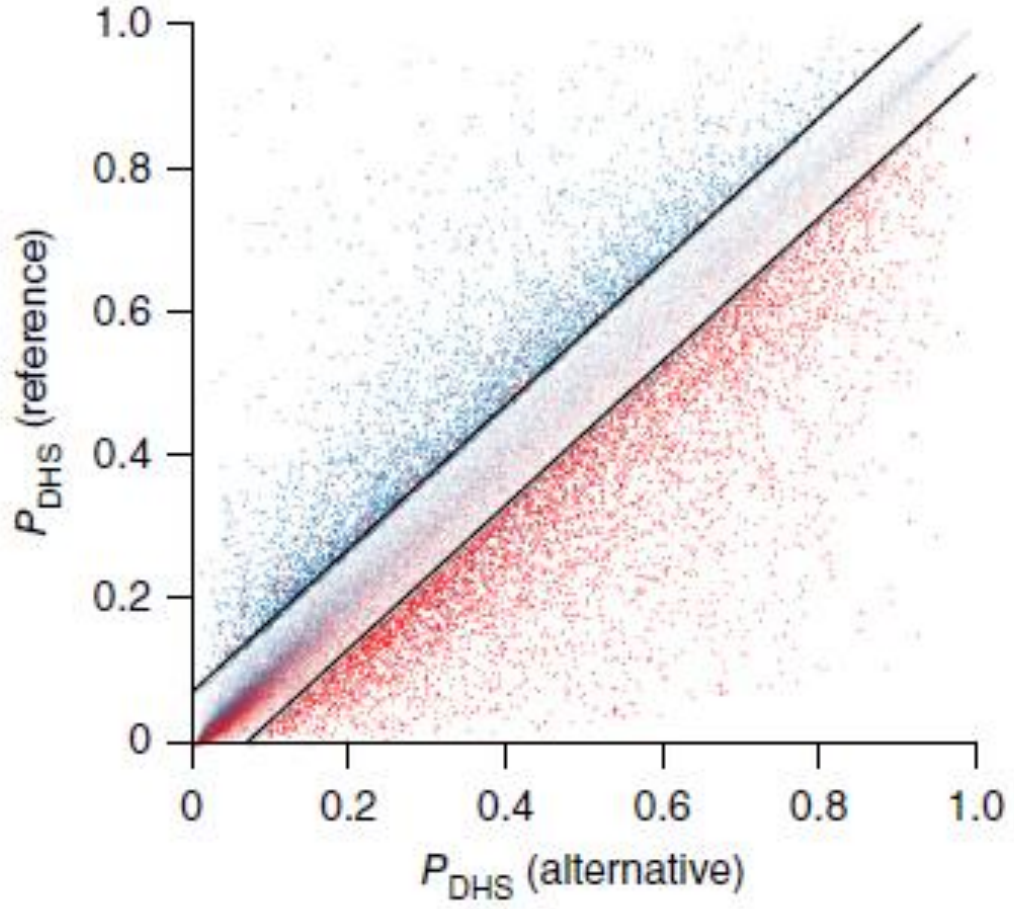


Axes = DeepSEA estimates
Color = Allelic imbalance

Validation of importance scoring: Allelic imbalance

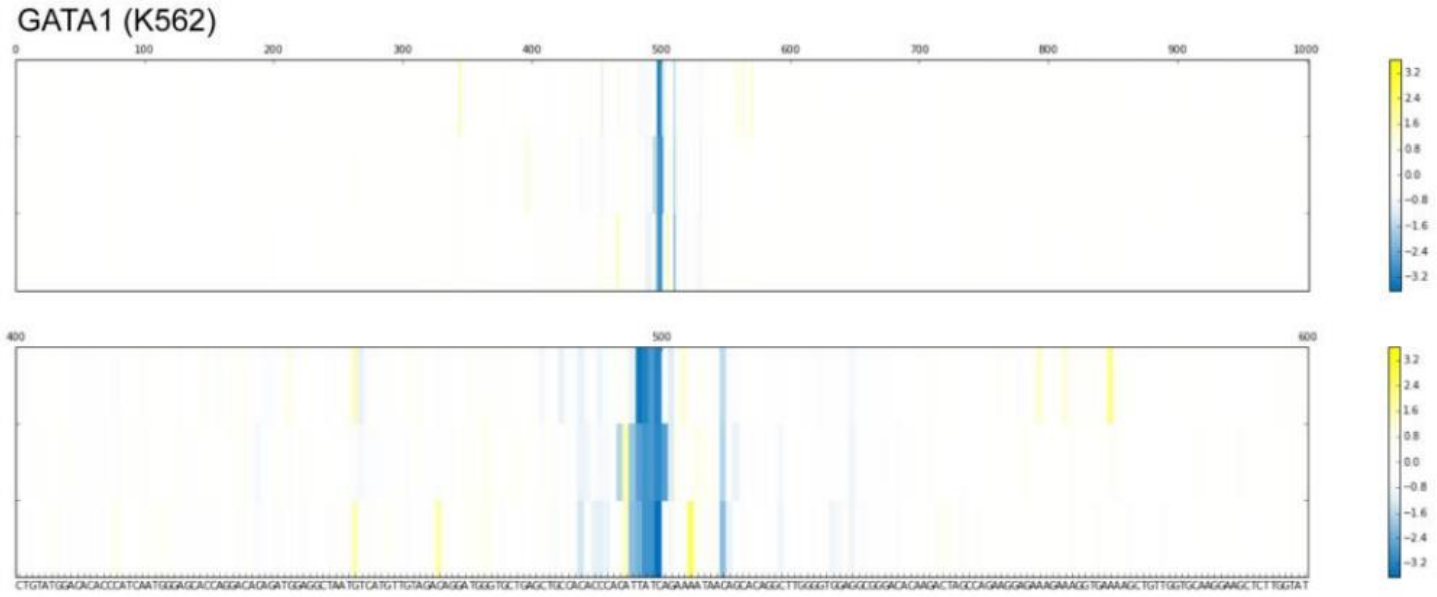


Validation of importance scoring: Allelic imbalance



Axes = DeepSEA estimates
Color = Allelic imbalance

Validation of importance scoring: positive controls



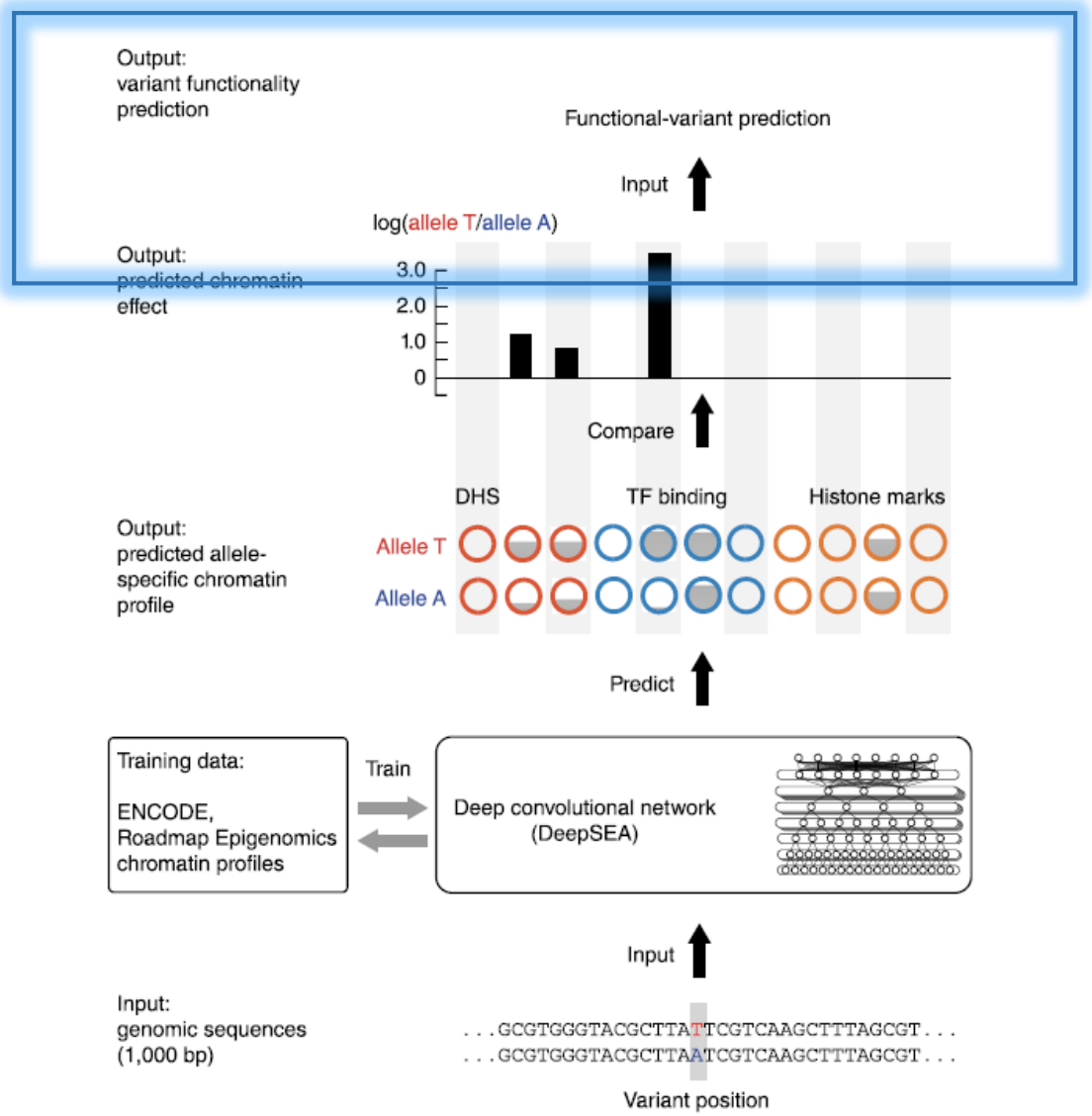
Blood disorder

Prioritization / overall effect of variant on function

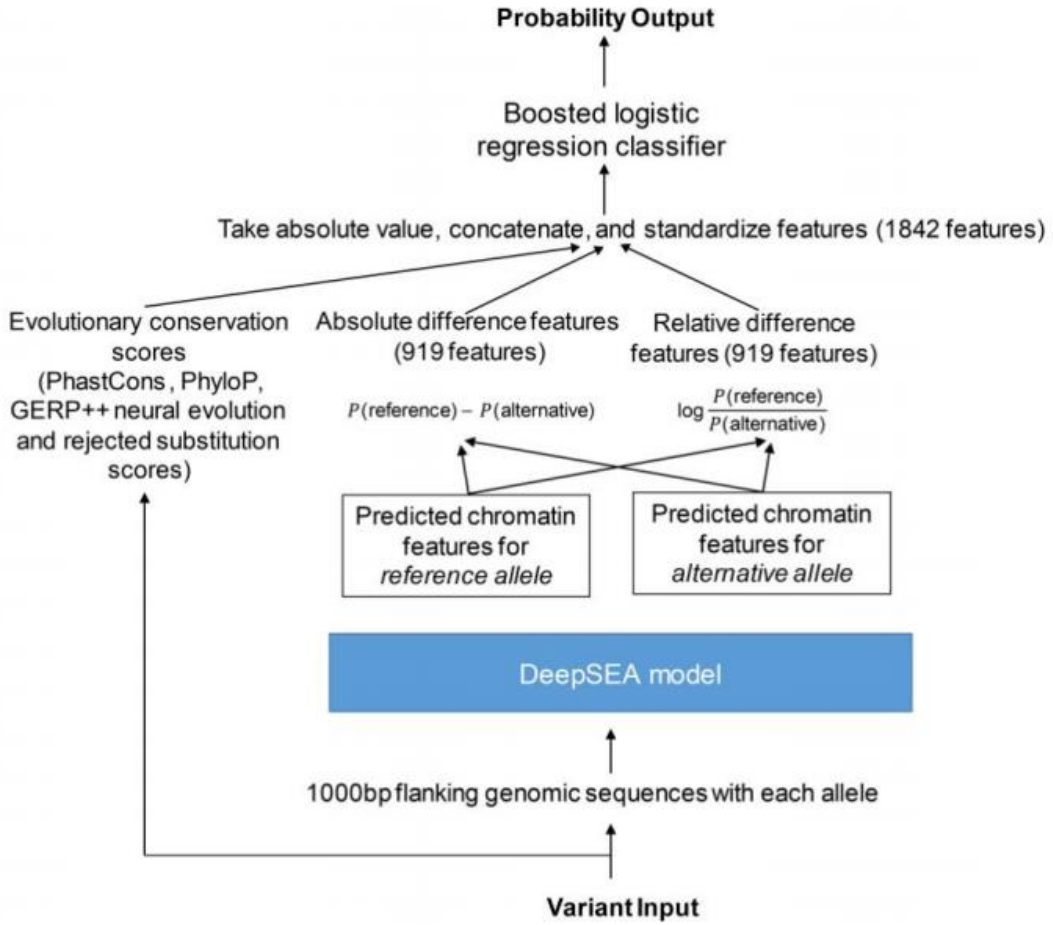
- Competing methods use a lot of high-throughput data as well, but virtually always include evolutionary conservation

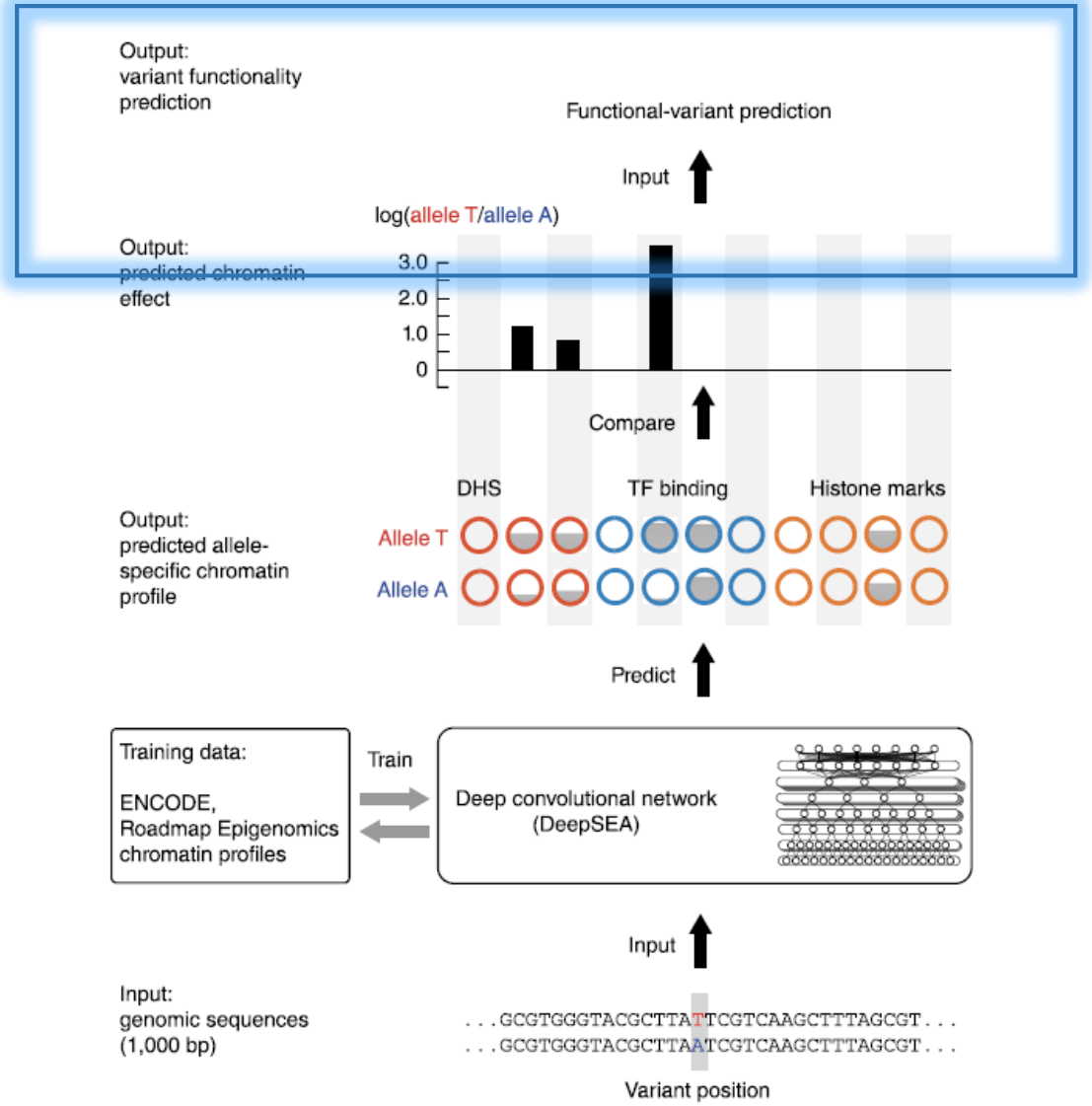
Logit(Probability variant is functional) ≈

$$\beta_0 + \overrightarrow{\beta_1} \cdot \overrightarrow{DeepSEA} + \overrightarrow{\beta_2} \cdot \overrightarrow{Evol.conservation}$$

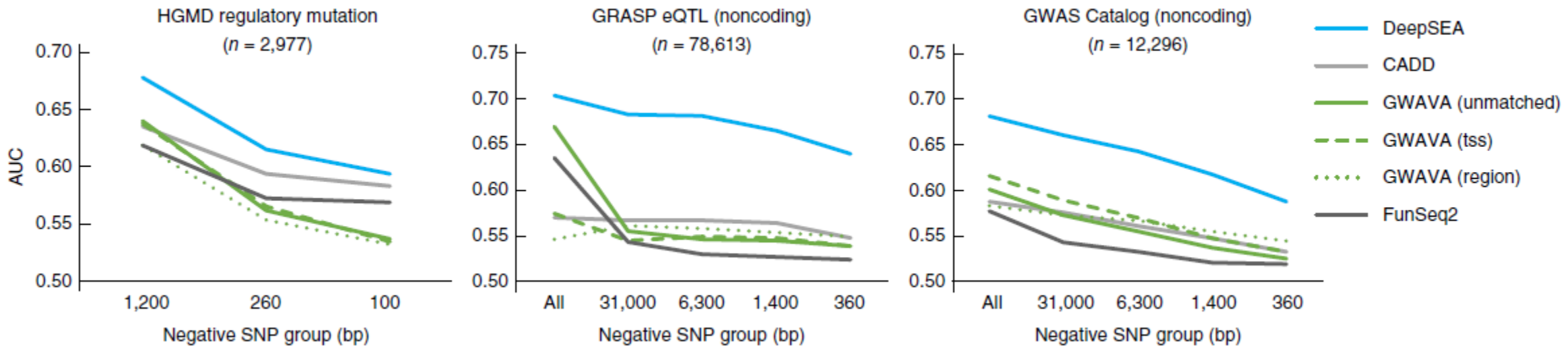


Prioritization / overall effect of variant on function

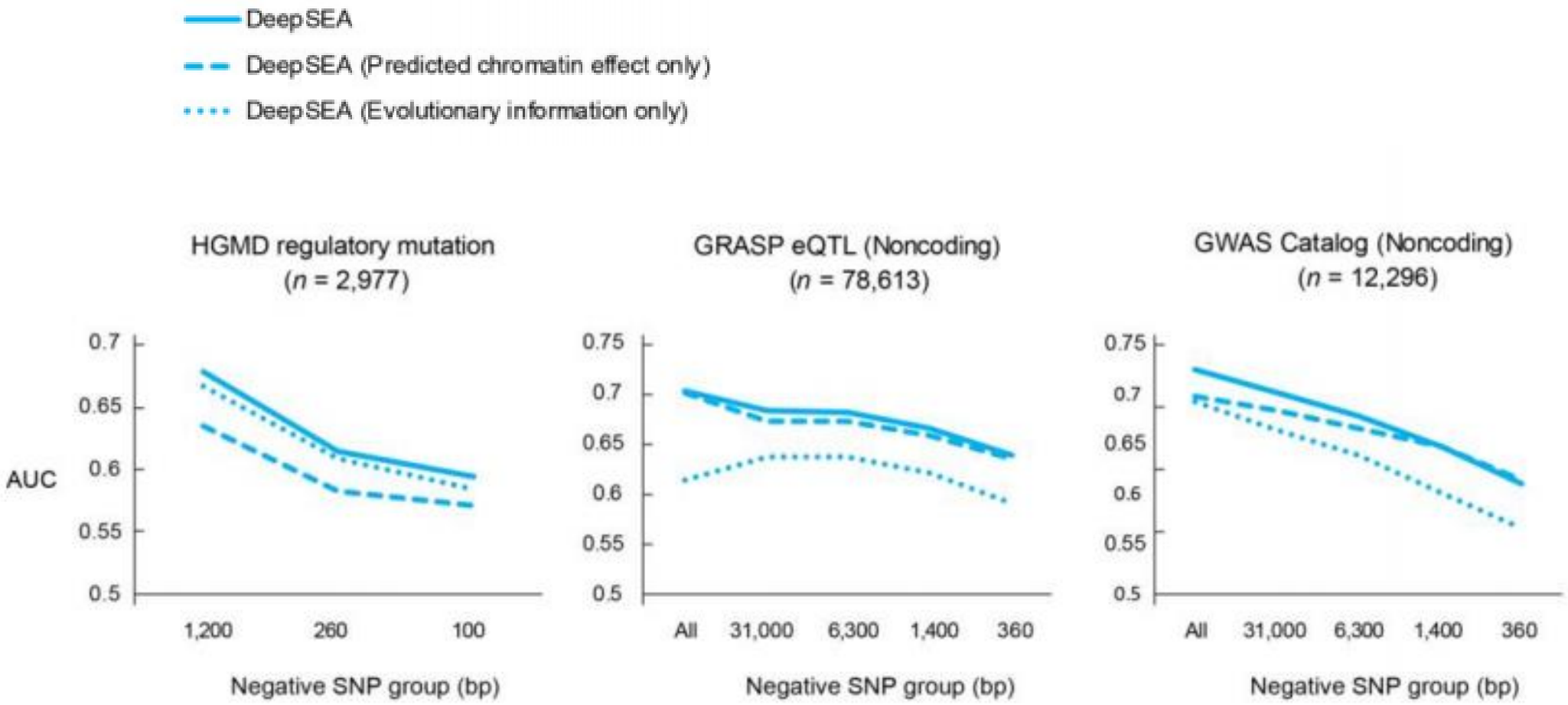




Performance of prioritization method



Prioritization / overall effect of variant on function



DeepSEA - Summary

(1) First wave of genomics CNN; predict functional effect from sequence

(2) Performance:

- Surprisingly good importance scoring
- Per task—could prob. be improved

(3) Some questionable choices e.g. objective function, learning rate, test performance metric, training and validation set choices

but…

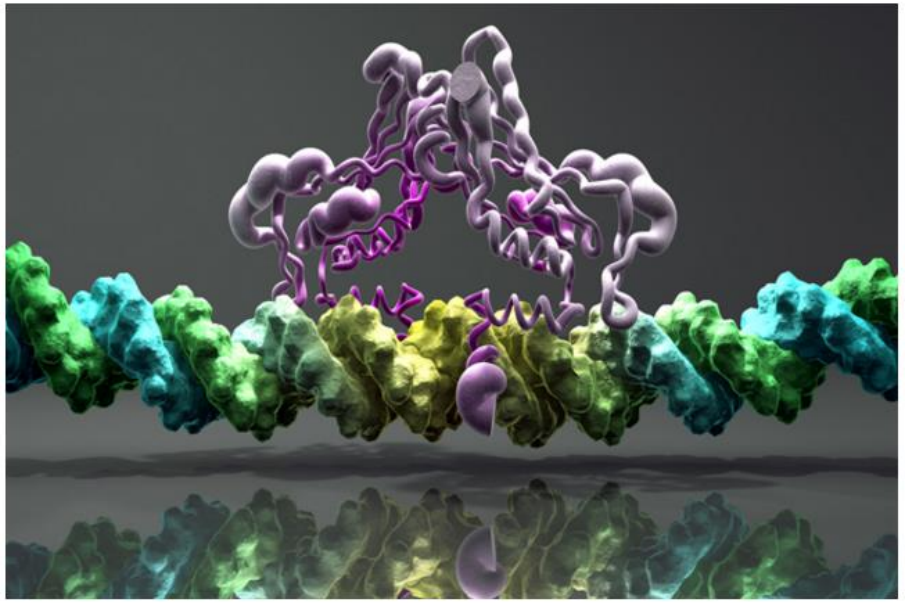
Flexible CNN + Heavy regularization

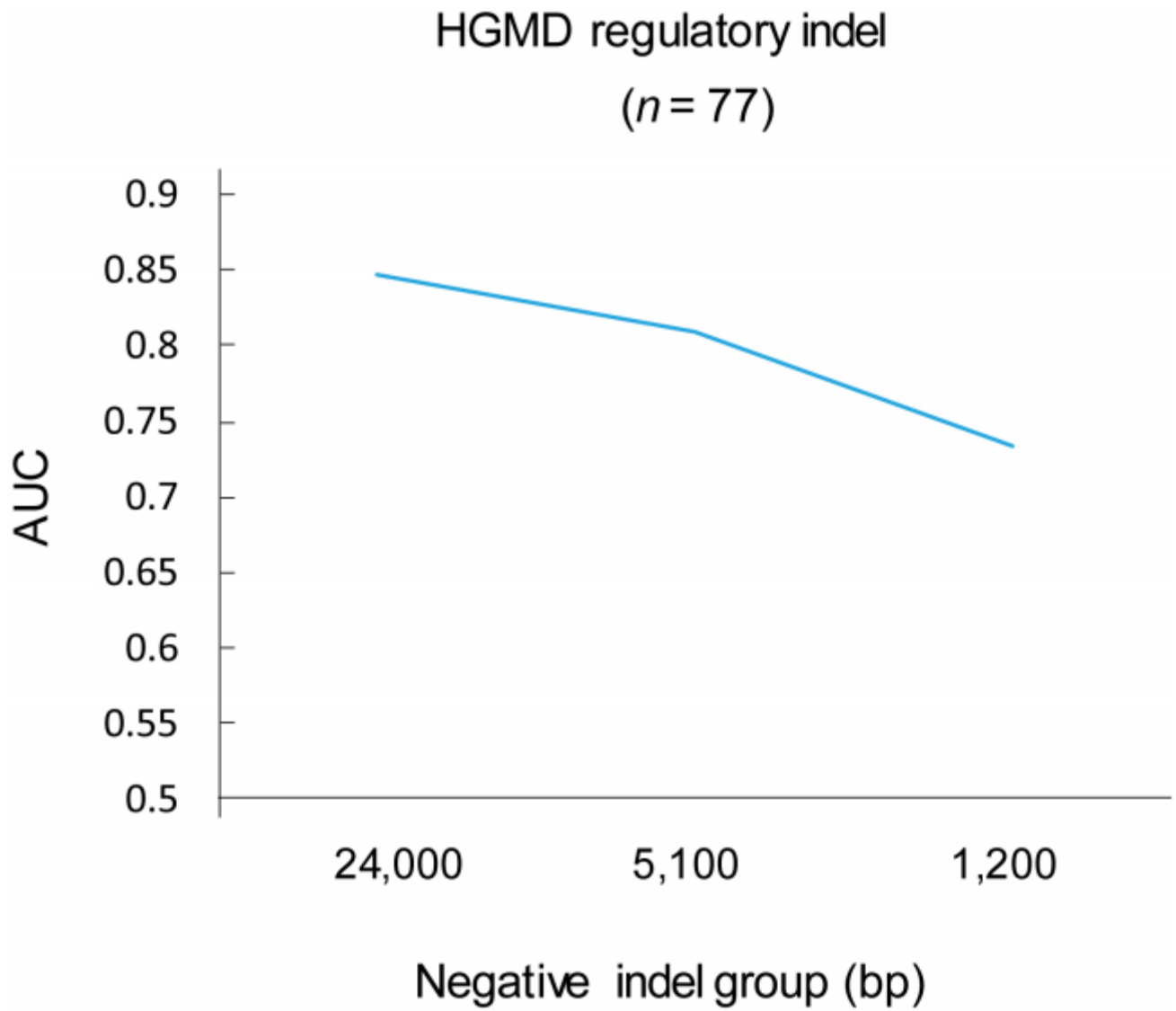
compensate for all crimes

(e.g. hyperparameters don't seem to matter much)

Software forecasts effects of mysterious mutations

BY KATE YANDELL / 26 AUGUST 2015




HGMD regulatory indel
(n = 77)
AUC
0.9
0.85
0.8
0.75
0.7
0.65
0.6
0.55
0.5
24,000
5,100
1,200
Negative indel group (bp)

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