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

Jian Zhou and Olga G. Troyanskaya

Nature Methods, 12, August 2015

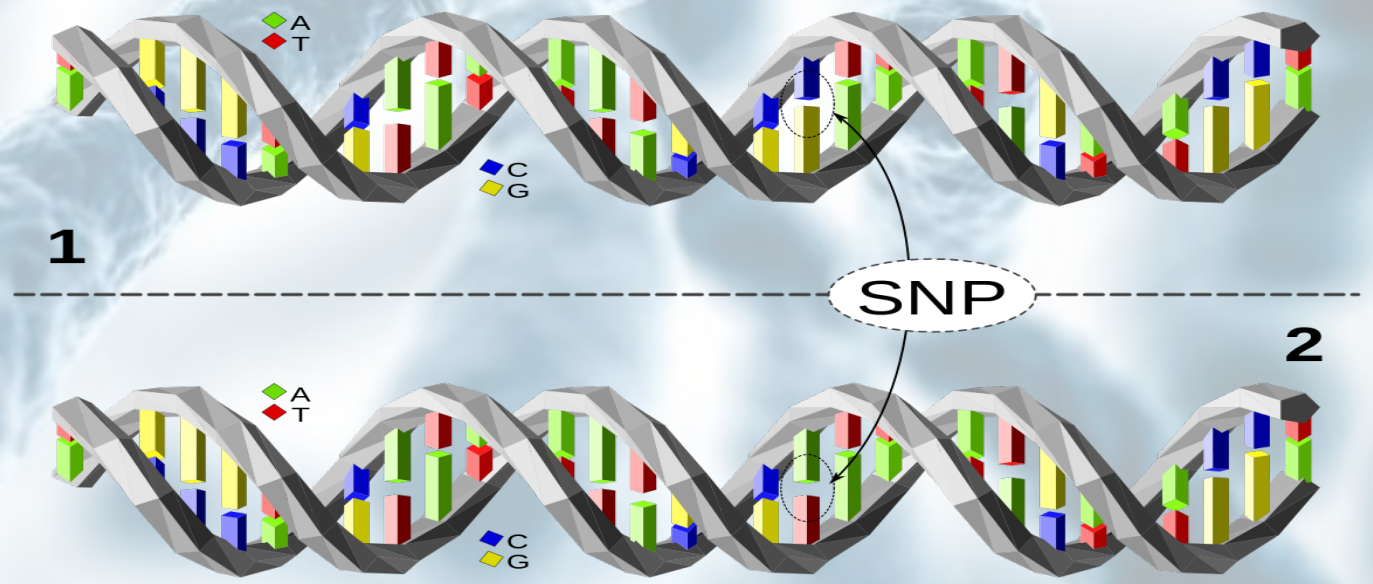
Journal Club | *June 15, 2017*

Outline

- Motivation
- Framework
  - Convolutional Neural Network (CNN)
  - Relative log-fold change
  - Regularized logistic regression
- Predictive Tasks
  - *In silico* mutagenesis
  - Chromatin effect prediction
  - SNP Functional prioritization
  - Indel prioritization
- Strengths & Weaknesses

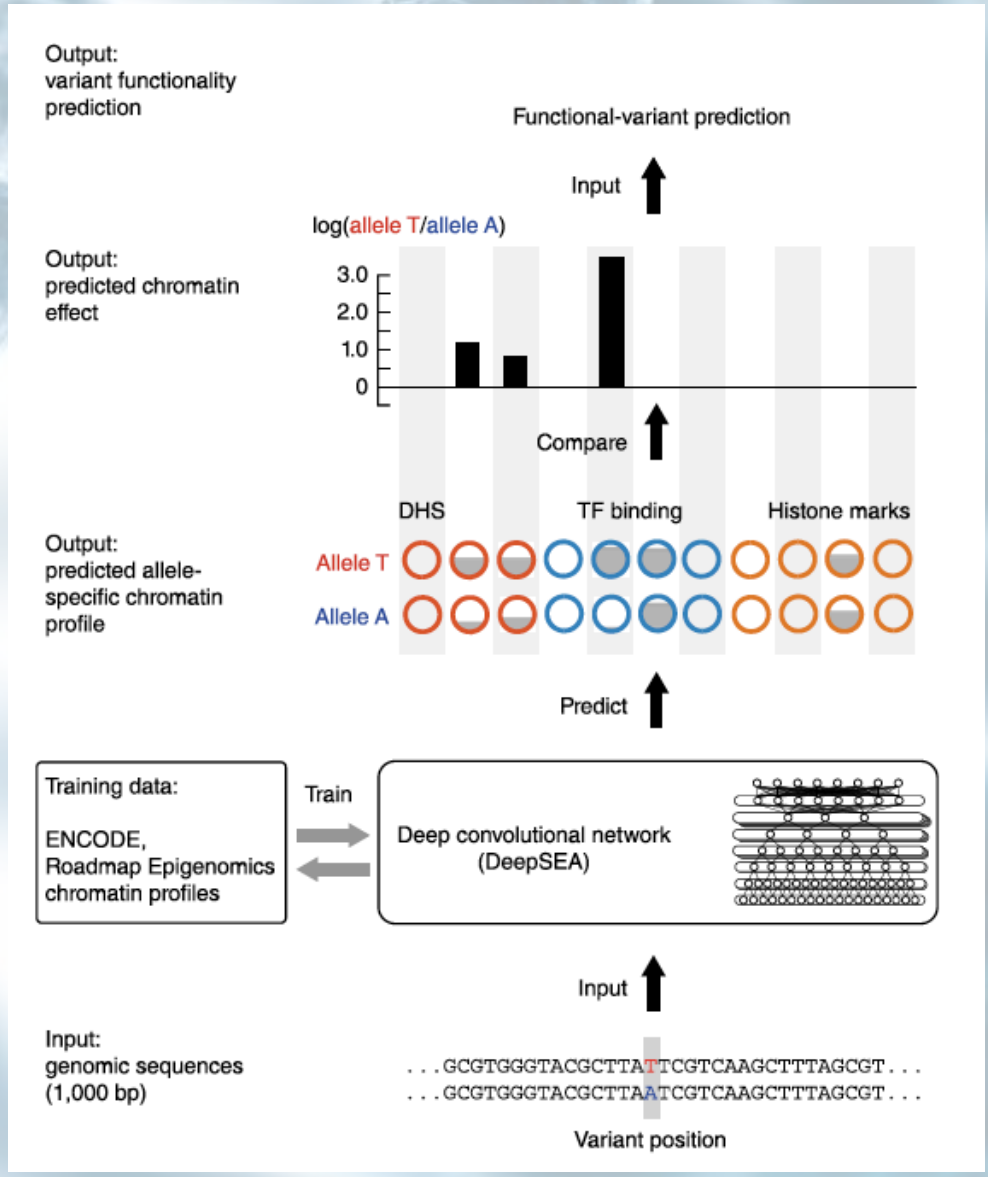
Motivation

- Most disease-related SNPs lie in noncoding regions
- Historically, coding regions have been given more attention
- *De novo* predictions help prioritize in regions with no or poor annotation



Framework

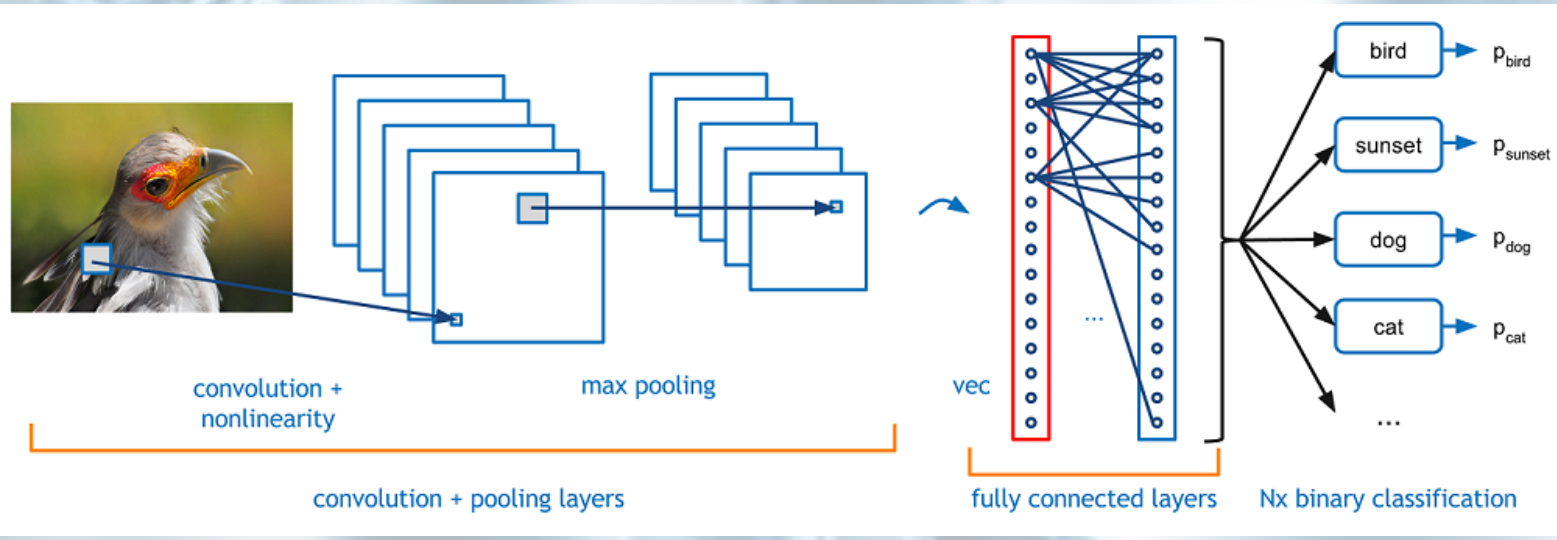
- Deep learning-based Sequence Analyzer (DeepSEA)
- DeepSEA Framework
  - ***Convolutional Neural Network (CNN)***
  - Relative log-fold change
  - Regularized logistic regression



Framework

Convolutional Neural Network

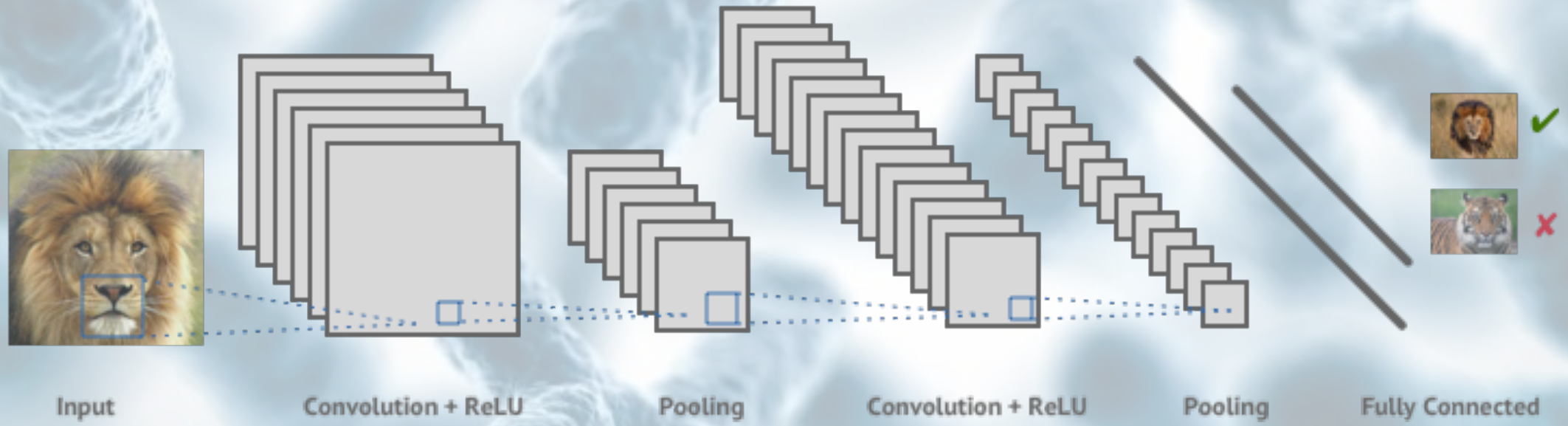
- Works on spatial features where order of features is important
- Inspired by receptive fields of animal visual cortex
- One of few approaches that revolutionized Deep Learning
- Popular for image classification



Framework

Convolutional Neural Network

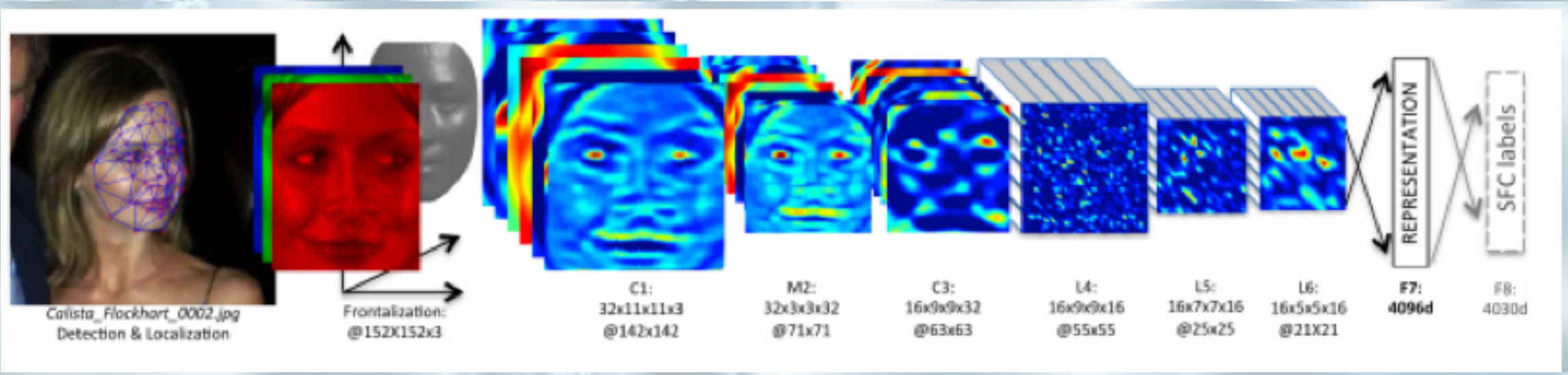
- Learning discriminative features automatically
- Output can be one or many values, depending on network architecture



Framework

Convolutional Neural Network

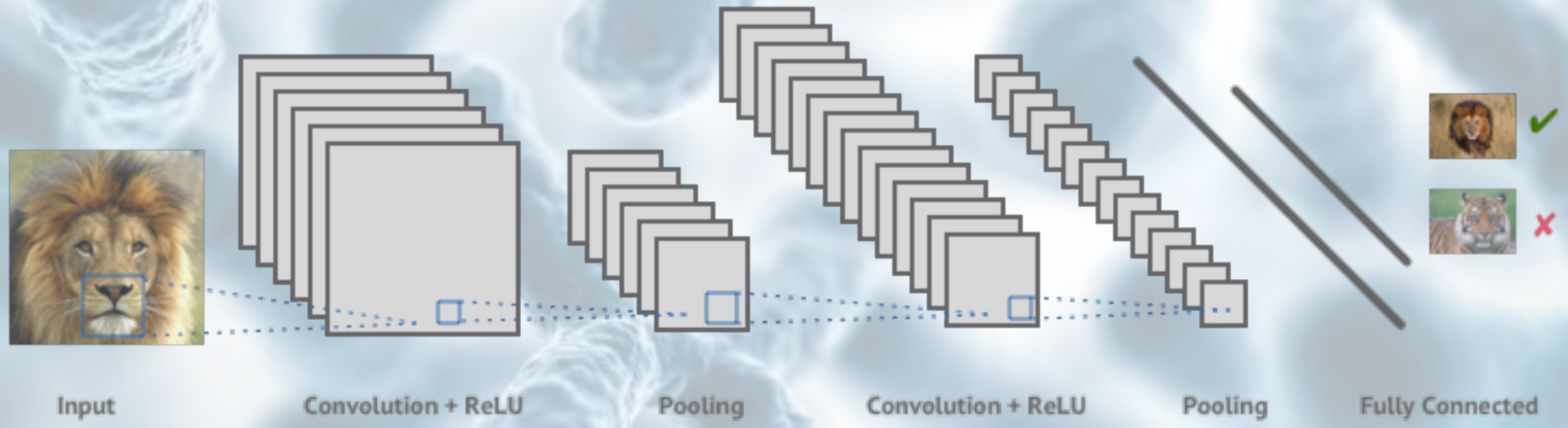
- Convolution Step
  - Scanning for each feature
- Pooling Step
  - Shrinking feature map (~ zooming in), also called subsampling



Framework

Convolutional Neural Network

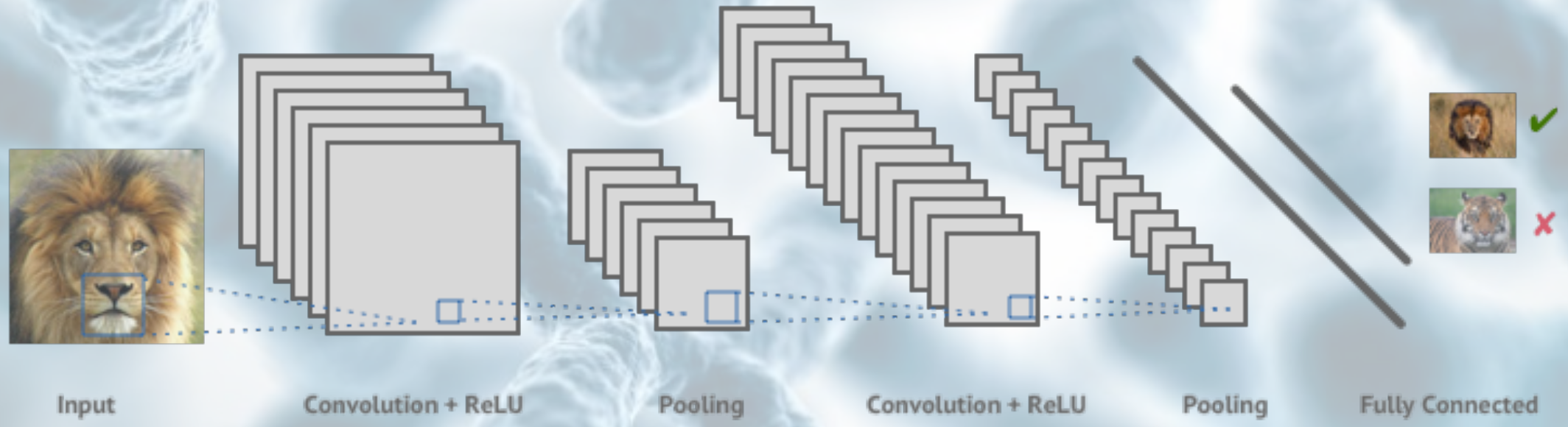
- Final layer is usually a fully connected neural network
- Can be any other classifier, such as SVM as in [Tang, 2013]



Framework

Convolutional Neural Network

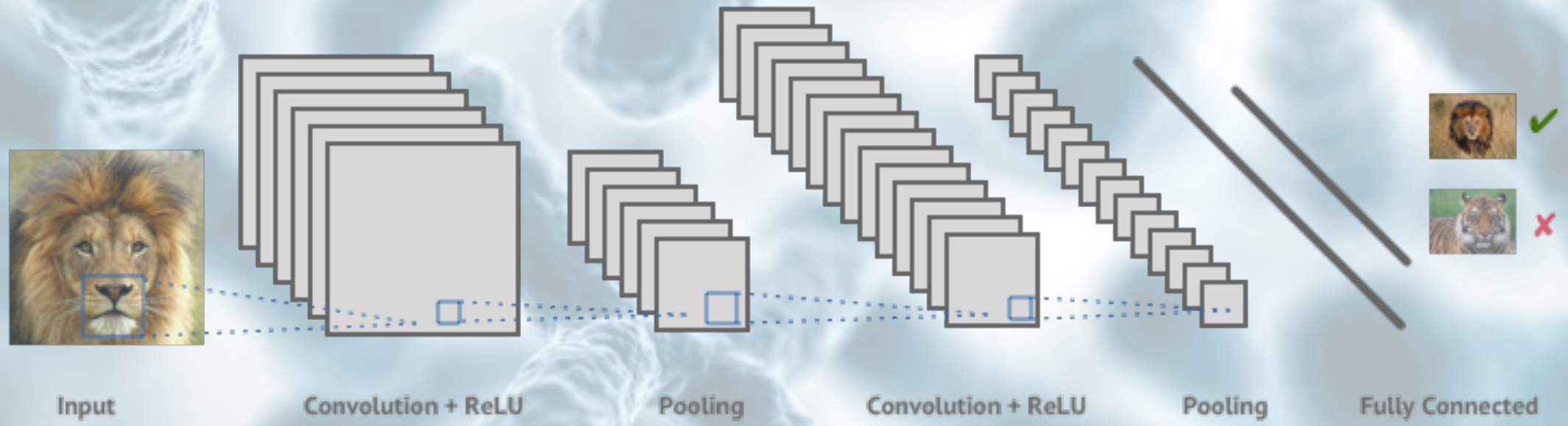
- Number of features, window size for scanning, and other parameters need to be optimized
- But why does it work on sequence data?



Framework

Convolutional Neural Network

- ***Works on spatial features where order of features is important***
  - ***DNA sequences, video frames, images, etc.***



Framework

Convolutional Neural Network

- ***Works on spatial features where order of features is important***
  - ***DNA sequences, video frames, images, etc.***
- Input: 1000-bp sequence

SEQUENCE:	A	T	C	T	G	G	A
$x_A(n)$	1	0	0	0	0	0	1
$x_C(n)$	0	0	1	0	0	0	0
$x_G(n)$	0	0	0	0	1	1	0
$x_T(n)$	0	1	0	1	0	0	0

- Outputs: Chromatin features
  - 975 values (670 TF binding, 125 DHS, and 104 histone modification values)
- Hundreds of features to scan for

Framework

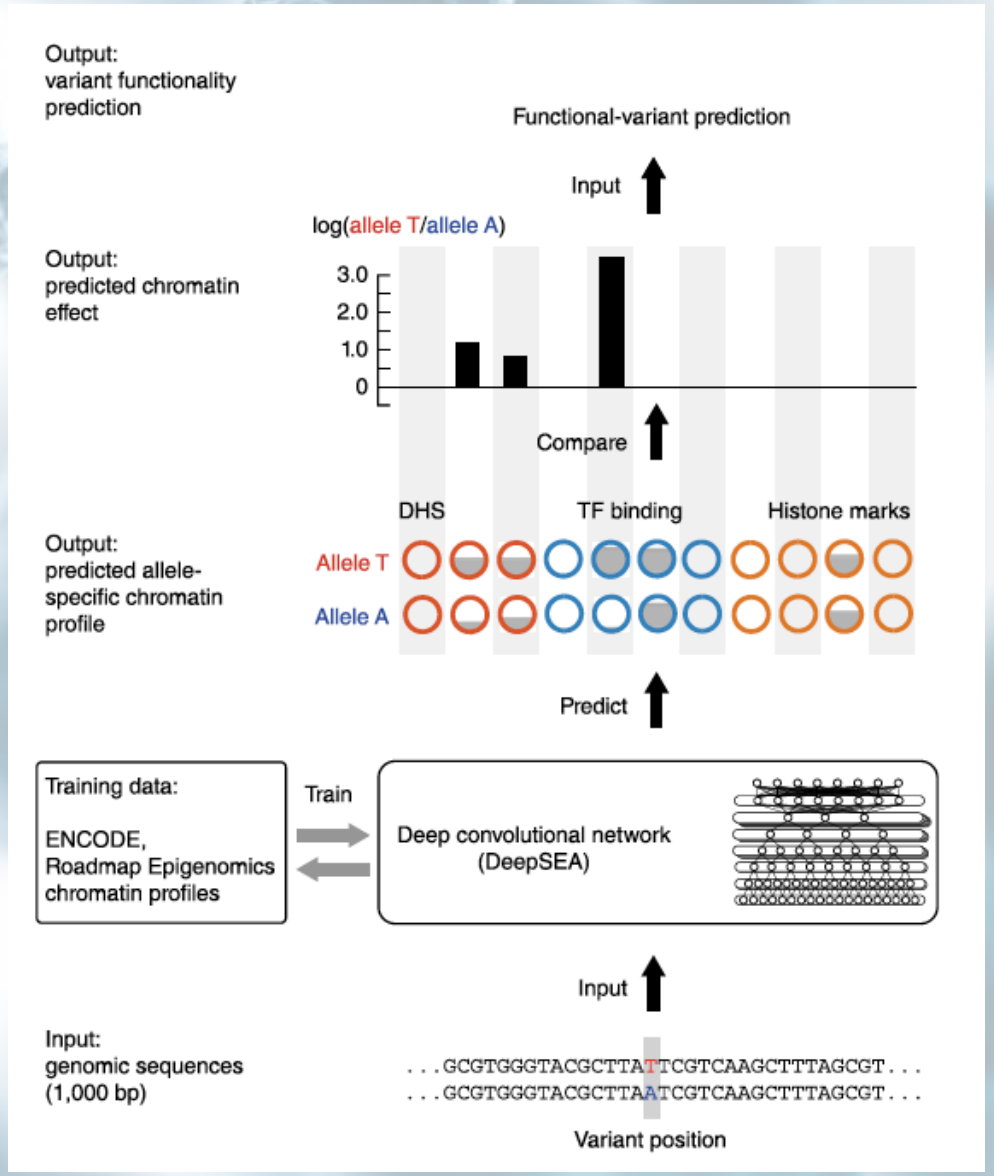
Convolutional Neural Network

- CNN Toy Example | MNIST Digit Classification via TensorFlow in Python [here]
- Setup on Farnam (~ 5 minutes) [here]
- Accuracy > 99%

Framework

Predictive Tasks

- Chromatin Feature Prediction
  - Training data
    - Genome wide chromatin profiles
    - 670 TF binding, 125 DHS, and 104 histone mark profiles
    - ENCODE and Roadmap Epigenomics
    - 521.6 Mbp (17%) of the genome bound 1+ of 160 chosen TFs
- Testing
  - Holdout sequences from the genome
  - 4,000 samples from chr7 region 30,508,751-35,296,850

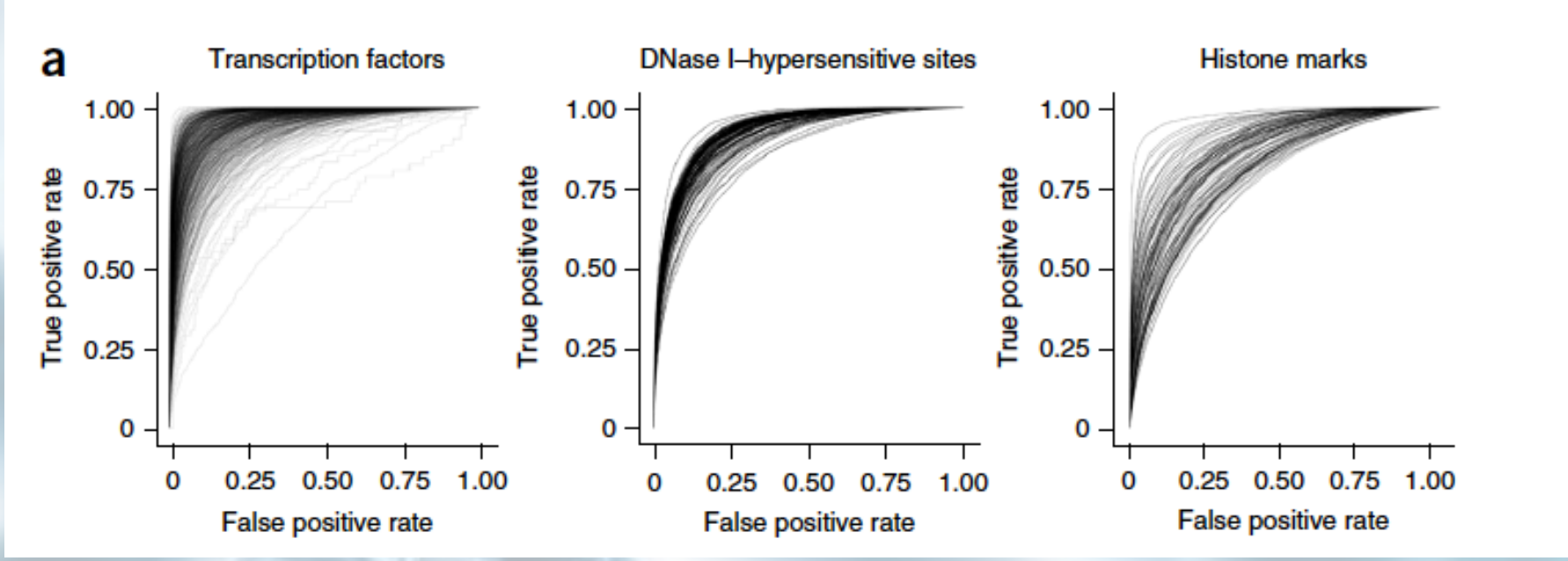


Framework

Predictive Tasks | Chromatin Feature Prediction

- Results
  - *TF binding sites* | Median AUC = 0.985
  - *DHS* | Median AUC = 0.923
  - *Histone modifications* | Median AUC = 0.865

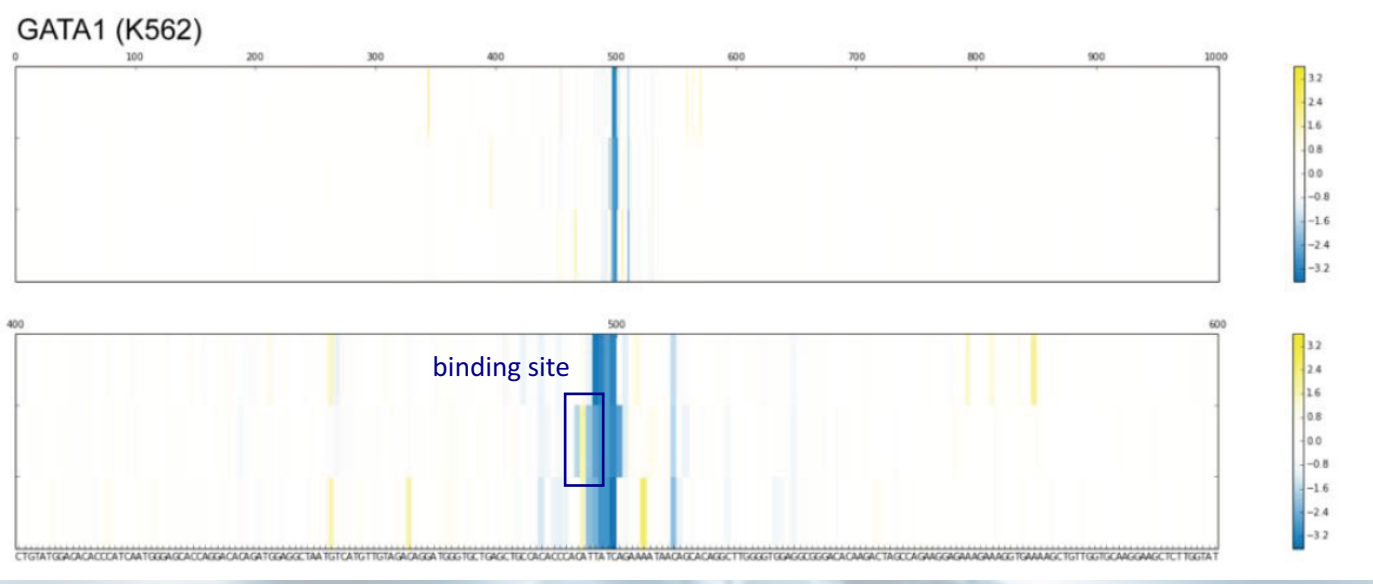
- SVM-based *gkm-SVM*
  - *TF binding sites* | Median AUC = 0.896
  - Two models: 300-bp & 1000-bp-based



Framework

Predictive Tasks | In Silico Mutagenesis

- Computational generation of all possible SNVs (3x1000 per 1KB input sequence)
- Validation against disease-related SNPs with experimental evidence
- Results
  - *Accurate prediction of TF binding effects on SNPs with experimentally validated known effects*
    - *Breast cancer risk locus C-to-T SNP rs4784227 in FOXA1*
    - *α-thalassemia T-to-C creates a binding site for GATA1*
    - *Pancreatic agenesis A-to-G mutation has deleterious effect on FOXA2 binding*

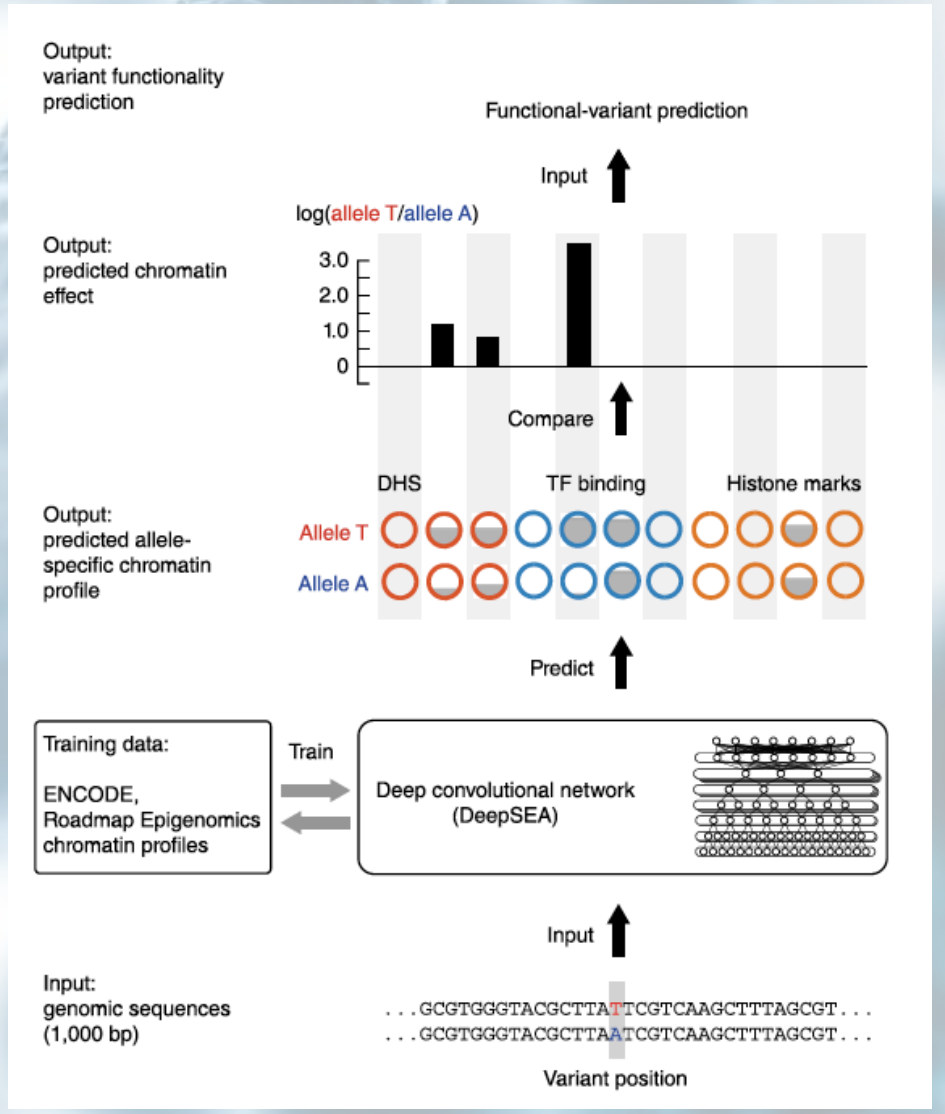


- A > C > G > T order
- Yellow increase in binding
- Blue decrease in binding

Framework

Predictive Tasks

- SNP Functional Prioritization
- CNN followed by regularized log-reg
- Sequence & evolutionary features (PhyloP & others)
  - Data
    - Human Gene Mutation Database (HGMD)
    - Noncoding eQTLs from Genome-Wide Repository of Associations between SNPs and Phenotypes
    - Noncoding SNPs from HGRI GWAS Catalog



Framework

Predictive Tasks | SNP Functional Prioritization

- Discriminating negative SNPs close to positive (functional) ones
- AUC (<0.7) lower on this task compared to all 3 previous chromatin effect prediction tasks
- Relatively low FPR

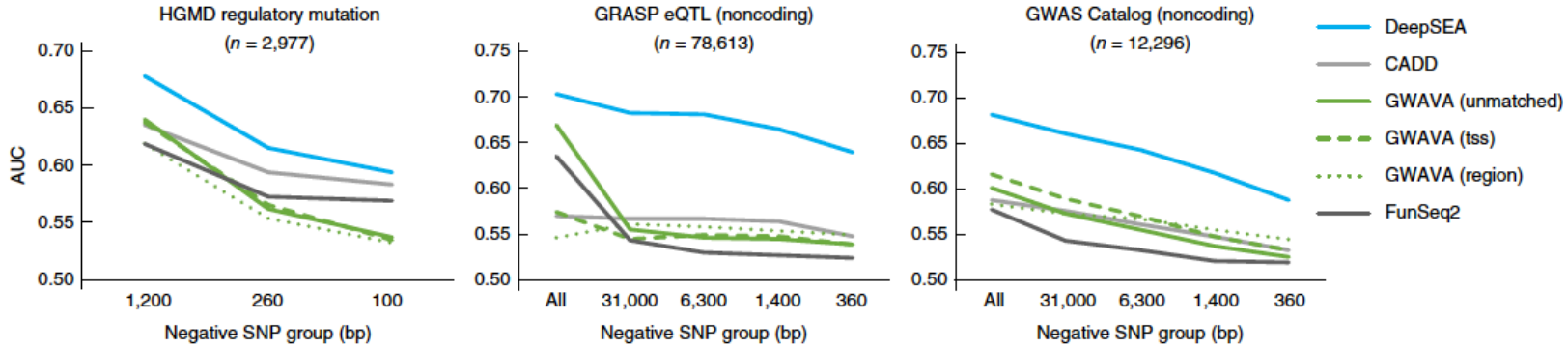
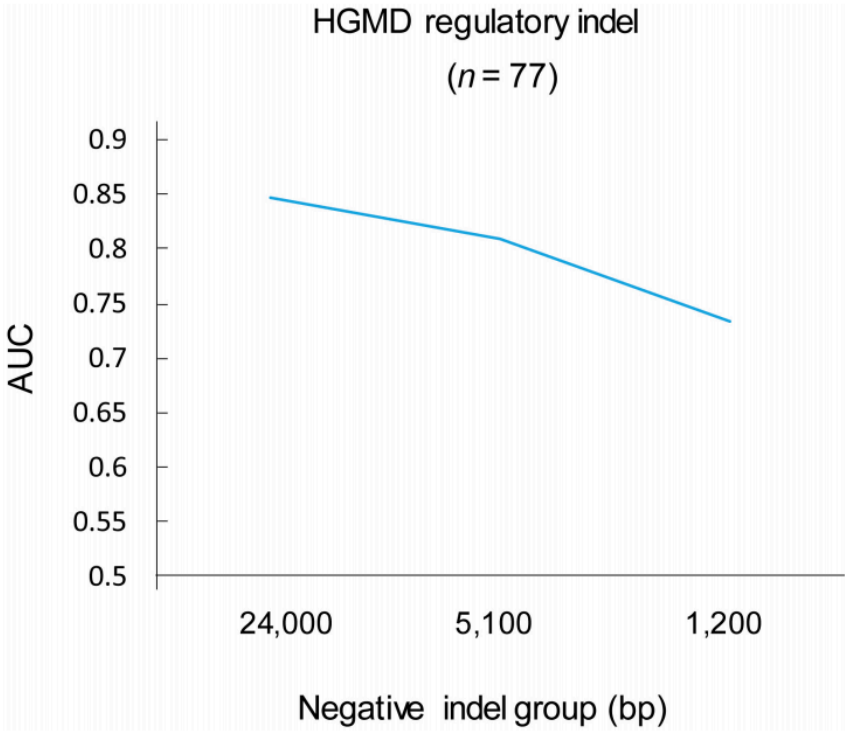


Figure 3 | Sequence-based prioritization of functional noncoding variants. Comparison of DeepSEA to other methods for prioritizing functionally annotated variants including HGMD annotated regulatory mutations, noncoding GRASP eQTLs and noncoding GWAS Catalog SNPs against noncoding 1000 Genomes Project SNPs (across multiple negative-variant groups with different scales of distances to the positive SNPs). The *x* axes show the average distances of negative-variant groups to a nearest positive variant. The "All" negative-variant groups are randomly selected negative 1000 Genomes SNPs. Because GWAVA was trained on the HGMD regulatory mutations, we filtered out GWAVA training positive-variant examples and closely located variants (within 2,000 bp) in evaluating its performance on HGMD regulatory mutations. Model performance is measured with area under the receiver operating characteristic curves (AUC).

Framework

Predictive Tasks | Indel Prioritization

- Data from HGMD
- 0.85 > AUC > 0.75



Supplementary Figure 8

DeepSEA-based classifier prioritized functionally annotated indels with high performance

HGMD regulatory indels prioritization performance was evaluated against negative 1000 Genomes indel groups with different distances to positive indels (average distance shown on the x-axis). The performance was measured by area under receiver operating characteristic (AUC). The prioritization model was trained with HGMD regulatory single nucleotide substitution mutations against 1200bp average distance negative variants.

Strengths & Weaknesses

- Strengths
  - First deployment of deep learning methods in variant prioritization
  - *De novo* predictions for multiple tasks
- Weaknesses
  - *gkb-SVM* optimized on 300-bp input sequences, not 1000-bp ones
  - N = 77 sequences only to test for indels
  - SNP functional prioritization is *de novo*, but not *de novo*
  - More focus on functionally negative rather than positive SNPs


END | THANK YOU